EMBO
reports

scientific report

Genomic expansion and clustering of ZAD-containing C2H2 zinc-finger genes in *Drosophila*

Ho-Ryun Chung+, Ulrich Schäfer, Herbert Jäckle & Siegfried Böhm[1,+]

Abteilung Molekulare Entwicklungsbiologie, Max-Planck-Institut für biophysikalische Chemie, Am Fassberg, D-37077 Göttingen and [1]Max Delbrück-Centrum für Molekulare Medizin, Department of Genetics, Bioinformatics and Structural Biology, Robert-Rössle-Strasse 10, D-13125 Berlin-Buch, Germany

Received August 8, 2002; revised September 27, 2002; accepted October 18, 2002

C2H2 zinc-finger proteins (ZFPs) constitute the largest family of nucleic acid binding factors in higher eukaryotes. *In silico* analysis identified a total of 326 putative ZFP genes in the *Drosophila* genome, corresponding to ~2.3% of the annotated genes. Approximately 29% of the *Drosophila* ZFPs are evolutionary conserved in humans and/or *Caenorhabditis elegans*. In addition, ~28% of the ZFPs contain an N-terminal zinc-finger-associated C4DM domain (ZAD) consisting of ~75 amino acid residues. The ZAD is restricted to ZFPs of dipteran and closely related insects. The evolutionary restriction, an expansion of ZAD-containing ZFP genes in the *Drosophila* genome and their clustering at few chromosomal sites are features reminiscent of vertebrate KRAB-ZFPs. ZADs are likely to represent protein–protein interaction domains. We propose that ZAD-containing ZFP genes participate in transcriptional regulation either directly or through site-specific modification and/or regulation of chromatin.

INTRODUCTION

C2H2 zinc-finger (ZF) motifs, which represent the most abundant nucleic acid binding motif in higher eukaryotes (Rubin *et al.*, 2000; Lander *et al.*, 2001; Venter *et al.*, 2001), are found in RNA-binding proteins (Joho *et al.*, 1990), transcription factors (Rosenberg *et al.*, 1986; Stanojevic *et al.*, 1989) and chromatin components (Reuter *et al.*, 1990). Lineage-specific subgroups of ZF proteins (ZFPs) can be found in the genomes of *Saccharomyces cerevisiae* (Böhm *et al.*, 1997), *Arabidopsis thaliana* (Riechmann *et al.*, 2000), *Caenorhabditis elegans* (Chervitz *et al.*, 1998), *Drosophila melanogaster* (Rubin *et al.*, 2000) and *Homo sapiens* (Lander *et al.*, 2001; Venter *et al.*, 2001) and are especially expanded in the higher eukaryotic species. In humans, this expansion includes ZFPs that contain the evolutionarily conserved BTB/POZ domains or SCAN and KRAB domains (reviewed by Collins *et al.*, 2001), which are restricted to vertebrates (Lander *et al.*, 2001). No corresponding expansion of ZFPs has been observed in the *C. elegans* genome. In *Drosophila*, as in humans, ZFPs were found to be associated with BTB/POZ domains and with a recently identified, but uncharacterized, C4DM domain (Lander *et al.*, 2001; Lespinet *et al.*, 2002). Here, we report a detailed *in silico* analysis of ZFPs in the *Drosophila* genome, showing that the C4DM domain is an N-terminal protein structure that is almost exclusively found in association with ZFPs. This ZF-associated C4DM domain (ZAD) characterizes the single largest subfamily of mostly clustered *Drosophila* ZFP genes and appears to be restricted to dipteran and closely related insect genomes.

RESULTS AND DISCUSSION

Characterization of ZFPs in the *Drosophila* genome

We identified a total of 326 C2H2 ZFP genes in the genome of *Drosophila*. This estimate differs from the previously published numbers of, for example, 352 (Rubin *et al.*, 2000) or 234 (Venter *et al.*, 2001). We propose that our estimate provides the most accurate assessment yet, as we did not rely solely on *in silico* methods but also performed a manual inspection of all identified ZF motifs (see Methods). Of all putative *Drosophila* ZFPs, 94 (~29%) are conserved in humans and/or *C. elegans,* an assignment based on the arrangement and sequence of the ZFs as well as sequence similarities outside the ZF domains of the proteins. The remaining 232 ZFPs appear to be *Drosophila*-specific or restricted to the insect lineage. The identified *Drosophila* ZFP genes and their chromosomal distribution are summarized in Table I (see also Supplementary data available at *EMBO reports* Online).

