the FEBS Journal

HYPOTHESIS

Death inducer obliterator protein 1 in the context of DNA regulation

Sequence analyses of distant homologues point to a novel functional role

Ana M. Rojas[1,2], Luis Sanchez-Pulido[1], Agnes Fütterer[2], Karel H. M. van Wely[2], Carlos Martinez-A[2] and Alfonso Valencia[1]

1 Protein Design Group, CNB/CSIC, Madrid, Spain
2 Department of Immunology and Oncology, CNB/CSIC, Madrid, Spain

Keywords
CGBP; DATF1; DIDO1; DIO; SPP1

Correspondence
A. Rojas, Centro Nacional de Biotecnología/CSIC, Darwin 3, Cantoblanco, E-28049, Madrid, Spain
Fax: +34 91/585 4506
Tel: +34 91/585 4669
E-mail: arojas@cnb.uam.es

(Received 5 April 2005, revised 6 May 2005, accepted 10 May 2005)

doi:10.1111/j.1742-4658.2005.04759.x

Death inducer obliterator protein 1 [DIDO1; also termed DIO-1 and death-associated transcription factor 1 (DATF-1)] is encoded by a gene thus far described only in higher vertebrates. Current gene ontology descriptions for this gene assign its function to an apoptosis-related process. The protein presents distinct splice variants and is distributed ubiquitously. Exhaustive sequence analyses of all DIDO variants identify distant homologues in yeast and other organisms. These homologues have a role in DNA regulation and chromatin stability, and form part of higher complexes linked to active chromatin. Further domain composition analyses performed in the context of related homologues suggest that DIDO-induced apoptosis is a secondary effect. Gene-targeted mice show alterations that include lagging chromosomes, and overexpression of the gene generates asymmetric nuclear divisions. Here we describe the analysis of these eukaryote-restricted proteins and propose a novel, DNA regulatory function for the DIDO protein in mammals.

Apoptosis has an important role in development, tissue homeostasis, and host defense, among other functions [1]. Death inducer obliterator 1 [DIDO1; also termed DIO-1, death-associated transcription factor 1 (DATF-1)] is a protein described in humans and mice. DIDO1 was initially identified by differential display in WOL-1 pre-B cells undergoing apoptosis following interleukin-7 starvation. Developmental studies in chicken models show that its misexpression disrupts limb development [2]. When overexpressed, DIDO1 translocates to the nucleus and subsequently triggers apoptosis [3]. DIDO1 is present in all tissues and its levels are up-regulated during apoptosis. The *DIDO1* gene comprises splice variants of various lengths; each variant encodes distinct protein domain architectures that share a canonical bipartite nuclear localization signal and a PHD domain (Zn finger) at the N-terminal region. The long isoform also contains transcription factor S domain II (TFS2M) (regulatory) and Spen paralog and ortholog C-terminal domain (SPOC) (protein interaction) [4] domains at the C-terminal region. Besides the functions attributed to these structural elements, very little is known about their role in DIDO activity.

The combination of domains (domain shuffling) is a driving force in Eukarya in the acquisition of

Abbreviations
CGBP, CpG binding protein; COMPASS, Complex proteins associated with Set1; DATF-1, death-associated transcription factor 1; DIDO, death inducer obliterator; HMMER, hidden Markov model profile; ING3, Inhibitor of growth protein 3; MLL, mixed lineage leukemia; PHD, plant homeodomain; SPEN, split ends domain; SPOC, Spen paralog and ortholog C-terminal domain; SPP1, suppressor of PRP protein 1; s-Zf, small zinc finger; TFS2M, transcription factor S domain II.

biological 'complexity' in terms of regulatory pathways. Indeed, breaking down a whole protein into its component domains to analyse its composition is a more appealing approach to infer functions [5,6] when data extracted from the whole protein are limited or inadequately informative. We therefore conducted domain analyses to obtain new insights from domain distribution, with the support of sequence and phylogenetic analyses, as well as experimental data.

By extensive sequence analysis, we found a set of *DIDO1* homologues in the context of DNA binding and chromatin regulation. These homologues include members of the CpG binding protein (CGBP) family of proteins that bind DNA at unmethylated CpG islands [7], which are always located to active chromatin. Other proteins identified are fungal homologues, of which suppressor of PRP protein (SPP1) (alias Cps40) is a well-characterized component of the Set1 complex (alias COMPASS) in *Saccharomyces cerevisiae* [8]. This complex is implicated in leukemia [9] by covalent modifications of chromatin.

