Heat-Shock Promoters: Targets for Evolution by P Transposable Elements in Drosophila

Jean-Claude Walser1, Bing Chen1, Martin E. Feder1,2*

1 Department of Organismal Biology and Anatomy, The College, The University of Chicago, Chicago, Illinois, United States of America, 2 The Committees on Evolutionary Biology, Genetics, and Molecular Medicine, The College, The University of Chicago, Chicago, Illinois, United States of America

Transposable elements are potent agents of genomic change during evolution, but require access to chromatin for insertion—and not all genes provide equivalent access. To test whether the regulatory features of heat-shock genes render their proximal promoters especially susceptible to the insertion of transposable elements in nature, we conducted an unbiased screen of the proximal promoters of 18 heat-shock genes in 48 natural populations of Drosophila. More than 200 distinctive transposable elements had inserted into these promoters; greater than 96% are P elements. By contrast, few or no P element insertions segregate in natural populations in a “negative control” set of proximal promoters lacking the distinctive regulatory features of heat-shock genes. P element transpositions into these same genes during laboratory mutagenesis recapitulate these findings. The natural P element insertions cluster in specific sites in the promoters, with up to eight populations exhibiting P element insertions at the same position; laboratory insertions are into similar sites. By contrast, a “positive control” set of promoters resembling heat-shock promoters in regulatory features harbors few P element insertions in nature, but many insertions after experimental transposition in the laboratory. We conclude that the distinctive regulatory features that typify heat-shock genes (in Drosophila) are especially prone to mutagenesis via P elements in nature. Thus in nature, P elements create significant and distinctive variation in heat-shock genes, upon which evolutionary processes may act.

Introduction

Genes may vary in evolvability for many reasons, including physical susceptibility to mutagenesis. Here we show that a class of genes with distinctive physical features—heat-shock genes—is remarkably prone to mutagenesis by insertion of a specific transposable element (TE), the P element of Drosophila.

TEs are mobile, repetitive DNA sequences and a structurally dynamic component of genomes [1]. TEs can cause gene and chromosome evolution in numerous ways, including insertional mutagenesis, retroposition, conveyance of regulatory elements to novel sites, and service as pivotal sites for ectopic recombination, and thus chromosomal rearrangements and gene duplication. For such evolution to occur, however, TEs must first insert into chromatin, which in turn requires that the target site be accessible to the transpositional machinery [2]. Indeed, insertion of Drosophila P elements, among the best-studied of TEs [3], into specific sites is associated with features of local chromatin architecture such as DNase I hypersensitivity, location in 5’-flanking sequence, presence of pre-existing TEs, and physical structure, but only weakly with insertion sites’ nucleotide sequence (e.g., [4–6]). These features vary widely and frequently throughout genomes [7], which is consistent with the irregular, but repeated, occurrence of TEs. Entire classes of genes also vary in TE frequency—and hence potentially evolvability via transposition—in laboratory studies [8], but for natural populations, neither the mechanistic basis for this variation nor its relevance for evolvability is clear.

In such experimental work with Drosophila, heat-shock genes (e.g., the major heat-shock gene Hsp70) stand out as a class receiving numerous TE insertions [8–10]. (By “gene,” we intend both the transcribed sequence and associated non-transcribed regulatory sequence.) This distinction is not unexpected from two perspectives. First, the local chromatin architecture of heat-shock proximal promoters is peculiar, incorporating constitutively decondensed chromatin and nucleosome-free regions [11,12], and constitutive engagement of the transcriptional machinery. In addition to the 5’ location of these promoters, such features should predispose these regions to TE insertion (see above). Second, TEs segregate at high frequency in natural populations in the 5’-flanking regions of the five genomic copies of Hsp70 [13–17]. This finding is remarkable given that TEs typically are at low allelic frequency in the Drosophila genome, presumably because they are deleterious [18–20]. The Hsp70 intragenic TEs are seemingly adaptive, exhibiting repeatable demographic variation in allelic frequency along natural thermal gradients and beneficial impacts on Hsp70 expression and components of fitness [14–17,21]. Nonetheless, TEs constitute 22% of the Drosophila genome [22] and are numerous (more...
Synopsis

Transposable elements can be a major source of evolutionary change. Their insertion can directly affect the genes into, or next to, which they insert. To insert, however, they must first gain access to the host gene. The authors reasoned that, because the DNA in the promoters (i.e., regulatory regions) of heat-shock genes is unusually accessible, these genes might harbor many transposable elements. With a technique that can detect any insertion into a gene, they discovered more than 200 distinctive transposable elements in the promoters regions of heat-shock genes in fruit flies from the wild—but few or none in the promoter regions of more typical genes. Surprisingly, out of the one hundred kinds of transposable elements in fruit flies, almost all were P elements. P elements are remarkable because they invaded the fruit fly genome only during the last century. These findings imply that the combination of accessible DNA and the recent invasion of P elements have left a distinctive imprint on the promoters of heat-shock genes.
are into identical sites in different populations (Figures 2, 4, and 5). Most (42 of 48) populations had a P element insertion in at least one investigated gene.

These findings prompt two methodological concerns, which are unwarranted. First, insertions common to multiple populations could be shared by descent, and hence tallying them as independent would overestimate insertion events; we exclude this possibility below. Second, the high proportion of P elements could stem from oversensitivity of the UFW screen to P elements. To address this concern, we re-screened a subset of genes known to contain both P elements and other TEs (the Hsp70 genes in each natural population) with a technique known to detect all these TEs. The re-screening used a reliable PCR with one primer complementary to Hsp70 and the other to each of the six TEs common in the Celera strain (rho, 1360, 297, Jockey, I, and Gypsy) [26]. This procedure re-detected each Jockey and Gypsy previously implicated by UFW in Hsp70, but revealed no additional TEs in any of the other genes. Two genes from each gene set (Gene Sets I–III) were likewise re-screened for four populations (F04, F40, F53, and F54), with the result entirely consistent with the previous screens. These results affirm that the method is universal, detects TEs when present, and does not favor P elements.