+Corresponding authors. H.-R.C. Tel: +49 551 201 1505; Fax: +49 551 201 1755. S.B. Tel: +49 30 94062478; Fax: +49 30 94062548.

Table I. Overview, conserved and unique *Drosophila* ZFPs and distribution on the four chromosomes

Chromosome, -arm	ZFPs	Conserved ZFPs	Unique ZFPs	ZAD-ZFPs	BTB-ZFPs
X	52	14	38	14	3
2L	60	22	38	12	3
2R	54	13	41	11	2
3L	59	22	37	10	2
3R	96	20	76	44	3
4	4	3	1	0	0
Unassigned	1	0	1	0	0
Σ	326	94	232	91	13

In order to place the 232 lineage-specific ZFPs into subgroups, we probed for associated protein motifs. We found 13 ZFPs (~4%) containing a BTB/POZ domain, a combination that has also been observed in human ZFPs (Lander *et al.*, 2001). In 91 ZFPs (~28%; Table I), we identified an N-terminal domain of >70 amino acids. This domain, which defines the single largest subfamily of *Drosophila* ZFPs, has recently been noted as a C4DM domain (Lander *et al.*, 2001; Lespinet *et al.*, 2002). In all but two cases, this domain is always ZFP-associated. The coding sequence of one of the two ZADs that are not associated with a ZFP coding region is found immediately upstream of a ZAD-ZFP-encoding gene (CG4639) and was not previously annotated. Thus, it is possible that this ZAD is included in an as-yet-unidentified splice variant of CG4639. The second ZFP-unrelated ZAD encoded by CG11371 is highly diverged and is part of a protein lacking any other significant protein domain or motif. For simplicity and to demonstrate the association, we refer to this motif as ZAD.

At present, mutant alleles have only identified for three ZAD-containing ZFP genes. These are *deformed wing/zeste-white5* (*dwg/zw5*; Fahmy and Fahmy, 1959), *grauzone* (Schupbach and Wieschaus, 1989) and *Serendipity* δ (Payre *et al.*, 1990). The functional characterization of these genes, as well as the results of biochemical studies, suggests that ZAD-containing ZFPs are involved in transcriptional control. *dwg/zw5* encodes a site-specific DNA-binding ZFP that promotes the formation of insulator complexes (Gaszner *et al.*, 1999), whereas *grauzone* and *Serendipity* δ encode transcription factors implicated in the activation of the genes *cortex* (Chen *et al.*, 2000; Harms *et al.*, 2000) and *bicoid* (Payre *et al.*, 1994), respectively. One additional ZAD-containing ZFP, termed DIP1, contains only a single ZF motif and can associate with the NFκB homologue Dorsal (Bhaskar *et al.*, 2000).

ZAD-encoding sequences were found in the available ESTs of the dipterans *Drosophila* spp., *Anopheles gambiae* and *Aedes aegyptii*, the hymenopteran *Apis mellifera* and the lepidopteran *Bombyx mori* (see Supplementary data). In contrast, not a single EST in over 7 million vertebrate samples (see Supplementary data) or in non-insect invertebrate such as *C. elegans* has been identified. These observations suggest that the ZAD is restricted to insects and has emerged during their evolution.

Classification of the ZAD

ZADs vary in length between 71 and 97 amino acid residues. A multiple sequence alignment of a representative subset of 32 ZADs (Figure 1; for a complete alignment of the *Drosophila* ZADs, see Supplementary data) shows that the domain consists of four conserved sequence blocks (blocks 1–4), which are linked by three variable regions (r1–r3) of different lengths (Figure 1). The most striking feature of ZADs is the occurrence of two invariant cysteine pairs in blocks 1 and 4, suggesting that they may coordinate the binding of a zinc ion to stabilize a distinct fold of the domain.