Results and Discussion

Sequence analyses of all splice variants link DIDO1 to other distantly related protein families involved in DNA binding and chromatin stability. Moreover, additional experimental data place the DIDO protein in the context of cell division. Our analyses strongly suggest that DIDO-related apoptosis occurs as a result of alterations in DNA regulation caused by chromatin instability.

Computational analyses

Although all splice variants share common domains, long regions of the protein are not identified using automatic searches (PFAM and SMART tools). To obtain further information regarding these regions, we thus surveyed for the existence of distant DIDO protein homologues, using these regions as queries for extensive searches. The murine DIDO PHD domain [residues 262–423; protein databank (pdb) code: 1WEM] was used as a query to retrieve several homologous sequences and to search databases after profile building (see Experimental procedures).

The searches found numerous members of different protein families in the first iteration of a PSI-BLAST search, with significant e-values of 2e-30, 1e-10 and 1e-08 for DATF-1, PHF3 and CGBP, respectively. Given the broad distribution and functional repertoire for the PHD domain in the databases, we studied its evolution (Fig. 1) in several representative PHD domain-containing sequences, including members of DIDO, CGBP, and fungal sequences. We conducted major phylogenetic analyses with more than 200 PHD domains retrieved in the initial PSI-BLAST search to locate the overall position of the DIDO PHD in the tree (data not shown). From these results, we extracted representatives of each family, obtained at significant PSI-BLAST e-values, to conduct more restricted phylogenetic analyses. We then selected three DIDO sequences, two representatives of the PHF family, six representatives of the CGBP family, and seven fungal representatives (Fig. 1). To root the tree, we used inhibitor of growth protein 3 (ING3) PHD, which is another domain involved in chromatin binding and clearly more divergent (e-value 0.019) from DIDO and the other sequences. As the alignment length is very short and divergent, we conducted probabilistic analyses, based on Bayesian inference, to perform phylogenies. As seen in the tree (Fig. 1), the branch of the

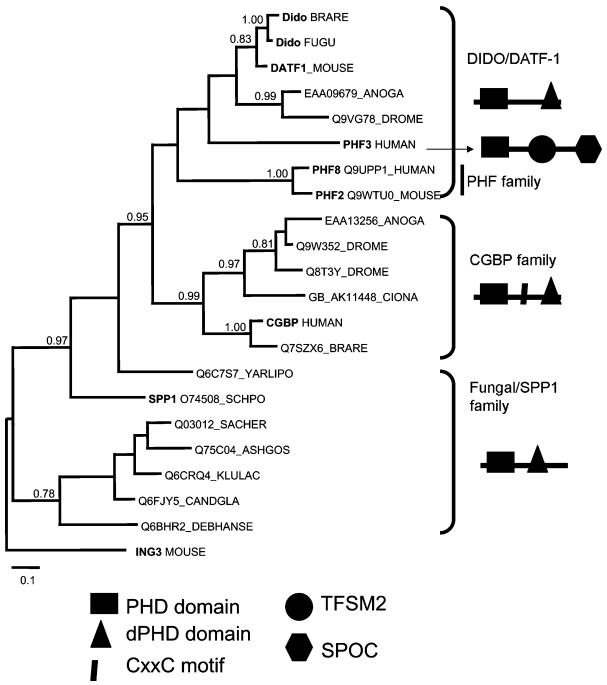

Fig. 1. Phylogenetic analyses of PHD domains. Representative PHD domains were aligned. Numbers indicate the frequency of clade probability values. Only values over 0.75 are shown. Black geometrical shapes are additional domains, as indicated. ANOGA, *Anopheles gambiae*; ASHGOS, *Ashbya gossypii*; BRARE, *Brachydanio rerio*; CANDGLA, *Candida glabrata*; CIONA, *Ciona intestinalis*; DEBHANSE, *Debaryomyces hansenii*; DROME, *Drosophila melanogaster*; KLULAC, *Kluyveromyces lactis*; SACHER, *Saccharomyces cerevisiae*; SCHPO, *Schizosaccharomyces pombe*; YARLIPO, *Yarrowia lipolytica*. ING3_MOUSE is the outgroup. DATF1_MOUSE is the SwissProt entry (DIDO1).

fungal sequences is clearly distant from the other branches, which was indeed predicted. The other main group, obtained 97% of the time in 20 740 explored trees, contained the DIDO PHD sequences clustered with the PHF and the CGBP proteins. As seen in the tree, two fungal PHD domains corresponded to an SPP1 cluster at the basal branch of the tree. Identical topologies were obtained by using distance methods and neighbor-joining trees (data not shown).