Abundance of Transposon Insertions in the Three Gene Sets

The UFW- and TE-specific PCR screens together initially detected 177 TE insertion sites (containing 171 P elements; Figures 2, 4, and 5), one Jockey inserted in a P element, and five Gypsy (one a Gypsy inserted into a Gypsy; Figure 2) in the 55 genes surveyed (Table 1) in the 48 natural Drosophila populations (Figure 1). Because the screens intentionally focused on the proximal promoter region of each surveyed gene, 167 of these insertion sites (containing 161 P elements, the Jockey, and the Gypsids) were in this region, often near the TATA box or initiator of the associated gene (e.g., Hsp22 and Hsp27). The screen also included more 3′ regions for some (but not all) genes, and detected ten P element insertion sites in coding sequence (e.g., for Hsp70, Hsp22, Hsp68, Hsp83, Hsr, su(s), act5C, and elf; Figures 2, 4, and 5) or in nearby genes (e.g., nearby bg25D and CG6296; Figure 5). Because these regions were surveyed only in those genes in which the UFW primers encompassed them, P elements in these regions were excluded from the following comparative analyses.

The data support three of our a priori expectations: (1) that novel TE insertions into Hsp70 genes should be readily discoverable (Figure 2), (2) that TE insertions should be numerous in Hsp genes other than Hsp70 (Figure 3), and (3) that TE insertions should be rare in genes dissimilar to Hsp genes (Figure 3). A total of 29 natural populations (with at least 44 distinct insertions) harbored P element insertions into at least one of the five Hsp70 gene copies (Figure 2), with 13 into the Hsp70aa gene, four into Hsp70Ab, and the balance not localized to a specific Hsp70 gene. Hsp70 genes also harbored all of the non-P transposable elements detected by the screen (one Jockey and five Gypsids). P element insertions were also numerous (in 30 natural populations with 37 distinct insertions) for Hsr (Figure 4), a single-copy sequence encoding a heat-inducible mRNA. Other heat-shock genes with insertions in more than five natural populations include Hsp22, Hsp23, Hsp26, Hsp27, Hsp68, and Hsp87 (Figure 3).

Excluding Hsp70, which presents an expanded target due to its multiple copies, one or more P elements were present in the proximal promoter of the first gene set (Gene Set I) in 88 samples (9.8%, 50 populations × 18 genes = 900 samples). In fact, 94% (152) of all 161 P insertions discovered in proximal promoters were located in heat-shock genes (Hsp70 and Gene Set I; Figure 1, inset; Figure 4). By contrast, in Gene Set III, selected for dissimilarity to Hsp promoters (see Introduction

Table 1. Genes Other Than Hsp70 Screened for Transposable Element Insertions in 5′-Flanking Sequence

<table>
<thead>
<tr>
<th>Gene Set I</th>
<th>SYM</th>
<th>ID</th>
<th>Gene Set II</th>
<th>SYM</th>
<th>ID</th>
<th>Gene Set III</th>
<th>SYM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG4460</td>
<td>Hsp22</td>
<td>CG12055</td>
<td>Gapdh1</td>
<td>CG7906</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG4463</td>
<td>Hsp23</td>
<td>CG6893</td>
<td>Gapdh2</td>
<td>CG7924</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG4183</td>
<td>Hsp26</td>
<td>CG9277</td>
<td>J tws56D</td>
<td>CG15634</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG4466</td>
<td>Hsp27</td>
<td>CG10798</td>
<td>dm</td>
<td>CG12011</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG5436</td>
<td>Hsp68</td>
<td>CG3481</td>
<td>Adh</td>
<td>CG4818</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG1242</td>
<td>Hsp83</td>
<td>CG12181</td>
<td>Sgs4</td>
<td>CG2916</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CR31400</td>
<td>Hsra</td>
<td>CG6222</td>
<td>su(s)</td>
<td>CG10591</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG10578</td>
<td>Dna1-1</td>
<td>CG4027</td>
<td>Act5C</td>
<td>CG11650</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG12240</td>
<td>Dna1-60</td>
<td>CG4550</td>
<td>ninaE</td>
<td>CG2961</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG5504</td>
<td>Dna1</td>
<td>CG7939</td>
<td>Rpl32</td>
<td>CG9506</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG9920</td>
<td>Hsp60</td>
<td>CG6382</td>
<td>Elf</td>
<td>CG7252</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG12101</td>
<td>Hsp60</td>
<td>CG8040</td>
<td>CG32061</td>
<td>CG4602</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG2830</td>
<td>Hsp60b</td>
<td>CG13340</td>
<td></td>
<td>CG9984</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG7756</td>
<td>Hsp60</td>
<td>CG4750</td>
<td></td>
<td>CG14023</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG4147</td>
<td>Hsp70-2</td>
<td>CG2512</td>
<td>fubh4D</td>
<td>CG10154</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG4264</td>
<td>Hsp70-4</td>
<td>CG1913</td>
<td>fubh4B</td>
<td>CG7724</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG8542</td>
<td>Hsp70-5</td>
<td>CG2238</td>
<td>Ef2b</td>
<td>CG2070</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG4167</td>
<td>Hsp67Ba</td>
<td>CG11181</td>
<td>cup</td>
<td>CG6296</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Gene Set I: Heat-inducible genes. Hsp genes, chaperone genes. Gene Set II: Genes resembling Hsp70 in regulation by polymerase pausing, DNase hypersensitivity, or high germline transcriptional activity. Gene Set III: Genes expressed only briefly in the life cycle and with negligible transcriptional activity in the germline. ID refers to the FlyBase ID; SYM indicates the standard abbreviation (gene symbol) for the name of the gene. For more details, see text or Table S1.

DOI: 10.1371/journal.pgen.0020165.t001
and Table S1), only 0.2% of the samples (50 populations × 18 genes = 900 samples) included a transposon (one insertion in one population each for CG6295 and CG14030; Figure 5). These transposons, furthermore, seem not to have inserted into the proximal promoter of their genes, but may have inserted into neighboring genes.

