Cell Death and Differentiation (2012) 19, 2040–2048

www.nature.com/cdd

Conservation of caspase substrates across metazoans suggests hierarchical importance of signaling pathways over specific targets and cleavage site motifs in apoptosis

ED Crawford[1,2], JE Seaman[1,2], AE Barber II[3], DC David[4,5], PC Babbitt[2,3], AL Burlingame[1] and JA Wells*[,1,2]

Caspases, cysteine proteases with aspartate specificity, are key players in programmed cell death across the metazoan lineage. Hundreds of apoptotic caspase substrates have been identified in human cells. Some have been extensively characterized, revealing key functional nodes for apoptosis signaling and important drug targets in cancer. But the functional significance of most cuts remains mysterious. We set out to better understand the importance of caspase cleavage specificity in apoptosis by asking which cleavage events are conserved across metazoan model species. Using N-terminal labeling followed by mass spectrometry, we identified 257 caspase cleavage sites in mouse, 130 in *Drosophila*, and 50 in *Caenorhabditis elegans*. The large majority of the caspase cut sites identified in mouse proteins were found conserved in human orthologs. However, while many of the same proteins targeted in the more distantly related species were cleaved in human orthologs, the exact sites were often different. Furthermore, similar functional pathways are targeted by caspases in all four species. Our data suggest a model for the evolution of apoptotic caspase specificity that highlights the hierarchical importance of functional pathways over specific proteins, and proteins over their specific cleavage site motifs.
Cell Death and Differentiation (2012) **19,** 2040–2048; doi:10.1038/cdd.2012.99; published online 24 August 2012

Apoptosis, a form of programmed cell death, is conserved across the metazoan lineage, and is crucial for removing harmful or unneeded cells from a healthy animal body. Apoptosis has been extensively studied in several model organisms, including mouse, *Drosophila*, and *Caenorhabditis elegans*. While some differences in the early stages of the pathway exist,[1] in all cases the later stages involve strong activation of one or more caspases, cysteine class proteases with exquisite specificity for cleavage after aspartic acid.

To better understand how caspase activation leads to cell death, there has been considerable interest in identifying protein substrates of apoptotic caspases in human cells[2–5] (reviewed in Crawford and Wells[6]). Most substrates are cleaved at a single site in a loop or disordered region, and the cut may cause a gain or loss of function, or may have no effect. Detailed molecular studies have established important roles of some substrates; many are key players, and sometimes good drug targets or biomarkers, in cancer and other diseases of apoptosis misregulation.[6] But for the vast majority of cleavages, the functional significance is unknown.

We have used a subtiligase N-terminal labeling strategy to identify caspase cleavages in three metazoan model organisms (mouse, *Drosophila*, and *C. elegans*) representing roughly 600 million years of evolution[7] (Figures 1a and b). We base our analysis of evolutionary conservation on a data set of 1444 human caspase substrates, generated in our lab. Though there are inevitably some gaps in this human list, we believe it is large enough to reveal important general patterns. Our model organism data sets are much smaller, so we are unable to evaluate the conservation of every known human caspase substrate. However, the majority of caspase cleavage sites discovered in mouse, roughly half in *Drosophila*, and one third in *C. elegans*, were in orthologs of human proteins known to be caspase substrates, suggesting that many cuts are indeed strongly conserved and thus likely to be functional. Notably, in many cases the cleavages occurred at different places in the two orthologs, often with the essential cleavage site Asp residue having been lost in humans. This is consistent with the tendency of short linear motifs to be lost over long evolutionary distances.[8]

[1]Department of Pharmaceutical Chemistry, University of California, San Francisco, CA, USA; [2]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, CA, USA; [3]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA, USA and [4]Department of Biochemistry and Biophysics, University of California, San Francisco, CA, USA
*Corresponding author: JA Wells, Department of Pharmaceutical Chemistry, UCSF Mission Bay, MC 2552 Byers Hall S504, 1700 4th Street, San Francisco, CA 94158-2330, USA. Tel: 415 514 4757; Fax: 415 514 4507; E-mail: Jim.Wells@ucsf.edu
[5]Present address: DZNE, German Center for Neurodegenerative Diseases, D-72076 Tübingen, Germany.
Keywords: caspase; apoptosis; PTM; proteomics; evolution
Abbreviations: ppm, part per million; GST, glutathione S-transferase; GO, gene ontology; IPA, Ingenuity Pathway Analysis; SDS, sodium dodecyl sulfate; FBS, fetal bovine serum; eIF, eukaryotic initiation factor; RPL, ribosomal protein, large subunit; PBS, phosphate buffered saline; MEF, mouse embryonic fibroblast; PTM, post-translational modification; PMSF, phenylmethanesulfonylfluoride; AEBSF, 4-(2-Aminoethyl) benzenesulfonyl fluoride hydrochloride; EDTA, Ethylenediaminetetraacetic acid; IAM, iodoacetamide
Received 13.4.12; revised 06.7.12; accepted 10.7.12; Edited by G Salvesen; published online 24.8.12

