CMLS, Cell. Mol. Life Sci. 60 (2003) 446–462
1420-682X/03/030446-17


CMLS Cellular and Molecular Life Sciences

Review

# Structure, function and evolution of haspin and haspin-related proteins, a distinctive group of eukaryotic protein kinases

**J. M. G. Higgins**

Division of Rheumatology, Immunology and Allergy, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Smith Building, Room 538D, One Jimmy Fund Way, Boston, Massachusetts 02115 (USA), Fax + 1 617 525 1010, e-mail: jhiggins@rics.bwh.harvard.edu

Received 16 August 2002; received after revision 31 August 2002; accepted 9 September 2002

**Abstract.** The haspins constitute a newly defined protein family containing a distinctive C-terminal eukaryotic protein kinase domain and divergent N termini. Haspin homologues are found in animals, plants and fungi, suggesting an origin early in eukaryotic evolution. Most species have a single haspin homologue. However, *Saccharomyces cerevisiae* has two such genes, while *Caenorhabditis elegans* has at least three haspin homologues and approximately 16 haspin-related genes. Mammalian haspin genes have features of retrogenes and are strongly expressed in male germ cells and at lower levels in some somatic tissues. They encode nuclear proteins with serine/threonine kinase activity. Murine haspin is reported to inhibit cell cycle progression in cell lines. One of the *S. cerevisiae* homologues, *ALK1,* is a member of the *CLB2* gene cluster that peaks in expression at M phase and thus may function in mitosis. Therefore, the haspins are an intriguing group of kinases likely to have important roles during or following both meiosis and mitosis.

**Key words.** Serine/threonine kinase; nucleus; testis; spermatogenesis; retrogene; nested gene; lineage-specific expansion; Gsg2.

## Introduction

The haspins are a recently described group of proteins with several interesting features. They contain a distinctive C-terminal kinase domain and form a newly defined subset within the eukaryotic protein kinase superfamily. The existence of at least one haspin homologue in all nearly complete sequenced eukaryotic genomes suggests an important function for these proteins, particularly since protein kinases are known to be critical regulators of a wide variety of cellular functions. Following an apparent origin early in eukaryotic evolution, the haspin family has undergone a marked expansion in at least one lineage, that of the nematode worm *Caenorhabditis elegans.* The haspin gene in mammals appears to have been retrotransposed into a preexisting gene to produce a rare example of a functional intronic retrogene. This review serves to highlight the existence of this new group of proteins, to describe their sequence features and to bring together information from various studies that have begun to address the expression patterns and functions of haspin homologues in a variety of species.

## Discovery of haspin

The first member of the haspin family to be characterized was from the mouse. Tanaka et al. [1] used subtractive hybridization to isolate cDNAs expressed in normal adult mouse testis but not in the testis of 4-month-old mutant

W/W$^v$ mice. Because the W/W$^v$ mice lacked germ cells [2], the transcripts isolated were those expressed in male germ cells but not in the supporting cells of the testis. One such clone was given the name Germ cell-specific gene 2 (Gsg2) [3]. Based on its expression pattern and analysis of the encoded protein (see below), the gene product was named haploid germ cell-specific nuclear protein kinase or *haspin*. Later, Higgins [4] and Tanaka et al. [5] cloned a human haspin homologue. Genes or transcripts encoding hypothetical haspin-like proteins have subsequently been identified in a wide variety of eukaryotes including various mammals (*Sus scrofa*, *Bos taurus*, *Rattus norvegicus*), the frog *Xenopus laevis*, fish (*Fugu rubripes*, *Tetraodon nigroviridis*, *Danio rerio*), invertebrates (*C. elegans*, *Drosophila melanogaster*, *Anopheles gambiae*), plants (*Arabidopsis thaliana*, *Oryza sativa*), fungi (*Schizosaccharomyces pombe*, *Saccharomyces cerevisiae*, *Aspergillus fumigatus*) and the microsporidia (*Encephalitozoon cuniculi*), but not in any of the sequenced prokaryotic or archaea genomes [6, 7; Higgins, unpublished data].

## Sequence features

The open reading frames of the murine and human haspin cDNAs encode proteins of 754 and 798 amino acids, respectively. The two proteins are 66% identical overall, but this homology is not evenly distributed along the polypeptide. The C-terminal portion is more highly conserved between the species: the two proteins are 83% identical from residue 484 to 798 and no insertions or deletions occur [4, 5]. Hypothetical genes in many other eukaryotic species encode proteins with homology to the mammalian haspins in the C-terminal domain, but with divergent N-terminal domains [6]. Various lines of evidence, discussed below, suggest that the relatively conserved C-terminal region comprises a eukaryotic protein kinase domain (fig. 1A).

## The kinase domain

Most known protein kinases in eukaryotes are members of the eukaryotic protein kinase superfamily (ePKs), although members of this class are also found in bacteria and archaea [8, 9]. The crystal structures of a number of ePK domains reveal a common bi-lobed structure [10, 11], and the conserved sequence elements of the 250- to 300-amino-acid domain have been labeled regions I–XI [12, 13]. The smaller N-terminal lobe (regions I–V) contains conserved residues required for nucleotide binding, while the larger C-terminal lobe (regions V–XI) is involved in substrate binding and initiating phosphotransfer. The cleft between the lobes contains further invariant amino acids that form the active site. Tanaka et al. [3] first noted the sequence similarity between murine haspin and regions I–III of cyclin-dependent kinases (fig. 1B). Later, comparison of haspin genes from a variety of species identified motifs corresponding to regions VIb, VII and IX of the consensus ePK domain in the haspin proteins. Secondary-structure predictions were consistent with the presence of structural components of the remaining regions IV, VIa, X and XI [6] (figs 1B, 2). Therefore, the C-terminal region of haspin likely folds to give the familiar bi-lobed structure of known protein kinases. This conclusion is compelling given the finding that phospatidylinositide kinases [14], phosphatidylinositol phosphate kinases [15] and aminoglycoside phosphotransferases [16] have three-dimensional structures similar to the ePKs even though they have barely detectable sequence homology. Of course, this prediction does not rule out the possibility of differences in secondary structure or orientation of some elements, particularly in the large lobe.

Importantly, there is evidence that the murine and human haspin proteins do in fact possess protein kinase activity. Kinase activity could be detected in an immunoprecipitate of haspin from a murine testis preparation [3], although in this case, the presence of other proteins in the sample meant that the activity could not be confidently ascribed to haspin itself. Furthermore, both human and murine haspin, when expressed as EGFP-fusion proteins in HEK293 cells, undergo apparent autophosphorylation in in vitro kinase assays [3, 5]. Phosphoamino acid analysis of hydrolyzed murine EGFP-haspin indicated that the phosphorylation occurred on serine and threonine residues. A mutant form of the murine EGFP-haspin protein with a 10-amino-acid deletion in region I of the kinase domain (that would be expected to abolish kinase activity) did not become phosphorylated, providing evidence that the detected serine/threonine kinase activity is intrinsic to the haspin protein [3].

The rapidly expanding genome sequence databases have allowed a number of new translations of hypothetical haspin-like genes even since previously published analyses. These new data allow a more refined multiple alignment of the haspin kinase domain sequences (fig. 2). This analysis suggests an alternative assignment of regions I–IV in the *S. cerevisiae* haspin homologues Alk1p and Ybl009wp. Because these proteins are the most distantly related proteins in the group (particularly in regions I–IV), the proposed alignment of Alk1p and Ybl009wp in particular must be regarded as tentative until three-dimensional structural information is available for one of these proteins. Overall, the alignment indicates that many of the residues that are essentially invariant in other ePKs, and known to be critical in forming the $Mg^{2+}$-ATP-binding and catalytic sites, are conserved in the majority of haspin proteins (see figs 1B, 2). Specifically, the G-x-G-x-x-G-x-V motif in region I, lysine in region II, glutamate