The sequences of DIDO1, CGBP and SPP1 were realigned (supplementary Fig. S1), and a cysteine-rich short motif spanning approximately 25 residues was detected downstream of the PHD domain. We termed this new region dPHD; it is well conserved among the sequences and always follows the PHD domain. Using PSI-BLAST, the dPHD of DIDO1 hits CGBP_MOUSE at the second iteration, with an e-value of 2e-06 (inclusion threshold of 0.03). In SPP1, the corresponding dPHD segment was used to obtain fungal sequences, and its profile hit the Q6PGZ4 protein (zebrafish CGBP) at an hidden Markov model profile (HMMER) e-value of 0.086 (Fig. 2). When using the combined profile of CGBP and fungal sequences, the murine DIO was hit at an HMMER e-value of 0.083. The statistical robustness is in agreement with the PHD phylogenetic distribution (Fig. 1), in which the SPP1 representatives are at the basal branch of CGBP and DIO.

The GCBP proteins contain a PHD domain, followed by a DNA-binding domain (the zf-CXXC) and the newly described dPHD region (supplementary Fig. S1). This family is involved in DNA binding at unmethylated CpG islands in active chromatin, and these proteins are essential for mammalian development [7]. In addition, CGBP subcellular distribution is identical to that of the human trithorax protein, suggesting that they may be components of a multimeric complex analogous to the *Saccharomyces* histone-methylating Set1 complex, which contains CGBP and trithorax homologues [10]. The members of the trithorax group encompass various subclasses of gene regulatory factors [11]; one subclass involves chromatin remodeling activity. Another subclass, the trxG, is poorly understood and includes trithorax itself, Ash1, and Ash2 [12]; these latter are homologues of components of the yeast COMPASS/SET1 complex. Some functional features have been reported [13,14] for the trithorax complex proteins in the context of domain composition. Recent analyses of trithorax/mixed lineage leukemia (MLL) and its relationship with COMPASS suggest a linkage between leukemogenesis and covalent modifications of chromatin [9]. Little is known, however, of the functional role of the remaining subclass members.

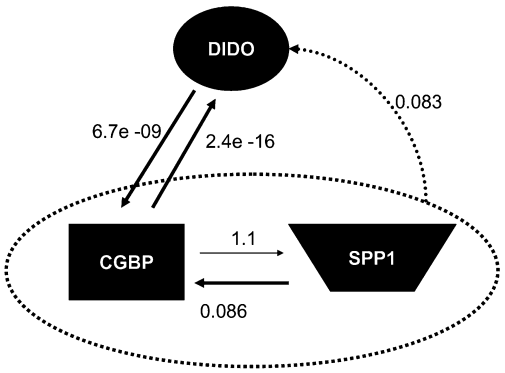

Fig. 2. HMMER e-values between the dPHD domain-containing families. Numbers correspond to HMMER e-values from global profile search results that connect the families. Arrows indicate profile search direction.

In this study, the yeast protein SPP1 (alias cps40), a member of the COMPASS complex [10], was statistically related for the first time to DIDO1 and CGBP proteins (Fig. 2). SPP1 also contains a PHD domain followed immediately by dPHD (Figs 1 and 3). The domain architectures concur with the phylogenetic distribution of PHD (Fig. 1) and the HMMER values of dPHD (Fig. 2). In both cases, the fungal sequences appear to be more closely related to CGBP than to DIDO1, although CGBP contains the CXXC signature between the two domains, probably as a result of recombination processes; this signature is missing in DIDO1 and SPP1. The absence of the CXXC signature distinguishes SPP1 from the CGBP family (Figs 1, 3 and S1); otherwise, the fungal sequences could well constitute CGBP homologues.

