Lineage-specific expansion of the Zinc Finger Associated Domain ZAD

Ho-Ryun Chung, Ulrike Löhr, and Herbert Jäckle

Max-Planck-Institut für biophysikalische Chemie, Abteilung Molekulare Entwicklungsbiologie, Am Fassberg 11, 37077 Göttingen, Germany

The zinc finger associated domain (ZAD), present in almost 100 distinct proteins, characterizes the largest subgroup of C2H2 zinc finger proteins in *Drosophila melanogaster* and was initially found to be encoded by arthropod genomes only. Here, we report that the ZAD was also present in the last common ancestor of arthropods and vertebrates, and that vertebrate genomes contain a single conserved gene that codes for a ZAD-like peptide. Comparison of the ZAD proteomes of several arthropod species revealed an extensive and species-specific expansion of ZAD-coding genes in higher holometabolous insects, and shows that only few ZAD-coding genes with essential functions in *Drosophila melanogaster* are conserved. Furthermore, at least 50% of the ZAD-coding genes of *Drosophila melanogaster* are expressed in the female germline, suggesting a function in oocyte development and/or a requirement during early embryogenesis. Since the majority of the essential ZAD coding genes of *Drosophila melanogaster* were not conserved during arthropod or at least during insect evolution, we propose that the LSE of ZAD-coding genes shown here may provide the raw material for the evolution of new functions that allow organisms to pursue novel evolutionary paths.

Introduction

C2H2 zinc finger proteins (ZFPs) represent the most abundant nucleic acid binding proteins in the eukaryotic kingdom (e.g. Böhm, Frishman, and Mewes 1997; Lander et al. 2001; Chung et al. 2002; Englbrecht, Schoof, and Böhm 2004). Many vertebrate ZFPs are characterized by associated domains, like the KRAB and SCAN domains (Collins, Stone, and Williams 2001). Recently, a novel zinc finger associated domain (ZAD) was found in the N-terminal portion of many ZFPs of *Drosophila melanogaster* (Chung et al. 2002). The structure of this domain, as revealed by X-ray crystallography, resembles that of a treble-clef fold. The ZAD structure is stabilized by zinc coordination via four cysteine residues found to be invariably present in all identified ZAD peptides (Jauch et al. 2003). Similar to the KRAB and SCAN domains, the ZAD can function as protein-protein interaction module as shown for the Grauzone protein (Jauch et al. 2003) and a ZAD protein called Serendipity-6 (Payre et al. 1997; Ruez, Payre, and Vincent 1998).

The zinc finger associated domains (ZAD, SCAN and KRAB) characterize protein families whose members independently proliferated in distinct lineages, a phenomenon referred to as lineage-specific expansion (LSE; Jordan et al. 2001; Lespinet et al. 2002). For example the genomes of mouse and humans encode a comparable number of KRAB domains (Huntley et al. 2006), but many of these KRAB domain-coding genes were independently generated by gene duplication in the lineages leading to either mouse or humans (e.g. Urrutia 2003; Huntley et al. 2006). It appears that KRAB-coding genes are prone to duplicate, suggesting that there is positive selection for these duplication events. However, despite their large numbers in mammals, only few KRAB-coding sequences have been found in other deuterostomes (Birtle and Ponting 2006 and references therein). Thus, the tendency to duplicate is not a unique feature of the genes per se but has to be viewed in the context of the lineage-specific organisinal constraints and demands.

Many aspects of the ZAD are very similar to the KRAB domain. Like the KRAB domain in mammals, it characterizes the largest subgroup of ZFPs in *Drosophila melanogaster* and has previously only been identified in arthropods (Chung et al. 2002). Here, we show that the ZAD was present in the last common ancestor of arthropods and vertebrates. Although present in vertebrates we were able to identify only a single gene encoding a ZAD-like peptide, suggesting that the ZAD-coding genes underwent LSE in the lineage leading to *Drosophila melanogaster*. Moreover, by comparing the ZAD sequences of six arthropod species, we show that ZAD-coding genes were subject to LSE especially in the higher holometabolous insects.

Only few *Drosophila melanogaster* ZAD-coding genes with known and essential functions are conserved in evolution. This observation indicates that some or even the majority of the ZAD-coding genes exert lineage-specific or even species-dependent functions. We speculate that at least some of the lineage-specific ZAD-coding genes may be involved in processes that lead to developmental differences. This hypothesis is supported by our observation that many ZAD-coding genes of *Drosophila melanogaster* are expressed in the female germline, implying that they are involved in some aspects of oogenesis and/or contribute to the maternal effect on early embryonic development. Consistent with this view, we find that the maternal effect gene *piita*, which plays a critical role during *Drosophila melanogaster* oogenesis (Laundrie et al. 2003), is conserved in all holometabolous insects examined.