The data do not support our last expectation, however, that TE insertions should be common in proximal promoter regions of non-Hsp genes similar to those of Hsp genes (Gene Set II). Only 0.8% of samples (50 populations × 18 genes = 900 samples) included a TE. Although ordinarily such ‘natural experiments’ do not permit replication, the FlyBase Database (http://flybase.bio.indiana.edu/) records anthropogenic insertions of natural and synthetic transposons. For example, the Berkeley Drosophila Gene Disruption Project (BDGDP) has undertaken genome-wide P element mutagenesis of laboratory stocks. Indeed, analysis of these data for the three gene sets (Gene Sets I–III) recapitulates the outcome of natural mutagenesis (Figure 3). As of 2004, 157 experimental transpositions into the same genes that we screened in the natural populations are on record, of which 69% were into Gene Set I (Hsp genes), 24% were into Gene Set II (Hsp-like non-Hsp genes), and 7% were into Gene Set III (negative control set with Hsp-dissimilar genes; Figure 6). Thus most genes from Gene Set I (14 out of 18; 78%) and II (13 out of 18, 72%) had at least one insertion, whereas only six out of 18 (33%) of Gene Set III had insertions (Figure 6). As in natural mutagenesis, Hsr is distinctive, receiving more than twice as many P element insertions as any other gene in the three sets. The two datasets are highly concordant when all genes surveyed are ranked according to number of transposon insertions in the proximal promoter for (1) the 48 natural populations and (2) the synthetic transposon mutant strains (Spearman rank correlation test; p < 0.001). In many cases this similarity extends to the specific insertion sites themselves (Figures 4 and 5).

According to the Fisher exact test, the natural P element insertions were more numerous in Gene Set I than in Gene Set II (p = 0.018) and in Gene Set III (p = 0.002). By contrast, experimental P element insertion in Gene Sets I and II were not significantly different (p = 1.0), but more numerous in each than in Gene Set III (p = 0.004 and p = 0.010, respectively).

Characteristics of P Elements within Promoters

The insertion sites of P elements in the Hsp promoters were themselves highly clustered, with up to eight populations putatively receiving different insertions at the same site (Figure 2). The elements also varied in orientation relative to the associated coding sequence, with no apparent orientation preference (Figure 7).
Re-screening Establishes that P Elements Are More Numerous than the UFW Screen Indicates

Of the transposon inserts discovered, 47 were clearly unique (i.e., no two populations shared the same transposon at the same location in the same gene). By contrast, many different populations exhibited identical insertion sites, with up to eight populations showing P element insertions at the same position (this, for example, in the proximal promoter of Hsp70; Figure 2). As cited in the methodological concerns (above), these could represent insertion(s) of a P element into a common ancestral population and its vertical transmission into daughter populations, rather than independent multiple insertions.

**Figure 2.** Locations of TEs Integrating into the Proximal Promoters of Hsp70 Genes

Six nearly identical Hsp70 genes are present in the sequenced Drosophila genome, but only five copies in natural populations. The locations of selected promoter elements and sites are indicated for all copies.

(A) Previously discovered TEs and experimental transpositions relative to the conserved Hsp70 sequence. a, Jockey element in Hsp70Ba [16]; b, c, and d, P elements in Hsp70Ba [14,15,21]. An S element is present between the oppositely oriented paralogs Hsp70Aa and Hsp70Bb [13], and is represented twice, corresponding to its location relative to each paralog, as are the HMS Beagle (e) [16] and ’56H8’ (f) [88] elements inserted within it. Triangles below the line indicate transgene insertion sites (FlyBase; http://flybase.bio.indiana.edu). (B) and (C) Bottom: newly discovered TEs, with the natural population in which they were discovered (F01–F54, exclusive of F06) indicated for each.

(B) TEs other than P elements. Four are Gypsy elements that have integrated into the S element in specific populations, the fifth is a Gypsy that has inserted into a Gypsy, and the sixth is a Jockey that has inserted into a P element. The Gypsies are arbitrarily plotted relative to Hsp70Ab and Hsp70Aa, respectively.

(C) Natural P elements in Hsp70. The arrows indicating the number of independent EPgy2 insertion sites recently described by Shilova et al. [10]. Except for the Gypsies, TEs were not mapped to a specific Hsp70 gene. Insertion sites localized within the Hsp70 region were all established by sequencing. For population codes, see Figure 1.

DOI: 10.1371/journal.pgen.0020165.g002
Figure 3. Number of Natural P Element Insertions (161 Total) Distinctive by Population and Location into the “Proximal Promoter Region” of Each of the Screened Genes (Table 1)

Genes without any such insertions are not represented in the main figure. These tallies and estimates are conservative in three ways: (1) P elements inserting within 1,000 bp of the transcription start site of Bsg25D and CG6396 are included although they actually insert into neighboring genes (see Figure 4); (2) The tally for Hsp70 excludes non-P elements and those previously discovered (Figure 2), and divides the remaining total (44, light gray bar in background) by five, the Hsp70 copy number for natural populations [17]; and (3) Re-screening of a subset of insertions implies an underestimation of the tally at the 161 P insertion sites (see Results and Figure 7). Inset: Frequencies of genes in each Gene Set (I, excluding Hsp70, II, and III) in which 0, 1, or >1 P elements had inserted.

DOI: 10.1371/journal.pgen.0020165.g003

insertion events at the same site, and hence overestimate P element insertions. Alternatively, at any insertion site detected by UFW, multiple but similarly sized P elements might have inserted and presently segregate in any natural population so that UFW screening might underestimate TE insertions.

To re-examine our above estimate of TE abundance in natural populations, 43 insertion sites (17 unique, with nine sites shared in two to six populations) in three different genes (Hsp23, Hsp27, and Hsr0) were chosen for re-analysis. We used a PCR-based technique that reports both the size and orientation of the P element (Figure 7).