Secondary structure analysis predicts that the variable regions 1–3, which contain preferentially small and polar amino acid residues (Figure 1), represent turns or unstructured spacers, whereas the conserved blocks 1–4 form β1β2α1β3α2-folds (with strong predictions except for β2; see Supplementary data), which are likely to represent the core of the ZAD structure (Figure 1). Within each of the blocks 1–4, most conserved amino acid residues are hydrophobic; the few exceptions include a highly conserved arginine residue (position 4; Figure 1) located between the cysteines of block 1. The importance of this conserved arginine residue, and of the domain itself, is supported by the finding that a point mutation that results in an arginine-to-glycine replacement in the *dwg/zw5* protein causes a lethal phenotype (Gaszner *et al.*, 1999). Furthermore, a point mutation in *Serendipity* δ that results in a tyrosine replacement of the second invariant cysteine of block 1 also causes a lethal phenotype (Crozatier *et al.*, 1992). These observations suggest that the core structure of the ZAD carries an essential function, at least in the case of *Serendipity* δ and *dwg/zw5*. Mutational analysis combined with biochemical studies showed that the ZAD-like domain of *Serendipity* δ functions as a protein–protein interaction domain (Payre *et al.*, 1997), a function that has been proposed for the ZAD of the *dwg/zw5* protein as well (Gaszner *et al.*, 1999). The experimental data therefore support the proposal that ZADs represent or contain protein–protein interaction surfaces that, with the possible exception of two out of 93 cases, are combined with arrays of putative DNA-binding ZFs.

Intron-based classification and clustering of ZAD-bearing ZFP genes

ZAD-containing ZFPs are not randomly distributed throughout the genome. The X chromosome, both arms of the second chromosome and the left arm of the third chromosome each contain between 10 and 14 ZAD-containing ZFP genes, whereas the right arm of the third chromosome contains 44 family members (Table II). Furthermore, nearly half of the ZAD-containing ZFP genes (41 of 91; see Supplementary data) are found in gene clusters (see below).

Based on the intron structure of the primary transcripts, ZADs can be divided into two large subsets. A total of 38 ZADs are encoded by a single exon (subset 1), whereas the open reading frames of 53 ZADs are split by an intron located in a conserved position in block 3 between the β-strand and the α-helix (subset 2). We could further place 45 ZAD-coding sequences into 10 sequence-related subgroups (Table II). Eight of these are distributed in a chromosome-specific manner, and each subgroup consists of either subset 1 or subset 2 ZADs. Another interesting