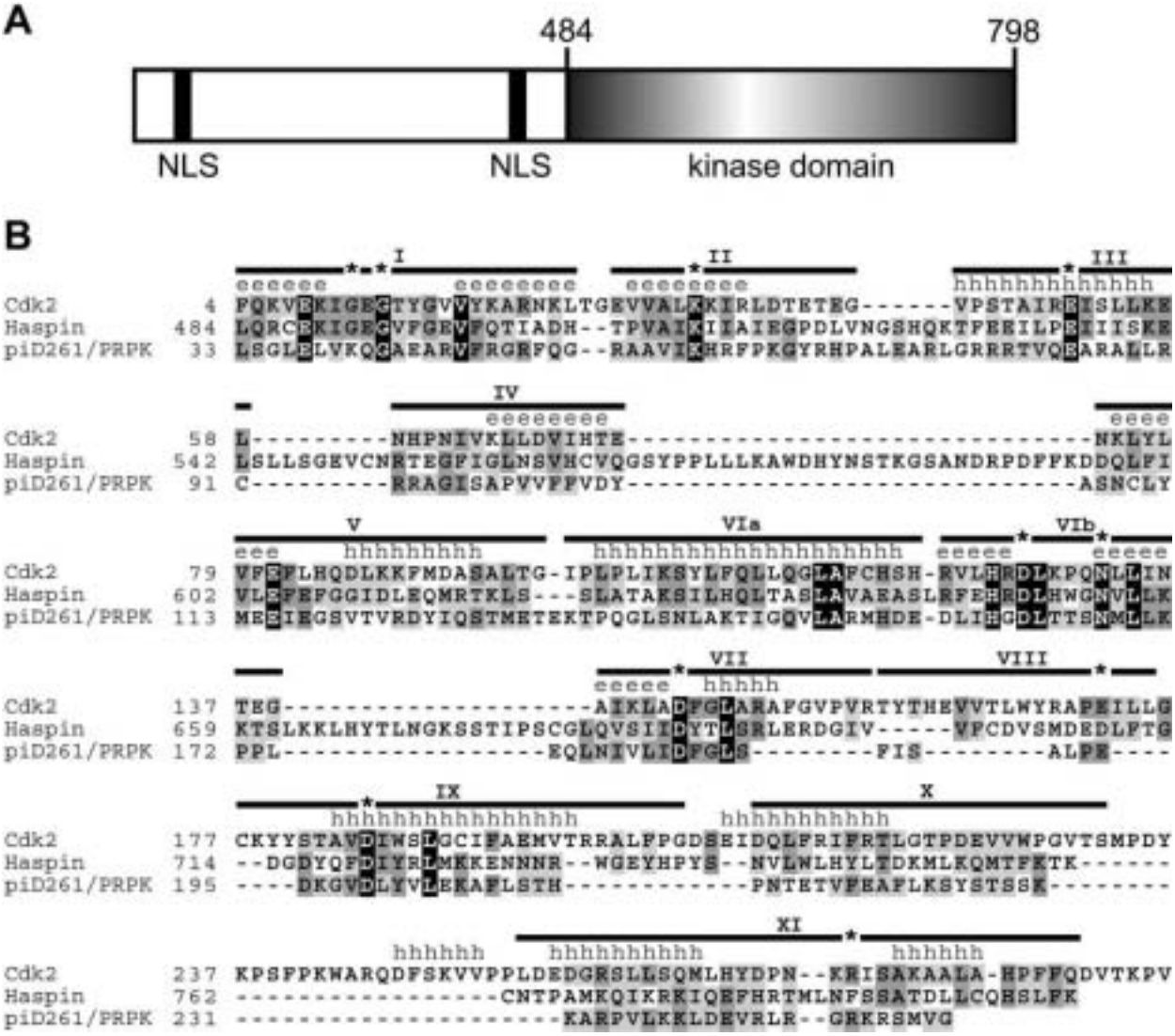



Figure 1. (*A*) Schematic structure of the haspin proteins. The C-terminal kinase domain is shaded, and two potential nuclear localization signals (NLS) found in mammalian haspins are shown in the N-terminal domain. The haspin and haspin-like proteins of other species have a similar overall organization, but with divergent N-terminal domains. Numbering is according to human haspin. (*B*) Sequence features of the haspin kinase domain. The kinase domain of human haspin is aligned with the most closely related human kinase of known three-dimensional structure, the cyclin-dependent kinase Cdk2 [92], and with a member of the divergent piD261 ePK family with known kinase activity, human piD261/p53-related protein kinase (PRPK)/Bud32 [18–20]. Sub-domains I–XI of the kinase fold are indicated with thick horizontal lines, and residues that are essentially invariant in previously identified kinases [12] are marked with asterisks. The known secondary structure of Cdk2 [11] is indicated, where $\alpha$ helices are marked h and $\beta$ strands as e. Completely conserved residues are shown as white letters on a black background, identical residues are dark-shaded and conservative substitutions (based on Gonnet Pam250 matrix positive values) are lightly shaded. Residues 484–798 of human haspin, residues 4–293 of human Cdk2, and residues 33–253 of human piD261 are shown. The alignment of haspin with Cdk2 is derived from previous work [4, 6] and that with piD261 is based on an alignment with protein kinase A supported by mutagenesis data [20].

in region III, aspartate and asparagine in region VIb, aspartate in region VII and aspartate or glutamate in region IX are present. In several cases, the glycines of the nucleotide-binding G-x-G-x-x-G-(V/A) motif are substituted with other small amino acids in the haspin proteins, particularly those from the lower eukaryotes. Indeed, in the *S. cerevisiae* proteins Alk1p and Ybl009wp, the G-x-G-x-x-G-x-(V/A) motif of region I is difficult to identify and, furthermore, the invariant catalytic aspartate residue of region VIb is replaced by asparagine. Of importance will be confirming the kinase activity of these proteins experimentally.

Together, several features distinguish the haspins from other ePKs. Region VIa contains a L-A-V-A-E motif in which the glutamate residue is invariant. The catalytic loop of region VIb is highly conserved and invariably has the sequence (F/L)-E-H-R-(D/N)-L-(H/T/N). In other ePKs, a highly conserved D-F-G triplet in region VII is

involved in orientation of the γ-phosphate of $Mg^{2+}$-ATP. In the haspins, a related sequence, D-(Y/F)-(S/T), is located within an (I/L)-I-D-(Y/F)-(S/T)-(L/C)-S-R motif in which the isoleucine-aspartate and arginine residues are invariant. The A-P-E motif found in region VIII of almost all ePK families is absent in the haspins, although an invariant phenylalanine residue is located in this region. Consistent with the lack of a conserved glutamate in an A-P-E motif is the corresponding absence of a conserved arginine residue in region XI. In other ePK families, these two residues form an ion pair that stabilizes the structure of the large lobe [10]. In region IX, a (E/D)-(I/V/T)-Y-R-x-M-(R/K) motif, in which the tyrosine and methionine are invariant, replaces the more usual D-x-W-S-x-G-x motif. A W-$x_6$-(T/S)-N-(V/I/L)-hydrophobic-W-L-x-Y-L in which the second tryptophan residue is invariant in region X is another conserved feature of the haspin proteins. Finally, the majority of the haspins have conserved inserted regions between regions IV–V and VIb–VII, although these additional sequences are more variably present in haspin proteins from the unicellular eukaryotes.

The mammalian haspins are clearly structurally divergent members of the ePK family. Indeed, phylogenetic analyses of kinases in both humans [6] and *C. elegans* [17] indicate that the haspin proteins do not fall in any of the previously described ePK families. Rather, the haspins form a distinctive group of ePKs that appeared early in eukaryotic evolution. In the extent of sequence divergence from 'classical' ePKs, haspin proteins are similar to members of the ABC1, RIO1, piD261 and AQ578 families of putative kinases that have been identified in archaea, eukaryotes and, in some cases, prokaryotes (fig. 1B) [9]. However, no haspin-like proteins have been identified so far in the sequenced genomes of non-eukaryotes [unpublished data]. Notably, human and *S. cerevisiae* piD261, members of one of these divergent ePK families, have demonstrable kinase activity [18–20].

A possible leucine zipper motif is present in the mammalian haspin proteins [3, 5] in the portion now identified as regions VIa and VIb of the kinase domain. However, this motif is incompletely conserved in the haspin proteins from other phyla. Also, algorithms that take into account features of these motifs in addition to the heptad repeat of leucine [21] do not consistently predict a coiled-coil in this region [unpublished data]. In fact, this sequence is likely to straddle the catalytic loop of the haspin proteins and to adopt a structure containing both α helices and β strands as in other kinases [10, 11] rather than a leucine zipper.

## Haspin-related proteins in *C. elegans*

A number of kinase families appear to be expanded in the nematode worm *C. elegans*. For example, the casein kinase I (CK1) family has at least 65 members in *C. elegans*, but only 4 in budding yeast and 6 in humans, and the FER family has over 42 members in the worm while mammals have just two homologues (FER and FES) [17, 22]. Similarly, *C. elegans* appears to have an enlarged family of haspin-related proteins [6]. The C01H6.9, Y18H1A.10 and F22H10.5 proteins shown in fig. 2 clearly fall within the haspin group. In addition, at least two additional loci with homology to parts of the haspin kinase domain are found in the *C. elegans* genome (W02H3.2, Y40A1A.1). Therefore, *C. elegans* may possess up to five clear haspin homologues (table 1).

Strikingly, approximately 16 additional hypothetical kinase genes encoding proteins more distantly related to haspin are found in *C. elegans*, distributed on various chromosomes (table 1). In some cases, identifying conceptual translations encoding complete kinase domains at these loci is difficult, perhaps due to ineffective prediction of exons, the existence of sequencing errors, or to the presence of non-functional pseudogenes. Because of this, these genes require definition at the mRNA and protein level before a full comparison with the other haspins can be made. However, some features of these proteins, which I will refer to as 'haspin-related' proteins to distinguish them from the 'canonical' haspins discussed so far, are

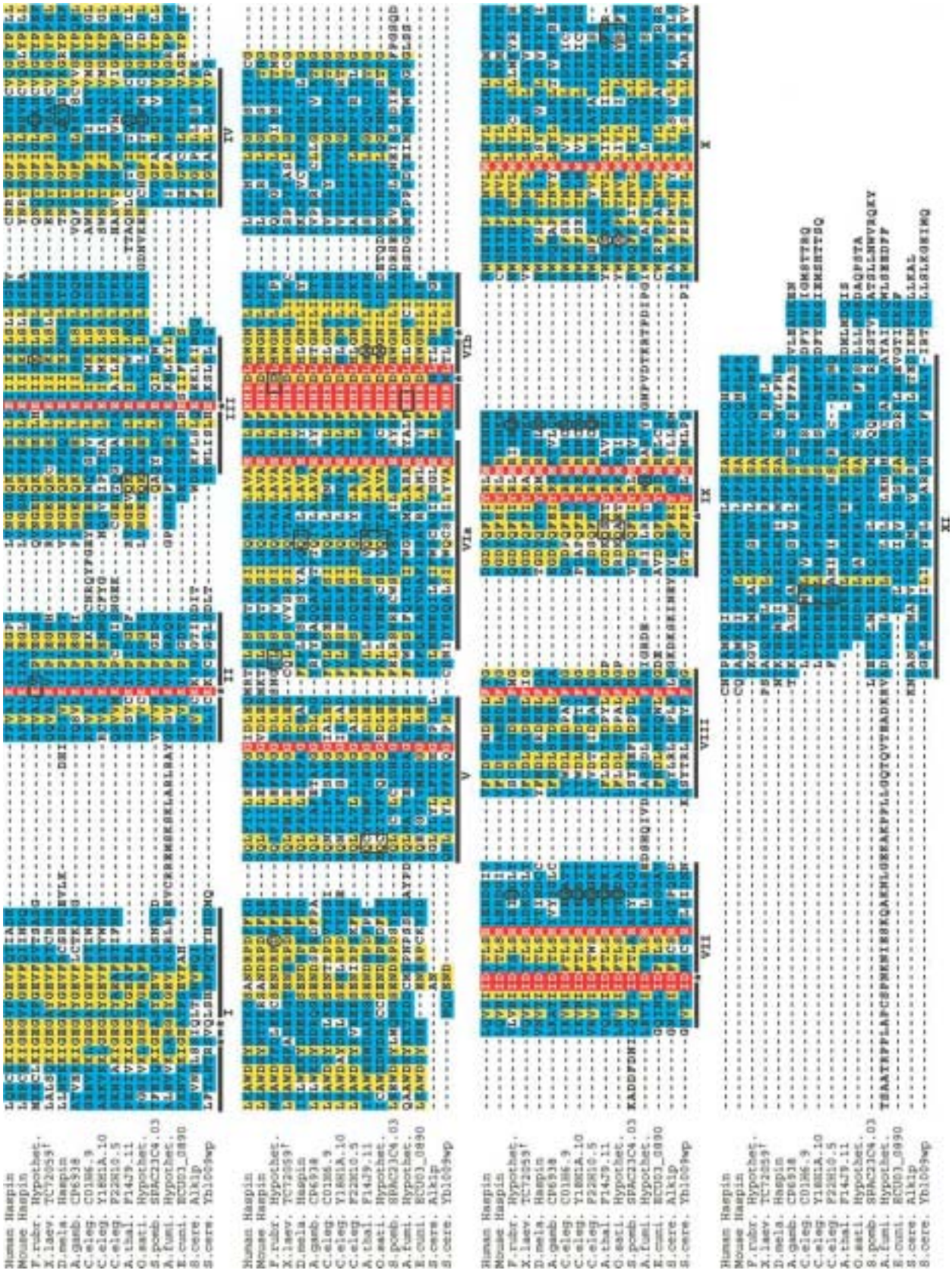
Table 1. Haspin-related genes of *C. elegans*.

| Gene[1] | Chromosome, location[2] (Mb) | Expression mountain[3] |
|---|---|---|
| Haspin family | | |
| C01H6.9 | I, 7.2 | ND |
| F22H10.5 | X, 16.7 | 0 |
| W02H3.2 | X, 11.1 | 0 |
| Y18H1A.10 | I, 0.68 | 0 |
| Y40A1A.1 | X, 2.0 | ND |
| Haspin-like genes | | |
| C04G2.10 | IV, 10.1 | 4 |
| C06E4.5 | IV, 7.3 | 3 |
| C26E6.1 | III, 5.0 | ND |
| C50H2.7 | V, 9.9 | 4 |
| F12A10.2 | II, 5.5 | 5 |
| F59E12.6 | II, 5.6 | 5 |
| H12I13.1 | II, 6.1 | 4 |
| K08B4.5 | IV, 6.1 | 0 |
| T05E8.2 | I, 5.5 | 9 |
| VY10G11R.1 | IV, 16.5 | ND |
| W03F9.3 | V, 0.13 | 4 |
| Y32G9A.11/ Y32G9A.3 | V, 1.9 | ND |
| Y40A1A.1 | X, 2.0 | ND |
| Y48B6A.10 | II, 14.2 | ND |
| Y73B6A.1 | IV, 6.7 | 4 |
| ZK177.2 | II, 5.5 | 5 |

[1] See figure 3 for details.