In the 17 unique sites are 30 P elements distinguishable by size and orientation. In the nine shared sites are 30 P elements distinguishable by size and orientation. Of these, five are present in two populations, and one is present in four populations (Figure 7). In all but one of these instances, the shared P element is singular in a first population and one of two to three forms segregating in a second population. In the remaining case, in Hsp27 in populations F51 and F54, two P elements segregate at the same insertion site in both populations. A similar re-screening of the Hsp70 genes at a “single” insertion site in each of five populations (Figure 7) detected one to six distinctive P element variants at each site (15 total). Each was localized relative to a sequence shared by the five Hsp70 genes in natural populations, but not to a specific Hsp70. Thus, each P element could be present at the same insertion site in one to five Hsp70 copies, with the tally of 15 under-representing the actual number of insertions.

In summary, excluding the re-screened P elements in Hsp70, 60 distinctive P element variants were found at the 43 sites re-analyzed, suggesting that the UFW screen undercounted distinctive P elements by nearly 30%. Corrected for this undercount, all 117 P element insertion sites in the three gene sets likely harbor 163 distinctive transposable elements in the natural populations, or 225 with those in the Hsp70s included.

These distinctive P elements may represent distinctive insertion events, distinctive evolution after common insertion events, or both. To estimate a lower bound for the frequency of insertion events, distinctive P elements at the same 43 insertion sites were re-tallied based on orientation only (and not size). Any two P elements with opposite orientation likely arose from independent insertion events rather than from evolution after a common insertion (but more than two independent insertions cannot be distinguished). On this basis, 43 distinctive P elements are distinguishable at the 43 sites. In other words, the P element tally based on UFW likely does not overestimate the number of independent insertion events, although it may underestimate this number.

Allelic Frequencies of the P Element Insertions

Individual P elements varied in both nucleotide sequence and allelic frequencies in populations. A prior study [14] suggested that transposition into the proximal promoters of Hsp genes can be selectively advantageous because of its impact on Hsp gene expression. Moreover, although deleterious TEs might be inactivated if not purged from populations [32,33], adaptive TEs might be maintained at high frequencies or modified. Surveys of several randomly selected populations (Table 2; Figure 7) are consistent with the simultaneous modification and maintenance of TEs. All P elements discovered whose size was determined were less than full length. The allelic frequencies of P elements at each site surveyed (Table 2) varied considerably among populations, ranging from very low (e.g., in F40 for Hsp22) to high (e.g., in F04 for Hsp27) or even fixation (e.g., in F51 for Hsr0). Populations also differed in the number of different variants of the P element inserted in a particular region of a gene or the number of insertion sites (Table 2), and frequency and number of insertion sites are not correlated. Population F40, for example, harbors three different P elements at two different sites in Hsr0. The allelic frequencies, however, of these insertions are low (6%), whereas that for a single P element in Hsp27 in population F04 is much higher (85%; Table 2; Figure 4).

In six (12%) natural populations (F09, F10, F11, F16, F41, and F42; Figure 1), no P elements had inserted in the proximal promoter regions of the genes under investigation according to UFW. Re-screening these populations with PCR revealed no P elements; P elements either are absent in these populations or too distant from one another to support PCR amplification. Interestingly, all six populations are geographically adjacent (Figure 1). Only about half (ten out of 21) of the African populations harbored two or more P element insertions, in contrast to 78% of the populations outside Africa. The population with the most numerous insertions (n = 14, F40 from Marrakech), however, is African. Although
Figure 4. Locations of P Elements Integrating into the Proximal Promoters of Heat-Shock Genes Other Than Hsp70 (Gene Set I)

Data are plotted as in Figure 2 except as follows: F-numbers in columns refer to natural populations with transposons integrating at identical sites. Primer sets used in the screens amplified regions 3' to transcription start site of different length; P elements discovered upstream of the initiator are plotted (pale), but not included in comparative analyses (i.e., in Hsp22 in population F31, Hsp68 in F05, and in Hsro in F14). In Hsro, numerous P elements were discovered in one region (box); a randomly chosen subset of these were localized (by sequencing) within that region (see enlargement). Several putative deletions were also discovered, and are plotted. For population codes see Figure 1. Table S1 provides additional information about Gene Set I and relevant sources for the organization of promoter regions.

DOI: 10.1371/journal.pgen.0020165.g004
specific mobile element insertions vary in frequency along geoclimatic gradients [14–17,21], in our data, the number of insertions bears no apparent relationship to latitude or climate.

**Discussion**

We have hypothesized that heat-shock genes as a class are distinctively evolvable because TEs integrate into their proximal promoter regions at unusually high rates, creating unique regulatory variation on which evolutionary processes such as natural selection can act. Here we test a major component of this hypothesis, that heat-shock proximal promoter regions are especially susceptible to the integration of TEs. TEs and other repetitive elements constitute more than 20% of the *D. melanogaster* genome [22], comprising 6,013 specific elements in more than 120 families [23,26]. Although these numbers are less than for other multicellular eukaryotes (e.g., ~45% in humans) [34], they nonetheless establish that TEs are numerous and diverse in *D. melanogaster*. TEs in *Drosophila*, moreover, are active, accounting for more than 50% of spontaneous mutation [35] (versus 0.2% in humans [36]). Accordingly, the discovery of more than 170 additional TEs in natural *Drosophila* populations is, in itself, unremarkable. What is striking is the predilection of these TEs for insertion in the proximal promoters of one class of genes, their persistence after insertion, and the fact that almost all are *P* elements.