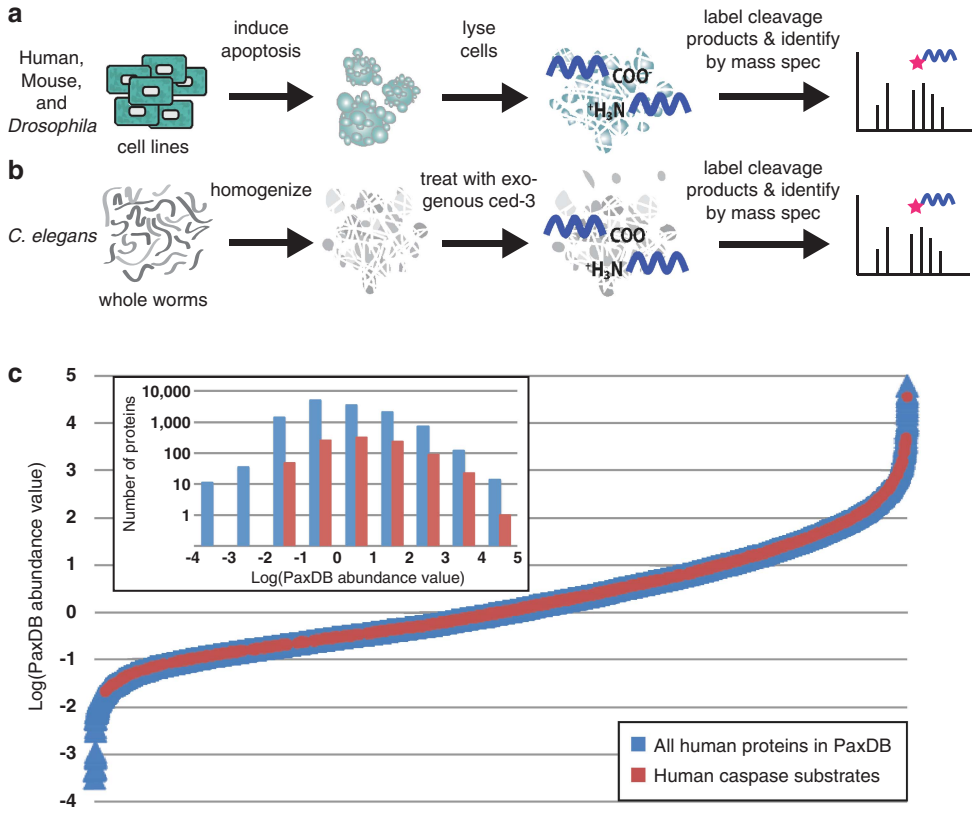

Figure 1 (**a**) For human, mouse, and *Drosophila* experiments, cell lines were grown under standard culture conditions, and induced to apoptose using various toxic agents. The apoptotic cells were lysed, and the subtiligase labeling method was used to enrich for N-termini generated during apoptosis. (**b**) For *C. elegans*, whole worms were homogenized, cuticles and other debris were spun out, and the resulting lysate was treated with exogenously expressed CED-3 for 2 h at room temperature. The subtiligase labeling method was again used to enrich for unblocked N-termini, in this case many of them derived from the added protease. (**c**) Abundance levels of all human proteins in the PaxDB database were plotted on a log scale in blue, and levels of the human proteins identified as caspase substrates in our human reference data set are plotted in red. Inset shows a histogram derived from the same data

Furthermore, pathway analysis showed that mouse, *Drosophila*, or *C. elegans* caspase substrates not conserved at the protein level often have a human ortholog with membership in a pathway known to be heavily targeted by human caspases. These findings suggest a hierarchical model of functional evolution: specific caspase cleavage motifs tend to be conserved only over short evolutionary distances; the set of protein caspase substrates is more conserved over longer distances; and the set of pathways targeted by caspases, even more conserved over even longer distances. Throughout this report, we refer to these three evolutionary modes as motif level, protein level, and pathway level conservation. A similar hierarchical conservation model has recently been proposed for the case of phosphorylation, another important post-translational modification.[9]

Results

Data collection

Human reference data set: We previously developed a technique for proteome-scale positive enrichment of N-termini liberated by proteolysis, using the enzyme subtiligase.[4] This method was employed using various combinations of eight different human cell lines treated with any of five different apoptosis-inducing compounds, leading to identification of 2021 Asp cleavage sites in 1444 proteins (published in Mahrus *et al.*,[4] Shimbo *et al.*,[10] and Crawford & Seaman *et al.*, in preparation). We are confident implicating caspases in these cuts, as caspases are virtually unique in having a strong preference for Asp at the P1 position (Schechter and Berger[11] protease notation; proteolysis occurs between the P1 and P1′ residues).

To determine how robust our human data set is to variations in protein abundance, we consulted the PaxDB database,[12,13] which lists relative protein abundance levels based on spectral counting data drawn from many mass spectrometry data sets. As Figure 1c shows, our coverage extends over more than six orders of magnitude of ppm, although representation is thinner at lower abundance levels.

At least 300 additional caspase substrates were not detected in our data, but have been described elsewhere. Many of these are listed in the MEROPS peptidase database.[5] As cleavage site location validation varies greatly among MEROPS's sources, we chose not to include these data in our analyses. However, in Supplementary Table 1 we note 20 human orthologs that are identified in MEROPS as substrates of one of the apoptotic effector caspases (caspases-3, -6, and -7). Only one MEROPS cleavage site

promotes an orthologous pair from not conserved to conserved at the motif level (IL16_HUMAN and IL16_MOUSE). Two additional orthologous pairs were promoted from not conserved (or pathway-level conserved) to protein-level conserved (BRCA1_HUMAN and BRCA1_MOUSE; TOP1_HUMAN and TOP1_DROME). The remaining 17 MEROPS sites either matched our human data, or matched an additional unaligned site.

Mouse and Drosophila data sets: Three mouse cell lines were used: A20, TK-1, and MEFs. Apoptosis was induced in multiple mouse experiments using bortezomib, doxorubicin, etoposide, or staurosporine. *Drosophila* S2 cells were treated with doxorubicin, cyclohexamide, or actinomycin D. For both species, Cell Titre Glo and Caspase Glo assays from Promega (Madison, WI, USA) were performed on separate experimental samples to confirm that the drugs, concentrations, and time points used induced caspase activation and cell death (data not shown). Cells were harvested at a series of time points surrounding the time of maximal caspase activity, from 4 to 24 h after treatment, and processed with the N-terminal labeling and mass spectrometry protocol described in Agard *et al.*[14] The data were searched with Protein Prospector (http://prospector.ucsf.edu/prospector/ms_home.htm). For mouse, the SwissProt database was searched; for *Drosophila*, the UniProtKB database was searched. The basic workflow is shown in Figure 1a, and results in Supplementary Table 1.

C. elegans data set: C. elegans is a classic apoptosis model organism, and has only one functional caspase, CED-3. No *C. elegans* immortalized cell lines have been established. In a gel-based proteomics study, Taylor *et al.*[15] identified 22 CED-3 substrates by treating an extract made from homogenized whole worms with recombinant CED-3. Following their protocol, we grew Bristol N2 worms in liquid culture, homogenized them, isolated the soluble protein fraction, and incubated it at room temperature for two hours with a CED-3 preparation made from *E. coli*. We then subjected the CED-3-digested extract to our *in vitro* N-terminal labeling protocol, as described[14] (Figure 1b). The data were searched with Protein Prospector, using the UniProtKB database. Results are listed in Supplementary Table 1.

Data analysis

Primary structure specificity: We used IceLogo[16] to generate visual representations of sequence specificity for all four species (Figure 2a). In each case, the appropriate SwissProt or UniProtKB database was used to establish background amino acid frequencies. Synthetic peptide-based experiments suggest a strong preference for Asp at P4 and Glu at P3 for human executioner caspases-3 and -7.[17] In contrast, physiological sets of whole human proteins cut by caspases consistently show weaker sequence specificity,[2,4,18] suggesting that sequence alone is not sufficient to explain caspase-substrate recognition in cells. Our data show that this holds true for mouse, *Drosophila*, and *C. elegans* as well. The issue of specificity is complicated somewhat by the mixture of different caspases that are active during apoptosis. However, studies focusing on single caspases (including the CED-3 experiment described here) have yielded similar sequence logos to the physiological data sets.[14]

As in other proteome-scale studies of caspase substrates,[4,19] the P1′ position shows a strong preference for a small amino acid. In all four species, $>75\%$ of P1′ residues are either Ala, Gly, or Ser. This is not a result of bias from the subtiligase N-terminal labeling procedure; subtiligase has a slight bias towards large residues on the labeled N terminus.[20]

Secondary structure specificity: Examining predicted secondary structure for the caspase cleavage sites in our data confirms a statistically significant preference for sites that occur in loops. Caspases, like most proteases, require substrate cleavage sites be in an extended conformation while binding to the active site.[21] These sites must therefore either be in loops or in structured regions that are dynamic enough to give caspases reasonable access to subpopulations in unstructured conformations. We used NetSurfP[22] to predict secondary structure for the P4–P4′ residues surrounding all cleavage sites, and for a background made up of all 8-mers in the same set of proteins with Asp at the P1 position. We observed enrichment in loop structures in all four species (Figure 2b).