[2] Chromosomal assignments and locations were obtained from WormBase (http://www.wormbase.org, release WS83: 03 Aug 2002 [91]).

[3] From Kim et al. [29]. See text for details. ND, not determined.



apparent (fig. 3). As in the other haspin homologues, the kinase-like sequence appears to be at the C terminus, but the kinase domain appears shorter than in the canonical haspins and it lacks motifs that identify regions VIII–X of the kinase structure. Because these elements form important core helices in the large lobe of the kinase domain, the structure of these haspin-related proteins is likely to differ significantly from that previously determined for other eukaryotic protein kinases [10, 11]. The G-x-G-x-x-G-(V/A) motif of region I, the invariant lysine, glutamate, aspartate/asparagine and aspartate residues of regions II, III, VIb and VII are well conserved, suggesting that many of these proteins are likely to have kinase activity. Additionally, the retention of the VIb–VII insert found in haspin proteins from all multicellular eukaryotes is striking.

The relationship between the haspin-related proteins of *C. elegans* and the canonical haspins is further supported by the finding that the two groups cluster together in two independent phylogenetic analyses using different methodologies and different subsets of kinases [6, 17]. Moreover, the intron-exon structure of the haspin-related genes is similar to that of the canonical haspins, supporting the existence of a common ancestor for all these genes (see fig. 3). A phase 1 intron in the region VII coding sequence and particularly the phase 0 intron in region XI are conserved in many or all members of the extended haspin family in *C. elegans*. Therefore, duplication of a haspin gene in the worm lineage was likely followed by deletion of a portion of the two exons encoding regions VIII, IX and X. This shortened haspin-related gene then multiplied and diversified to yield the large family now found in *C. elegans*. While most eukaryotic species appear to have only one haspin homologue, or two in the case of *S. cerevisiae*, *C. elegans* has developed a diverse array of haspin-related proteins. As is the case for the CK1, KIN-15 and FER ePK families, the reason for this intriguing worm-specific expansion of the haspin family is unclear. Other lineage-specific expansions (LSEs) seem to involve genes that play a role in responses to pathogens, xenobiotics and stress or in determination of particular morphological adaptations [22].

## The N-terminal domain

In contrast to the C-terminal kinase domain, the N-terminal 483 amino acids of human haspin contain a number of deletions and an insertion when compared to murine haspin, and the two proteins share only 53% identity in this region (fig. 4) [4, 5]. This N-terminal region does not have obvious homology to any known domain type or to other proteins in the sequence databases. The identity of 13 of 19 amino acids from murine haspin (residues 144–162) to the transcription factor murine MEF2B has been highlighted [3]. However, these residues are not particularly highly conserved in MEF2B or haspin proteins from different species and the significance of this short region of homology remains unknown.

Interestingly, the N-terminal regions of the mammalian haspins do not have any apparent homology with the corresponding regions of conceptual haspin translations from non-mammalian species. The N-terminal domains of many of the proteins shown in fig. 2 do share some general features, including a preponderance of serine and lysine or arginine residues. Because the open reading frames of these hypothetical genes have not yet been confirmed experimentally, some caution must be exercised in interpretations drawn from these sequences. However, while the C-terminal kinase domain of haspin has been relatively well conserved through evolution, the N-terminal domain has apparently diverged considerably. This relative divergence is clear when comparing haspin pro-

Figure 2. Multiple sequence alignment of haspin kinase domains. Residues that are completely conserved in all haspin proteins are shown in white on a red background. Residues that are identical (yellow background) or similar (cyan background) in 50% or more sequences are indicated. Introduced gaps are shown as dashes. Sub-domains I–XI of the kinase fold are indicated with thick horizontal lines, and residues that are essentially invariant in previously identified kinases [12] and that are identifiable in the haspins are marked with asterisks. The positions of phase 0 introns are shown by boxes, phase 1 introns by circles, and phase 2 introns by hexagons. All translations are derived from genomic DNA except for *Xenopus laevis* TC72059, marked †, which is derived from cDNA, and thus introns are not shown in this case. The alignment was produced using CLUSTAL W [93] with default settings at the University of California San Diego Supercomputer Center Biology Workbench (http://workbench.sdsc.edu), and adjusted manually according to Hanks and Quinn [12] and by comparison to protein kinase alignments at the Protein Kinase Resource (http://www.sdsc.edu/ Kinases). The alignment includes residues 484–798 of human haspin (GenBank NP_114171), 440–754 of murine haspin (GenBank AAK30301), and the C-terminal regions of hypothetical proteins from *F. rubripes* (hypothetical translation of FS:S005536 from the Fugu Genomics project [94]), *X. laevis* (translation of TC72059, an EST assembly from The Institute for Genomic Research (TIGR) Xenopus Gene Index [26]), *D. melanogaster* (residues 248–566 of GenBank P83103 and [6]), *A. gambiae* (residues 83–404 of hypothetical protein agCP6938, GenBank EAA05110, from the International Anopheles Genome project, a collaboration between Celera Genomics, Genoscope, University of Notre Dame, EBI/Sanger Institute, EMBL, Institut Pasteur, IMBB and TIGR), *C. elegans* (hypothetical proteins C01H6.9, Y18H1A.10 and F22H10.5; see fig. 3 for details), *A. thaliana* (residues 287–599 of hypothetical protein F14J9.11, GenBank AAC33205 [95]), *O. sativa* (hypothetical translation from GenBank AAAA01006721 [96]) *S. pombe* (residues 156–488 of hypothetical protein SPAC23C4.03, GenBank NP_593176 [97]), *A. fumigatus* (hypothetical translation of preliminary genomic sequence data obtained from the TIGR website at http://www.tigr.org), *E. cuniculi* (residues 183–454 of hypothetical protein ECU03_0890, GenBank NP_597598 [87]) and *S. cerevisiae* (residues 487–759 of Alk1p/Ygl021wp, GenBank NP_011494; and 401-676 of hypothetical protein Ybl009wp, GenBank NP_009544 [98]).

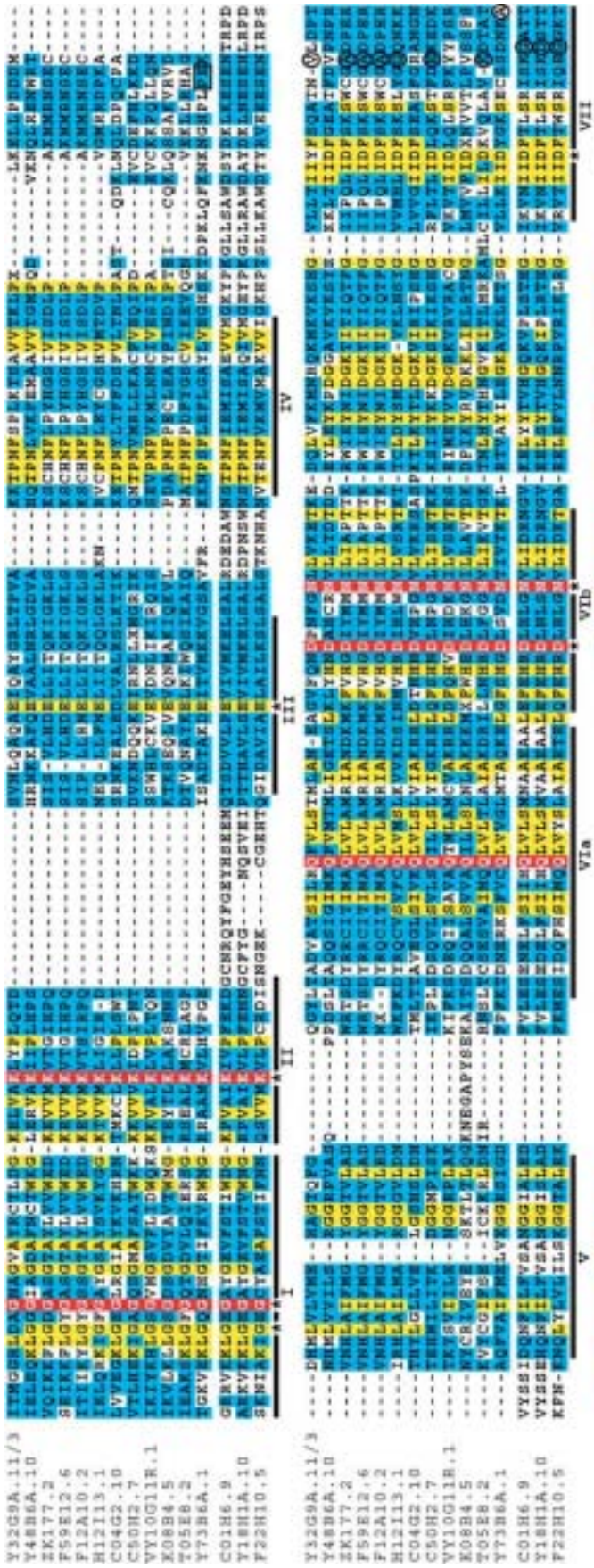


Figure 3. Multiple sequence alignment of haspin-related kinase domains from *C. elegans*. The alignment was generated and is coloured and annotated as described for figure 2. The upper panel shows the haspin-related proteins and the lower panel the canonical haspin proteins also shown in figure 2. Note that translations of the genes shown are named according to the Genefinder predictions [99] but are revised hypothetical translations based on the intron-exon structure of C01H6.9 and F59E12.6, and sequence similarity to these proteins. Specifically, sequences Y18H1A.10 and F22H10.5 were identified based on similarity to residues 589 – 920 of hypothetical protein C01H6.9 (GenBank CAA95786) and the rest on a revised translation of hypothetical gene F59E12.6 supported by EST evidence at WormBase (http://www.wormbase.org, release WS83: 03 Aug 2002 [91]). Note that the two predicted genes Y32G9A.11 and Y32G9A.3 are joined to produce a single hypothetical coding sequence. In the cases of Y32G9A.11/3, F12A10.2 and K08B4.5, shifts in reading frame were required at positions marked X to give the translations shown. These regions require resequencing to confirm that the translations shown here are legitimate.