**Methodological Issues**

Deducing this susceptibility from compilations of insertion sites is prone to bias unless (1) all TEs are detected, and (2)
Promoter Evolution in Heat-Shock Genes

The abundance of P elements in Drosophila heat-shock promoters may arise from the species’ distinctive (but common) ecological niche. Like many organisms, D. melanogaster often undergoes hyperthermia in nature (e.g., [39,40]). Drosophila infests necrotic fruit, wherein eggs, larvae, and pupae are prone to heat stress. At the biochemical level, this hyperthermia is deleterious to proteins and membranes, and for the former may initiate a cytotoxic cascade of denaturation and aggregation of proteins [41]. At the organismal level, this heat stress compromises development, reproduction, and survival. A primary and important response to heat stress is therefore the expression of heat-inducible molecular chaperones, which can deter protein aggregation, target damaged proteins for degradation, help non-native proteins refold in the cell, and/or remove proteins from aggregates for refolding or degradation [42]. Given these essential roles, heat-inducible molecular chaperones are poised for rapid and massive accumulation upon heat shock [41]—hence their original name, “heat-shock protein.” Indeed, several distinctive features of the heat-shock genes of complex eukaryotes appear to facilitate heat-induced expression: constitutively decondensed chromatin, nucleosomes positioned outside the proximal promoter, a pre-assembled (but paused) polymerase apparatus, a pre-expressed, inactive, but readily activatable transcription factor (HSF), and the absence of introns needing splicing (e.g., [12,43–48]). Each feature can be viewed as an elimination or minimization of a time-consuming step in gene expression, and is thus an appropriate component of an emergency response to rapid and unpredictable thermal damage.

As explained in the Introduction, these same features give TEs accessibility to chromatin, which could facilitate insertion [2]. Because Hsp genes share some of these features and represent extreme manifestations of others, Lerman et al. [15] suggested that proximal promoters of Hsp genes in general were natural “hotspots” for TE integration. Although this suggestion was consistent with the discovery of naturally occurring TEs in Hsp70 genes’ promoters [15], (1) these TEs were few, (2) with few exceptions [49], naturally occurring TEs had not been discovered in other Hsp genes, and (3) TEs were not known to be comparatively rare in the proximal promoter regions of non–heat-shock genes. Three independent lines of evidence now establish that insertions of one TE, P elements, are common, not only in Hsp70 promoters, but also in other (single copy) heat-shock promoters:

The present screen of natural populations documents (1) numerous P elements inserted into the proximal promoters of Hsp70 genes (Figure 2), (2) numerous P elements inserted into the proximal promoters of Hsp genes other than Hsp70 (Figure 3), and (3) few or no P elements inserted in the proximal promoters of Hsp-dissimilar genes (Figure 4). That P elements have inserted in non-transcribed sequence is not unexpected [50], but our comparison is of identical regions in each gene set. If transposition into a region is solely a function of that region’s representation in the euchromatin, then each gene set should accumulate equal numbers of P elements. The UFW and P element screens, by contrast, detect 152 (94.4%) in the heat-shock genes, seven (4.4%) in Gene Set II, and only two (1.2%) in Gene Set III (Figure 1, inset; Figure 3). As explained, the two in Gene Set III actually reside in nearby genes (Figure 5) and could therefore be excluded (although we have not done so).

The accompanying survey of transgene insertion experiments (Figure 6), mainly from a genomic mutagenesis scheme that relied on P elements [8], has largely the same outcome (in number and position), at least for heat-shock genes (Hsp70 and Gene Set I) and the negative gene set (Gene Set III).

Shilova et al. [10] mobilized P transposon constructs (EFG2) adjacent to the Hsp70A gene cluster and documented numerous new insertions into Hsp70 via local transposition (Figure 2C). They recovered 46 independent insertions of
which, remarkably, 50% were into the same two nucleotides, $-96$ and $-97$, which is in correspondence with our results (Figure 2C). These two nucleotides harbor 12 (27%) of the 44 P elements we discovered in Hsp70 genes in natural populations, and represent a remarkable conjunction of DNase I hypersensitivity, adjacency to GAGA factor binding sites, absence of nucleosomes, and insertion site preference [10].

This largest number of P element insertions, both natural and transgenic, is into Hsro, a highly conserved single-copy

---

**Figure 7.** Distinctive P Elements Revealed by Re-screening a Random Sample of P Element Insertion Sites in Natural Populations for Four Genes, Hsp23, Hsp27, Hsro, and Hsp70

The P element insertion sites were selected from Gene Set I. A plus sign (+) indicates successful PCR amplification with one primer complementary to the focal gene and another complementary to a unique sequence in the P element (top), and thus reports the size and orientation of the P element; a minus sign (−) indicates no amplification. Table S7 provides sequences of these primers. At each insertion site in a population, one to six distinctive P elements segregated; these are designated a−f. For Hsp23, Hsp27, and Hsro, nine insertion sites shared by two or more natural populations (indicated by boxes) and 17 unique insertion sites were re-screened. Amplicons that share a symbol (filled square [●], filled triangle [▲], filled circle [●], etc.) occurred at the same integration site in different populations and were indistinguishable by size or orientation. For Hsp70, a five-copy gene in natural populations [17], the specific gene of insertion was not determined; thus, each distinctive amplicon (a−f) could represent insertion(s) at the same site in one to five of the Hsp70 genes. For population codes see Figure 1.

ORF, open reading frame.

DOI: 10.1371/journal.pgen.0020165.g007
gene that encodes a heat-inducible mRNA, but no protein [51]. Like many non-coding RNAs, Hsrw transcripts play diverse regulatory roles [51], including regulation of distribution of RNA binding proteins [52] and the suppression of non-Hsp gene expression upon heat shock [53], and suppression of polyglutamine neurotoxicity [54]. In 30 (62.5%) of the 48 natural populations surveyed, a P element has integrated in Hsrw, with 70% distinctive integration events into nucleotides −89 to −161 upstream of the transcription start site. Why this region is so susceptible to P element integration is not apparent, but these insertions are not into any major regulatory element [55], and most of this region is not conserved within D. melanogaster.