Establishing orthologous relationships: We next determined the level of conservation of caspase substrates using the strategy shown in Figure 3. The EggNOG database[23] was used to determine orthologous relationships between caspase substrates in human and those in mouse, *Drosophila* or *C. elegans*. We looked mainly at orthologous groups on the metazoan level (called meNOGs), as caspases are restricted to this lineage.[24]

The mouse, *Drosophila*, and *C. elegans* caspase substrate sets are remarkably highly enriched in proteins whose human orthologs are also known caspase substrates. Table 1 summarizes the results described here, and the full data, including the meNOG and human orthologs associated with each cleavage, are shown in Supplementary Table 1. A total of 216 mouse (84%), 64 *Drosophila* (49%), and 17 *C. elegans* (34%) caspase cleavage sites occur in proteins that have human orthologs that are caspase substrates. The sizes of the meNOG group overlaps between all four species are shown in Figure 4a. For 38 mouse, 36 *Drosophila*, and 22 *C. elegans* proteins, human orthologs were present in corresponding meNOGs, but these human proteins are not known to be caspase substrates. The remaining 2 mouse, 30 *Drosophila*, and 11 *C. elegans* proteins were either not present in any meNOG or were present in a meNOG containing no human members. Considering that the human caspase substrate set makes up only about 7% of the SwissProt human proteome, it is clear that the nonhuman data sets are highly enriched for proteins with human caspase substrates as orthologs.

Alignments: For meNOGs that contained a *Drosophila* or *C. elegans* caspase substrate and at least one human caspase substrate, we created multiple sequence alignments using the EINS-i algorithm from MAFFT[25] and examined them using Jalview 2.7[26] to determine relative placement of all caspase cleavage sites. The same procedure was used for mouse substrates, except in cases where the mouse substrate's meNOG contained only one human ortholog

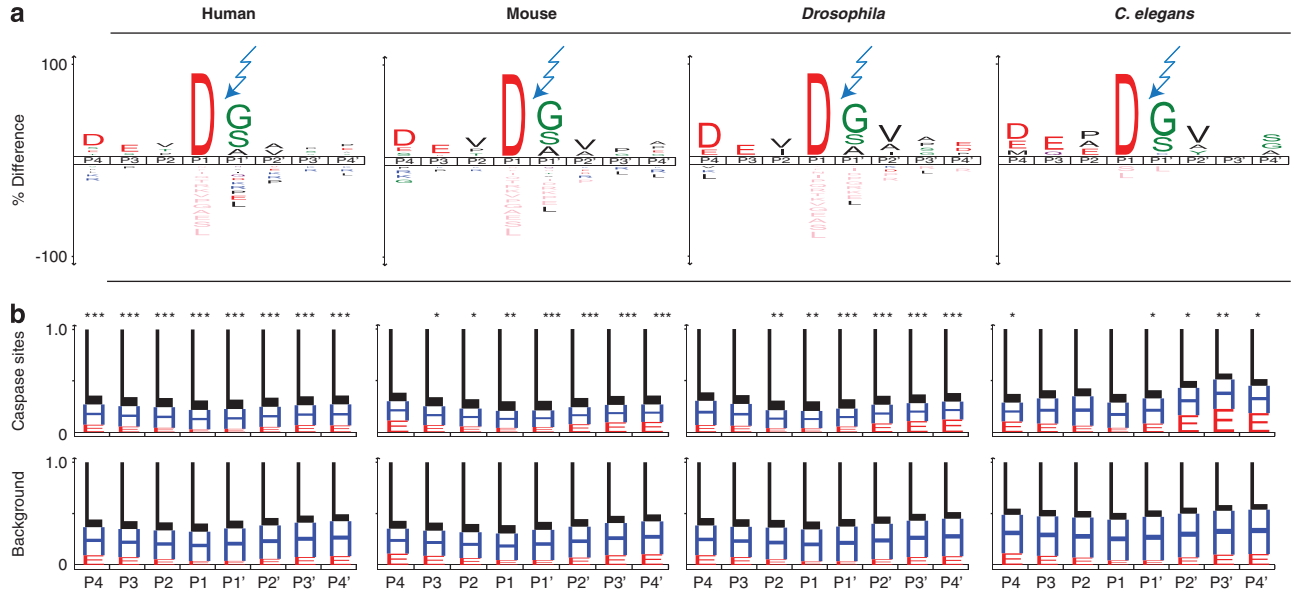

Figure 2 (**a**) IceLogo diagrams depicting primary structure preferences for the P4–P4′ residues for cleavages following aspartic acid in each of the four species. Letters above the axis indicate residues enriched over background, and letters below the axis indicate residues depleted with respect to background. (**b**) Secondary structure predictions for the P4–P4′ residues for cleavages following aspartic acid in each of the four species. L = loop, H = alpha helix, and E = beta sheet. The height of the letter indicates the fraction of sites with the corresponding secondary structure prediction, based on predictions using the NetSurfP server.[18] Top row represents caspase cleavage sites; bottom row represents all 8-mers with D at the P1 position in the same proteins. Asterisks indicate statistically significant enrichment of loops in caspase sites compared with background. *$P<0.05$. **$P<0.01$. ***$P<0.001$

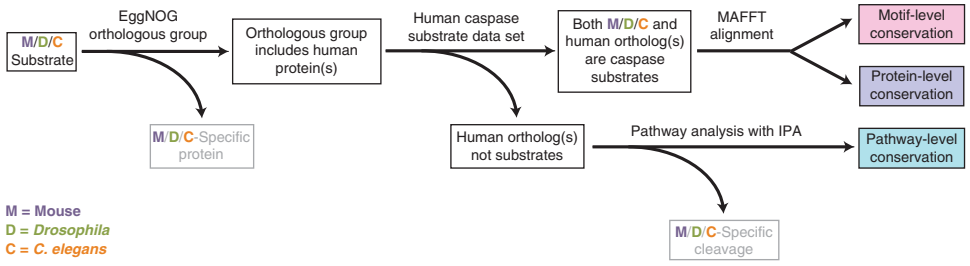

Figure 3 Data analysis pipeline. For each caspase cleavage observed in mouse, *Drosophila*, or *C. elegans*, we searched the EggNOG database for human orthologs. If the human ortholog found was also present in our human caspase substrate database, an alignment was created to determine whether the orthologs were cleaved at the same site or different sites. If the human ortholog found was not known to be a caspase substrate, we searched IPA's list of 'Canonical Pathways' to determine whether it functioned in any pathway(s) known to be enriched for caspase substrates. Each mouse, *Drosophila*, or *C. elegans* protein is thus assigned to one of five categories: (1) no human ortholog (species-specific protein), (2) human ortholog is not known to be a substrate (species-specific caspase cleavage), (3) pathway-level conservation, (4) protein-level conservation, or (5) motif-level conservation. Supplementary Table 1 shows all mouse, *Drosophila*, and *C. elegans* data, organized into these categories