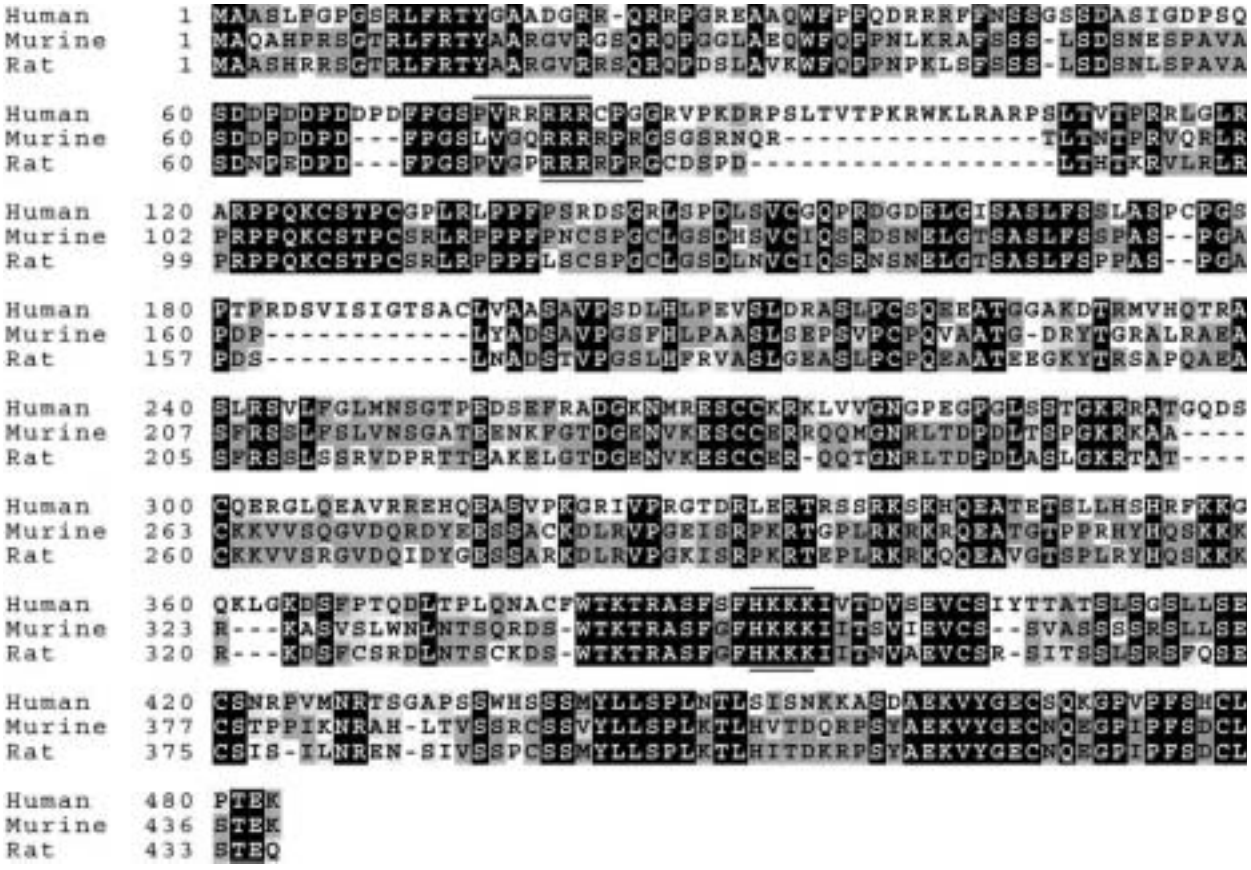



Figure 4. Alignment of the N-terminal domains of human, murine and rat haspin proteins. The alignment is shaded as described for figure 1B and was generated as described in fig. 2. Potential nuclear localization signals found in all three proteins are indicated with horizontal lines. The rat sequence is a hypothetical translation from the rat HTGS database, GenBank AC119593.

teins from different mammals, and has apparently proceeded to such an extent that detectable homology between the N-terminal regions of haspins from more distantly related eukaryotes has been lost. This suggests that the evolutionary pressure to maintain the integrity of the N-terminal domain of haspin has been less than that upon the kinase domain, or that these pressures have been different in different species. The N-terminal domain possibly serves a function such as intracellular targeting of the haspin protein that does not have such stringent structural requirements as the catalytic kinase domain. Alternatively, the N-terminal domain may have assumed different functions in different species. For example, changes in the subcellular localization, molecular interactions or regulation of the kinase activity of haspin might be brought about by alterations in the N-terminal domain. Indeed, a variety of different domain types are found in the N-terminal regions of the haspin-related proteins of *C. elegans*, suggesting that members of this family might serve in diverse signaling pathways in the worm. Resolution of these questions requires further investigation of haspin function in various organisms.

Some limited insight into the role of the N-terminal domain has been obtained from analysis of the mammalian haspins. Two potential nuclear localization signals (NLS) are found in similar locations in the N-terminal domains of human, murine and rat haspin (see figs 1A, 4) [4], although the functionality of these motifs remains to be tested. When expressed as an EGFP fusion protein in HEK293 or COS cells, both murine and human haspin are localized exclusively to the nucleus. Indeed, EGFP-haspin is found in discrete foci within the nuclei in a pattern that is consistent with nucleoli, although this has yet to be confirmed [3, 5]. Immunohistochemical analysis also confirms the localization of murine haspin in the nuclei of haploid germ cells [3]. A third possible NLS in murine and rat haspin is not found in the human protein and deletion of this motif apparently does not prevent murine haspin localization to the nucleus [3].

## Tissue distribution

As described above, a murine haspin cDNA was originally isolated in a differential screen for transcripts expressed in male germ cells, but not in the supporting tis-

sue of the testes. Northern analysis confirmed that haspin mRNA is abundantly expressed in the testes of wild-type mice, but not in mutant strains lacking germ cells. During prepubertal development, haspin mRNA was undetectable at 4, 10 and 16 days of age, and was first observed at 24 days. The timing was similar to, or slightly preceded that of protamine-1 mRNA expression, and coincided with spermatids becoming the predominant spermatogenic cells in mouse testis [1, 3]. In contrast, no haspin mRNA could be detected in the ovary or in a variety of somatic tissues examined. Similar results were reported for human haspin [3, 5].

Another study confirmed the high expression in murine and human testis, but also detected haspin mRNA in some somatic tissues, albeit at much lower levels. In particular, sites of leukocyte development such as the bone marrow, thymus, spleen and fetal liver contained haspin mRNA. Moreover, all proliferating cell lines examined possessed haspin mRNA, with the highest levels found in Ramos B cells and Jurkat T cells. Similar results were found in both human and murine tissues, perhaps suggesting functional significance for this expression in somatic cells [4]. Indeed, this type of expression pattern is not unique and many transcripts that are known to encode functional proteins in somatic cells are greatly overexpressed in early spermatids [23]. For example, the ubiquitous TATA-binding protein mRNA is found at levels up to 200 times greater in total testis than in spleen and liver [24]. Therefore, haspin is likely not a truly testis-specific gene, and it may play a role in some somatic cells, particularly lymphocytes and proliferating cells. This view is further supported by the presence in the sequence databases of numerous haspin mRNA-derived expressed sequence tags (ESTs) isolated from human and murine lymphoid and embryonic tissues and cancer cells, as well as from the testis [25, 26]. In contrast, ESTs from the classical testis-specific protamine genes are found exclusively in the testis (and, curiously, the brain). The apparent low level of haspin mRNA in lymphoid tissues might reflect expression by a subset of lymphocytes, a common finding in this diverse cell population.

Transcription in early spermatids has often been described as ‘promiscuous’ [27]. The reason for this broad and high-level gene expression is unclear, as many mRNAs in these cells seem to be inefficiently translated [23]. Perhaps the promiscuity is a result of opening the chromatin structure to facilitate recombination during meiosis, or a byproduct of the need to produce high levels of transcripts for proteins such as protamines that are translated at later stages of spermiogenesis when transcription has ceased [23]. Importantly, haspin protein is detected in mouse, human and rat testicular lysates using antisera raised to specific peptides of murine haspin [3, 5]. Immunohistochemical localization studies in mouse testis show that haspin protein is expressed in some germ cells but not in surrounding cells such as Sertoli or Leydig cells. Cells at early stages of spermatogenesis (spermatogonia, leptotene and zygotene spermatocytes) did not stain with anti-haspin antisera. The highest levels of haspin were detected in the nuclei of round spermatids, but were lower in late spermatids with greater nuclear condensation [3]. Although no haspin protein was detected in somatic tissues by Western blot analysis [3], this method is probably not sensitive enough to detect haspin expressed in a subset of cells within these tissues, and the somatic tissues with the highest levels of haspin expression (thymus and fetal liver) were not examined. The identification of the somatic cells that contain haspin protein, if any, will be an important goal in the future.

Limited information is available regarding the tissue distribution of haspin in species other than mammals. *Drosophila* and *Xenopus* ESTs for haspin have been isolated from embryos, oocytes and from the ovary [26, 28]. Recently, a gene expression map for *C. elegans* was developed in which genes with a similar expression pattern in a variety of experimental systems are grouped together into one of 43 different ‘mountains’ [29]. Information for 14 of the haspin and haspin-related proteins is included. Interestingly, 5 of these genes map to a mountain (mount 4) that contains many sperm-enriched transcripts and a large number of protein kinases. However, the other haspin-related genes, including the canonical haspin homologues Y18H1A.10 and F22H10.5, map to a variety of other mountains with less clear functional correlates (table 1).

## Gene structure

In humans and mice, the haspin gene has several notable features. Perhaps most striking is its location, in both species, within an intron of the gene encoding the integrin $\alpha$E chain [4, 5, 7]. Integrin $\alpha$E pairs with a $\beta$7 chain to form the E-cadherin-binding $\alpha$E$\beta$7 integrin that is expressed predominantly on intra-epithelial T cells [30–32]. The function and expression pattern of integrin $\alpha$E therefore does not seem to be closely related to that of haspin [4, 33]. The nested organization of the two genes, shown in fig. 5, is unusual but not unprecedented for group III introns. The human factor VII, neurofibromatosis type I and thyroglobulin genes all contain one or more unrelated genes within large introns [34–36] and the *D. melanogaster* dunce gene contains no fewer than three other genes in two introns [37]. Remarkably, the 2.8-kb transcript of haspin is encoded within an intron of only 4.4 kb.