Results and Discussion

Based on initial searches in expressed sequence tag databases the ZAD was proposed to be restricted to arthropods (Chung et al. 2002). Meanwhile, however, we found a ZAD-like domain among the zinc finger proteins reported by the database of protein families Pfam (Finn et al. 2006). This zinc finger protein, termed ZFP276, initially described in mouse and subsequently in human, was previously reported to contain C2H2 zinc fingers only (Wong et al. 2007).
The Zinc Finger Associated Domain

2000; Wong et al. 2003). Additional and intensive analysis did not reveal any other instance of the ZAD in vertebrate protein, EST and genome sequence databases. Thus, ZFP276 and its orthologs are the only ZAD-coding genes present in vertebrate genomes. Orthologs of ZFP276 can be traced back to fish such as the zebrafish Danio rerio. This finding suggests that the ancestral form of the ZAD was present before the split of the fish and tetrapod vertebrates.

Figure 1 shows a multiple sequence alignment of the ZAD of ZFP276 orthologs and the ZAD of the Drosophila melanogaster transcription factor Grauzone (ZAD grau; Jauch et al. 2003). The sequences forming secondary structure elements in ZAD grau can be aligned without gaps or insertions. In loop regions, small sequence gaps and insertions are observed. The alignment also shows that most residues implicated to form the hydrophobic core of ZAD grau (Jauch et al. 2003) maintain their hydrophobic character and its orthologs are the only ZAD-coding genes present in all vertebrate species. As only a single ZAD-coding gene is present in the genomes of basal deuterostomes such as ascidians (Ciona intestinalis and Ciona savignyi) and echi- noderms (Strongylocentrotus purpuratus), the majority of ZAD sequences should be conserved in all arthropod genomes may have evolved convergently, since neither ZAD nor ZAD-like sequences are present in the genomes of basal deuterostomes such as ascidians (Ciona intestinalis and Ciona savignyi) and echi- noderms (Strongylocentrotus purpuratus).

Lineage specific expansion of ZAD-coding genes in higher insects

The high number of ZAD proteins in Drosophila melanogaster suggests that the ZAD proteome has been shaped by an evolutionary history that was dominated by expansion of ZAD-coding genes. As only a single ZAD-coding gene was found in vertebrates, the very high number of ZADs found in Drosophila melanogaster must have been generated soon after the divergence of deuterostomes and protostomes and/or only very recently in the evolutionary history of Drosophila melanogaster. We reasoned that we could distinguish between these possibilities by comparing the ZADs of Drosophila melanogaster to ZADs of other protostomes.

In order to get a broad overview of the set of ZADs in each species, we concentrated our efforts on species for which whole genome sequences are available. These species all belong to the arthropod phylum, including the dipteran mosquito Anopheles gambiae, the lepidopteran silk worm Bombyx mori, the coleopteran red flour beetle Tribolium castaneum, the hymenopteran honeybee Apis mellifera and the crustacean water flea Daphnia pulex (see also Figure 2). If ZAD coding genes expanded very early after the split of the deuterostomes and protostomes, the majority of ZAD sequences should be conserved in these arthropod species and orthologs should be readily identifiable. Conversely, in case of a recent expansion, the number of ZAD sequences conserved in all arthropod species should be rather limited, indicating that the majority of ZADs were generated by lineage-specific expansions (LSEs).

We were able to identify ZAD-coding sequences in each genome of the six above listed species. However, the numbers of ZADs were very different (Figure 2). In Daphnia pulex, we found four ZADs. This low number increases to 29 in the most basal holometabolous insect Apis mellifera (Savard et al. 2006) and further increases to 75 in the coleopteran species Tribolium castaneum, 86 in the lepidopteran species Bombyx mori, 98 in the dipteran species Drosophila melanogaster and peaks with 147 in the second dipteran species Anopheles gambiae. These numbers suggest that the expansion occurred predominantly in insects, after the divergence of the crustacean and insect lineage, since a more basal expansion can only be explained by a subsequent massive gene loss in Daphnia pulex.

In order to infer the characteristics of the LSE in higher insects, we applied the logic described above. Thus, it was necessary to assign potential orthologous and species-specific paralogous groups. The classical approach for identifying orthologs as well as paralogs involves phylogenetic analysis and a procedure referred to as tree reconciliation (e.g. Page and Charleston 1997). This approach tries to relate the topology of a gene tree to a chosen species tree employing the parsimony principle, i.e. a minimal number of duplications and gene losses in the evolution of the gene tree. Thus, it appears as if this approach is the method of choice for our task. However, we observed that gene trees built from a multiple sequence alignment of all ZADs are very unreliable in most portions of the tree suggesting that they contain many uncertainties and artifacts. We concluded that using such an unreliable gene tree as input for tree reconciliation could lead to many artifacts, which in turn would render an interpretation of the results difficult. Therefore, we developed an alternative approach that enables us to focus on the reliable parts of the gene tree. Briefly, we searched for ZAD peptide sequences that in the case of
ZAD-coding genes present before the speciation events (see main text for details). Note the high number of ZADs in hexapodes as compared to vertebrates and crustaceans. Numbers in black circles denote the minimal number of ZAD-coding genes present before the speciation events (see Materials and Methods for details). ZAD sequence. In both cases we tested whether the similarity is significant (see Materials and Methods for details).

