

Research

# Phylogenetic and structural analysis of centromeric DNA and kinetochore proteins

Patrick Meraldi<sup>✉\*</sup>, Andrew D McAINSH<sup>✉\*\*‡</sup>, Esther Rheinbay\* and Peter K Sorger\*

Addresses: \*Department of Biology, Massachusetts Institute of Technology, Massachusetts Ave., Cambridge, MA 02139, USA. †Institute of Biochemistry, ETH Zurich, Schafmattstr., 18 CH-8093 Zurich, Switzerland. ‡Chromosome Segregation Laboratory, Marie Curie Research Institute, The Chart, Oxted, Surrey RH8 0TL, UK.

✉ These authors contributed equally to this work.

Correspondence: Peter K Sorger. Email: psorger@mit.edu

Published: 22 March 2006

Genome **Biology** 2006, **7**:R23 (doi:10.1186/gb-2006-7-3-r23)

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2006/7/3/r23>

Received: 19 October 2005

Revised: 19 December 2005

Accepted: 24 February 2006

© 2006 Meraldi et al.; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## Abstract

**Background:** Kinetochores are large multi-protein structures that assemble on centromeric DNA (*CEN* DNA) and mediate the binding of chromosomes to microtubules. Comprising 125 base-pairs of *CEN* DNA and 70 or more protein components, *Saccharomyces cerevisiae* kinetochores are among the best understood. In contrast, most fungal, plant and animal cells assemble kinetochores on *CENs* that are longer and more complex, raising the question of whether kinetochore architecture has been conserved through evolution, despite considerable divergence in *CEN* sequence.

**Results:** Using computational approaches, ranging from sequence similarity searches to hidden Markov model-based modeling, we show that organisms with *CENs* resembling those in *S. cerevisiae* (point *CENs*) are very closely related and that all contain a set of 11 kinetochore proteins not found in organisms with complex *CENs*. Conversely, organisms with complex *CENs* (regional *CENs*) contain proteins seemingly absent from point-*CEN* organisms. However, at least three quarters of known kinetochore proteins are present in all fungi regardless of *CEN* organization. At least six of these proteins have previously unidentified human orthologs. When fungi and metazoa are compared, almost all have kinetochores constructed around Spc105 and three conserved multi-protein linker complexes (MIND, COMA, and the NDC80 complex).

**Conclusion:** Our data suggest that critical structural features of kinetochores have been well conserved from yeast to man. Surprisingly, phylogenetic analysis reveals that human kinetochore proteins are as similar in sequence to their yeast counterparts as to presumptive *Drosophila melanogaster* or *Caenorhabditis elegans* orthologs. This finding is consistent with evidence that kinetochore proteins have evolved very rapidly relative to components of other complex cellular structures.

## Background

Kinetochores are eukaryote-specific structures that assemble on centromeric (*CEN*) DNA and perform three crucial functions: they bind paired sister chromatids to spindle microtubules (MTs) in a bipolar fashion compatible with chromatid disjunction; they couple MT (+)-end polymer dynamics to chromosome movement during metaphase and anaphase [1]; and they generate the spindle checkpoint signals linking anaphase onset to the completion of kinetochore-MT attachment [2]. Despite the conservation of these functions, and of MT structure and dynamics, *CEN*s in closely related organisms are highly diverged in sequence, as are *CEN*s on different chromosomes in a single organism [2,3]. The simplest known *CEN*s, those in the budding yeast *Saccharomyces cerevisiae*, consist of 125 base-pairs (bp) of DNA and three protein-binding motifs (CDEI, CDEII and CDEIII) that are present on all 16 chromosomes [4]. These short *CEN* sequences, often called 'point' *CEN*s, are structurally similar to enhancers and transcriptional regulators in that their assembly is initiated by highly sequence-selective DNA-protein interactions [5]. In contrast, *CEN* DNA in fungi such as the budding yeast *Candida albicans* and fission yeast *Schizosaccharomyces pombe*, plants such as *Arabidopsis thaliana*, and metazoans such as *Drosophila melanogaster* and *Homo sapiens*, are longer and more complex and exhibit poor sequence conservation [6-10]. These regional *CEN*s range in size from 1 kb in *C. albicans* [6], to several megabases in *H. sapiens* [8] and typically contain long stretches of repetitive AT-rich DNA. *CEN* organization is particularly divergent in nematodes such as *Caenorhabditis elegans*, which contain holocentric *CEN*s with MT-attachment sites distributed along the length of chromosomes [11]. Sequence-selective DNA-protein interactions have not been identified in regional *CEN*s and it is thought that kinetochore position is determined by a specialized chromatin domain whose formation at one site on each chromosome is controlled by epigenetic mechanisms [2,12].