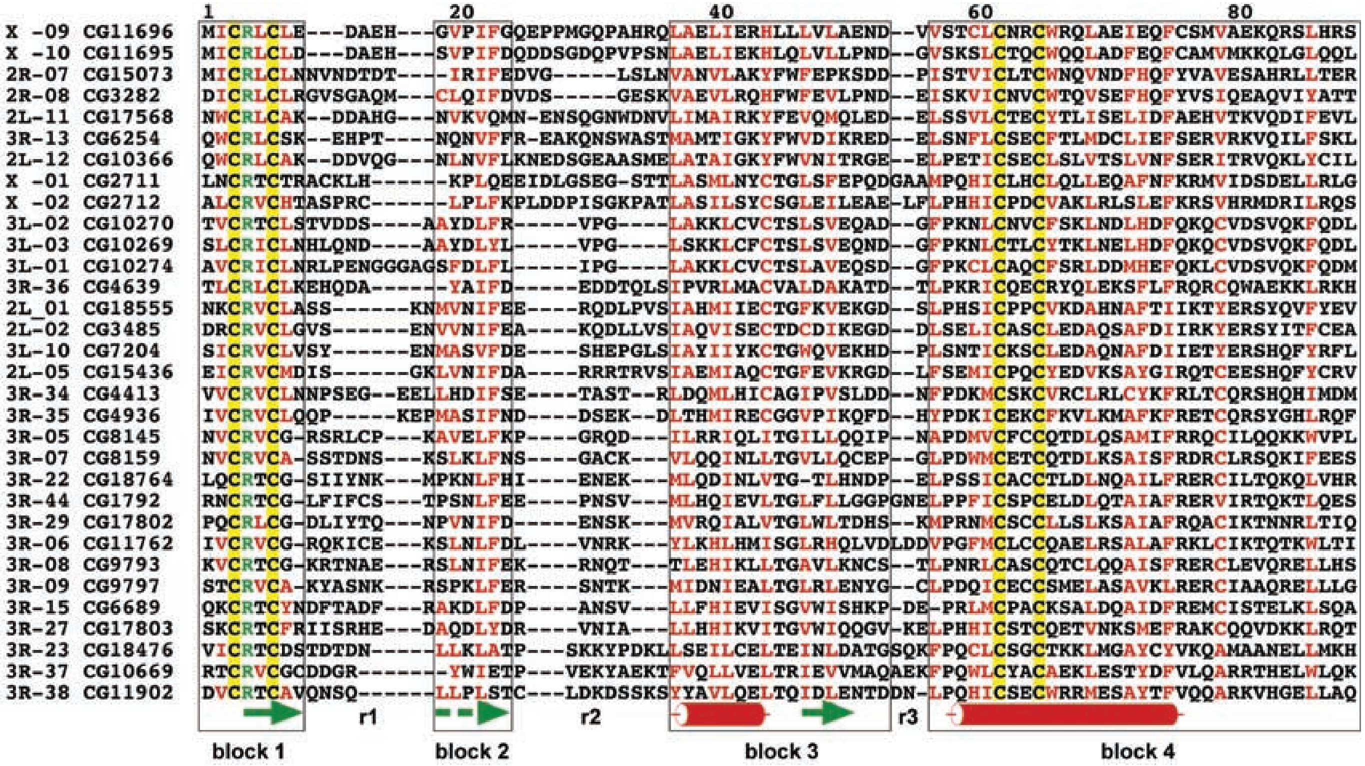
1
20
40
60
80
X -09 CG11696 MICRLCLE---DAEH---GVPIFGQEPPMGQPAHRQLAELIERHLLLVLAEND--VVSTCLCNRCWRQLAEIEQFCSMVAEKQRSLHRS
X -10 CG11695 MICRLCLD---DAEH---SVPIFDQDDSGDQPVPSNLAELIEKHLQLVLLPND--GVSKSLCTQCWQQLADFEQFCAMVMKKQLGLQQL
2R-07 CG15073 MICRLCLNNVNDTDT----IRIFEDVG------LSLNVANVLAKYFWFEPKSDD--PISTVICLTCWNQVNDFHQFYVAVESAHRLLTER
2R-08 CG3282 DICRLCLRGVSGAQM---CLQIFDVDS------GESKVAEVLRQHFWFEVLPND--EISKVICNVCWTQVSEFHQFYVSIQEAQVIYATT
2L-11 CG17568 NWCRLCAK--DDAHG---NVKVQMN-ENSQGNWDNVLIMAIRKYFEVQMQLED--ELSSVLCTECYTLISELIDFAEHVTKVQDIFEVL
3R-13 CG6254 QWCRLCSK--EHPT----NQNVFFR-EAKQNSWASTMAMTIGKYFWVDIKRED--ELSNFLCSECFTLMDCLIEFSERVRKVQILFSKL
2L-12 CG10366 QWCRLCAK--DDVQG---NLNVFLKNEDSGEAASMELATAIGKYFWVNITRGE--ELPETICSECLSLVTSLVNFSERITRVQKLYCIL
X -01 CG2711 LNCRTCTRACKLH------KPLQEEIDLGSEG-STTLASMLNYCTGLSFEPQDGAAMPQHICLHCLQLLEQAFNFKRMVIDSDELLRLG
X -02 CG2712 ALCRVCHTASPRC-------LPLFKPLDDPISGKPATLASILSYCSGLEILEAE-LFLPHHICPDCVAKLRLSLEFKRSVHRMDRILRQS
3L-02 CG10270 TVCRTCLSTVDDS----AAYDLFR------VPG----LAKKLCVCTSLSVEQAD--GFPKNLCNVCFSKLNDLHDFQKQCVDSVQKFQDL
3L-03 CG10269 SLCRICLNHLQND----AAYDLYL------VPG----LSKKLCFCTSLSVEQND--GFPKNLCTLCYTKLNELHDFQKQCVDSVQKFQDL
3L-01 CG10274 AVCRICLNRLPENGGGAGSFDLFL------IPG----LAKKLCVCTSLAVEQSD--GFPKCLCAQCFSRLDDMHEFQKLCVDSVQKFQDM
3R-36 CG4639 TLCRLCLKEHQDA-------YAIFD------EDDTQLSIPVRLMACVALDAKATD--TLPKRICQECRYQLEKSFLFRQRCQWAEKKLRKH
2L_01 CG18555 NKCRVCLASS-------KNMVNIFEE----RQDLPVSIAHMIIECTGFKVEKGD--SLPHSICPPCVKDAHNAFTIIKTYERSYQVFYEV
2L-02 CG3485 DRCRVCLGVS-------ENVVNIFEA----KQDLLVSIAQVISECTDCDIKEGD--DLSELICASCLEDAQSAFDIIRKYERSYITFCEA
3L-10 CG7204 SICRVCLVSY-------ENMASVFDE----SHEPGLSIAYIIYKCTGWQVEKHD--PLSNTICKSCLEDAQNAFDIIETYERSHQFYRFL