Two additional aspects favoring successful P element integration in the germline are expression of the host gene, and the occurrence of this expression in germ cells prior to meiosis, an embryonic process in D. melanogaster [9,56,57]. Gene Set III, essentially negative controls, was selected for restriction to narrow developmental windows and low levels of expression (Table S1). Genes in this set also exhibit neither constitutive chromatin decondensation, nor positioned nucleosomes, nor regulation via polymerase pausing; i.e., none of the attributes hypothesized to favor TE integration in Hsp promoters. Accordingly, the near absence of TEs discovered in this set and their complete absence from the proximal-promoter region is consistent with the above conclusions.

The paucity of transposons discovered in Gene Set II, the “Hsp-like” genes, is more difficult to reconcile, however. One possible explanation is that the “true” Hsps are prone to massive environmentally induced expression during the embryonic period in which germline transformation is possible, whereas the Hsp-like genes are not. As we discuss below, an alternative, but non-exclusive, explanation is that natural selection more effectively eliminates TEs from Hsp-like genes than from Hsp genes.

**Why Have P Elements Persisted in the Promoters of Hsp Genes in Nature?**

A common view is that TE insertions into genes are generally deleterious because mutations in general are usually deleterious and are therefore eliminated from the genome by purifying selection. Evidence includes the deleterious phenotypes that commonly result when TEs integrate into genes (for P elements, reviewed by [58]), the comparative rarity of TE insertions into coding or regulatory regions of genes, the low frequencies of intragenic TEs segregating in natural populations [18,19], and the rapid evolution of suppression of transposition in natural populations (e.g., trans-acting RNAs, DNA methylation, and specific antisense RNA). For Drosophila, instances in which intragenic TEs are beneficial are so rare that accounts of such instances are newsworthy [59–62]. Why, then, are P elements in the Hsp promoters of naturally occurring Drosophila so numerous?

One explanation is that P elements are distinctive transposable elements in several respects. First, unlike other TEs (for example, the Ty elements of Saccharomyces, which insert in specific sites of specific genes [2]), the information content of P element insert sites is relatively low [63]. P element insertions thus are general and robust reporters of exposed chromatin. Second, P elements have invaded the D. melanogaster genome relatively recently. Biogeographic patterns of P element occurrence in D. melanogaster and the recent near-disappear-

### Table 2. Frequency of P Element Insertions in the Promoter Region of Nine Heat-Shock Genes in Different Populations

<table>
<thead>
<tr>
<th>Population</th>
<th>Hsp22</th>
<th>Hsp23</th>
<th>Hsp26</th>
<th>Hsp27</th>
<th>Hsp68</th>
<th>Hsp83</th>
<th>Hsrw</th>
<th>DnaI-1</th>
<th>Hsc70-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>F04</td>
<td>.</td>
<td>47</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>F05</td>
<td>.</td>
<td>35</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>F12</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>F40</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>F51</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

First digit in superscript refers to the number of insertion sites found for this gene in this population and the second digit refers to the number of distinct gene alleles. A dash (—) indicates not tested because screening did not indicate insertion. For population codes see Figure 1.

### Table 3. P Element Insertions into Genes in Regions of High and Low Recombination

<table>
<thead>
<tr>
<th>Gene Set</th>
<th>High Recombination</th>
<th>Low Recombination</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P Element(s) Inserted</td>
<td>No Insertion</td>
</tr>
<tr>
<td>I</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>II</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>III</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td>15</td>
</tr>
</tbody>
</table>

Indicated are the numbers of genes in each gene set that are in regions of high and low recombination according to [38], and with one or more, or zero P elements in their proximal promoter region.

DOI: 10.1371/journal.pgen.0020165.t003
ance of D. melanogaster populations without P elements together suggest that P elements invaded the D. melanogaster genome within the last century [29–31]. Since then, P elements spread widely and effectively throughout the D. melanogaster genome, likely due to self-regulation and hybrid dysgenesis [29,64].

TEs typically undergo excision or degeneration with time. Indeed, whereas full-length P elements encode a transposase and are thus autonomous, every P element discovered in the UFW screen whose size we have determined is less than full length. Thus, the occurrence of P elements in the Hsp proximal promoters could be transient, representing a successful genomic invasion that has not yet been purged. Although this may be so, populations we have surveyed average two to three P elements in their Hsp promoters. Given that natural populations typically carry no more than 50 P elements distributed throughout their more than 13,000 genes [65], their abundance in Hsp promoters is remarkably high.

Importantly, in natural Drosophila populations TEs in Hsp promoters may be advantageous and therefore persist via positive selection (e.g., [21]). Although Hsps typically encode proteins that can function as molecular chaperones, these proteins have numerous other functions including intra- and extra-cellular signaling (small Hsps, Hsp60, and Hsp70) [66], regulation of cell cycle and apoptosis [67,68], and maturation and regulation of nuclear receptors, among others [66,69]. Hsp90 (encoded by Hsp83 in Drosophila) alone interacts with numerous (>200) client proteins in the cell [70]. Thus, whereas massive Hsp accumulation often may be advantageous for temperature tolerance and mitigation of thermal damage, in the absence of heat stress, stringent regulation of Hsp levels may be essential. Indeed, unbridled expression of Hsp genes in the absence of heat stress is often deleterious (reviewed by [15]). The typical phenotype of TE insertions into Hsp70 promoters is decreased gene expression [14], and natural P insertions have similar phenotype in Hsp26 in population F32 (unpublished data). Accordingly, selection may favor the retention of TEs in Hsp promoters because these elements reduce expression of a situationally harmful protein. Natural and experimental selection can either increase or decrease Hsp expression depending on which outcome a given thermal environment favors (reviewed by [15]). P element insertion may thus simply be an opportunistic way to suppress Hsp activity in the wild.