Fig. 2.—Overview of the identified ZAD-peptides. On the left a species tree: on the right the numbers of identified ZAD-coding sequences. Note the high number of ZADs in hexapodes as compared to vertebrates and crustaceans. Numbers in black circles denote the minimal number of ZAD-coding genes present before the speciation events (see main text for details; hsap: Homo sapiens; ggal: Gallus gallus; xlae: Xenopus laevis; dder Drosophila melanogaster; dmel: Drosophila melanogaster; dmel: Drosophila melanogaster; agam: Anopheles gambiae.

Species-specific paralogous groups were more similar to each other than to any other ZAD sequence of another species, or, in the case of orthologous groups, to any other ZAD sequence. In both cases we tested whether the similarity is significant (see Materials and Methods for details).

We obtained a total of 27 species-specific paralogous groups under these rigorous criteria (Table 1). Seven groups including 40 out of 98 ZAD sequences were found in Drosophila melanogaster, eight groups containing 84 of the 147 genes in Anopheles gambiae, five in Bombyx mori containing 33 of the 86 genes, seven groups in Tribolium castaneum with 25 of the 75 genes, and, finally, no group at all in Apis mellifera and Daphnia pulex. Each group contained between two and 51 members (Table 1). Thus, with the exception of Apis mellifera and Daphnia pulex, between one third and more than half of the ZADs can be assigned to species-specific paralogous groups. This finding indicates that many ZAD-coding genes of the holometabolous insects, possibly excluding Hymenoptera, have been recently generated, or to be more explicit are the result of LSE. The result is consistent with the observation that many members of the paralogous groups in Drosophila melanogaster can be found in neighboring positions in the genome (data not shown).

In order to further test this conclusion, we focused on the orthologs. We identified a total of 15 putative orthologous groups. As shown in Figure 3, ZADs of Drosophila melanogaster (Figure 3 B – I) and Anopheles gambiae (Figure 3 A - C and E – I) were found in eight of these orthologous groups, Bombyx mori ZADs in seven groups (Figure 3 B, C E, F and J – L), Tribolium castaneum ZADs in nine (Figure 3 A – D, J, K and M – O), Apis mellifera ZADs in ten (Figure 3 A – D and J – O) and finally a single Daphnia pulex ZAD in one group (Figure 3 A). None of the 15 orthologous groups contained sequences from all six species. But two groups contained sequences from five holometabolous insects (Figure 3 B and C) and one group also contained sequences of Daphnia pulex, lacking sequences of Bombyx mori and Drosophila melanogaster (Figure 3 A).

We used the parsimony principle to infer the number of ZAD-coding genes present before the speciation events that led to the recent species. The results of this analysis is summarized in Figure 2. They indicate that at least one ZAD-coding gene was present in the last common ancestor of Daphnia pulex and the insect species. There were at least ten ZAD-coding genes present in the last common ancestor of Apis mellifera and the other holometabolous insects. This number stays the same in the last common ancestor of Tribolium castaneum and the remaining holometabolous insects; there is one instance where we find conservation between Apis mellifera and Bombyx mori, but fail to identify an ortholog in Tribolium castaneum (Figure 3 L). The last common ancestor of Bombyx mori and the dipteran insects had at least nine ZAD-coding genes, with three losses (Figure 3 M – O) and two gains (Figure 3 E and F). Finally, we find that at least nine ZAD-coding genes were present prior to the divergence of Anopheles gambiae and Drosophila melanogaster, with three losses (Figure 3 J – L) and three gains (Figure 3 G – I).

Thus, we conclude that a first burst of expansion occurred after the split of the crustacean and insect lineages, before the divergence of Hymenoptera and the other holometabolous insects. Thereafter, only few ZAD-coding genes have been fixed in evolution. Hence, our results indicate that only few ZAD-coding genes have been fixed in the early evolutionary history of the analyzed species. This in turn suggests that the majority of the holometabolous ZADs was recently generated long after the speciation events that formed the major holometabolous insect orders (and families, in the case of the two dipteran species), which is consistent with our observation that many ZADs can be found in species-specific paralogous groups (see above). This interpretation appears not to be true for Apis mellifera, in which we find that one third of the genes are conserved in evolution and fail to identify any species-specific paralogous group. Here, it is more parsimonious to assume that the majority of the Apis mellifera ZAD-coding genes was fixed very early after the split from the other holometabolous insects, such that we cannot find any significant sequence similarity between potential species-specific paralogous proteins. In support of this hypothesis, we find that 20 of the 29 Apis mellifera ZAD-genes have orthologs in a second hymenopteran species, a parasitic wasp Nasonia vitripennis (data not shown).