2L-05 CG15436 EICRVCMDIS-------GKLVNIFDA----RRRTRVSIAEMIAQCTGFEVKRGD--LFSEMICPQCYEDVKSAYGIRQTCEESHQFYCRV
3R-34 CG4413 VVCRVCLNNPSEG--EELLHDIFSE----TAST--RLDQMLHICAGIPVSLDD--NFPDKMCSKCVRCLRLCYKFRLTCQRSHQHIMDM
3R-35 CG4936 IVCRVCLQQP-----KEPMASIFND----DSEK--DLTHMIRECGGVPIKQFD--HYPDKICEKCFKVLKMAFKFRETCQRSYGHLRQF
3R-05 CG8145 NVCRVCG-RSRLCP---KAVELFKP----GRQD---ILRRIQLITGILLQQIP--NAPDMVCFCCQTDLQSAMIFRRQCILQQKKWVPL
3R-07 CG8159 NVCRVCA-SSTDNS---KSLKLFNS----GACK---VLQQINLLTGVLLQCEP--GLPDWMCETCQTDLKSAISFRDRCLRSQKIFEES
3R-22 CG18764 LQCRTCG-SIIYNK---MPKNLFHI----ENEK---MLQDINLVTG-TLHNDP--ELPSSICACCTLDLNQAILFRERCILTQKQLVHR
3R-44 CG1792 RNCRTCG-LFIFCS---TPSNLFEE----PNSV---MLHQIEVLTGLFLLGGPGNELPPFICSPCELDLQTAIAFRERVIRTQKTLQES
3R-29 CG17802 PQCRLCG-DLIYTQ---NPVNIFD-----ENSK---MVRQIALVTGLWLTDHS--KMPRNMCSCCLLSLKSAIAFRQACIKTNNRLTIQ
3R-06 CG11762 IVCRVCG-RQKICE---KSLNLFDL----VNRK---YLKHLHMISGLRHQLVDLDDVPGFMCLCCQAELRSALAFRKLCIKTQTKWLTI
3R-08 CG9793 KVCRTCG-KRTNAE---RSLNIFEK----RNQT---TLEHIKLLTGAVLKNCS--TLPNRLCASCQTCLQQAISFRERCLEVQRELLHS
3R-09 CG9797 STCRVCA-KYASNK---RSPKLFER----SNTK---MIDNIEALTGLRLENYG--CLPDQICECCSMELASAVKLRERCIAAQRELLLG
3R-15 CG6689 QKCRTCYNDFTADF---RAKDLFDP----ANSV---LLFHIEVISGVWISHKP-DE-PRLMCPACKSALDQAIDFREMCISTELKLSQA
3R-27 CG17803 SKCRTCFRIISRHE---DAQDLYDR----VNIA---LLHHIKVITGVWIQQGV-KELPHHICSTCQETVNKSMEFRAKCQQVDKKLRQT
3R-23 CG18476 VICRTCDSTDTDN-----LLKLATP---SKKYPDKLLSEILCELTEINLDATGSQKFPQCLCSGCTKKLMGAYCYVKQAMAANELLMKH
3R-37 CG10669 RTCRVCGCDDGR-------YWIETP---VEKYAEKTFVQLLVELTRIEVVMAQAEKFPQWLCYACAEKLESTYDFVLQARRTHELWLQK
3R-38 CG11902 DVCRTCAVQNSQ-------LLPLSTC---LDKDSSKSYYAVLQELTQIDLENTDDN-LPQHICSECWRRMESAYTFVQQARKVHGELLAQ
r1
r2
r3
block 1
block 2
block 3
block 4