Intragenic TEs segregating in natural Drosophila populations are typically at very low frequencies, consistent with their usually deleterious phenotypes (see above). On the other hand, intragenic TEs considered advantageous, by contrast, are typically at much higher frequencies, if not fixed, in natural populations [13,20,71]. In our survey, although frequencies of P elements at specific sites varied considerably in natural populations, these sometimes were much higher than those invoked as evidence for deleterious or neutral phenotypes [18,19]. Ten of the 16 P element insertions we arbitrarily chose for examination in detail were at population frequencies ≥15%, seven at ≥35%, and four at ≥50% (Table 2). This outcome resembles that for prior studies of TEs in Hsp70 promoters in natural populations, in which frequencies were high and varied inversely with Hsp70 protein levels [14–17,21]. A testable prediction is that experimental evolution in contrasting thermal regimes should be capable of altering P element frequencies in Hsp genes according to the phenotypes of P element insertions; i.e., increasing P element allelic frequency when beneficial and decreasing frequency when deleterious. The same prediction ought to be applicable to other instances in which TE-derived sequences modify a host function and have been assimilated by the host genome [1,20,49,72,73].

Interestingly, in the transgene lines, laboratory strains that are intentionally isolated from natural selection, the Hsp and Hsp-like genes included in our survey received transposon insertions at similar frequencies (Figure 6, inset) and positions. In nature, by contrast, the proximal promoters of the Hsp-like genes (Gene Set II) were relatively depauperate P elements. This contrast further implicates positive or balancing selection for maintenance of P elements in the Hsp genes (Gene Set I and Hsp70). In other words, in nature, selection may routinely purge P elements from the proximal promoters of non-Hsp genes and/or favor insertions in Hsp genes. It suggests that accessibility of chromatin to P element integration may be necessary, but not sufficient, to generate stable P element insertions in the proximal promoters of these genes in nature because of negative or purifying selection.

Heat-shock genes are a component of an ancient, but effective, response to acute thermal stress in the natural environment, and include features that facilitate rapid and massive gene expression (e.g., [40,42,74]). These same features, however, may facilitate the integration of P elements into Hsp promoters, which in turn affect gene expression. These intragenic P elements thereafter may segregate in natural populations and are a form of genetic variation upon which natural selection and other evolutionary processes may act. The present study and another (unpublished data) establish that this scenario is not specific for a single multicytokine gene (Hsp70) in a small number of natural populations as previously described, but generally applicable to Drosophila populations worldwide and to the entire heat-shock genome. From this perspective, intragenic TEs in natural Drosophila populations are less newsworthy exceptions [59–62] than a normal expectation. The near-exclusive involvement of P elements and the frequencies of these elements establish, moreover, that the TEs in Hsp promoters are not the remnant of an ancient event, but a manifestation of active and ongoing microevolution in natural populations.

Heat-shock genes have been posited to make a special contribution to evolvability [75]. The capacity of their products—molecular chaperones—for conformation-specific recognition of diverse client proteins has led to their involvement in diverse regulatory processes and their assignment to diverse structural roles. Importantly, molecular chaperones may transiently suppress conformational mutations in client proteins, thereby protecting such variation from selection under routine conditions, but exposing it during stress [76]. TEs have likewise been posited to make a special contribution to evolvability [1]. They are the most widespread and effective of natural mutagens, can redistribute genetic material throughout the genome, and form the recombinatorial substrate for much gene duplication, reposition, and creation of hybrid genes. Here we show that these two components of evolvability intersect: heat-shock genes as a class are distinctively prone to the integration of at least one TE, the P element, and that, accordingly, the
expression of heat-shock genes may evolve distinctively from the rest of the genome.

Materials and Methods

*Drosophila* strains. The *Drosophila* stocks used in this survey were derived from wild-caught flies and were maintained as isofemale lines with the exception of F32, F51, and F52. Flies were from 51 different worldwide locations (Figure 1), plus a reference strain. Most lines were obtained from Dr. Jean David (Centre Nationale de la Recherche Scientifique, France), and have been the subject of previous investigations [77–80]. In addition, strains F01 (14021-0231.21), F02 (14021-0231.22), F03 (14021-0231.23), F04 (14021-0231.25), and F05 (14021-0231.26) were obtained from the Drosophila Species Stock Center, Tucson, Arizona, whose reference numbers are in parentheses. The strain whose reference name is in parentheses, “Lower Nahal Oren, Israel” [82], Additional strains and origins include: F49, Dr. Arne Bäcker (University of California, Irvine) and is one of the lines reared on a yeast, cornmeal, molasses, and agar medium at 25 °C for 2 min prior to DNA isolation. The washing buffer was removed and another 180-μl PBS was added for grinding. Total DNA was extracted according to [28].

Genes. Genes (Tables 1 and S1) were selected and grouped into sets for analysis a priori according to the following criteria: (1) the nearly identical Hsp70 genes [17]. These were included for re-analysis because of the prior discovery of several TE insertions in their 5′-flanking regions (see Introduction). These genes are arranged in two clusters (Hsp70A at 87A7, comprising Hsp70Aa and Hsp70Ab, and Hsp70B at 87C1, comprising Hsp70Ba, Hsp70Bb, and Hsp70Bb [Hb present], and Hsp70Bc); (2) other heat-shock genes (Gene Set I) [74,85]. Although all genes included in this group of genes increase in expression upon heat shock or other proteotoxic stresses, most (unlike Hsp70) [86] are expressed constitutively. Also, although several subsets of these genes share similarities in sequence, they are not multicycopy genes in the same sense as the Hsp70s, which encode proteins of identical sequence [17]; (3) genes resembling heat-shock genes in regulation of expression, chromatin configuration, associated promoter elements, etc. (see Introduction) (Gene Set II). These were identified from a list of TE superfamilies deposits (Table S1); and (4) genes dissimilar to heat-shock genes or with no known features similar to those of heat-shock genes (Gene Set III). These were selected from the database of Atkinos et al. [87], available online (http://flygenes.yale.edu), according to least expression throughout the *Drosophila* life cycle. All else equal, genes with limited embryonic expression were preferred. Initially selected genes were discarded if a literature search disclosed characteristics that might qualify them for inclusion in Gene Set II. In most cases, genes included in this set had not been studied in detail when the set was compiled, or had hypothetical functions according to sequence homology with better-studied genes (Table S1).