In Drosophila melanogaster nine ZAD coding genes possess essential functions, namely grauzone (grau; Schipbach and Wieschaus 1989), Serendipity-δ (Sry-δ; Payre et al. 1990), deformed wings (dvg; Fahmy and Fahmy 1959), pita (pita; Laundrie et al. 2003), weckle (wek; Luschnig et al. 2004), hangover (hang; Scholz, Franz, and Heberlein...
2005), phyllopod (phyt; Chang et al. 1995; Dickson et al. 1995), determiner of breaking down of Ci activator (debra; Dai, Akimaru, and Ishii 2003) and pois au dos (pad; Gibert et al. 2005). Notably, we find that only three of these genes are conserved in evolution: the gene pita in all five holometabolous insects examined (Figure 3 C); the gene hang in Apis mellifera and Tribolium castaneum (Figure 3 D); the gene pad in Anopheles gambiae (Figure 3 G). The remaining six genes with essential functions appear to be specific for Drosophila melanogaster. It is possible that we failed to identify orthologs for these genes for several reasons, which include the possibilities that the genome sequences may still contain gaps and uncertainties and/or that the orthologous sequences have diverged so much such that we cannot identify them by means of sequence similarity. Genome sequence quality is certainly an issue. However, it appears to be rather unlikely that gaps and uncertainties in the genome sequences consistently involved loci containing orthologs of these genes in all five species. Fast sequence divergence clearly limits our ability to identify orthologs. The fast decline of sequence similarity implies that the genes in question underwent a period of relaxed selective pressure, which is not compatible with the assumption that they fulfilled essential functions in, for example, the last common ancestor of Drosophila melanogaster and Anopheles gambiae. It is much more likely that they acquired these functions in the lineage leading to Drosophila melanogaster after the split between the two dipteran flies by

### Table 1

Paralogous Groups of ZADs

<table>
<thead>
<tr>
<th>group number</th>
<th>number of members</th>
<th>member IDS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Drosophila melanogaster</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>CG18555, CG3485, CG15436, CG7204, CG17612, CG31782a, CG31782b, CG2678, CG4318</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>wek, CG17568, CG10366, CG6254</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>CG15073, gran</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>CG10274, CG10270, CG10269, CG7386</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>CG1145, CG11762, CG1819, CG7979, CG9779, CG31388, CG18764, CG17806, CG17802, CG7357,</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>CG1792, CG6689, CG6813, CG31441, CG17803</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>CG12219, CG2889, CG11695, CG11696</td>
</tr>
<tr>
<td><strong>Σ</strong></td>
<td><strong>40</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Anopheles gambiae</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16</td>
<td>003_agam, 004_agam, 005_agam, 006_agam, 122_agam, 173_agam, 172_agam, 121_agam, 133_agam,</td>
</tr>
<tr>
<td>2</td>
<td>51</td>
<td>134_agam, 135_agam, 137_agam, 138_agam, 085_agam, 139_agam, 086_agam</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>008_agam, 009_agam, 010_agam, 011_agam, 012_agam, 013_agam, 014_agam, 015_agam, 165_agam,</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>016_agam, 017_agam, 166_agam, 018_agam, 019_agam, 020_agam, 058_agam, 080_agam, 081_agam,</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>118_agam, 145_agam, 146_agam, 147_agam, 148_agam, 149_agam, 150_agam, 151_agam, 157_agam,</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>167_agam, 158_agam, 159_agam, 160_agam, 161_agam, 162_agam, 171_agam, 142_agam, 032_agam,</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>031_agam, 070_agam, 079_agam, 143_agam, 055_agam, 061_agam, 125_agam, 156_agam, 075_agam,</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>076_agam, 126_agam, 123_agam, 041_agam, 066_agam, 102_agam</td>
</tr>
<tr>
<td><strong>Σ</strong></td>
<td><strong>84</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Bombyx mori</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16</td>
<td>001_bmor, 002_bmor, 016_bmor, 025_bmor, 026_bmor, 030_bmor, 031_bmor, 041_bmor, 042_bmor,</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>005_bmor, 043_bmor, 008_bmor, 036_bmor, 089_bmor, 090_bmor, 093_bmor</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>006_bmor, 007_bmor</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>017_bmor, 054_bmor, 082_bmor, 084_bmor, 057_bmor, 018_bmor</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>085_bmor, 044_bmor</td>
</tr>
<tr>
<td><strong>Σ</strong></td>
<td><strong>33</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Tribolium castaneum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>001_tcas, 040_tcas, 038_tcas, 041_tcas, 044_tcas, 045_tcas, 048_tcas, 039_tcas</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>009_tcas, 010_tcas, 074_tcas, 075_tcas</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>013_tcas, 014_tcas, 054_tcas</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>015_tcas, 018_tcas, 026_tcas</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>020_tcas, 025_tcas</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>056_tcas, 057_tcas</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>069_tcas, 070_tcas, 071_tcas</td>
</tr>
<tr>
<td><strong>Σ</strong></td>
<td><strong>25</strong></td>
<td></td>
</tr>
</tbody>
</table>
means of positive selection. We conclude that it is possible that the six non-conserved essential ZAD-coding genes of Drosophila melanogaster carry out functions that are specific for the lineage leading to Drosophila melanogaster. This in turn implies that some of the ZAD-coding genes are likely to carry lineage-specific or even species-specific functions.