Fig. 1. Multiple sequence alignment of a representative subset of 32 ZAD-containing ZFPs. Yellow boxes, invariant cysteine pairs; green characters, highly conserved arginine residues; red characters, conserved (>60%) hydrophobic amino acid residues. Blocks 1–4 are framed; r1–r3 denote the variable regions 1–3. Green arrows point to putative β-strands, red cylinders putative α-helical structures. The dashed arrow points to a weakly predicted β-strand (see Supplementary data for an alignment of the 91 *Drosophila* ZADs and a comparison with ZADs identified within the genome of *A. gambiae*).

Table II. Overview, ZAD-containing ZFPs, classification into subgroups and subsets and chromosomal clustering

Chromosome, -arm	ZAD-ZFPs	ZADs with EST(s)	Subset 1 exon	Subset 2 exons	Subgroups subset 1		Subgroups subset 2	
X	14	11	8	6	None		a	4 (2)
2L	12	8	5	7	A	4 (2)	b	3 (–)
2R	11	11	6	5	None		c	2 (–)
3L	10	7	10	0	B	5 (4)	None	
					C	2 (2)		
					A	1		
3R	44	34	9	35	None		d	15 (13)
							e	4 (2)
							f	2 (2)
							g	2 (–)
							b	1
Σ	91	71	38	53		12 (8)		33 (19)

Number in brackets denotes number of members found in cluster.

finding is that members of most subgroups represent clustered genes (27 of 45; Table II) and that their sequence similarity includes not only the ZADs but also the associated array of ZFs.

A comparative tree (see Supplementary data) containing all 91 ZADs of *Drosophila* and 71 newly identified ZAD-containing ZFPs encoded by the *A. gambiae* genome shows that the members of the 10 subgroups occupy neighbouring positions in the tree and that in most cases the ZADs of the two species are located on distinct branches. In only a few instances are direct neighbours in the tree derived from the two species. This indicates that the majority of the ZADs of both species underwent species-specific expansions. In *Drosophila*, these findings suggest that (i) the duplication events occurred after the intron-containing ZADs had separated from those lacking the intron and (ii) the expansion and clustering of the ZAD-containing ZFPs involved multiple local duplication events of the ancestral founder genes.

To examine whether both individual and clustered ZAD-containing ZFP genomic sequences are transcribed, we searched for ESTs corresponding to the individual transcripts (Table II). We found 466 ESTs corresponding to 71 ZAD-coding sequences, implying that the majority of ZAD-containing ZFP genes is transcribed. The remaining 20 ZAD sequences, for which no ESTs could be identified, may either be expressed at very low levels and/or only in a few cells or may represent non-functional pseudogenes.

Speculation

Enrichment of lineage-specific ZAD-containing ZFPs and their clustering at distinct chromosomal locations suggest a recent expansion of this ZFP subfamily. An analogous lineage-specific expansion of transcription factors has been observed for nuclear hormone receptors in the *C. elegans* genome (Ruvkun and Hobert, 1998; Sluder and Maina, 2001) and KRAB-containing ZFPs in humans (Lander *et al.*, 2001). The finding that most ZAD-containing ZFPs are expressed suggests that the expansion has been accompanied by stabilizing partially redundant functions of newly generated transcription units in the genomes or allowed them to adopt novel functions that were subsequently maintained. Alternatively, the expansion has occurred only very recently in the evolutionary history of *Drosophila*. If so, most members of the sequence-related subgroups may still carry largely redundant functions, explaining why the majority of the *Drosophila* ZAD-containing ZFPs has escaped functional detection by mutagenesis screens (e.g. Nüsslein-Volhard and Wieschaus, 1980; Spradling *et al.*, 1999; Peter *et al.*, 2002). This explanation would also be consistent with the finding that most ZAD-coding sequences of *A. gambiae* show only modest sequence similarity with the *Drosophila* counterparts (see Supplementary data). Expanded ZAD-containing ZFPs could therefore provide an important source for the emergence of novel protein–protein and/or protein–DNA interactions that contribute to a species-specific regulatory diversity in the control of transcription and/or chromatin structure and function. Since ZAD and the analogous KRAB domain participate in a lineage-specific expansion of ZFPs in insect and vertebrate genomes, respectively, the results described here may constitute an example of convergent evolution at the level of transcriptional regulation, the significance of which remains to be addressed experimentally.

METHODS

Identification of C2H2 ZFPs and ZFP-associated protein motifs in the *Drosophila* genome. In order to identify C2H2 ZFPs in the *Drosophila* proteome (GadFly release 2), we used the Pfam domain PF00096 (Bateman *et al.*, 2002) and the Pfam search tool. As a threshold, we assigned a minimal score of 0.0. The identified ZF motifs were subsequently manually inspected to eliminate false-positives. This was done by checking for overlaps with other protein motifs in Pfam or SMART (Letunic *et al.*, 2002); putative C2H2 motifs that overlap other more significant hits to protein domains or motifs were eliminated. The identified ZFPs were analysed with Pfam and SMART to find additional domains.