A combination of genetics and mass spectrometry in *S. cerevisiae* has yielded a fairly detailed view of the composition and architecture of its simple kinetochores. *S. cerevisiae* kinetochores contain upwards of 70 protein subunits organized into 14 or more multi-protein complexes that together have a molecular mass in excess of 5 to 10 MDa [5]. *S. cerevisiae* kinetochore proteins can be assigned to DNA-binding, linker, MT-binding and regulatory functions. While 'linker

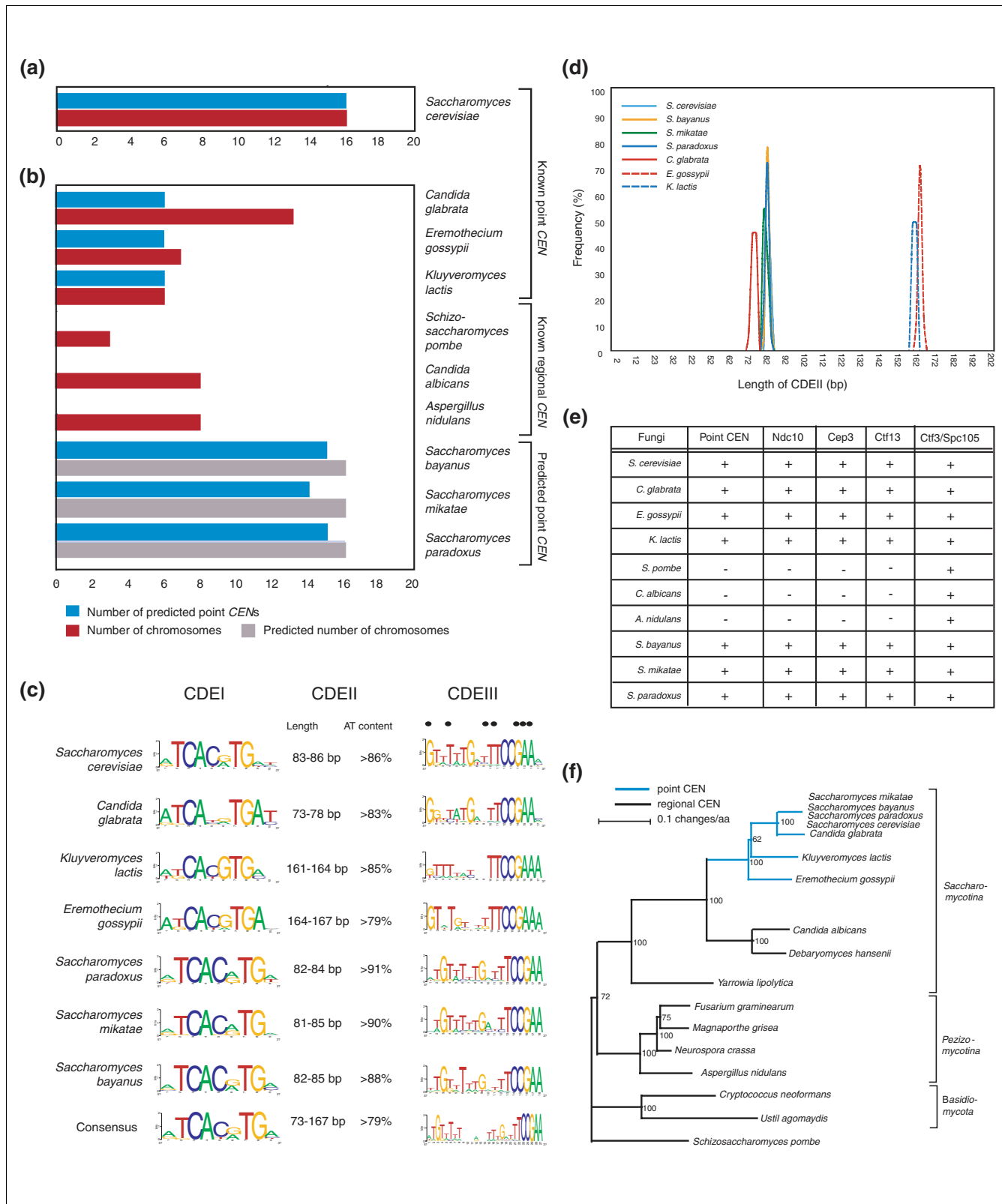
protein' is used rather loosely, all linkers exhibit a clear hierarchical relationship with respect to DNA and MT-binding proteins: linker proteins require DNA binding proteins, and possibly also other linker proteins, for *CEN* DNA binding but not MTs or MT-associated proteins (MAPs).