In summary, the results show that many ZAD-coding genes have been recently generated by LSE in four out of five holometabolous insects. Only the most basal holometabolous insect order, Hymenoptera (Savard et al. 2006), showed no evidence for such a recent burst of expansions of ZAD-coding genes. In general, it appears that ZAD-coding genes are prone to duplicate, suggesting that there is positive selection for duplication of ZAD-coding genes. Furthermore, the failure to identify orthologs for the members of the paralogous groups suggests that they have diversified by means of positive selection, which in turn implies that they have acquired novel functions. It is possible that these functions contributed to the establishment of novel processes, leading to novel phenotypic traits, or substituted for the function of other proteins in conserved processes. We note, however, that more neutral mechanisms as outlined in the model of orphan genes in Drosophila (Domazet-Loso and Tautz 2003) or the common neutral mechanisms of subfunctionalization after gene duplication (Force et al. 1999) could result in an inability to detect the orthologous relationship.

ZFPs in general seem to be prone to LSE (see for example Lander et al. 2001; Chung et al. 2002; Englbrecht, Schoof, and Böhm 2004). LSEs of ZFPs are especially often found in proteins that also contain additional domains. In arthropods these additional domains include the ZAD (Chung et al. 2002) and the BTB domain (Lander et al. 2001), while in vertebrates these include in particular the SCAN and KRAB domains (Collins, Stone, and Williams 2001; Lander et al. 2001; Huntley et al. 2006). All four domains are thought to be protein-protein interaction domains, suggesting that a combination of C2H2 zinc fingers with an additional protein-protein interaction domain represents a versatile platform, which can be used to adopt novel functionalities. These may include the recruitment to the regulation of target genes or entirely different functions, as exemplified by wet that functions as an adaptor, binding to the Toll receptor at the plasma membrane (Chen et al. 2006).

The KRAB domain defines the largest group of ZFPs in vertebrates. The LSE of KRAB-ZFPs appears to be restricted to the tetrapod vertebrates (Urrutia 2003), although the domain itself can be traced back to the base of the deuterostomes (Birtle and Ponting 2006). If we compare that to the ZAD, we find striking similarities, i.e. the LSE of the ZAD is restricted to the higher holometabolous insects, but the domain itself was probably present in the last common ancestor of deuterostomes and protostomes. Although present in vertebrates, we find evidence for only one instance of the ZAD in the protein encoded by ZFP276 (see above), suggesting that in vertebrates the ZAD-ZFPs are not prone to duplication as has been inferred for the higher holometabolous insects (Chung et al. 2002 and this study). Thus, the differential expansion of ZAD-ZFPs in higher holometabolous insects and KRAB-ZFPs in tetrapod vertebrates may reflect distinct evolutionary constraints and demands that are specific for the lineages.

Drosophila melanogaster ZADs are expressed in the female germline

It is possible that lineage-specific functions of ZAD-coding genes are required in processes that evolved in response to changing ecological conditions. Alternatively, ZAD-coding genes may have been involved in processes
that led to developmental changes. Based on the observation that four of the nine known ZAD-coding genes of *Drosophila melanogaster* are expressed in the female germ-line, we reasoned that ZADs may be involved in either oogenesis and/or are maternally required during early embryogenesis. The functions of the four aforementioned genes are consistent with this view. *grau* encodes a transcription factor that is required for the completion of meiosis (Chen et al. 2000; Harms et al. 2000) and *pita* is involved in the formation of egg-chambers (Laundrie et al. 2003). The protein encoded by *Sry-δ* activates the expression of the anterior determinant *bicoid* (Payre, Crozatier, and Vincent 1994), while *wek* function is required for the establishment of the dorso-ventral axis (Luschnig et al. 2004; Chen et al. 2006).

In order to test whether female germline expression is common among ZAD-coding genes, we examined two independent microarray datasets (Manak et al. 2006; Hooper et al. 2007) that contain gene expression time series of developing *Drosophila melanogaster* embryos. We find that of the 98 ZAD-coding genes, 46 genes are maternally expressed. These 46 genes include 5 of the 9 previously described *Drosophila melanogaster* ZAD-coding genes *wek*, *pita*, *hang*, *pad* and *dbr*. Though *grau* and *Sry-δ* are not included in this dataset, maternal expression was previously observed. This indicates that the 46 genes identified in the dataset represent a minimum estimate of maternally expressed ZAD-coding genes.