Profile construction and searches with the ZAD. An initial ClustalW 1.81 (Thompson *et al.*, 1994) alignment of the identified ZADs was used to construct a profile hidden Markov model (HMM) using the HMMER package 2.1.1 (Eddy, 1998). We performed a search against the genomic regions of the identified ZFPs using the Wise package 2.2.0 (Birney *et al.*, 1996). The genomic structure of the identified ZAD-containing ZFPs was determined (if possible) using the Gene2EST package (Gemünd *et al.*, 2001) in combination with BLAST 2.2.2 (Altschul *et al.*, 1997). The verified protein sequences encoding the ZAD were aligned using ClustalW 1.81. This alignment was used to construct an enhanced profile HMM, with which we performed searches against the publicly available EST database (NCBI DbEST, downloaded May 2002) and the set of all annotated fly proteins (Gadfly release 2). All searches against nucleotide databases were performed using the Wise 2.2.0 package; searches against protein databases were performed using the HMMER 2.1.1 package.

Classification of ZADs into subgroups and tree construction. To subgroup the ZADs, we calculated a distance matrix with PROT-DIST of the Phylip 3.5c package (Felsenstein, 1993) from a multiple sequence alignment. The resulting distance matrix was used to construct a tree using the neighbour-joining algorithm provided by Neighbor (Phylip). Sequence-related subgroups were defined: (i) all members of the sequence-related subgroups form distinct branches of the tree and no non-member is part of this branch; and (ii) the average distance between all members plus the standard deviation is smaller than the averaged distances to all non-members (in the case of subgroups containing only two members, the maximal distance between these has been arbitrarily set to 1.4).

We used the ZAD-HMM in conjunction with Wise 2.2.0 (as described above) to identify ZAD or ZAD-like motifs in the genomic sequences of *A. gambiae* ZFPs extracted from EnsEMBL 8.1b.1 (Hubbard *et al.*, 2002). The identified *A. gambiae* ZADs and the *Drosophila* ZADs were aligned and a tree was constructed as described above.

Secondary structure prediction. Secondary structure prediction was carried out with ALB (Ptitsyn and Finkelstein, 1989). A consensus prediction was calculated from the prediction of all ZADs in all alignment positions which have <10% gaps. The secondary structure prediction was verified using PHD (Rost, 1996). Since the predictions of the two programs did not differ significantly, we show the result obtained by ALB.

Supplementary data. Supplementary data are available at *EMBO reports* Online.

ACKNOWLEDGEMENTS

We thank our colleagues for help and critical discussions. This work was supported by the German Human Genome Project (grant 01 KW 9632/9; to H.J.). H.-R.C. thanks the Boehringer Ingelheim Fonds for a predoctoral fellowship.

REFERENCES

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

Bateman, A. *et al.* (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.

Bhaskar, V., Valentine, S.A. and Courey, A.J. (2000) A functional interaction between dorsal and components of the Smt3 conjugation machinery. *J. Biol. Chem.*, **275**, 4033–4040.

Birney, E., Thompson, J.D. and Gibson, T.J. (1996) PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucleic Acids Res.*, **24**, 2730–2739.

Böhm, S., Frishman, D. and Mewes, H.W. (1997) Variations of the C2H2 zinc finger motif in the yeast genome and classification of yeast zinc finger proteins. *Nucleic Acids Res.*, **25**, 2464–2469.

Chen, B., Harms, E., Chu, T., Henrion, G. and Strickland, S. (2000) Completion of meiosis in *Drosophila* oocytes requires transcriptional control by Grauzone, a new zinc finger protein. *Development*, **127**, 1243–1251.

Chervitz, S.A. *et al*. (1998) Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science*, **282**, 2022–2028.

Collins, T., Stone, J.R. and Williams, A.J. (2001) All in the family: the BTB/POZ, KRAB, and SCAN domains. *Mol. Cell. Biol.*, **21**, 3609–3615.

Crozatier, M., Kongsuwan, K., Ferrer, P., Merriam, J.R., Lengyel, J.A. and Vincent, A. (1992) Single amino acid exchanges in separate domains of the *Drosophila* Serendipity δ zinc finger protein cause embryonic and sex biased lethality. *Genetics*, **131**, 905–916.

Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Fahmy, O.G. and Fahmy, M. (1959) New mutants report. *Dros. Inf. Serv.*, **33**, 82–94.

Felsenstein, J. (1993) PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle, WA.