Table 1 Summary of results

Species	Number of cell lines	Number of experiments	Proteins cut by caspases	Total caspase cuts	No human ortholog	Human ortholog not known substrate	Level of conservation		
							Pathway	Protein	Motif
Human	8	>50	1444	2021	—	—	—	—	—
Mouse	3	7	221	257	2	31	8	41	175
Drosophila	1	7	118	130	30	20	16	49	15
C. elegans	N/A	1	46	50	11	13	9	11	6

substrate. In these cases, most of which had extremely high sequence identity and aligned cleavage sites, we used pairwise ClustalW alignments performed on the UniProtKB website (www.uniprot.org).[27] In only a few of these cases did the cleavage sites not align; for these we repeated the alignment using MAFFT's EINS-i algorithm.

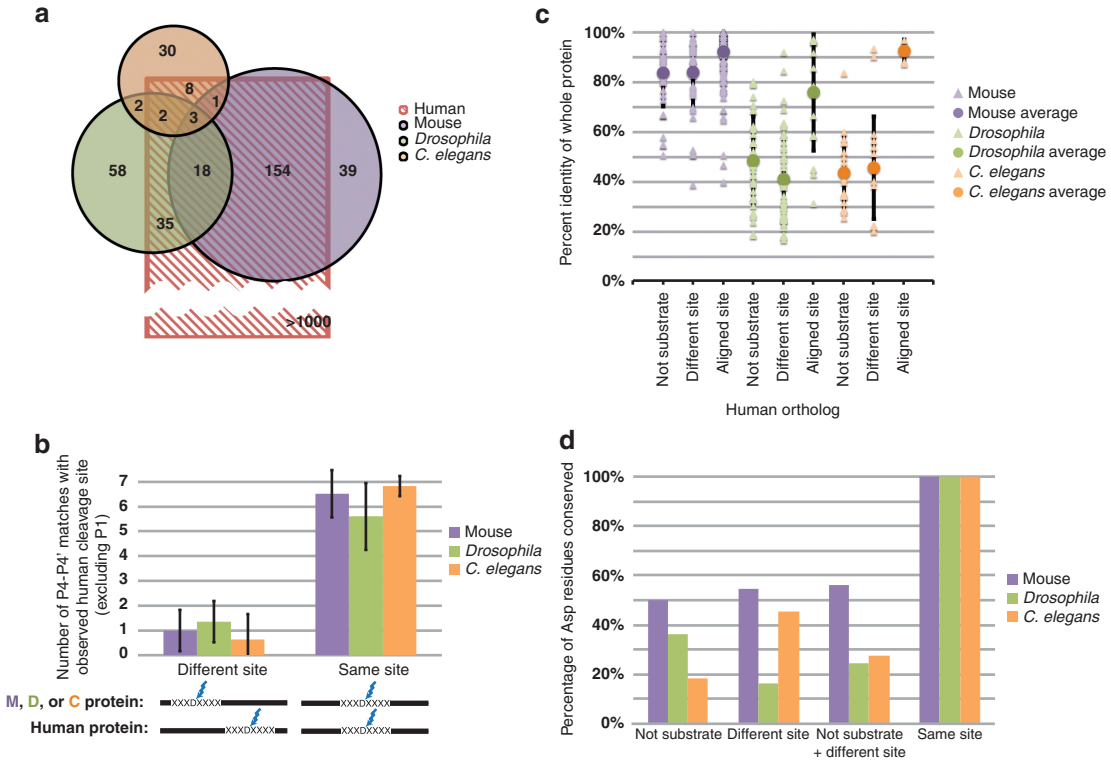

Figure 4 (**a**) Venn diagram indicating the overlap in the meNOGs associated with caspase substrates in human, mouse, *Drosophila*, and *C. elegans*. (**b**) When Mouse, *Drosophila*, or *C. elegans* caspase cleavage sites aligned with known cleavage sites in human orthologs, the aligned sites shared an average of between 5 and 7 identical residues (considering the P4, P3, P2, P2′, P3′, and P4′ positions, as the P1 position is fixed to Asp). In contrast, when a mouse, *Drosophila*, or *C. elegans* protein has a human ortholog with a cleavage site at a different location, the two observed cleavage sites share only an average of between zero and two identical residues (considering the same seven positions). Error bars represent S.D. (**c**) The protein %ID was calculated for pairwise alignments of each mouse, *Drosophila*, or *C. elegans* protein with its closest human ortholog. Pairwise alignments were calculated in Jalview. Triangles represent individual pairwise alignments, while large circles represent the average %ID for each category, with error bars representing the S.D. In some cases, datapoints in the 'not substrate' category may be false negatives (resulting from our human data set being less than fully complete). (**d**) In cases where the human orthologs were either not known substrates, or were cut at different sites, the Asp residue was conserved between 15 and 65% of the time

We found that 176 of 217 mouse sites, 15 of 64 *Drosophila* sites, and 6 of 17 *C. elegans* sites aligned perfectly with cleavage sites observed in their human orthologs, indicating motif-level conservation. In the remaining cases, the human site was not aligned. For both groups, we compared the P4-P2 and P1′-P4′ sequences for the observed mouse, *Drosophila*, or *C. elegans* cleavage site with that of the observed (either aligned or unaligned) human cleavage site. We found that for substrates with motif-level conservation, pairs of cleavage sites shared an average of more than 5.5 out of 7 identical residues (Figure 4b). In contrast, for substrates with different sites, pairs shared an average of less than 1.5 out of 7 identical residues, suggesting that the majority of sites that are not aligned evolved separately and are not simply the result of one site shifting in the sequence.

We also looked at overall protein sequence similarity between our mouse, *Drosophila*, and *C. elegans* substrates and their human orthologs, using pairwise alignments calculated with Jalview. Not surprisingly, proteins with higher percent identity were more likely to share an aligned cleavage site (Figure 4c).

In cases where the human ortholog either was not a substrate or was cleaved at a different site from the mouse, *Drosophila*, or *C. elegans* site, we asked whether the non-cleaved aligned site contained the required Asp at the P1 position (Supplementary Table 1 and Figure 4d). The Asp was conserved in less than 30% of cases for both *Drosophila* and *C. elegans*, and less than 60% for mouse. For these cases, we cannot rule out the possibility that the human orthologs are true caspase substrates, but have never been observed (perhaps due to low expression levels in the cell lines used, poor detectability of the peptide, or other factors). But in cases where the essential P1 Asp residue is not conserved in human we are confident in categorizing these cleavage events as having been lost or moved during the course of evolution.