The presence of a TFS2M domain is also indicative of a possible function for the DIDO1 long isoform (Figs 3 and S2). It is the second domain of the elongation factor, TFSII [15], that stimulates RNA polymerase II to transcribe through regions of DNA. The structure of domain III of this elongation factor is also solved structurally [16], as a Zn-binding domain. Both domains II and III constitute the minimal transcriptionally active fragment and are required simultaneously to maintain transcription.

The PHF3 protein was recovered in initial analyses and also contains a TFS2M domain showing the same architecture as the DIDO1 long isoform, although lacking the dPHD. The PHF3 protein is expressed ubiquitously in normal tissues, and its expression is dramatically reduced or lost in human glioblastoma (a malignant astrocytic brain tumor) [17,18]. Downstream of TFS2M, a very small motif was detected containing two histidine and two cysteine residues

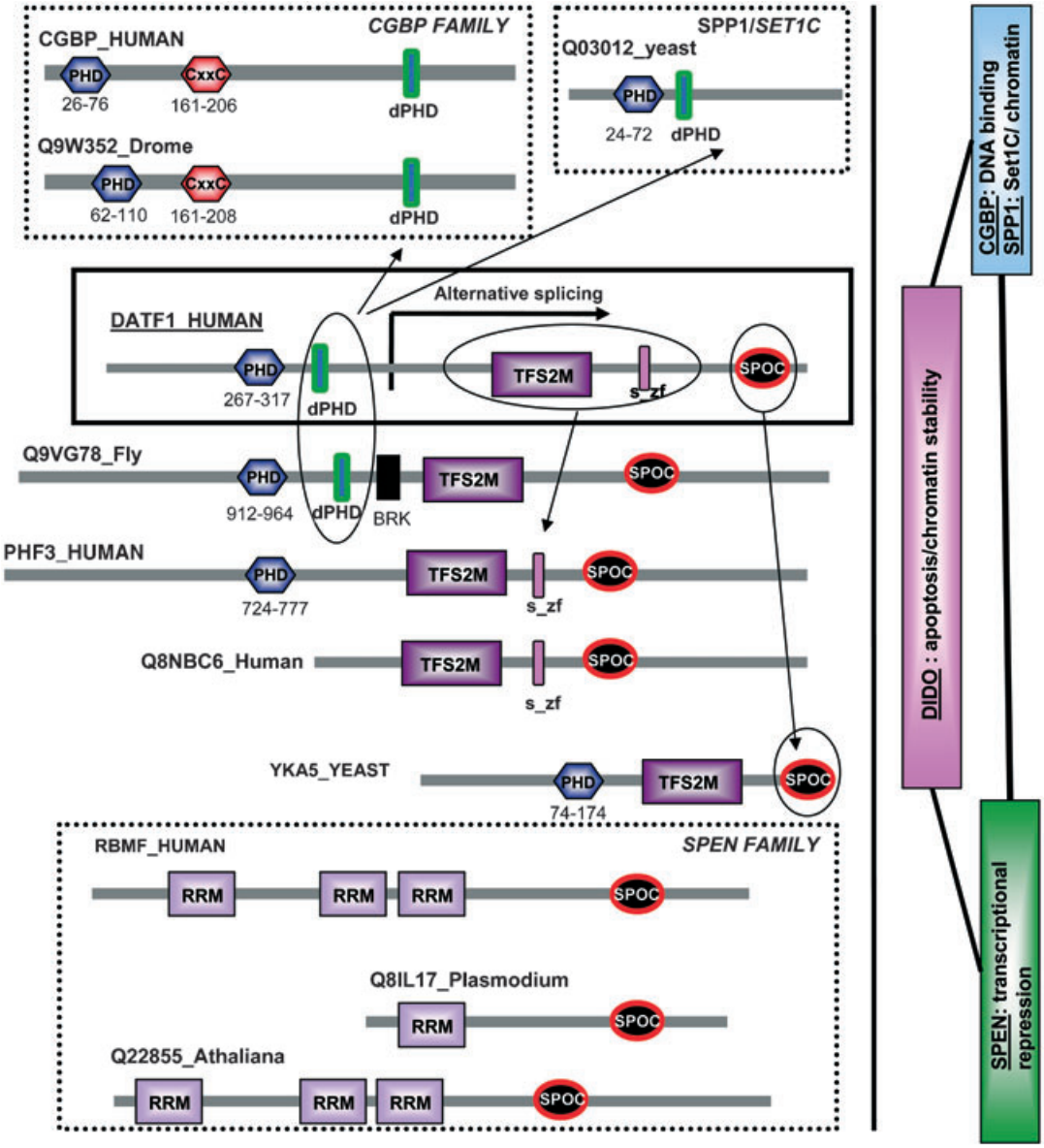

Fig. 3. Domain dissection of death inducer obliterator protein (DIDO) proteins. Sequence names are SwissProt/TrEMBL identifiers. PHD (dark blue), dPHD (green/blue), CXXC (orange), TFS2M (purple), sZf (pink), SPOC (black/red), RRM (mauve). The central boxed protein is DIDO (DATF_HUMAN). The dPHD region connects DIDO with CGBP and SPP1 families (upper panel), and with a fly homolog, Q9VG78 (below the DIDO box). TFS2M–sZf links DIDO with PHF3_HUMAN (center panel). The SPOC domain connects to the SPEN family (lower panel) [4]. CXXC, CGBP-specific; BRK, fly specific; and RRM, SPEN-specific. DATF_HUMAN is the SwissProt identifier for DIDO.

(Fig. S2), for which we propose the name small Zinc finger (s-Zf). This region is too small to assess with any confidence, based on statistical terms. Nonetheless, further searches using other methods (such as pattern matching) were conducted in databases, from which no conclusive results were obtained (data not shown). This region is present in PHF3, in another protein (Q8NBC6), and in the DIDO1 long isoform, however, and appears to be restricted to mammals. This architecture in some way resembles the distribution of TFSII domains II and III.