Transposable element screening. We executed two screens with differing characteristics. The first, a UFW screen reported size polymorphism in amplicons between the reference strain Celera (F06) and natural populations. We modified UFW as originally described by Myrick and Gelbart [27] and exemplary gel images are in Myrick and Gelbart [27] or Walser et al. [28].

About 15% of all UFW reactions initially failed or produced only modest amplification even after the nested PCR step, and were therefore repeated with a different set of primers. Populations F18, F50, and F52 were removed from the dataset after the reaction failed for multiple genes and primer sets. Presumptive positives (defined as extra bands in the UFW footprint compared to the reference strain) were cloned and then sequenced for identification (at least one per gene). We screened 15 of the 30 isofemale lines for primer design, and exemplary gel images are in Myrick and Gelbart [27] or Walser et al. [28].

When the vast majority of insertions (98%) that UFW screening revealed were P elements, a second PCR-based screen was designed to confirm the UFW results and discover further insertions. In this second screen, one primer was chosen in the 5′-flanking region of the associated gene, and the other in the 31-bp (base pair) inverted terminal repeat region of the P element. The P-specific PCR screen also included a further population from Israel (F51) and two populations from Japan (F53 and F54). Although primers for the second screen were gene specific, all positive samples (samples amplified in the second screen) were re-screened with another primer specific to the conserved region of the associated gene. PCR reactions with only one primer, the P element–specific primer, served as a control for inadvertent amplification of sequence between multiple P element insertions. Additionally, we sequenced 75% of all positive samples, including all fragments bigger then 450 nt. For amplicons smaller then 450 nt, a 6-FAM–labeled P element–specific primer was used, and the PCR products were sized on an ABI 3730 DNA sequencer with LIZ-500 as internal size standard using Genemapper 3.0 (Applied Biosystems, Foster City, California, United States).

Frequency of insertion. For selected genes and five populations (F04, F05, F12, F40, and F51) the frequency of *P* element insertions was estimated by analyzing 35–48 individual flies per population (Figure 1). DNA was purified from single flies according to the manufacturer’s recommendations (Puregene DNA Purification Kit; Gentra System, Minneapolis, Minnesota, United States).

Distinguishing P element insertion events. *P* elements discovered at the same site in multiple populations might have been inherited from a common ancestral *P* element insert, or have inserted independently. Also, any *P* element discovered at a site might either be singular or represent multiple elements segregating in a population. To distinguish among these alternatives, 43 *P* element insertions (at different sites with populations sharing zero to six *P* transposons at each site) from three UFW screen (Hsp70, Hsp70, and Hsp90) were re-screened with a PCR-based technique that reports both the size and orientation of the *P* element (Figure 7; Table S7). This screen exploited the tendency of the internal regions of *P* elements (but not
their termini) to truncate during evolution [3], but in highly variable fashion. From an alignment of full-length and truncated \(P\) element sequences deposited in the NCBI GenBank database, six forward and seven reverse primer sites were chosen. The corresponding amplicons (indicated by ‘+’ in Figure 7) are indicative of the size and orientation of the corresponding \(P\) element.

**Transgenic insertion sites.** The flanking region of all experimental \(P\) element transpositions thus far reported by the FlyBase Database (http://flybase.bio.indiana.edu) were used to characterize insertion site for the genes of the different gene sets (Gene Sets I–III). In addition, we also included locations of \(EPg2\) element insertions for \(Hsp70\) recently described by Shilova et al. [10] for comparison with the naturally occurring \(P\) elements. Spearman rank correlation was used to test for a monotonic relationship between the natural and the transgene insertion sites for the different genes. The Fisher exact test for count data was used to compare the number of gene with element inserts in the three gene sets.

**Other polymorphisms.** The UFW screen also detected several insertions/deletions, presumably deletions by parsimony (Figure 4). In population F40, a 1,381-nt deletion is in the 5′ region of \(Hsp27\). The deletion occurs upstream of the TATA box and removes all five heat-shock elements (HSE). Another deletion of 365 nt occurs in \(Hsp68\) in population F21. This removes all four HSES, the TATA box, and the initiator. Population F01 exhibited a pair of deletions in \(Hsro\), in which 17 nt and 5 nt of the 5′ and 3′ regions flanking the TATA box were removed. Screens of individual \(Drosophila\) (\(n = 50\)) suggest that these deletions segregate in populations at very low frequencies (<2%).

**Supporting Information**

**Table S1.** Supplementary Information on Genes in the Three Gene Sets

**Table S2.** UFW Primers for Gene Set I

**Table S3.** UFW Primers for Gene Set II

**Table S4.** UFW Primers for Gene Set III

**Table S5.** Transposable Element–Specific Primers for Additional Screens

**Table S6.** Primers for Intergenic Regions

**Table S7.** \(P\) Element–Specific Primers for Orientation and Length Determination

References


4. Berg CA, Spradling AC (1991) Studies on the rate and site-specificity of \(P\) element transpositions thus far reported by the FlyBase Database.


9. Gilmour, Ping Zhang, and Margaret Kidwell for helpful discussion.

Acknowledgments

The authors thank Marcos Antezana, Dimitri A. Petrov, Anthony V. Furano, and Michael Evgenev for critical reading of the manuscript, Robert Harmon for help in the lab, and Allen Spradling, David Gilmour, Ping Zhang, and Margaret Kidwell for helpful discussion. We are most grateful to the numerous investigators listed in Materials and Methods, who supplied \(Drosophila\) samples. The extremely helpful comments by Cedric Feschotte and the suggestion of two other anonymous referees improved the manuscript substantially.

Author contributions. J CW and MEF conceived and designed the experiments. J CW and BC performed the experiments. J CW and MEF analyzed the data. J CW and BC contributed reagents/materials/analysis tools. J CW and MEF wrote the paper.

Funding. Work was supported by National Science Foundation Grant 03-16027.

Competing interests. The authors have declared that no competing interests exist.