Pathway and function analysis: Our finding that caspase substrate sets are more broadly conserved at the protein level than the motif level led us to ask whether we could detect even broader conservation at the level of functional pathways (Figure 3). It is known that caspase substrates are more likely to occur in protein–protein complexes with each other than are a randomly selected set of proteins.[4] We reasoned that in many cases caspases could achieve the same functional effect by cleaving one (or more) of several proteins in the same complex or pathway. We used the highly

curated IPA (Ingenuity Pathway Analysis, Ingenuity Systems, Redwood City, CA, USA) analysis software to generate a list of predicted 'Canonical Pathways' that were significantly enriched in our human caspase substrate data set (Supplementary Table 2). We asked whether any of the non-substrate human orthologs of mouse, *Drosophila*, or *C. elegans* substrates were members of these pathways. Eight out of 39 mouse, 16 out of 36 *Drosophila*, and 9 out of 22 *C. elegans* substrates did indeed have non-substrate human orthologs that were members of one or more of these pathways (Supplementary Table 1). Although these proteins had no direct human orthologs known to be substrates, they are conserved at the pathway level. Many of these pathways are highly conserved and involved in essential processes such as translation or proteasomal degradation.

One example is the highly conserved EIF2 Signaling pathway (Supplementary Figure 1). In many cases within this pathway, caspases in more than one species cut the same protein; in other cases, different proteins in the same complex are cut in different species. For example, one component of the eIF4γ complex, EIF4G2, is a substrate in *Drosophila*; the same protein, plus two other members of this complex (EIF4G1 and EIF4G3) are also known human substrates. In the 60 S ribosomal subunit, RPL27 and RPL4 are substrates in *Drosophila*, but are not known substrates in human. Instead, three different 60S ribosomal subunit proteins, RPL5, RPL17, and RPLP2, are known substrates in human, and RPL5 is a known substrate in mouse.

To examine the important biological functions of caspase-cleaved proteins, we looked at GO term assignments for the caspase substrates of each organism. In all four species, there was enrichment in GO terms related to the nucleus, the cytoskeleton, transcription, and nucleic acid binding, consistent with previous GO term analysis on human caspase substrates[4] (Supplementary Table 3).

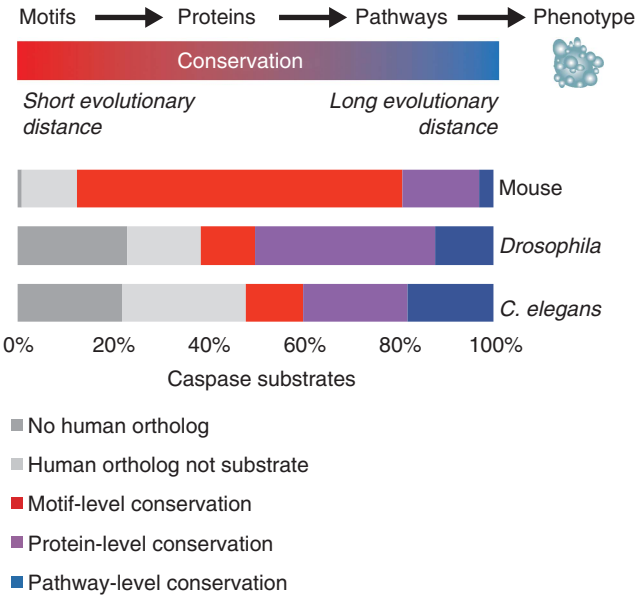

Figure 5 Results of this study show that caspases tend to recognize and cleave particular motifs only across short evolutionary distances (represented by the human–mouse comparison, spanning $<$100 million years of evolution[37]), but that the same proteins will remain targets across longer distances, and the same pathways over even longer distances (represented by the human-*Drosophila* and human-*C. elegans* comparisons, both represented by roughly 600 million years of evolution[7]). Ultimately, the phenotype of apoptosis is conserved across the whole metazoan lineage. A small number of false negatives resulting from our human data set being incomplete would likely only serve to shift all trends depicted in this figure towards higher overall conservation

Discussion

The apoptosis pathway appears to be evolutionarily conserved, both phenotypically and molecularly, across all metazoans studied to date, indicating that its basic features were likely present in the last common metazoan ancestor.[6,24] While the triggers for apoptosis vary depending on species, cell type, and environmental conditions, all apoptotic pathways in all species studied lead eventually to caspase activation, followed by dismantling of the apoptotic cell in a non-inflammatory manner. Here, we begin to globally determine apoptotic caspase cleavage sites in three non-human model metazoans and investigate the evolutionary conservation of the common caspase stage of apoptosis.

Our human data set, collected over six years, contains 1444 protein targets; the mouse, *Drosophila*, and *C. elegans* data sets are roughly an order of magnitude smaller, reflecting a smaller number of experiments on these species. Furthermore, eight different human cell lines were examined, compared with three mouse cell lines and one *Drosophila* cell line. Despite the incomplete sampling of the non-human metazoans, we can make meaningful positive comparisons between these smaller sets against the much more complete human data set because the low false positive rate ($<$3%, data not shown) provides high confidence identification.

All comparative proteomics studies come with a risk of bias towards high-abundance proteins. High abundance tends to be positively correlated with high conservation,[28] so sampling bias towards high abundance proteins might yield a false result of high conservation. However, our experimental technique reduces the impact of this problem. By enriching samples at the tryptic peptide level (leaving only one or a few peptides per protein), our N-terminal labeling method decreases the complexity of the peptide mixture further than traditional methods, diminishing the chance of low abundance proteins being masked by extremely high abundance proteins like actin. In fact, studies using our method in human serum detected proteins with known abundance values spanning six orders of magnitude.[29] Using the PaxDB database,[12,13] we have established high coverage of the abundance range of the caspase substrates identified in our cell line-based human reference data set (Figure 1c). Therefore, we are confident that abundance bias does not have a large role in our results.

Our results suggest a hierarchical arrangement of conservation, depicted in Figure 5. Caspases target a specific set of pathways across metazoans in order to complete apoptosis. In some cases, caspases in different species target different proteins within the same pathways. In other cases, the same protein remains a caspase substrate across a broad lineage, and in still other cases, caspases even target the same cleavage site motif in different species. Our data are

Table 2 Subset of Drosophila and *C. elegans* Caspase Substrates with no Human Ortholog

Species	Acc #	SwissProt ID	Description	Function
Drosophila	Q95RU0	CUE_DROME	Protein cueball	Spermatogenesis
	Q8MRY4	—	SD13619p	Spermatogenesis
	Q8T044	—	LD29665p	Spermatogenesis
	Q9VFE6	RRP15_DROME	RRP15-like protein	Neurogenesis
	Q9VAF4	—	Dim gamma-tubulin 1	Neurogenesis
	Q7K126	—	LD13864p	Neurogenesis
	Q09024	IMPL2_DROME	Neural/ectodermal development factor IMP-L2	Embryogenesis
	Q9VYV4	—	CG2446, isoform A	Development, various
	Q9VI56	—	CG1943, isoform A	Wing disc development
	Q24478	CP190_DROME	Centrosome-associated zinc finger protein CP190	Gypsy chromatin insulator complex
	P08970	SUHW_DROME	Protein suppressor of hairy wing	Gypsy chromatin insulator complex
	Q86B87	MMD4_DROME	Modifier of mdg4	Gypsy chromatin insulator complex
C. elegans	P05690	VIT2_CAEEL	Vitellogenin 2	Embryogenesis (extracellular)
	P06125	VIT5_CAEEL	Vitellogenin 5	Embryogenesis (extracellular)
	Q18823	LAM2_CAEEL	Laminin-like protein lam-2	Embryogenesis (extracellular)
	Q17796	—	hgrs-1	Embryogenesis
	Q22469	—	Protein T13H5.4	Embryogenesis

consistent with previous observations that short linear motifs are poorly conserved over evolution,[8] but also show that loss of the short linear motif recognized by caspases is not likely to stop the same protein from being targeted.