Furthermore, we examined the microarray datasets if the eight *Drosophila melanogaster* genes, which had at least one ortholog in other species, were expressed maternally. Significantly, we find that this is the case for seven of these eight ZAD-coding genes. The single conserved ZAD-coding gene that appeared to have no maternal expression corresponds to *CG31109*. *CG31109* encodes a protein with an isolated ZAD, lacking additional C2H2 zinc finger domains. This gene is conserved in all five holometabolous insects examined. Moreover, we were able to identify potential orthologs in the more ancestral, hemimetabolous orthopteran species *Laupala kohalensis*, suggesting that it was present in the last common ancestor of Orthoptera and Holometabola (Figure 4 A).

Given the high-throughput nature of the microarray technology, it was likely that some of the maternally expressed genes were missed. For this reason we examined whether *CG31109* is maternally expressed by *in situ* hybridization of RNA probes to whole mounted *Drosophila melanogaster* ovaries (see Materials and Methods). Figure 4 B shows the expression pattern of *CG31109*. It was observed that maternal *CG31109* expression starts during stage 8 of oogenesis in the nurse cells and that the transcript accumulates evenly in the oocyte. Thus, all eight conserved *Drosophila melanogaster* ZAD-coding genes are expressed in the female germline. If we add *CG31109* to the list of maternally expressed 48 ZAD-coding *Drosophila melanogaster* genes, we conclude that
at least 50% of the 98 ZAD-coding genes found in *Drosophila melanogaster* are maternally expressed. Thus, it appears that female germline expression of ZADs is a rather common phenomenon. Therefore, we speculate that many ZAD-coding genes are involved either in some aspects of oogenesis and/or contribute to the maternal effect on early embryonic development. The observation that all conserved *Drosophila melanogaster* ZADs are expressed in the female germline indicates that maternal expression of ZAD genes is an ancestral property of ZAD-coding genes. This hypothesis suggests that ancestral ZAD-coding genes might have been involved in either oogenesis and/or early embryogenesis.

An ancestral function of ZADs in holometabolous insects

We found that the eight conserved *Drosophila melanogaster* ZAD-coding genes are expressed in the female germline. Two of them, namely *pita* and *CG31109*, are present in all holometabolous insects examined. Currently, only *pita* has been studied, while the function of *CG31109* remains unknown. Orthologs of *pita* can be traced back to *Apis mellifera* (see above). Given that our orthology assignment is based only on the ZAD peptide sequence, we tried to extend the protein sequence to the C2H2 zinc fingers. We identified the full-length sequence of *pita* for four species examined with the exception of *Bombyx mori*, in the NCBI database. A multiple sequence alignment of these sequences showed that in addition to the ZAD sequences, a region including the C2H2 zinc fingers is also conserved (see Supplementary Figure S1), suggesting that these sequences are indeed orthologous. Thus, we conclude that *pita* was present in the last common ancestor of Hymenoptera and the other holometabolous insects.

*pita* is required during oogenesis, as mutations assayed in germline clones lead to defects in the development of egg chambers. The observed defects include degeneration of the egg chambers and abnormal or absent nurse cell nuclei (Laundrie et al. 2003). Apart from its role during oogenesis, it has been found that *pita* function is generally required in proliferating tissues as well as cells undergoing endoreplication during S-phase. There it acts as a sequence-specific transcription factor that activates the expression of target genes, including the *Orc4* gene, which is essential for initiation of DNA replication (Page et al. 2005). Thus, the oogenesis defects seen in *pita* mutant germline clones can be attributed to a failure of the division cycles in the germarium that generate the germ cell cysts and/or a failure of the endoreplication cycles during nurse cell differentiation (Laundrie et al. 2003; Page et al. 2005).

Both, formation of germ cell cysts and endoreplication of nurse cells are characteristics of the meroistic type of oogenesis. Meroism evolved independently in several groups of insects, but it appears to be of monophyletic origin in the holometabolous and the paraneopteran insects, as has been inferred from several common features such as formation of branched germ cell cysts (Bünig 1996). In the meroistic ovary, only one cell of the cyst becomes the oocyte while the remaining cells differentiate to nurse cells, whereas the more basal panoistic ovaries lack nurse cells. The nurse cells produce most of the components that are deposited in the oocyte. Nurse cells become polyploid by means of endoreplication in order to cope with the high metabolic burden (Bünig 1996).

Given the conservation of *pita* in all holometabolous insects we speculate that *pita* has acquired its function during oogenesis before the radiation of holometabolous insects. In this view, the results suggest that *pita* may have been involved in the establishment of a novel phenotypic trait, the meroistic ovary. It is also possible that a more general function of *pita* involving DNA replication was established first and its specific role in oogenesis followed after the divergence of the major orders of the holometabolous insects.