Our surveys provided statistically significant e-values connecting the PHD domain-containing families DIDO1, CGBP, and the fungal family that we term SPP1. All these proteins contain a small, previously undetected domain downstream of the PHD, which we called dPHD, that allows connection of these families (despite the small length of alignment) at reliable

e-values (Fig. 2). It is noteworthy that the retrieved proteins all appear to have a role in DNA regulation, in the context of chromatin stability, and to form part of higher complexes linked to active chromatin.

The *DIDO1* gene has been linked to the split ends domain (SPEN) family of proteins [4], involved in transcriptional repression via their C-terminal domain, SPOC. Here we show that additional protein families, CGBP and SPP1, are linked to this gene by two domains (Fig. 3).

Experimental analyses

The localization of the *DIDO1* gene in the context of chromatin stability was addressed experimentally. Although the experiments conducted were preliminary, they indicated the involvement of this gene in cell division. This is consistent with our observations, that the *DIDO1* gene shares two domains with CGBP and SPP1, both proteins being bound to active chromatin and involved in DNA regulation; in addition, the yeast protein, SPP1, is well characterized by tandem affinity-purification experiments.

Ectopically expressed DIDO1 associated with chromatin throughout the cell cycle (Fig. 4A,B), causing a high incidence of asymmetric divisions. Cells from DIDO1-targeted mice show a notable incidence of lagging chromosomes (10 of 237; 4.2%) during anaphase (Fig. 4C,D), which was not observed in cultures of wild-type cells (0 of 140; 0.0%). Although merotelic kinetochore attachment to centromeres is generally considered to be a major cause of lagging chromosomes, they can also be caused by changes in chromatin composition [19,20]. As DIDO1 associates with chromatin in general, and not only in centromeric regions, chromatin instability is the most probable explanation for the lagging chromosomes in DIDO1-targeted cells.