Kinetochore assembly in *S. cerevisiae* is initiated by association of the essential four-protein CBF3 complex with the CDEIII region of *CEN* DNA. CBF3-CDEIII association then recruits several additional DNA binding proteins, including scCse4, a specialized histone H3 found only at *CEN*s (CenH3). CenH3-containing nucleosomes are thought to be core components of all kinetochores [13]. When *CEN* associated, the DNA binding subunits of *S. cerevisiae* kinetochores recruit four essential multi-protein linker complexes, the NDC80 complex (four proteins), COMA (four proteins), MIND (four proteins) and the SPC105 complex (two proteins). These complexes, in turn, recruit a multiplicity of motor proteins and MAPs to form a fully functional MT-attachment site (P De Wulf and PK Sorger, unpublished observation) [14-16].

A key question in the study of kinetochores is whether architectural features currently being elucidated in *S. cerevisiae* are conserved in higher cells. Some *S. cerevisiae* proteins have been shown to have orthologs in one or more metazoa. These metazoan orthologs include CenH3, CENP-C<sup>Mif2</sup>, Mis6<sup>Ctf3/CENP-1</sup>, Spc105<sup>KNL-1/Kia1570</sup>, members of the NDC80 and MIND complexes as well as MT-associated proteins such as EB1<sup>Bim1</sup> and CLIP170<sup>Bik1</sup>, Mad-Bub spindle checkpoint proteins and some regulatory kinases [2,17-26]. To date, however, only CenH3 and CENP-C have been carefully compared at a sequence level in a wide range of organisms [27]. Here we report a systematic analysis of sequence relationships among a set of approximately 50 fungal, plant and metazoan kinetochore proteins with the overall aim of exploring their structural and evolutionary relationships. Our analysis supports the conclusion that the four linkers at the core of *S. cerevisiae* kinetochores, the NDC80 complex, MIND, COMA, and the SPC105 complex, have been conserved through eukaryotic evolution. A subset of kinetochore proteins, perhaps 20% of the total in *S. cerevisiae*, seems to be specific to point *CEN*s, all of which are very closely related. A second set of kinetochore proteins is found only on regional *CEN*s. It appears, therefore, that all kinetochores have a single ancestor, proba-

### Figure 1 (see following page)

Point centromeres are derived from regional centromeres and appeared only once during evolution. **(a)** The 16 *CEN*s from *S. cerevisiae* were used to train a HMM. The blue bar indicates the number of predicted point *CEN*s in the genome and the red bar represents the number of known chromosomes. **(b)** HMM from (a) was used to search the genome of fungi with known point *CEN*s, known regional *CEN*s and predicted point *CEN*s. Blue and red bars are as described in (a) except gray bars, which indicate the predicted number of chromosomes, based on synteny within other *Saccharomyces* species. **(c)** Sequence comparison of the CDEI, CDEII and CDEIII elements from budding yeast with point centromeres. **(d)** Frequency distribution of the CDEII length (measured in bp) in each budding yeast with point centromeres. **(e)** Evolutionary conservation of CBF3 subunits in fungi with point and regional *CEN*s. **(f)** Phylogenetic analysis of 17 different fungi, including the 7 budding yeast with point centromeres and the 3 budding yeast with regional centromeres using 3 highly conserved reference proteins ( $\alpha$ -tubulin, the signal recognition protein SRP54 and the DNA replication factor PCNA). Blue branches represent fungi with point centromeres and black branches those with regional centromeres.