Our study has yielded similar conservation patterns to those observed for phosphorylation and transcription factors. Tan *et al.*[9] used mass spectrometry to identify phosphorylation sites in *Drosophila*, *C. elegans*, and yeast, and compared them with known human sites. They found that the set of modifications conserved at the motif level was small, but networks defining relationships between kinases and substrates were conserved more broadly. Additional computational phosphorylation studies have also revealed that motif-level conservation is weaker than protein-level conservation.[30,31] Studies of transcription factor binding in yeast show a somewhat similar pattern: while DNA binding sites are only modestly conserved, overall regulatory networks and the functions they impact are retained across hundreds of millions of years of evolution.[32–34] Notably, three studies of single *Drosophila* caspase substrates have also demonstrated protein level (but not motif level) conservation.[35–37]

Caspases, like all enzymes that catalyze specific post-translational modifications, face the challenge of recognizing their particular targets in the complex cellular milieu. How caspases achieve this is not well understood. Executioner caspases, like human caspase-3, are functionally polyspecific, meaning that they recognize several otherwise unrelated substrates, and yet are restrained from targeting all Asp sites in all proteins in a digestive manner.[6] The degeneracy of the physiological cleavage site consensus sequence seen in all four organisms (Figure 2a), plus the higher conservation at the protein level than the motif level, suggest that other factors, such as exosite binding,[38,39] may influence what is cleaved in cells.

We examined UniProtKB, Flybase,[40] and Wormbase (www.wormbase.org, release WS229) annotations for the 62 *Drosophila* and 19 *C. elegans* caspase substrate proteins that are specific to those species, having either no human orthologs, or human ortholog(s) not known to be substrates. Many have little or no functional annotation, but several in both species are either known or predicted to be involved in development (Table 2). In *Drosophila*, three of these proteins function in spermatogenesis, three in neurogenesis, and three in other types of development. Five *C. elegans*-specific proteins are either known or predicted to be involved in embryo development. Apoptosis is an important aspect of development in all species, but further studies are needed to determine whether these cleavages are relevant to developmental apoptosis. Three of the five *C. elegans* development proteins are known to be extracellular. It is possible that the caspase cleavages of these proteins are not physiologically relevant, but result from the *C. elegans* extract being made from whole bodies, rather than a cell line.

Another interesting finding is that three proteins cleaved in *Drosophila* are all members of the gypsy chromatin insulator complex. This complex regulates accessibility of certain chromosomal regions,[41] so these cleavages are in accordance with the GO analysis showing the tendency for caspases to cleave chromatin-associated complexes (Supplementary Table 3).

In aggregate, these studies represent a unique systematic comparison of apoptotic caspase substrates across metazoans, shedding light on what is functionally most important. In all three non-human species, at least half of the substrates identified are conserved with human substrates on either the motif, protein, or pathway level, and given that our human caspase substrate data set is likely not comprehensive, this could be an underestimate. This finding supports the view that a substantial fraction of caspase cleavages do serve important functions that have been conserved by natural selection across 600 million years of metazoan evolution.

Materials and Methods

Cell lines. Mouse Embryonic Fibroblasts (MEFs) were a kind gift from Dr. Richard Flavell. TK-1 and A20 cells were purchased from ATCC (Manassas, VA, USA) and cultured in the recommended media with 10% FBS at 37 °C. When

cells were between 1×10^6 and 2×10^6 cells/ml, one of four apoptosis-inducing agents was added (bortezomib, doxorubicin, etoposide or staurosporine). Cells were harvested by centrifugation after 4–24 h, washed once in 1 × PBS, and then frozen at −80 °C as pellets.

Drosophila S2 cells were purchased from ATCC and cultured in Schneider's *Drosophila* Medium from Gibco (Carlsbad, CA, USA) with 10% FBS at 26 °C. At a cell density of 2×10^6 cells/ml, one of three apoptosis-inducing agents was added (doxorubicin, cycloheximide, or actinomycin D). Cells were harvested by centrifugation after 5–24 h, washed once in 1 × *Drosophila* PBS (pH 6.7), and then frozen at −80 °C as pellets.

Both mouse and *Drosophila* cell lysates were formed by resuspending pellets in an SDS solution containing protease inhibitors (including the pan-caspase inhibitor z-VAD-fmk, to prevent any post-harvest caspase activity) and then sonicating. The N-terminal labeling reaction and mass spectrometry prep was performed as described for apoptotic cells.[14]

***C. elegans*.** Bristol N2 worms were obtained from Dr. Cynthia Kenyon at UCSF. They were grown in 250 ml volumes of S Complete Media liquid culture containing streptomycin and the fungicide carbendazim, with daily additions of streptomycin-resistant OP50 *E. coli*. When worm density reached 100 worms/μl (without regard to worm size or age), and the *E. coli* was visibly depleted from the culture flasks, the cultures were harvested by centrifugation, washed with M9 buffer, and separated from left over *E. coli* by centrifuging in 30% sucrose. They were immediately removed from the sucrose, washed three times with M9 buffer, and incubated on a nutator at room temperature for 30 min to allow for digestion of residual bacteria present in the intestine. They were then drip-frozen in liquid nitrogen, and stored at −80 °C. The frozen worms were homogenized with a large mortar and pestle kept cold with liquid nitrogen. When substantial breakage of worm bodies was observed under a light microscope, the homogenate was collected and mixed with just enough M9 buffer (containing protease inhibitors PMSF, AEBSF, IAM and EDTA) to form a slurry. The slurry was sonicated at 4C, and then passed through low and then high gauge needles. Finally, the cuticles and other insoluble debris were removed by centrifugation, and the extract frozen at −80 °C until needed. Later, the extract was labeled and processed for mass spectrometry as described for *in vitro* cleavage assays.[14]

The *ced-3* gene, with the prodomain (residues 1–220) removed, was cloned from a *C. elegans* cDNA library (kindly provided to us by Dr. Aimee Kao). The construct was inserted into various vectors with different configurations of His6, GST, and MBP tags at either end. We found in all cases that CED-3 solubility was extremely limited, with the majority of the enzyme lost in the *E. coli* pellet after lysis by microfulidization or sonication. The best yields came from expression in pPAL7 vector with N-terminal HIS and GST tags, and a single HIS column purification. As further purification methods, and efforts at concentration, led to increased losses, we chose to use a low concentration, semi-pure CED-3 extract for our proteomics experiments. As we did not produce enough enzyme to complete an active site titration, we estimated the concentration of active CED-3 by relating its activity on the peptide substrate DEVD-afc to the activity of a pure human caspase-3 sample of known concentration. The CED-3 extract had a V_{max} of 0.075 μmol/s and a K_M of 8.4 μM, which is equivalent to a 200 nM sample of caspase-3 purified in our lab (with k_{cat}/K_M of 0.3 $M^{-1}s^{-1}$). The CED-3 extract was added to the worm preparation at a 1:10 volume ratio, so the final concentration of CED-3 was equivalent to 20 nM human caspase-3 by activity.