Collectively, we have provided evidence that the ZAD-coding genes are subject to an ongoing LSE, which is most pronounced in the higher holometabolous insects. The LSE of these genes can be explained by the versatile functions that are adopted by the ZAD-containing proteins. ZAD-coding genes in general appear to be involved in developmental processes, suggesting that the frequent duplications of ZAD-coding genes provide the raw material to evolve functionalities that are employed during development. This in turn implies that ZAD-coding genes may have been involved in the establishment of novel morphological characters, such as the meroistic ovary. The relationship between ZAD genes and meroistic ovary development could be experimentally verified once the genomic data from paraneopteran insects, which have also meroistic ovaries, are available.

### Materials and Methods

#### Genomic Sequence Data

We used the set of all predicted protein sequences of *Drosophila melanogaster* Release 3.2 (Celniker et al. 2002) and extracted all ZAD coding sequences using hmmsearch of the HMMer package (Eddy 1998) with the ZAD profile hidden markov model described in Chung et al. 2002. We downloaded the genome sequences of *Anopheles gambiae* from Ensembl (Assembly AgamP3; Holt et al. 2002; Mongin et al. 2004), *Bombyx mori* from NCBI (Accession numbers BAAB01000001-BAAB01213289; Mita et al. 2004 and Accession numbers AADK0100001-AADK01066482; Xia et al. 2004), *Apis mellifera* (Version 4.0) from http://www.hgsc.bcm.tmc.edu/projects/honeybee/, *Tribolium castaneum* (Version 2.0) from http://www.hgsc.bcm.tmc.edu/projects/tribolium/, *Daphnia pulex* from http://wfleabase.org/prerelease and *Nasonia vitripennis* from http://www.hgsc.bcm.tmc.edu/projects/nasonia/.

In order to identify ZAD coding sequences we employed tblastn of the BLAST package (Altschul et al. 1997) using the Drosophila melanogaster ZAD peptide sequences as query. All positive contigs were further analyzed with geneiewise of the Wise2 package (Binney, Thompson, and Gibson 1996) using the ZAD profile hidden markov model. All ZAD peptide sequences were extracted and manually inspected, i.e., we discarded sequences with STOP codons, partial and nearly identical sequences and introduced introns if necessary. All identified ZAD peptide sequences are reported in the Supplementary File S2.
Detection of Orthologs and Paralogs

In order to detect orthologous and paralogous groups of proteins we extended the Inparanoid algorithm (Remm, Storm, and Sonnhammer 2001). We constructed a distance matrix for all ZAD peptide sequences of all six examined species plus the ZAD found in the human ortholog of ZFP276. The construction of the distance matrix involved the following steps: (i) an all-against-all run of search of the FASTA package (fasta34 series; Pearson and Lipman 1988) using the parameters “-n # -E 1e100 -m 9 -H -s BL80”; (ii) the reported bit scores were converted to raw species plus the ZAD found in the human ortholog of proteins we extended the Inparanoid algorithm (Remm, D.

\[ D(i,j) = \frac{\ln(2) + \ln(K) \mid D(i,j) \mid}{\lambda}, \] with \( K = 0.071 \) and \( \lambda = 0.299; \) (iii) the raw scores of a match between sequences \( i \) and \( j \) were converted to symmetrical relative similarity scores by \( S_{raw} = \frac{D_{raw}(i,j)}{D_{raw}(j,i)} = 0.5 \frac{[S_{raw}(i,j) - S_{rand}]}{[S_{raw}(i,j) - S_{rand}]} \) where \( D_{raw}(j,i) \) with \( S_{rand} = \frac{1}{\lambda} \frac{\ln(\kappa \cdot L(i) \cdot L(j))}{\ln(1.0e5)} \) otherwise.

Next, we searched for every sequence \( i \) of species \( A \) the most similar sequence \( k \) of species \( X \neq A \). This sequence corresponded to a potential ortholog of sequence \( i \). We then searched for sequence \( j \) of species \( B = A \) whose distance \( D(i,j) \) was smaller than the distance \( D(i,k) \) to the potential ortholog. For every sequence \( j \) that fulfilled this criterion we conducted a statistical test whether the sequences \( i \) and \( j \) were more closely related than to sequence \( k \). We employed the statistical test proposed by Nei, Stephens, and Saitou 1985 in UPGMA trees that tests whether a branch that separates sequences \( i \) and \( j \) from sequence \( k \) is significantly longer than zero. Every sequence \( j \) that passed this test was assigned to be a species-specific paralog. The so assigned groups were combined using single linkage clustering, i.e. groups that contained at least one overlap were merged. The members of the resulting clusters were then reported to be potential species-specific paralogs.