**Figure 1** (see legend on previous page)

**Table 1****Sequence similarities among selected fungal kinetochore proteins of point *CEN***

Location	Complex	Protein	Ubiquitous	Point <i>CEN</i> specific	Similarity*	Identity*	
DNA-binding	Monomer	Mif2	+		65%	23%	
	?	Sgt1	+		74%	28%	
	CBF3	Cep3			+	65%	14%
		Ctf13			+	53%	9%
		Ndc10			+	48%	10%
Linker layer	COMA	Mcm21	+		45%	7%	
		Ctf19	+		47%	7%	
		Amel		+	45%	9%	
		Okp1		+	51%	7%	
	MIND	Nnf1	+		67%	14%	
		Nsl1	+		69%	15%	
	NDC80	Ndc80	+		73%	20%	
		Spc24	+		63%	6%	
	SPC105	Spc105	+			48%	5%
		Ydr532C			+	52%	6%
		?	Chl4	+		52%	11%
		?	Ctf3	+		51%	6%
		?	Nkp1†		+	55%	6%
		?	Nkp2†		+	63%	6%
		?	Mcm16		+	52%	7%
		?	Mcm22		+	53%	4%
?	Iml3‡	+		24%	6%		
?	Cnn1			+	40%	4%	
MT-binding	DASH	Ask1	+		43%	11%	
		Dam1	+		54%	6%	
Regulatory	?	Bub3	+		65%	18%	
	?	Mad2	+		98%	54%	

\*As determined from the proteins in *S. cerevisiae*, *C. glabrata*, *E. gossypii* and *K. lactis*. †Instead of *E. gossypii* the sequences were derived from the very closely related *S. kluyveri*. ‡Similarity was determined from proteins of the point *CEN* containing *S. cerevisiae*, *S. kudriavzevii*, *K. waltii* and *S. kluyveri*.

bly based on a regional *CEN*, from which contemporary kinetochores diverged rapidly while conserving key structural features.

## Results

### Point centromeres have a common origin

As a first step in determining relationships among kinetochores in different organisms, we searched fungal genomes for point *CEN*s similar in structure to those in *S. cerevisiae*. Three such examples are already known, *C. glabrata*, *E. gossypii* and *K. lactis* [28], but a significant number of newly sequenced genomes have not yet been analyzed. Finding new *CEN*s with a CDEI-CDEII-CDEIII structure is not trivial because the number of identical bases in CDEI and CDEIII is relatively small, even among chromosomes in *S. cerevisiae*. Moreover, CDEII is not conserved in sequence but, rather, is characterized by high AT content and alternating runs of poly-A and poly-T. To capture this information we constructed a tri-partite computational model based on profiles

for CDEI and CDEIII, a hidden Markov model (HMM) for CDEII (Figure 1a), and *S. cerevisiae* *CEN*s as a training set. When the model was tested on *C. glabrata*, *E. gossypii* and *K. lactis*, organisms whose genomes are fully annotated, 6/13 centromeres in *C. glabrata*, 6/7 centromeres in *E. gossypii* and 6/6 in *K. lactis* were identified correctly (Figure 1b). Conversely, no point-*CEN* sequences were found in *S. pombe*, *C. albicans* or *A. nidulans*, organisms known to have regional *CEN*s (Figure 1b). With a success rate of >70% and a false positive rate of <5%, we conclude that our computer model is effective at finding point *CEN*s.

When unannotated genomes were analyzed using the tri-partite computational model, 15 CDEI-II-III sequences were found in *S. bayanus*, 14 in *S. mikatae* and 15 in *S. paradoxus* (Figure 1b) [29]. *S. bayanus*, *S. mikatae* and *S. paradoxus* contigs have not yet been fully assembled, but sequence similarity and synteny suggest that all 3 have 16 chromosomes, close to the number of putative *CEN* sequences identified computationally in each organism. When these newly identi-

fied point *CENs* were combined with those in the literature, 85 CDEI-II-III sequences from 7 organisms became available. These yielded a clear consensus for CDEI and CDEIII and revealed that, within a single organism, CDEII can vary in sequence from one chromosome to the next but that length distributions are very narrow ( $\pm 3\%$ ; Figure 1c, d). Most fungi have 84 bp CDEII sequences but *E. gossypii* and *K. lactis* have 164 bp CDEIIs, suggesting the presence of two copies of an underlying approximately 84 bp CDEII module (Figure 1d). To a first approximation, the extent of conservation among CDEI and CDEIII sequences on different chromosomes within a single organism was not much greater than the extent of conservation among syntenic *CENs* in different organisms (Figure 1c). Together, these data strongly imply that all organisms with CDEI-II-III point *CENs* arose from a relatively recent common ancestor.