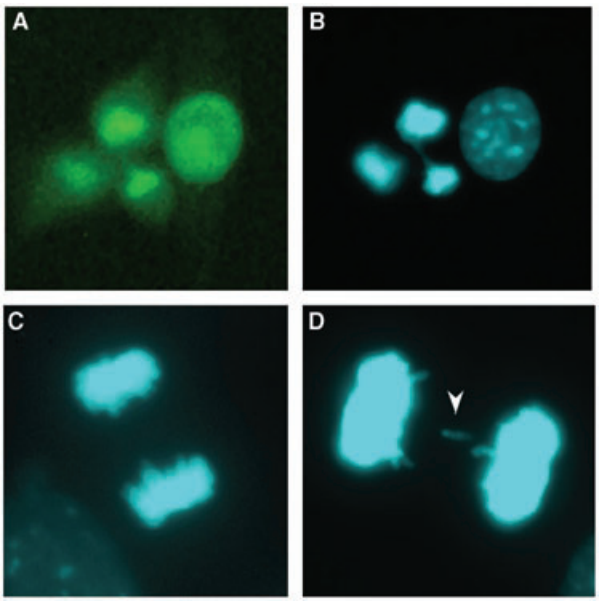

Fig. 4. Chromosomal instability in death inducer obliterator protein (DIDO)-overexpressing and DIDO-targeted cells. (A, B) Ectopic expression of DIDO1. (C) Normal anaphase of a wild-type mouse cell. (D) Mitosis in DIDO-targeted cells, showing lagging chromosomes during anaphase (arrowhead).

Targeting of the *DIDO1* locus leads to genomic instability, as shown by the occurrence of lagging chromosomes in mitosis. The domains targeted in mice are PHD and dPHD, which are domains shared with CGBP and SPP1. Ectopic DIDO1 expression leads to a high incidence of asymmetric divisions. The reported DIDO-induced apoptosis could thus be a consequence of alterations in DNA regulation or chromatin stability. A similar case is the Suv39h histone methyltransferase, in which both deletion and overexpression lead to alterations in pericentric chromatin, chromosome missegregation in mitosis and meiosis, and apoptosis [19,21].

Concluding remarks

Computational analyses, combined with some basic and preliminary experimental assays (some of which we show here), enable us to hypothesize an additional functional role for DIDO1. In this case, apoptosis induction should be explained within the context of DNA regulation, especially considering that the apoptosis induced by DIDO1 requires protein translocation to the nucleus. Although functional analyses of individual domains have not been addressed, the global context of the domain analyses allows us to draw a more general picture of the involvement of this gene in nuclear processes. Nonetheless, the data presented here build a hypothesis that should be experimentally addressed in detail.

Experimental procedures

Computational analyses

The complete protein sequences of human DIDO isoforms 1 and 2 were searched against PFAM [22,23] and SMART [24] databases to automatically detect domains; they were further used as queries against NCBI databases using PSI-BLAST [25,26]. Complete gene sequences were found only in mouse or human; only partial fragments were found in other organisms. We first performed BLAST searches against unfinished genomes from NBCI [27], then against

EST databases, with further EST assembly of reliable hits. Any new sequence was incorporated into profiles to improve profile quality. Profile-based sequence searches were performed against the nonredundant and Uniref90 protein databases with the corresponding global hidden Markov models [28] (HMMer version 2.3.2 PVM). Alignments were generated by using T-COFFEE and checked manually [29].

Phylogenies of the PHD domain were obtained by using probabilistic approaches [30] (Mr Bayes 3 version), which run for 1 000 000 generations in four independent Markov chains. When convergence was reached, a total of 20 740 trees were explored to further construct a consensus tree. Numbers indicate the frequency of clade probability values.

Green fluorescent protein (GFP)–DIDO-expressing cell lines

To construct the GFP–DIDO1 fusion, human DIDO1 cDNA was transferred from pGEMT (Promega, Madison, WI, USA) to pEGFP-C1 (Clontech, Mountain View, CA, USA) by using unique *Spe*I and *Apa*I sites. This yields a plasmid expressing DIDO1 in-frame with GFP under control of the cytomegalovirus (CMV) promoter. NIH 3T3 cells were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% (v/v) fetal bovine serum and antibiotics. To generate stable cell lines, 10^5 cells were seeded in each well of a six-well plate (BD Falcon, San Jose, CA, USA) and transfected with plasmid DNA (1 μg) and FuGene 6 (3 μl; Roche, Indianapolis, IN, USA), as recommended by the manufacturer. Cells were selected by incubation with 500 $\mu g \cdot mL^{-1}$ G418 for 2 weeks, and clones with efficient expression, as judged by fluorescence microscopy, were used for further experiments. To visualize GFP-DIDO1, $\approx 5 \times 10^4$ cells were seeded on glass coverslips. After 48 h, cells were formaldehyde fixed, mounted in Vectashield containing 4,6-diamino-2-phenylindole (DAPI) (Vector Laboratories, Burlingame, CA, USA), and studied by standard fluorescence microscopy.