Additional complexity at this locus results from the existence of an alternatively spliced form of the integrin $\alpha$E gene that originates on the opposite DNA strand, less than 70 base pairs upstream of the haspin transcription start

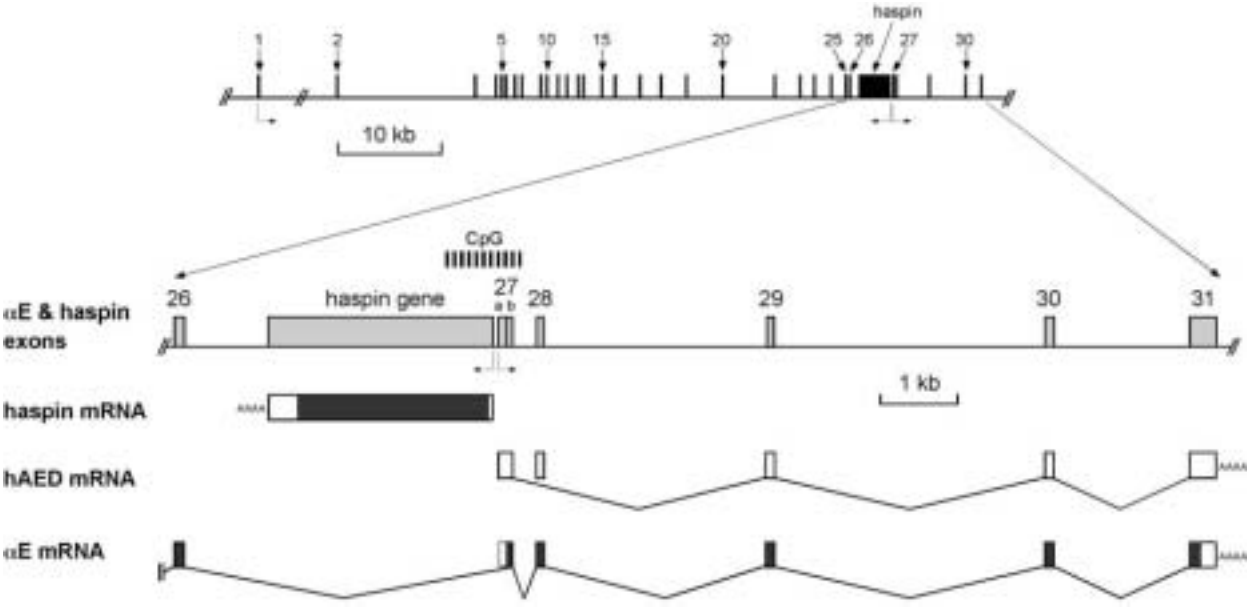



Figure 5. Outline structure of the genomic region containing the integrin $\alpha$E, hAED and haspin genes. The top line shows the intron-exon structure of the integrin $\alpha$E gene, and within it the haspin gene. On the second line, the 3′ region of the $\alpha$E gene containing the haspin gene is shown in more detail. Intronic regions are shown as horizontal lines and exons as boxes. The $\alpha$E exons are numbered. Bent arrows represent points of transcription initiation, and the dashed line indicates the location of a CpG island. The lower three lines show the three transcribed products of this genomic region. In each case, boxes represent exonic portions of RNA, thin lines indicate introns removed by splicing, dark shading indicates protein-coding regions, and AAAA indicates the poly(A) tail. Reprinted from Higgins [4], copyright 2001, with permission from Elsevier Science.

site (fig. 5). The function, if any, of this 900-bp polyadenylated RNA, named human or mouse alpha-E-derived transcript (hAED or mAED) is unclear. It appears not to encode a protein related to integrin $\alpha$E, and its short potential open reading frame may not be translated [4]. Such a head-to-head arrangement of genes, transcribed from a bi-directional promoter, can allow co-ordinate expression of related genes, as for collagen IV $\alpha$1 and $\alpha$2 [38]. However, although both hAED/mAED and haspin are abundantly transcribed in the testis, the two RNAs are not always co-expressed in cell lines [4]. The hAED/mAED transcript could be simply a non-functional product of 'leaky' transcription, thought to be particularly common in the testis, and perhaps enhanced by the nearby transcription of the haspin gene. Indeed, a number of other 'aberrant' mRNAs that do not encode functional proteins are produced from intronic promoters in the testis [39]. An alternative suggestion arises from the finding that the common promoter region of hAED/mAED and haspin is encompassed by a CpG island. These regions are often associated with the 5′ ends of housekeeping genes and with many tissue-specific genes [40]. Hypomethylation of CpG islands is usually critical for efficient promoter activity. Since transcription may be an important component of the mechanism that maintains hypomethylation at CpG islands [41], the process of hAED/mAED transcription itself may be important to maintain the island in a state that will allow normal haspin expression. Another possibility is that hAED/mAED plays an as yet undefined role as a non-coding RNA. Finally, the role of antisense transcripts in the regulation of a number of genes is becoming increasingly apparent [42, 43]. The haspin mRNA is antisense to the unspliced integrin $\alpha$E mRNA and vice versa, but the significance of this remains to be investigated.

Another distinctive feature of the haspin gene in humans and mice is its lack of introns [4, 5, 7]. The *S. cerevisiae*, *S. pombe* and *E. cuniculi* haspin genes also have no introns, but this is consistent with the absence of introns in the majority of genes in these species [44]. In contrast, the vast majority of vertebrate genes have many introns and that mammalian haspin genes lack them is puzzling. Indeed, the haspin genes in other higher eukaryotes do contain introns. At least some of these introns are precisely conserved in both position and phase in the kinase-encoding regions of haspin genes from animals and plants. In particular, an intron in the region VII coding sequence is found in the *F. rubripes*, *C. elegans*, *O. sativa* and *A. thaliana* genes (see fig. 2), suggesting that it is an old intron that predates the divergence of animals and plants. The complete absence of ancestral or more recently inserted introns from the mammalian haspin genes is therefore remarkable.

One mechanism by which a gene may lose all introns is retro-insertion of a processed mRNA into the genome. This process usually yields processed pseudogenes be-

cause it is unlikely that the insertion event will generate a suitable promoter to drive gene expression. Even so, various other features of the mammalian haspin gene have led to the suggestion that it is a rare example of an expressed retro-gene [4, 7]. The mechanism of retrotransposition is thought to include a staggered cut in genomic DNA at the site of insertion, leading to the generation of short direct repeats flanking the inserted sequence. In addition, a polyadenine tract derived from the 3′ end of the integrated mRNA is often incorporated with the inserted sequence. Because retro-transposition of the haspin gene must have occurred prior to the divergence of the primate and rodent lineages, extensive nucleotide substitution and other remodeling in the area of the genes would have occurred since that time. So, confidently identifying the short direct repeats is likely to be difficult, and the polyadenine tract has probably degenerated substantially [45]. In fact, multiple Alu elements have clearly been inserted into the 3′ end of the human haspin gene, including one in the 3′ untranslated region [4, 5]. Nevertheless, both human and mouse haspin genes contain an adenine-rich sequence at the 3′ end that might be a remnant of the polyadenine portion of the inserted mRNA [4, 7]. Also, a direct repeat of seven nucleotides (GAGCC[A/T]T) has been identified at the 5′ and 3′ ends of the human and murine haspin genes [5, 7] that could conceivably have been generated during the insertion process. Finally, the location of the gene in an $\alpha$E intron that has the same phase and position as introns in other integrin genes suggests that the haspin gene was generated by insertion into a preexisting intron [4]. Together, the evidence points to a retro-genic origin of the mammalian haspin genes. One should note that the remodeling of the 3′ end of the human gene may be responsible for the generation of a 4.4-kb haspin transcript in addition to the apparently intronless 2.8-kb mRNA in humans. This larger transcript may be generated from an intron-containing precursor [4].

Many functional retro-genes are expressed principally in meiotic and haploid spermatogenic cells, and haspin fits neatly into this pattern. A number of these retro-genes, including the human and mouse phosphoglycerate kinase-2 (PGK-2) and pyruvate dehydrogenase E1$\alpha$ (PDHA2) genes, appear to be autosomal copies of intron-containing genes found on the X chromosome [45, 46]. This has led to the suggestion that the evolutionary maintenance of these autosomal retro-genes stems from the need to express critical genes after loss or inactivation of the X chromosome in spermatogenic cells [45]. This may be the case for some retro-genes while others, such as the murine S-adensylmethionine decarboxylase-2 (Amd-2) and poly(A)-binding protein-2 (Pabp2) genes, do not appear to originate from a parent gene on the X chromosome [47]. Transcriptional promiscuity in spermatogenic cells has been suggested to increase the chance of a gene becoming retro-transposed. This, coupled with broad translational repression in these cells, might provide a favourable environment for the generation of expressed retro-genes [47]. Also, because retro-insertion events must occur in the germline in order to be inherited, parent genes must be transcribed there to allow generation of retro-genes. Retro-insertion into chromosomal loci that are open and active in germ cells is likely to be favoured. These two factors may also be important contributors to the apparent predominance of retro-genes expressed in spermatogenic cells.

Can a candidate parent gene for haspin be identified? On chromosome 16 in mice is a sequence closely related to the 3′ end of the haspin gene (starting at approximately nucleotide 1960 of the murine haspin cDNA) [48]. This partial haspin sequence contains shifts in the haspin open reading frame and premature stop codons, suggesting that it is a pseudogene unlikely to encode a functional protein product. It contains no introns and has a polyadenine tract at precisely the position that polyadenylation occurs in the bona fide murine haspin mRNA. Immediately following this is a 16-nucleotide sequence (AAGAAAATAAGATTGG) that is a direct repeat of the sequence found just 5′ of the pseudogene. These findings, combined with the fact that the murine pseudogene is more closely related to the murine haspin gene than to either human or rat haspin [unpublished data], suggest that it was created by a second retro-transposition of the haspin retro-gene relatively recently, after the split of the mouse and rat lineages. This locus does not appear to be the parent gene of haspin. Southern analysis of murine genomic DNA with a probe to nucleotides 1–844 of the murine haspin cDNA is consistent with the presence of only a single full-length haspin gene in this species [7] and no other candidate parent genes can be identified in the murine or near-complete human genome databases. Therefore, the parent haspin gene has probably been lost since the creation of the haspin retrogene.