In order to assign orthologous groups we employed a similar scheme as outlined above. We merged all members of a paralogous group to form a single entry and we added all singleton sequences. Thus, every sequence \( i \) was represented only once in the new gene list, either as member of a paralogous group or as a singleton. For each entry \( i \) of species \( A \) we searched for the closest entry \( j \) of species \( X \neq A \). There were now four possible scenarios: (i) entry \( i \) is a singleton and entry \( j \) is a singleton (one-to-one); (ii) entry \( i \) is a singleton and entry \( j \) is a paralogous group (one-to-many); (iii) entry \( i \) is a paralogous group and entry \( j \) is a singleton (many-to-one); and (iv) entry \( i \) is a paralogous group and entry \( j \) is a paralogous group (many-to-many). As a distance measure between the entries \( i \) and \( j \) we used the average distance between all members of entry \( i \) and all members of entry \( j \). In order to assign putative orthologs, we required that entries \( i \) and \( j \) form symmetrical best hits. For each pair \( i \) of species \( A \) and \( j \) of species \( B \) we searched for the closest sequence (of the original set) \( k \) of either species \( A \) or \( B \) whose distance was greater than the (average) distance between entries \( i \) and \( j \) (this was necessary in order to exclude putative species-specific paralogs that did not pass the test described above). We tested whether the sequences in entry \( i \) were more closely related to the sequences in entry \( j \) than to the sequence \( k \) using the test described above. Every entry \( j \) that passed this test was assigned to be a potential ortholog of entry \( i \). The so-derived groups were combined if they contained at least one common member (single linkage clustering). The members of the resulting clusters were reported to be orthologous groups.

Making of DIG labeled RNA in situ Probes

Drosophila genomic DNA was isolated from females using standard methods. The longest exons of ZAD-coding genes were isolated by PCR on genomic DNA using Taq polymerase (Fermentas) and standard methods. Primers (MWG) used to amplify the longest exons of CG31109 (forward: 5’- gcggattcagctcctgatc-3’; reverse: 5’- gctgtagatgctcagcatcag-3’; fragment size 300 bp) contained a 5’ EcoRI and a 3’ XbaI site. Addition of cloning sites into the PCRragments via the primers allowed directional cloning into the Bluescript SK vector. DIG-labeled in situ probes were made following standard protocols, using T3 (Fermentas) to make antisense probes and T7 (Fermentas) to make sense probes.

in situ Hybridizations on Drosophila Ovaries

Ovaries were dissected from 2-day-old females in Grace’s medium and placed on ice. Ovaries were fixed in 5% formaldehyde (Sigma) in PBSTX (PBS + 0.1% Triton-X 100 (Sigma)) for 20 min at RT. Ovaries were rinsed 3 times and washed 3 times for 5 min in PBSTX. Proteinase K digestion (final concentration: 5μg/μl in PBSTX) was conducted at RT for 10 min. Ovaries were rinsed 3 times with PBSTX and post-fixed with 5% formaldehyde in PBSTX for 20 min at RT. For pre-hybridization ovaries were first incubated in 1:1 PBSTX/HybBTX (5x SSC; 50% formamide (Sigma); 25 mg/ml torula yeast RNA (Sigma); 0.1% Triton-X 100) for 20 min at RT. The solution was substituted for HybBTX and ovaries were placed at 57 °C for at least one hour. In this time HybBTX was replaced at least three times. Probes were diluted 1:100 in HybBTX, placed at 80°C for 2 min and cooled on ice. HybBTX was removed from the ovaries and replaced with the probe mix. Hybridization was conducted at 57°C overnight. Probe was removed and ovaries were washed 3 times for 20 min with HybBTX at 55°C and once with 1:1 PBSTX/HybBTX at RT. Ovaries were rinsed 3 times and washed 3 times for 20 min with PBSTX at RT. Ovaries were incubated in a 1:1000 dilution of an alkaline phosphatase coupled sheep anti-DIG antibody (Roche) for 1 h at RT. Ovaries were rinsed 3 times and washed 3 times for 20 min with PBSTX. Ovaries were rinsed three times in staining buffer (0.1 M TRIS pH 9.5; 50 mM MgCl2; 0.1 M NaCl; 0.1% Triton-X 100), 10 μl NBT/BCIP solution (Roche) was added to ovaries in 1 ml staining buffer. Staining progress was monitored under a bifocal microscope. Staining was stopped by rinsing three times and washing several times with PBSTX. Ovaries were completely dehydrated in an ethanol dilution series and mounted in Canada Balsam (Sigma). Photographs were taken using a Zeiss ariophot microscope with
1942

Nomarski optics and a ProgRES 3012 camera using ProgRes 4.0 software (Kontron Elektronik). Images where further processed using Photoshop 7.0 (Adobe).

Supplementary Material

Supplementary figure S1 and S2 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank the anonymous reviewers for the helpful comments. We thank our colleagues, especially Ralf Pflanz, for help and critical discussions. H.-R.C. thanks the Boehringer Ingelheim Fonds for a predoctoral fellowship. Work was supported by the Max-Planck-Gesellschaft.

Literature Cited