Mass spectrometry. All mass spectrometry data were collected in the UCSF Mass Spectrometry Facility on a QStar Elite from Applied Biosystems (Carlsbad, CA, USA), with the exception of one mouse data set, which was collected on a QExactive instrument kindly made accessible to us by Thermo Scientific (Waltham, MA, USA). Data were searched with Protein Prospector (UCSF Mass Spectrometry Facility). For the human and mouse data, the search library contained only proteins from the SwissProt database. For *Drosophila* and *C. elegans*, the highly curated SwissProt database is limited in size, so the broader UniProtKB library was searched instead. For these two organisms, if one peptide could be mapped to either a SwissProt entry or a non-SwissProt UniProtKB entry, then only the SwissProt entry was considered.

Abundance analysis. Data on protein abundance were taken from the integrated human data set from PaxDB version 2.0 (available in the archive section at www.pax-db.org).[13]

Primary and secondary structure analyses. Primary structure logos were generated using IceLogo[16] with the appropriate SwissProt or UniProtKB database as background. Secondary structure predictions were determined using NetSurfP.[22] Secondary structure background was based on all 8-mers with Asp in the fourth position (equivalent to P1 position) found in the same set of proteins. The logos for these data were created using the 'Filled Logo' option on the IceLogo server, and significance was determined by χ^2 tests.

Orthology and alignments. The EggNOG database was built using non-supervised clustering methods to assign proteins from 1133 species to over 700 000 orthologous groups. Closely related paralogs in the same species are all represented in the same orthologous group, meaning that in some cases a single protein in one species has more than one ortholog in another species. EggNOG's metazoan orthologous groups (meNOGs) containing the *Drosophila* and *C. elegans* proteins found in this study were retrieved from the downloads area of the EggNOG 3.0 website.[23] Each meNOG, if it contained a human protein known to be a caspase substrate, was aligned using the EINS-i algorithm from MAFFT.[25] Trees derived from these alignments were examined by eye to confirm that sequences were well distributed across the metazoan phylogeny. Alignments were then displayed and analyzed using Jalview. For each mouse, *Drosophila* or *C. elegans* cleavage site, we assessed four points: (1) whether it aligned with a caspase cleavage site in a human ortholog, (2) the number of residues in common between it and the most similar human caspase cleavage site in the P4–P2 and P1′–P4′ positions, (3) if the sites did not align, whether the P1 Asp of the mouse, *Drosophila*, or *C. elegans* cleavage site was conserved in any human ortholog, and finally (4) the pairwise percent identity between the mouse, *Drosophila* or *C. elegans* protein and its closest human ortholog. This was calculated with the pairwise alignment tool in Jalview 2.7, which uses the BLOSUM62 matrix and gap opening and extending penalties of 12 and 2, respectively.

In some cases, the mouse, *Drosophila*, or *C. elegans* peptide discovered in our mass spectrometry experiments was matched to more than one protein by the Protein Prospector program. We checked each of these peptides individually, and determined that in all cases, the results presented here were the same no matter which protein was chosen.

Pathway analysis. The set of 1444 human caspase substrates was uploaded to Ingenuity Pathway Analysis (IPA) software. A 'Core Analysis' was performed to generate a list of all 'Canonical Pathways' whose genes were significantly over-represented in the human caspase substrate data set, with P values < 0.05 (Supplementary Table 2). We then took the list of non-substrate human orthologs for each of the three other species and checked the IPA database to see which, if any, canonical pathways they were associated with.

GO term enrichment was determined using the GO::TermFinder software.[42] Lists of unique proteins were created for each species based on the discovered apoptotic peptides. The data sets were uploaded to the database and tested for enrichment against a background of the organism's SwissProt (mouse) or UniProtKB (*Drosophila* and *C. elegans*) database. The mouse, *Drosophila* and *C. elegans* tests used all GO evidence codes, and the human test used all evidence codes except ND (No biological data available) and IEA (Inferred from Electronic Annotation). Enriched terms in human sets were defined as those with corrected *P*-values less than 0.01. The human set was then filtered to remove terms that were not statistically significantly different from untreated background (data not shown, manuscript in preparation).

The significant GO terms in mouse, *Drosophila* and *C. elegans* were compared with the top human terms in each ontology. The fold enrichment was calculated for each significant term; the percentage of proteins annotated with each term in the experimental set was divided by the percentage of proteins annotated with each term in the proteome background.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements. We thank Sami Mahrus, David Wildes, Nicholas Agard, Hikari Yoshihara, Huy Nguyen, Gerald Hsu, and Kazutaka Shimbo for contributing to the human caspase substrate data set. We also thank Cynthia Kenyon and Aimee Kao for assistance with the *C. elegans* work, and David Maltby, Jonathan Trinidad, and Shenheng Guan and the rest of the UCSF Mass Spectrometry Facility, plus

Iman Mohtashemi and Yan Chen of Thermo Scientific, for mass spectrometry assistance. We thank Bent Petersen for assistance with the NetSurfP server. This project was supported by a grant from the UCSF Stephen and Nancy Grand Multiple Myeloma Translational Initiative. It was also supported by the National Science Foundation GRFP (EDC), NIH Training Grant T32 GM007175 (JES and PCB), the Pharmaceutical Research and Manufacturers of America (AEB), National Institutes of Health R01 GM60595 (PCB), and National Institutes of Health R01 GM081051 (JAW). Mass spectrometry was performed at the Bio-Organic Biomedical Mass Spectrometry Resource at UCSF (ALB, Director), which is supported by grants from the National Center for Research Resources (5P41RR001614) and the National Institute of General Medical Sciences (8 P41 GM103481) from the National Institutes of Health.