## Function

### Studies in mammalian cells

Functional data regarding the haspin family are beginning to emerge. Most closely examined is murine haspin. Overexpression of EGFP-haspin in HEK-293 or COS-7 cell lines by transient transfection stopped the cells from proliferating [3]. While the untransfected population showed the usual distribution of cells in G1, S and G2/M phases, an increased proportion of the transfected cells were in G1 (or possibly G0) with a corresponding decrease in G2/M phase cells. Cells transfected with EGFP-haspin containing a 10-amino-acid deletion in region I of the kinase domain, and lacking kinase activity (see above), showed an even more profound loss of G2/M

phase cells. After transfection of this mutant haspin, essentially all cells were in G1/G0 2 days after transfection, while approximately 6 days were required for wild-type EGFP-haspin to have a similar effect [3]. This result is intriguing because it suggests that this capacity of haspin cannot be ascribed directly to its kinase activity. EGFP is known to have toxic effects on cells, but EGFP transfection alone was stated to have no such effect on cell growth. The data are consistent with the suggestion that haspin overexpression causes cell cycle arrest in G1, but alternative explanations are possible. For example, increased cell death in the S or G2/M phases might lead to similar results. Furthermore, a toxic effect of the EGFP fusion protein remains possible. While confirming these findings with non-tagged forms of haspin and with human haspin will be important, the effects described are remarkable and warrant further investigation.

Haspin can interact with nucleic acid, and this activity may be regulated by the kinase domain. A proportion of EGFP-haspin was found to bind to a calf thymus DNA column in half-physiological saline. In contrast, EGFP-haspin containing the deletion in the kinase domain completely bound to the column under the same conditions [3]. This led to the proposal that the kinase activity has a negative impact on the DNA-binding capacity of haspin, possibly due to autophosphorylation. Of interest is that the increased binding of the mutated haspin to DNA correlates with its increased ability to inhibit cell growth. The nucleic acid-binding capacity of haspin is likely to reside in the basic N-terminal region but whether it is sequence specific or a general affinity for DNA or whether haspin might also bind RNA is unknown. The physiological significance of this activity remains unclear. Taken together, these data led to the notion that haspin is a 'transcription factor or a cell cycle regulatory factor of haploid germ cells', and that it might function to maintain suspension of the cell cycle in spermatids after meiosis [3, 5].

## Genetic analysis

A possible functional role for haspin in reproduction has been suggested based on genetic mapping studies. The murine haspin/integrin $\alpha$E locus maps to chromosome 11 [49, 50], and to the syntenic location on human chromosome 17p13.3 [51]. Matsui and co-workers have pointed out that the murine haspin gene is close to the ovum (*Om*) mutant locus [3, 49]. This mutation is responsible for the remarkable reproductive properties of the DDK mouse strain. When DDK females are mated to inbred non-DDK males, few F1 embryos reach the blastocyst stage. However, this reduced fertility is not observed when DDK males are bred with non-DDK females [52]. This 'DDK syndrome' is thought to result from an incompatibility between a maternal component in DDK eggs and a paternal factor in non-DDK sperm [52, 53]. However, more recent fine mapping of the *Om* locus suggests that the haspin gene falls outside the minimum interval containing the responsible gene(s) [54], reducing the likelihood that mutation of the haspin gene accounts for this phenotype. Other phenotypes that map to chromosome 11 near the haspin locus include quantitative trait loci affecting sex-specific alcohol preference in C57BL/6 mice [55] and the incubation time of prion disease [56], and a region responsible for a reduction in postnatal female viability in the offspring of C57BL/6 × DBA/2J mating that may or may not be the same as the *Om* locus [57]. Haspin does not seem to be an immediately compelling candidate for these traits, however.

In humans, numerous studies of various cancers, including lung, breast and ovarian cancer and leukaemia, have mapped one or more potential tumour suppressor genes to the p13.3 region of chromosome 17 [58–63]. Haspin probably lies outside a small region most often showing loss of heterozygosity (LOH) in a variety of tumours [61, 63]. However, large deletions of this region of chromosome 17 are common and the patterns of allelic imbalance or LOH at this region are complex and may vary between cancer types [59, 64]. The haspin gene lies between the 14-3-3$\varepsilon$ [65] and the hypermethylated in cancer-1 (Hic1) [66] and OVCA2 [62, 67] genes that have been proposed as candidate tumour suppressor genes. Coupled with the finding that haspin may be expressed in somatic cells and the proposal that it plays a role in regulating cell proliferation, haspin could be considered another such candidate gene.

## Non-mammalian haspins

What is known about the function of haspin-like proteins in other species? An increasing number of studies have documented changes in the levels of multiple *S. cerevisiae* transcripts in response to stress or during progression of the cell cycle, including those of the haspin homologues *ALK1* and *YBL009W*. *ALK1* mRNA levels are strikingly periodic during the mitotic cell cycle of *S. cerevisiae*, with a peak in expression early in M phase [68, 69]. In fact, the pattern of *ALK1* follows very closely that of the cell cycle regulator cyclin *CLB2* [70], and *ALK1* is robustly induced by Clb2p expression. These properties unquestionably place *ALK1* in the 33-gene '*CLB2* cluster' that contains many genes important for mitosis [69]. Like other similarly regulated genes, the region upstream of *ALK1* contains a sequence motif that is a potential binding site for the Mcm1p/SFF transcription factor that is activated by Clb2p-Cdc28p [69, 70]. More recently, confirmation was obtained that the oscillation of *ALK1* expression, as for other members of the *CLB2* cluster, is abolished in yeast strains lacking the Fkh1p and Fkh2p forkhead transcription factors that bind to a consensus site similar to that of the SFF factor [71]. These findings

strongly suggest a role for Alk1p during mitosis. In contrast, *ALK1* transcript levels do not change to a great extent during meiosis upon sporulation of *S. cerevisiae* [72]. Expression of a lacZ gene insertion near the *ALK1* gene is induced by pheromone exposure [73]. However, the increase seen in *ALK1* mRNA in the presence of mating factor alpha in a microarray experiment is less marked [74].

In its GenBank entry, *S. cerevisiae ALK1* is described as a novel DNA damage-responsive gene [75], although no data supporting this conclusion have been published. This is intriguing considering the apparent DNA-binding capacity of murine haspin. Other DNA-binding kinases include DNA-PK, ATM and ATR. These proteins are known to play a role in the response to DNA damage [76]. The expression of mammalian haspin mRNA in germ cells and lymphoid cells that utilize the DNA recombination machinery during meiosis, antigen receptor rearrangement, and somatic hypermutation, and more generally in proliferating cells that are replicating their DNA [76–78] is consistent with an involvement of haspin in these activities. However, in recent studies, the levels of *ALK1* transcripts were found to decrease in response to treatments that induce DNA damage [79]. Because these treatments also cause cell cycle arrest in S phase or at the G2/M boundary, many genes that are expressed at M phase, such as *CLB2*, show similar declines in expression. Therefore, the observed decrease in *ALK1* mRNA is probably due to inhibition of cell cycle progression [79]. Nevertheless, the pattern of *ALK1* expression is clearly not typical of DNA damage-responsive genes [69].

In contrast to *ALK1*, expression of the second *S. cerevisiae* haspin gene *YBL009W* peaks in G1, although the peak-to-trough height of this oscillation is lower than that of *ALK1* [68, 69]. Unlike *ALK1*, *YBL009W* is significantly induced during sporulation [72]. The timing of expression follows the so-called 'early-middle' pattern. Such genes are first expressed in the 2 h following induction of sporulation, and undergo a second increase at 5–7 h. This pattern is consistent with a role for Ybl009wp at some point during meiosis, as genes involved in processes such as sister chromatid cohesion and exit from meiosis are expressed in a similar manner [72]. During responses to stress and DNA damage, *YBL009W* transcripts show only moderate changes in level. Interestingly though, genes that have similar expression patterns to *YBL009W* in these studies include nucleolar proteins and ribosome subunits [79, 80]. Given the suggestion that mammalian haspins localize to the nucleolus (see above), haspin proteins possibly play a role in a cell cycle-regulated process such as ribosome synthesis that takes place in this structure.

The data described above suggest that Alk1p may be important in the mitotic cell cycle during M phase, and Ybl009wp may play a role during both the mitotic cell cycle and meiosis. Neither protein is required for vegetative growth, however, since disruption of the *ALK1* or *YBL009W* gene has no effect on this process in *S. cerevisiae* (http://genome-www.stanford.edu/Saccharomyces/). This information allows some speculation on the role of these proteins, but expression patterns alone often cannot provide precise assignments of function [68, 81]. A second wave of genome-wide experiments in yeast may provide more clues. These studies involve the analysis of yeast two-hybrid protein-protein interactions between almost all possible combinations of yeast proteins. In such a study, an interaction between Alk1p and a single protein, the *S. cerevisiae* choline kinase Cki1p, was detected [81]. This enzyme is involved in the synthesis of phosphatidylcholine, the most abundant phospholipid in *S. cerevisiae*, and is a known target of protein kinase A [82, 83]. A number of genes involved in lipid synthesis are upregulated at M phase in yeast [69]. Since the expression level of the *CKI1* choline kinase gene does not change significantly during the cell cycle [69], the increased levels of Alk1p during this period might possibly provide an alternative means to upregulate Cki1p activity, perhaps to aid cell membrane synthesis. In the same study, Ybl009wp was identified among 69 different proteins that had a detectable interaction with the Soh1p protein. Soh1p has limited homology to RNA polymerases and interacts with factors involved in DNA repair (e.g. Rad5p) and with components of the RNA polymerase II complex [84]. While this provides an interesting connection to DNA damage responses, confirming both these two-hybrid interactions by independent methods will be important. This is particularly true for Soh1p because it binds so many proteins in the two-hybrid screen, a finding that could be produced by 'noise' in the screening system [81].

Given the possibility of functional redundancy, the presence of multiple haspin and haspin-like proteins in *C. elegans* may complicate analysis of their function in this species. Targeting of Y18H1A.10 (Y18H1A_68.g) and the haspin-related gene T05E8.2 by RNA interference did not alter brood size or yield detectable phenotypes upon visual examination of *C. elegans* larvae or adults [85]. Similar targeting of the C26E6.1 haspin-like gene did not cause a detectable effect upon mitotic or meiotic cell divisions [86]. However, studies using this technology are likely to provide interesting insights in the future.

## Concluding remarks

Much has yet to be learned regarding the haspin family but several features suggest that they deserve further attention. In particular, haspin proteins are conserved in eukaryotic evolution. Indeed, at the time of writing, a haspin homologue is present in all 'finished' eukaryotic geno-

mes, including that of the microsporidian *E. cuniculi*. This organism has a genome of only 2.9 million base pairs, smaller than in many bacteria, and encoding only about 2000 genes [87]. *E. cuniculi* is thus highly simplified and has jettisoned many features that are common to other eukaryotes, even detectable mitochondria. This is reflected in the dramatic reduction in gene number, coding sequence length and intervening sequence length in its genome [88]. The retention of a haspin gene in this species therefore suggests an important and perhaps fundamental function in eukaryotic life.

The haspin proteins are clearly related to the ePKs, although they form a distinctive and divergent group. Despite this divergence, the mammalian haspins appear to have kinase activity. If the haspin homologues from other species also prove to have this capacity, we may learn interesting lessons regarding the minimal structural requirements for phosphotransferase activity in the ePK superfamily. Because protein phosphorylation is a critical mechanism for modulating an enormous variety of substrates, including other kinases, metabolic enzymes, transcription factors and adapter molecules, haspin proteins will likely have important regulatory functions in eukaryotic cells.