Gaszner, M., Vazquez, J. and Schedl, P. (1999) The Zw5 protein, a component of the scs chromatin domain boundary, is able to block enhancer–promoter interaction. *Genes Dev.*, **13**, 2098–2107.

Gemünd, C., Ramu, C., Altenberg-Greulich, B. and Gibson, T.J. (2001) Gene2EST: a BLAST2 server for searching expressed sequence tag (EST) databases with eukaryotic gene-sized queries. *Nucleic Acids Res.*, **29**, 1272–1277.

Harms, E., Chu, T., Henrion, G. and Strickland, S. (2000) The only function of Grauzone required for *Drosophila* oocyte meiosis is transcriptional activation of the *cortex* gene. *Genetics*, **155**, 1831–1839.

Hubbard, T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.

Joho, K.E., Darby, M.K., Crawford, E.T. and Brown, D.D. (1990) A finger protein structurally similar to TFIIIA that binds exclusively to 5S RNA in *Xenopus*. *Cell*, **61**, 293–300.

Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lespinet, O., Wolf, Y.I., Koonin, E.V. and Aravind, L. (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.*, **12**, 1048–1059.

Letunic, I. *et al.* (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.*, **30**, 242–244.

Nüsslein-Volhard, C. and Wieschaus, E. (1980) Mutations affecting segment number and polarity in *Drosophila*. *Nature*, **287**, 795–801.

Payre, F., Noselli, S., Lefrere, V. and Vincent, A. (1990) The closely related *Drosophila* Sry β and Sry δ zinc finger proteins show differential embryonic expression and distinct patterns of binding sites on polytene chromosomes. *Development*, **110**, 141–149.

Payre, F., Crozatier, M. and Vincent, A. (1994) Direct control of transcription of the *Drosophila* morphogen Bicoid by the Serendipity δ zinc finger protein, as revealed by *in vivo* analysis of a finger swap. *Genes Dev.*, **8**, 2718–2728.

Payre, F., Buono, P., Vanzo, N. and Vincent, A. (1997) Two types of zinc fingers are required for dimerization of the Serendipity δ transcriptional activator. *Mol. Cell. Biol.*, **17**, 3137–3145.

Peter, A. *et al.* (2002) Mapping and identification of essential gene functions on the *X* chromosome of *Drosophila*. *EMBO rep.*, **3**, 34–38.

Ptitsyn, O.B. and Finkelstein, A.V. (1989) Prediction of protein secondary structure based on physical theory. Histones. *Protein Eng.*, **2**, 443–447.

Reuter, G., Giarre, M., Farah, J., Gausz, J., Spierer, A. and Spierer, P. (1990) Dependence of position-effect variegation in *Drosophila* on dose of a gene encoding an unusual zinc-finger protein. *Nature*, **344**, 219–223.

Riechmann, J.L. *et al.* (2000) *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.

Rosenberg, U.B., Schröder, C., Preiss, A., Kienlin, A., Côte, S., Riede, I. and Jäckle, H. (1986) Molecular genetics of *Krüppel*, a gene required for segmentation of the *Drosophila* embryo. *Nature*, **313**, 336–339.

Rost, B. (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.*, **266**, 525–539.

Rubin, G.M. *et al.* (2000) Comparative genomics of the eukaryotes. *Science*, **287**, 2204–2215.

Ruvkun, G. and Hobert, O. (1998) The taxonomy of developmental control in *Caenorhabditis elegans*. *Science*, **282**, 2033–2041.

Schupbach, T. and Wieschaus, E. (1989) Female sterile mutations on the second chromosome of *Drosophila melanogaster*. I. Maternal effect mutations. *Genetics*, **121**, 101–117.

Sluder, A.E. and Maina, C.V. (2001) Nuclear receptors in nematodes: themes and variations. *Trends Genet.*, **17**, 206–213.

Spradling, A.C., Stern, D., Beaton, A., Rhem, E.J., Laverty, T., Mozden, N., Misra, S. and Rubin G.M. (1999) The Berkeley *Drosophila* Genome Project gene disruption project: single P-element insertions mutating 25% of vital *Drosophila* genes. *Genetics*, **153**, 135–177.

Stanojevic, D., Hoey, T. and Levine, M. (1989) Sequence-specific DNA-binding activities of the gap proteins encoded by *hunchback* and *Krüppel* in *Drosophila*. *Nature*, **341**, 331–335.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Venter, J.C. *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

DOI: 10.1093/embo-reports/kvf243