1. Zmasek CM, Zhang Q, Ye Y, Godzik A. Surprising complexity of the ancestral apoptosis network. *Genome Biol* 2007; **8**: R226.
2. Van Damme P, Martens L, Van Damme J, Hugelier K, Staes A, Vandekerckhove J *et al.* Caspase-specific and nonspecific in vivo protein processing during Fas-induced apoptosis. *Nat Methods* 2005; **2**: 771–777.
3. Dix MM, Simon GM, Cravatt BF. Global mapping of the topography and magnitude of proteolytic events in apoptosis. *Cell* 2008; **134**: 679–691.
4. Mahrus S, Trinidad JC, Barkan DT, Sali A, Burlingame AL, Wells JA. Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. *Cell* 2008; **134**: 866–876.
5. Rawlings ND, Barrett AJ, Bateman A. MEROPS: the peptidase database. *Nucleic Acids Res* 2010; **38**: D227–D233.
6. Crawford ED, Wells JA. Caspase substrates and cellular remodeling. *Annu Rev Biochem* 2011; **80**: 1055–1087.
7. Erwin DH, Davidson EH. The last common bilaterian ancestor. *Development* 2002; **129**: 3021–3032.
8. Neduva V, Russell R. Linear motifs: evolutionary interaction switches. *FEBS Lett* 2005; **579**: 3342–3345.
9. Tan C, Bodenmiller B, Pasculescu A. Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. *Sci Signal* 2012; **2**: ra39.
10. Shimbo K, Hsu GW, Nguyen H, Mahrus S, Trinidad JC, Burlingame AL *et al.* Quantitative profiling of caspase-cleaved substrates reveals different drug-induced and cell-type patterns in apoptosis. *Proc Nat Acad Sci USA* 2012; **109**: 12432–12437.
11. Schechter I, Berger A. On the size of the active site in proteases. I. Papain. *Biochem Biophys Res Commun* 1967; **27**: 157–162.
12. Schrimpf S, Weiss M, Reiter L, Ahrens C. Comparative functional analysis of the Caenorhabditis elegans and Drosophila melanogaster proteomes. *PLoS Biology* 2009; **7**: 616–627.
13. Weiss M, Schrimpf S, Hengartner M. Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome. *Proteomics* 2010; **10**: 1297–1306.
14. Agard NJ, Mahrus S, Trinidad JC, Lynn A, Burlingame AL, Wells JA. Global kinetic analysis of proteolysis via quantitative targeted proteomics. *Proc Nat Acad Sci USA* 2012; **109**: 1913–1918.
15. Taylor RC, Brumatti G, Ito S, Hengartner MO, Derry WB, Martin SJ. Establishing a blueprint for CED-3-dependent killing through identification of multiple substrates for this protease. *J Biol Chem* 2007; **282**: 15011–15021.
16. Colaert N, Helsens K, Martens L, Vandekerckhove J, Gevaert K. Improved visualization of protein consensus sequences by iceLogo. *Nat Methods* 2009; **6**: 786–787.
17. Thornberry N, Rano T, Peterson E, Rasper D, Timkey T, Garcia-Calvo M *et al.* A combinatorial approach defines specificities of members of the caspase family and granzyme B. *J Biol Chem* 1997; **272**: 17907.
18. Lüthi AU, Martin SJ. The CASBAH: a searchable database of caspase substrates. *Cell Death Differ* 2007; **14**: 641–650.
19. Schilling O, Overall C. Proteome-derived database-searchable peptide libraries for identifying protease cleavage sites. *Nat Biotechnol* 2008; **26**: 685–694.
20. Chang TK, Jackson DY, Burnier JP, Wells JA. Subtiligase: a tool for semisynthesis of proteins. *Proc Natl Acad Sci USA* 1994; **91**: 12544–12548.
21. Fontana A, de Laureto PP, Spolaore B, Frare E, Picotti P, Zambonin M. Probing protein structure by limited proteolysis. *Acta Biochim Pol* 2004; **51**: 299–321.
22. Petersen B, Petersen T, Andersen P, Nielsen M, Lundegaard C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct Biol* 2009; **9**: 51.
23. Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J *et al.* eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res* 2012; **40**: D284–D289.
24. Lamkanfi M, Declercq W, Kalai M, Saelens X, Vandenabeele P. Alice in caspase land. A phylogenetic analysis of caspasesfrom worm to man. *Cell Death Differ* 2002; **9**: 358–361.
25. Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* 2008; **9**: 286–298.
26. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 2009; **25**: 1189–1191.
27. Consortium U. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* 2012; **40**: D71–D75.
28. Pál C, Papp B, Lercher MJ. An integrated view of protein evolution. *Nat Rev Genet* 2006; **7**: 337–348.
29. Wildes D, Wells JA. Sampling the N-terminal proteome of human blood. *Proc Nat Acad Sci USA* 2010; **107**: 4561–4566.
30. Beltrao P, Trinidad JC, Fiedler D, Roguev A, Lim WA, Shokat KM *et al.* Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. *PLoS Biol* 2009; **7**: e1000134.
31. Holt LJ, Tuch BB, Villén J, Johnson AD, Gygi SP, Morgan DO. Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science* 2009; **325**: 1682–1686.
32. Tsong AE, Tuch BB, Li H, Johnson AD. Evolution of alternative transcriptional circuits with identical logic. *Nature* 2006; **443**: 415–420.
33. Tuch BB, Galgoczy DJ, Hernday AD, Li H, Johnson AD. The evolution of combinatorial gene regulation in fungi. *PLoS Biol* 2008; **6**: e38.
34. Booth LN, Tuch BB, Johnson AD. Intercalation of a new tier of transcription regulation into an ancient circuit. *Nature* 2010; **468**: 959–963.
35. Creagh E, Brumatti G, Sheridan C, Duriez P, Taylor R, Cullen S *et al.* Bicaudal is a conserved substrate for drosophila and mammalian caspases and is essential for cell survival. *PLoS One* 2009; **4**: 3.
36. Yokoyama H, Mukae N, Sakahira H, Okawa K, Iwamatsu A, Nagata S. A novel activation mechanism of caspase-activated DNase from Drosophila melanogaster. *J Biol Chem* 2000; **275**: 12978–12986.
37. Amarneh B, Matthews KA, Rawson RB. Activation of sterol regulatory element-binding protein by the caspase Drice in Drosophila larvae. *J Biol Chem* 2009; **284**: 9674–9682.
38. Fuentes-Prior P, Salvesen GS. The protein structures that shape caspase activity, specificity, activation and inhibition. *Biochem J* 2004; **384**(Pt 2): 201–232.
39. Boucher D, Blais V, Denault J-B. Caspase-7 uses an exosite to promote poly(ADP ribose) polymerase 1 proteolysis. *Proc Nat Acad Sci USA* 2012; **109**: 5669–5674.
40. McQuilton P St, Pierre SE, Thurmond J, Consortium F. FlyBase 101–the basics of navigating FlyBase. *Nucleic Acids Res* 2012; **40**: D706–D714.
41. Wallace JA, Felsenfeld G. We gather together: insulators and genome organization. *Curr Opin Genet Dev* 2007; **17**: 400–407.
42. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM *et al.* GO TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* 2004; **20**: 3710–3715.

Supplementary Information accompanies the paper on Cell Death and Differentiation website (http://www.nature.com/cdd)