Without a doubt the most interesting outstanding question regarding haspin biology is: what are these functions? In mice, the high level of haspin expression in postmeiotic spermatids and its absence in more mature spermatogenic cells is consistent with a role at the later stages of meiosis, or the complex differentiation process that begins soon afterwards. This might include a part in regulating changes in nucleolar structure or chromosomal packaging that occur at this time [89, 90]. Haspin expression patterns in both mammals and yeast also imply a function during the mitotic cell cycle. The cell cycle phase-dependent oscillation in expression of the *S. cerevisiae ALK1* and *YBL009W* genes, and the possible block in G1 found upon overexpression of haspin in mammalian cell lines are suggestive of a role in cell cycle regulation, or at least of a function that is required at certain stages of cell growth and division. Again, a role for haspin in nucleolar events, including ribosome biogenesis, or in the response to DNA damage is an intriguing possibility, but has yet to be substantiated. Identifying proteins that bind to, and regulate, or are substrates for haspin will be informative. Indeed, some candidates have now been discovered in *S. cerevisiae*. Studies of protein-protein interactions and gene 'knockouts' in model organisms such as *Drosophila* and the yeasts should complement similar studies in mammalian systems.

*Acknowledgements*. I thank Dr David Lee for his comments on the manuscript. Sequencing of *Aspergillus fumigatus* was funded by the National Institute of Allergy and Infectious Disease U01 AI 48830 to David Denning and William Nierman. Sequencing of *Anopheles gambiae* was funded by the National Institute of Allergy and Infectious Disease to Celera Genomics. Rice WGS sequences were produced by a publicly funded group led by the Beijing Genomics Institute.

1 Tanaka H., Yoshimura Y., Nishina Y., Nozaki M., Nojima H. and Nishimune Y. (1994) Isolation and characterization of cDNA clones specifically expressed in testicular germ cells. FEBS Lett. **355:** 4–10
2 Coulombre J. L. and Russell E. S. (1954) Analysis of the pleiotropism at the W-locus in the mouse – the effects of W and $W^v$ substitution upon postnatal development of germ cells. J. Exp. Zool. **126:** 277–296
3 Tanaka H., Yoshimura Y., Nozaki M., Yomogida K., Tsuchida J., Tosaka Y. et al. (1999) Identification and characterization of a haploid germ cell-specific nuclear protein kinase (haspin) in spermatid nuclei and its effects on somatic cells. J. Biol. Chem. **274:** 17049–17057
4 Higgins J. M. G. (2001) The haspin gene: location in an intron of the integrin $\alpha$E gene, associated transcription of an integrin $\alpha$E-derived RNA and expression in diploid as well as haploid cells. Gene **267:** 55–69
5 Tanaka H., Iguchi N., Nakamura Y., Kohroki J., Egydio de Carvalho C. and Nishimune Y. (2001) Cloning and characterization of human *haspin* gene encoding haploid germ cell-specific nuclear protein kinase. Mol. Hum. Reprod. **7:** 211–218
6 Higgins J. M. G. (2001) Haspin-like proteins: a new family of evolutionarily conserved putative eukaryotic protein kinases. Prot. Sci. **10:** 1677–1684
7 Yoshimura Y., Tanaka H., Nozaki M., Yomogida K., Yasunaga T. and Nishimune Y. (2001) Nested structure of haploid germ cell specifc haspin gene. Gene **267:** 49–54
8 Hardie G. and Hanks S. (1995) The Protein Kinase FactsBook, Academic Press, London
9 Leonard C. J., Avarind L. and Koonin E. V. (1998) Novel families of putative protein kinases in bacteria and archaea: evolution of the 'eukaryotic' protein kinase superfamily. Genome Res. **8:** 1038–1047
10 Knighton D. R., Zheng J. H., Ten Eyck L. F., Ashford V. A., Xuong N. H., Taylor S. S. et al. (1991) Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. Science **253:** 407–414
11 De Bondt H. L., Rosenblatt J., Jancarik J., Jones H. D., Morgan D. O. and Kim S.-H. (1993) Crystal structure of cyclin-dependent kinase 2. Nature **363:** 595–602
12 Hanks S. and Quinn A. M. (1991) Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members. Methods Enzymol. **200:** 38–62
13 Hanks S. K. and Hunter T. (1995) The eukaryotic protein kinase family: kinase (catalytic) domain structure and classification. FASEB J. **9:** 576–596
14 Walker E. H., Perisic O., Ried C., Stephens L. and Williams R. L. (1999) Structural insights into phosphatidylinositide 3-kinase catalysis and signalling. Nature **402:** 313–320
15 Rao V. D., Misra S., Boronenkov I. V., Anderson R. A. and Hurley J. H. (1998) Structure of type II$\beta$ phosphatidylinositol phosphate kinase: a protein kinase fold flattened for interfacial phosphorylation. Cell **94:** 829–839
16 Hon W. C., McKay G. A., Thompson P. R., Sweet R. M., Yang D. S. C., Wright G. D. et al. (1997) Structure of an enzyme required for aminoglycoside antibiotic resistance reveals homology to eukaryotic protein kinases. Cell **89:** 887–895
17 Plowman G. D., Sudarsanam S., Bingham J., Whyte D. and Hunter D. (1999) The protein kinases of *Caenorhabditis elegans*: a model for signal transduction in multicellular organisms. Proc. Natl. Acad. Sci. USA **96:** 13603–13610
18 Stocchetto S., Marin O., Carignani G. and Pinna L. A. (1997) Biochemical evidence that Saccharomyces cerevisiae

YGR262c gene, required for normal growth, encodes a novel Ser/Thr-specific protein kinase. FEBS Lett. **414:** 171–175
19 Abe Y., Matsumoto S., Wei S., Nezu K., Miyoshi A., Kito K. et al. (2001) Cloning and characterization of a p53-related protein kinase expressed in interleukin-2-activated cytotoxic T-cells, epithelial tumor cell lines, and the testes. J. Biol. Chem. **276:** 44003–44011
20 Facchin S., Lopreiato R., Stocchetto S., Arrigoni G., Cesaro L., Marin O. et al. (2002) Structure-function analysis of yeast piD261/Bud32, an atypical protein kinase essential for normal cell life. Biochem. J. **364:** 457–463
21 Lupas A. (1997) Predicting coiled-coil regions in proteins. Curr. Opin. Struct. Biol. **7:** 388–393
22 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
23 Kleene K. C. (2001) A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. Mech. Dev. **106:** 3–23
24 Schmidt E. E. and Schibler U. (1995) High accumulation of components of the RNA polymerase II transcription machinery in rodent spermatids. Development **121:** 2373–2783
25 Schuler G. D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. J. Mol. Med. **75:** 694–698
26 Quackenbush J., Cho J., Lee D., Liang F., Holt I., Karamycheva S. et al. (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res. **29:** 159–164
27 Schmidt E. E. (1996) Transcriptional promiscuity in testes. Curr. Biol. **6:** 768–769
28 Consortium T. F. (2002) The FlyBase database of the *Drosophila* genome projects and community literature. Nucleic Acids Res. **30:** 106–108
29 Kim S. K., Lund J., Kiraly M., Duke K., Jiang M., Stuart J. M. et al. (2001) A gene expression map for *Caernorhabditis elegans*. Science **293:** 2087–2092
30 Cepek K. L., Shaw S. K., Parker C. M., Russell G. J., Morrow J. S., Rimm D. L. et al. (1994) Adhesion between epithelial cells and T lymphocytes mediated by E-cadherin and the $\alpha_E\beta_7$ integrin. Nature **372:** 190–193
31 Karecla P. I., Bowden S. J., Green S. J. and Kilshaw P. J. (1995) Recognition of E-cadherin on epithelial cells by the mucosal T cell integrin $\alpha_{M290}\beta_7$ ($\alpha_E\beta_7$). Eur. J. Immunol. **25:** 852–856
32 Higgins J. M. G., Mandlebrot D. A., Shaw S. K., Russell G. J., Murphy E. A., Chen Y.-T. et al. (1998) Direct and regulated interaction of integrin $\alpha_E\beta_7$ with E-cadherin. J. Cell Biol. **140:** 197–210
33 Shaw S. K., Cepek K. L., Murphy E. A., Russell G. J., Brenner M. B. and Parker C. M. (1994) Molecular cloning of the human mucosal lymphocyte integrin $\alpha_E$ subunit: unusual structure and restricted RNA distribution. J. Biol. Chem. **269:** 6016–6025
34 Levinson B., Kenwrick S., Gamel P., Fisher K. and Gitschier J. (1992) Evidence for a third transcript from the human factor VIII gene. Genomics **14:** 585–589
35 Xu G., O'Connell P., Viskochil D., Cawthon R., Robertson M., Culver M. et al. (1990) The neurofibromatosis type I gene encodes a protein related to GAP. Cell **62:** 599–608
36 Meijerink P. H. S., Yanakiev P., Zorn I., Grierson A. J., Bikker H., Dye D. et al. (1998) The gene for the human Src-like adaptor protein (hSLAP) is located within the 64-kb intron of the thyroglobulin gene. Eur. J. Biochem. **254:** 297–303
37 Furia M., Digilio F. A., Artiaco D., Giordano E. and Polito L. C. (1990) A new gene nested within the dunce genetic unit of *Drosophila melanogaster*. Nucleic Acids Res. **18:** 5837–5841
38 Pöschl E., Pollner R. and Köhn K. (1988) The genes for the $\alpha$1(IV) and $\alpha$2(IV) chains of human basement membrane collagen type IV are arranged head-to-head and separated by a bidirectional promoter of unique structure. EMBO J. **7:** 2687–2695
39 Ivell R. (1992) 'All that glisters is not gold' – common testis gene transcripts are always what they seem. Int. J. Androl. **15:** 85–92
40 Gardiner-Garden M. and Frommer M. (1987) CpG islands in vertebrate genomes. J. Mol. Biol. **196:** 261–282
41 Macleod D., Ali R. R. and Bird A. (1998) An alternative promoter in the mouse major histocompatibility complex class II I-A$\beta$ gene: implications for the origin of CpG islands. Mol. Cell. Biol. **18:** 4433–4443
42 Lee J. T. and Lu N. (1999) Targeted mutagenesis of *Tsix* leads to nonrandom X inactivation. Cell **99:** 47–57
43 Brantl S. (2002) Antisense-RNA regulation and RNA interference. Biochim. Biophys. Acta **1575:** 15–25
44 Logsdon J. M. Jr (1998) The recent origins of spliceosomal introns revisited. Curr. Opin. Genet. Dev. **8:** 637–648
45 McCarrey J. R. and Thomas K. (1987) Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. Nature **326:** 501–505
46 Dahl H. M., Brown R. M., Hutchinson W. M., Maragos C. and Brown G. K. (1990) A testis-specific form of the human pyruvate dehydrogenase E1$\alpha$ subunit is coded for by an intronless gene on chromosome 4. Genomics **8:** 225–232
47 Kleene K. C., Mulligan E., Steiger D., Donohue K. and Mastrangelo M. A. (1998) The mouse gene encoding the testis-specific isoform of poly(A) binding protein (Pabp2) is an expressed retroposon: intimations that gene expression in spermatogenic cells facilitates the creation of new genes. J. Mol. Evol. **47:** 275–281
48 Tanaka H. and Tanaka H. (2000) *Mus musculus* haspin pseudogene, GenBank AB044156
49 Matsui M., Ichibara H., Kobayashi S., Tanaka H., Tsuchida J., Nozaki M. et al. (1997) Mapping of six germ cell-specific genes to mouse chromosomes. Mamm. Genome **8:** 873–874
50 Schön M. P., Arya A., Murphy E. A., Adams C. M., Strauch U. G., Agace W. W. et al. (1999) Mucosal T lymphocyte numbers are selectively reduced in integrin alpha E (CD103)-deficient mice. J. Immunol. **162:** 6641–6649
51 Touchman J. W., Anikster Y., Dietrich N. L., Braden Maduro V. V., McDowell G., Scholelersuk V. et al. (2000) The genomic region encompassing the nephropathic cystinosis gene (CTNS): complete sequencing of a 200-kb segment and discovery of a novel gene within the common cystinosis-causing deletion. Genome Res. **10:** 165–173
52 Wakasugi N. (1974) A genetically determined incompatibility system between spermatozoa and eggs leading to embryonic death in mice. J. Reprod. Fertil. **41:** 85–96
53 Renard J. P., Baldacci P., Richoux-Duranthon V., Pournin S. and Babinet C. (1994) A maternal factor affecting mouse blastocyst formation. Development **120:** 797–802
54 Cohen-Tannoudji M., Vandormael-Pournin S., Le Bras S., Coumailleau F., Babinet C. and Baldacci P. (2000) A 2-Mb YAC/BAC-based physical map of the ovum mutant (Om) locus region on mouse chromosome 11. Genomics **68:** 273–282
55 Melo J. A., Shendure J., Pociask K. and Silver L. M. (1996) Identification of sex-specific quantitative trait loci controlling alcohol preference in C57BL/ 6 mice. Nat. Genet. **13:** 147–153
56 Lloyd S. E., Onwuazor O. N., Beck J. A., Mallinson G., Farrall M., Targonski P. et al. (2001) Identification of multiple quantitative trait loci linked to prion disease incubation period in mice. Proc. Natl. Acad. Sci. USA **98:** 6279–6283
57 Shendure J., Melo J. A., Pociask K., Derr R. and Silver L. M. (1998) Sex-restricted non-Mendelian inheritance of mouse chromosome 11 in the offspring of crosses between C57BL/6J and (C57BL/6J × DBA/2J)F1 mice. Mamm. Genome **9:** 812–815
58 Stack M., Jones D., White G., Liscia D. S., Venesio T., Casey G. et al. (1995) Detailed mapping and loss of heterozygosity analysis suggests a suppressor locus involved in sporadic breast