Analysis of anaphases in embryonic fibroblasts

To determine the occurrence of lagging chromosomes during anaphase, low-passage mouse embryonic fibroblast cultures from targeted (PHD and dPHD regions) and wild-type mice were fixed with methanol and acetic acid, mounted in Vectashield containing DAPI, and studied by standard fluorescence microscopy. Lagging chromosomes were scored only when no attachment whatsoever to either of the chromosome pools could be detected in anaphase.

Acknowledgements

We thank Catherine Mark for editorial assistance. This work was financed, in part, by the 6th EU Framework Program Project IMPAD QLGI-CT-2001-01536, MEC and GenFun LSHG-CT-2004-503567. The Department of Immunology and Oncology was founded and is supported by the Spanish Council for Scientific Research (CSIC) and by Pfizer.

References

1 Aravind L, Dixit VM & Koonin EV (1999) The domains of death: evolution of the apoptosis machinery. *Trends Biochem Sci* **24**, 47–53.
2 Garcia-Domingo D, Leonardo E, Grandien A, Martinez P, Albar JP, Izpisua-Belmonte JC & Martinez AC (1999) DIO-1 is a gene involved in onset of apoptosis in vitro, whose misexpression disrupts limb development. *Proc Natl Acad Sci USA* **96**, 7992–7997.
3 Garcia-Domingo D, Ramirez D, Gonzalez de Buitrago G & Martinez AC (2003) Death inducer-obliterator 1 triggers apoptosis after nuclear translocation and caspase upregulation. *Mol Cell Biol* **23**, 3216–3225.
4 Sanchez-Pulido L, Rojas AM, van Wely KH, Martinez AC & Valencia A (2004) SPOC: a widely distributed domain associated with cancer, apoptosis and transcription. *BMC Bioinformatics* **5**, 91.
5 Copley RR, Doerks T, Letunic I & Bork P (2002) Protein domain analysis in the era of complete genomes. *FEBS Lett* **513**, 129–134.
6 Ponting CP & Dickens NJ (2001) Genome cartography through domain annotation. *Genome Biol* **2**, Comment 2006.
7 Voo KS, Carlone DL, Jacobsen BM, Flodin A & Skalnik DG (2000) Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. *Mol Cell Biol* **20**, 2108–2121.
8 Miller T, Krogan NJ, Dover J, Erdjument-Bromage H, Tempst P, Johnston M, Greenblatt JF & Shilatifard A (2001) COMPASS: a complex of proteins associated with a trithorax-related SET domain protein. *Proc Natl Acad Sci USA* **98**, 12902–12907.
9 Tenney K & Shilatifard A (2005) A COMPASS in the voyage of defining the role of trithorax/MLL-containing complexes: Linking leukemogensis to covalent modifications of chromatin. *J Cell Biochem* **95**, 429–436.
10 Roguev A, Schaft D, Shevchenko A, Pijnappel WW, Wilm M, Aasland R & Stewart AF (2001) The *Saccharomyces cerevisiae* Set1 complex includes an Ash2 homologue and methylates histone 3 lysine 4. *EMBO J* **20**, 7137–7148.
11 Kennison JA (1995) The Polycomb and trithorax group proteins of *Drosophila*: trans-regulators of homeotic gene function. *Annu Rev Genet* **29**, 289–303.