cancer within a distal region of chromosome band 17p13.3. Hum. Mol. Genet. **4:** 2047–2055
59 Konishi H., Takahashi T., Kozaki K., Yatabe Y., Mitsudomi T., Fujii Y. et al. (1998) Detailed deletion mapping suggests the involvement of a tumor suppressor gene at 17p13.3, distal to p53, in the pathogenesis of lung cancers. Oncogene **17:** 2095–2100
60 Sankar M., Tanaka K., Kumaravel T. S., Arif M., Shintani T., Yagi S. et al. (1998) Identification of a commonly deleted region at 17p13.3 in leukemia and lymphoma associated with 17p abnormality. Leukemia **12:** 510–516
61 Hoff C., Seranski P., Mollenhauer J., Korn B., Detzel T., Reinhardt R. et al. (2000) Physical and transcriptional mapping of the 17p13.3 region that is frequently deleted in human cancer. Genomics **70:** 26–33
62 Schultz D. C., Vanderveer L., Berman D. B., Hamilton T. C., Wong A. J. and Godwin A. K. (1996) Identification of two candidate tumor suppressor genes on chromosome 17p13.3. Cancer Res. **56:** 1997–2002
63 Phillips N. J., Ziegler M. R., Radford D. M., Fair K. L., Steinbrueck T., Xynos F. P. et al. (1996) Allelic deletion on chromosome 17p13.3 in early ovarian cancer. Cancer Res. **56:** 606–611
64 Hoff C., Mollenhauer J., Waldau B., Hamann U. and Poustka A. (2001) Allelic imbalance and fine mapping of the 17p13.3 subregion in sporadic breast carcinomas. Cancer Genet. Cytogenet. **129:** 145–149
65 Konishi H., Nakagawa T., Harano T., Mizuno K., Saito H., Masuda A. et al. (2002) Identification of frequent G(2) checkpoint impairment and a homozygous deletion of 14-3-3epsilon at 17p13.3 in small cell lung cancers. Cancer Res. **62:** 271–276
66 Makos Wales M., Biel M. A., Deiry W. el, Nelkin B. D., Issa J. P., Cavenee W. K. et al. (1995) p53 activates expression of HIC-1, a new candidate tumour suppressor gene on 17p13.3. Nat. Med. **1:** 570–577
67 Phillips N. J., Zeigler M. R. and Deaven L. L. (1996) A cDNA from the ovarian cancer critical region of deletion on chromosome 17p13.3. Cancer Lett. **102:** 85–90
68 Cho R. J., Campbell M. J., Winzeler E. A., Steinmetz L., Conway A., Wodicka L. et al. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell **2:** 65–73
69 Spellman P. T., Sherlock G., Zhang M. Q., Iyer V. R., Anders K., Eisen M. B. et al. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol. Biol. Cell **9:** 3273–3297
70 Mendenhall M. D. and Hodge A. E. (1998) Regulation of Cdc28 cyclin-dependent protein kinase activity during the cell cycle of the yeast *Saccharomyces cerevisiae*. Microbiol. Mol. Biol. Rev. **62:** 1191–1243
71 Zhu G., Spellman P. T., Volpe T., Brown P. O., Botstein D., Davis T. N. et al. (2000) Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. Nature **406:** 90–94
72 Chu S., DeRisi J., Eisen M., Mulholland J., Botstein D., Brown P. O. et al. (1998) The transcriptional program of sporulation in budding yeast. Science **282:** 699–705
73 Erdman S., Lin L., Malczynski M. and Snyder M. (1998) Pheromone-regulated genes required for yeast mating differentiation. J. Cell Biol. **140:** 461–483
74 Roberts C. J., Nelson B., Marton M. J., Stoughton R., Meyer M. R., Bennett H. A. et al. (2000) Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. Science **287:** 873–880
75 Moen C., Lindstedt B. A., Berdal K. G., Rognes T. and Seeberg E. C. (1995) A novel DNA damage-response gene from *Saccharomyces cerevisiae* with homology to protein kinase, GenBank X87672
76 Smith G. C. and Jackson S. P. (1999) The DNA-dependent protein kinase. Genes Dev. **13:** 916–934
77 Papavasiliou F. N. and Schatz D. G. (2000) Cell-cycle-regulated DNA double-stranded breaks in somatic hypermutation of immunoglobulin genes. Nature **408:** 216–221
78 Flores-Rozas H. and Kolodner R. D. (2000) Links between replication, recombination and genome instability in eukaryotes. Trends Biochem. Sci. **25:** 196–200
79 Gasch A. P., Huang M., Metzner S., Botstein D., Elledge S. J. and Brown P. O. (2001) Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. Mol. Biol. Cell **12:** 2987–3003
80 Gasch A. P., Spellman P. T., Kao C. M., Carmel-Harel O., Eisen M. B., Storz G. et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol. Biol. Cell **11:** 4241–4257
81 Ito T., Chiba T., Ozawa R., Yoshida M., Hattori M. and Sakaki Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA **98:** 4569–4574
82 Hosaka K., Kodaki T. and Yamashita S. (1989) Cloning and characterization of the yeast CKI gene encoding choline kinase and its expression in *Escherichia coli*. J. Biol. Chem. **264:** 2053–2059
83 Kim K. H. and Carman G. M. (1999) Phosphorylation and regulation of choline kinase from *Saccharomyces cerevisiae* by protein kinase A. J. Biol. Chem. **274:** 9531–9538
84 Fan H. Y., Cheng K. K. and Klein H. L. (1996) Mutations in the RNA polymerase II transcription machinery suppress the hyperrecombination mutant hpr1 delta of *Saccharomyces cerevisiae*. Genetics **142:** 749–759
85 Fraser A. G., Kamath R. S., Zipperlen P., Martinez-Campos M., Sohrmann M. and Ahringer J. (2000) Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. Nature **408:** 325–330
86 Gönczy P., Echeverri C., Oegema K., Coulson A., Jones S. J. M., Copley R. R. et al. (2000) Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III. Nature **408:** 331–336
87 Katinka M. D., Duprat S., Cornillot E., Metenier G., Thomarat F., Prensier G. et al. (2001) Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature **414:** 450–453
88 Keeling P. J. (2001) Parasites go the full monty. Nature **414:** 401–402
89 Schultz M. C. and Leblond C. P. (1990) Nucleolar structure and synthetic activity during meiotic prophase and spermiogenesis in the rat. Am. J. Anat. **189:** 1–10
90 Wouters-Tyrou D., Martinage A., Chevaillier P. and Sautiere P. (1998) Nuclear basic proteins in spermiogenesis. Biochimie **80**: 117–128
91 Stein L., Sternberg P., Durbin R., Thierry-Mieg J. and Spieth J. (2001) WormBase: network access to the genome and biology of *Caenorhabditis elegans*. Nucleic Acids Res. **29:** 82–86
92 Tsai L. H., Harlow E. and Meyerson M. (1991) Isolation of the human cdk2 gene that encodes the cyclin A- and adenovirus E1A-associated p33 kinase. Nature **353:** 174–177
93 Thompson J. D., Higgins D. G. and Gibson T. J. (1994) CLUSTAL W: improving the sensitivity of progressive sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680
94 Aparicio S., Chapman J., Stupka E., Putnam N., Chia J. M., Dehal P. et al. (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science **297:** 1301–1310
95 The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796–815

96 Yu J., Hu S., Wang J., Wong G. K., Li S., Liu B. et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science **296:** 79–92
97 Wood V., Gwilliam R., Rajandream M. A., Lyne M., Lyne R., Stewart A. et al. (2002) The genome sequence of *Schizosaccharomyces pombe*. Nature **415:** 871–880
98 Goffeau A., Barrell B. G., Bussey H., Davis R. W., Dujon B., Feldmann H. et al. (1996) Life with 6000 genes. Science **274:** 546, 563–547
99 The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science **282:** 2012–2018