12 Shearn A (1989) The ash-1, ash-2 and trithorax genes of *Drosophila melanogaster* are functionally related. *Genetics* **121**, 517–525.
13 Mazo AM, Huang DH, Mozer BA & Dawid IB (1990) The trithorax gene, a trans-acting regulator of the bithorax complex in *Drosophila*, encodes a protein with zinc-binding domains. *Proc Natl Acad Sci USA* **87**, 2112–2116.
14 Tripoulas N, LaJeunesse D, Gildea J & Shearn A (1996) The *Drosophila* ash1 gene product, which is localized at specific sites on polytene chromosomes, contains a SET domain and a PHD finger. *Genetics* **143**, 913–928.
15 Morin PE, Awrey DE, Edwards AM & Arrowsmith CH (1996) Elongation factor TFIIS contains three structural domains: solution structure of domain II. *Proc Natl Acad Sci USA* **93**, 10604–10608.
16 Qian X, Jeon C, Yoon H, Agarwal K & Weiss MA (1993) Structure of a new nucleic-acid-binding motif in eukaryotic transcriptional elongation factor TFIIS. *Nature* **365**, 277–279.
17 Fischer U, Struss AK, Hemmer D, Michel A, Henn W, Steudel WI & Meese E (2001) PHF3 expression is frequently reduced in glioma. *Cytogenet Cell Genet* **94**, 131–136.
18 Struss AK, Romeike BF, Munnia A, Nastainczyk W, Steudel WI, Konig J, Ohgaki H, Feiden W, Fischer U & Meese E (2001) PHF3-specific antibody responses in over 60% of patients with glioblastoma multiforme. *Oncogene* **20**, 4107–4114.
19 Peters AH, O'Carroll D, Scherthan H, Mechtler K, Sauer S, Schofer C, Weipoltshammer K, Pagani M, Lachner M, Kohlmaier A *et al.* (2001) Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability. *Cell* **107**, 323–337.
20 Cimini D, Mattiuzzo M, Torosantucci L & Degrassi F (2003) Histone hyperacetylation in mitosis prevents sister chromatid separation and produces chromosome segregation defects. *Mol Biol Cell* **14**, 3821–3833.
21 Shen WH & Meyer D (2004) Ectopic expression of the NtSET1 histone methyltransferase inhibits cell expansion, and affects cell division and differentiation in tobacco plants. *Plant Cell Physiol* **45**, 1715–1719.
22 Sonnhammer EL, Eddy SR, Birney E, Bateman A & Durbin R (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* **26**, 320–322.
23 Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M & Sonnhammer EL (2002) The Pfam protein families database. *Nucleic Acids Res* **30**, 276–280.
24 Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP & Bork P (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res* **32**, Database issue, D142–144.
25 Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.
26 Altschul SF & Koonin EV (1998) Iterated profile searches with PSI-BLAST – a tool for discovery in protein databases. *Trends Biochem Sci* **23**, 444–447.
27 Cummings L, Riley L, Black L, Souvorov A, Resenchuk S, Dondoshansky I & Tatusova T (2002) Genomic BLAST: custom-defined virtual databases for complete and unfinished genomes. *FEMS Microbiol Lett* **216**, 133–138.
28 Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* **14**, 755–763.
29 Notredame C, Higgins DG & Heringa J (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205–217.
30 Ronquist F & Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574.

Supplementary material

The following material is available online.

Fig. S1. Multiple alignment of DIDO, CGBP and SPP1 proteins. Additional proteins were included. Names are SwissProt or sptrembl identifiers, with added common species name: Chick, *Gallus gallus*; Fugu, *Fugu rubripes*; Brare, *Danio rerio*; Anoga, *Anopheles gambiae*; Drome, *Drosophila melanogaster*; Sacce, *Saccharomyces cerevisiae*; Schizo, *Schizosaccharomyces pombe*; Ciona, *Ciona intestinalis*; Caeel, *Caenorhabditis elegans;* Caebri, *Caenorhabditis briggsae*. The DIDO1 EST consensus sequence was reconstructed manually by assembling ESTs. Boxed, vertebrate-restricted. Red-boxed sequence names are CGBP, showing a specific CXXC motif absent in other sequences (a solid red box above the alignment). A solid dark blue box indicates the PHD domain, where rectangles and boxes indicate secondary structural elements from the DIDO1 mouse structure (pdb code: 1WEM); a solid green/blue box identifies the newly identified dPHD domain. DATF1_MOUSE and DATF_HUMAN are the SwissProt identifiers for DIDO.

Fig. S2. Multiple alignment of death inducer obliterator protein (DIDO), PHF3 and yeast proteins. Additional proteins were included. Naming conventions are as in Fig. S1. The solid purple box indicates the TFS2M domain. Pink-boxed sequence names are proteins containing the newly identified s-Znf signature (solid pink box above the alignment). The solid black/red box identifies the SPOC domain.