### Kinetochores proteins specific to organisms with point centromeres

Does the existence of *CENs* with similar CDEI-II-III structures imply the existence of similar DNA-binding kinetochore proteins? In addressing this question, the CDEI-binding Cbf1 protein is not very useful because it functions not only as a kinetochore subunit but also as a transcription factor for a set of highly conserved biosynthetic genes [30], implying conservation of non-kinetochore function. We therefore concentrated on components of the CBF3 complex, three of whose subunits are thought to function only in CDEIII-binding (the fourth subunit, scSkp1, is also a component of the SCF ubiquitin ligase complex [31] and, like Cbf1, has conserved non-kinetochore functions). When PSI-BLAST was used to search predicated open reading frames in 17 fungal genomes for orthologs of scCtf13, scCep3 and scNdc10, all 3 CBF3 subunits were found in the organisms with point *CENs* (7 in total), but not in organisms with regional *CENs* (Figure 1e). As a positive control for the PSI-BLAST search, orthologs of scMis6<sup>Ctf3</sup> and scSpc105 could be found in all fungi examined (Figure 1e). Importantly, Mis6<sup>Ctf3</sup> and Spc105 have approximately the same degree of sequence divergence in point-*CEN* containing fungi (51% and 48% similarity, respectively) as Ndc10 (48% similarity; Table 1). We provisionally conclude that CBF3 proteins are present only in fungi with CDEI-II-III *CEN* DNA whereas other kinetochore proteins (such as Spc105 and Ctf3) are ubiquitous. Moreover, when organisms with point *CENs*

and CBF3 subunits are mapped on a phylogenetic tree (constructed using the highly conserved reference proteins  $\alpha$ -tubulin, the signal recognition particle subunit SRP54 and PCNA) they were found to cluster closely together (Figure 1f). While recognizing the possibility for false-negative findings in cross-species sequence searching, we conclude that CDEI-II-III *CENs* and CBF3 *CEN*-binding proteins are probably found only in a subset of closely related budding yeasts and, thus, may have co-evolved. Intriguingly, the apparent common ancestor of point-*CEN* and regional-*CEN* organisms appears to be a fungus containing regional *CENs*, implying that simple point *CENs* arose from complex regional *CENs* and not the other way round.

To delineate further which kinetochore proteins are specific to point *CENs*, and which are more widely distributed, we analyzed all known *S. cerevisiae* kinetochore proteins for sequence conservation. As a starting point we examined scMis12<sup>Mtw1</sup> and scNdc80<sup>Hec1</sup>, kinetochore proteins first identified in yeast and subsequently shown to have human orthologs (hsMis12 and hsNdc80<sup>Hec1</sup>) that localize to kinetochores and play a role in chromosome segregation [20,25]. Experimental and sequence data establish that yeast and higher cell Ndc80<sup>Hec1</sup> and Mis12<sup>Mtw1</sup> proteins represent true orthologs [20,32-34]. Nonetheless, the overall degree of similarity among Ndc80<sup>Hec1</sup> and Mis12<sup>Mtw1</sup> proteins across eukaryotes was found to be relatively modest (approximately 15% to 30%) as compared to proteins involved in DNA replication (PCNA, approximately 75%) or protein translocation (SRP54, approximately 60%). Multiple protein sequence alignments of fungal, plant, and metazoan Ndc80<sup>Hec1</sup> and Mis12<sup>Mtw1</sup> showed that sequence similarity is confined to 30 to 100 residue blocks interspersed by stretches of non-homology, many of which correspond to coiled coils (Figure 2a, b). This pattern of block-by-block similarity was also observed with five other kinetochore proteins for which orthology has been established experimentally, and is consistent with previous proposals that kinetochore proteins have evolved rapidly [35] (Figure 2c). Importantly, for our purposes, data obtained from known kinetochore orthologs suggests that it is necessary to use conserved blocks, rather than complete sequences, when searching kinetochore proteins for patterns of sequence conservation.

### Figure 2 (see following page)

Sequence similarity between kinetochore proteins is restricted to short stretches between orthologs. Multiple sequence alignments of the (a) Mis12<sup>Mtw1</sup> and (b) Ndc80<sup>Hec1</sup> families. Schematic drawing above the alignment indicate the length of the *S. cerevisiae* proteins and the percentages denote the degree of similarity of successive sequence blocks (black boxes) within fungi (red letters) or fungi, metazoa and plantae (green letters). The schematic drawing above the Ndc80 multiple sequence alignment also indicates the relative position of the globular and coiled-coil domain of Ndc80, as determined by electron-microscopy [32,33]. White letters on black denote identical residues, white letters on green, identical residues in  $\geq 80\%$  of the organisms and black letters on green, similar residues in  $\geq 80\%$  of the organisms. (c) Schematic drawings indicating the percentage similarity of successive sequence blocks (black boxes) within fungi (red letters) or fungi, metazoa and plantae (green letters) based on multiple sequence alignments of the Nuf2, Spc25, Spc24, CENP-C<sup>Mif2</sup> and Mis6<sup>Ctf3/CENP-1</sup>, PCNA and SRP54 protein families

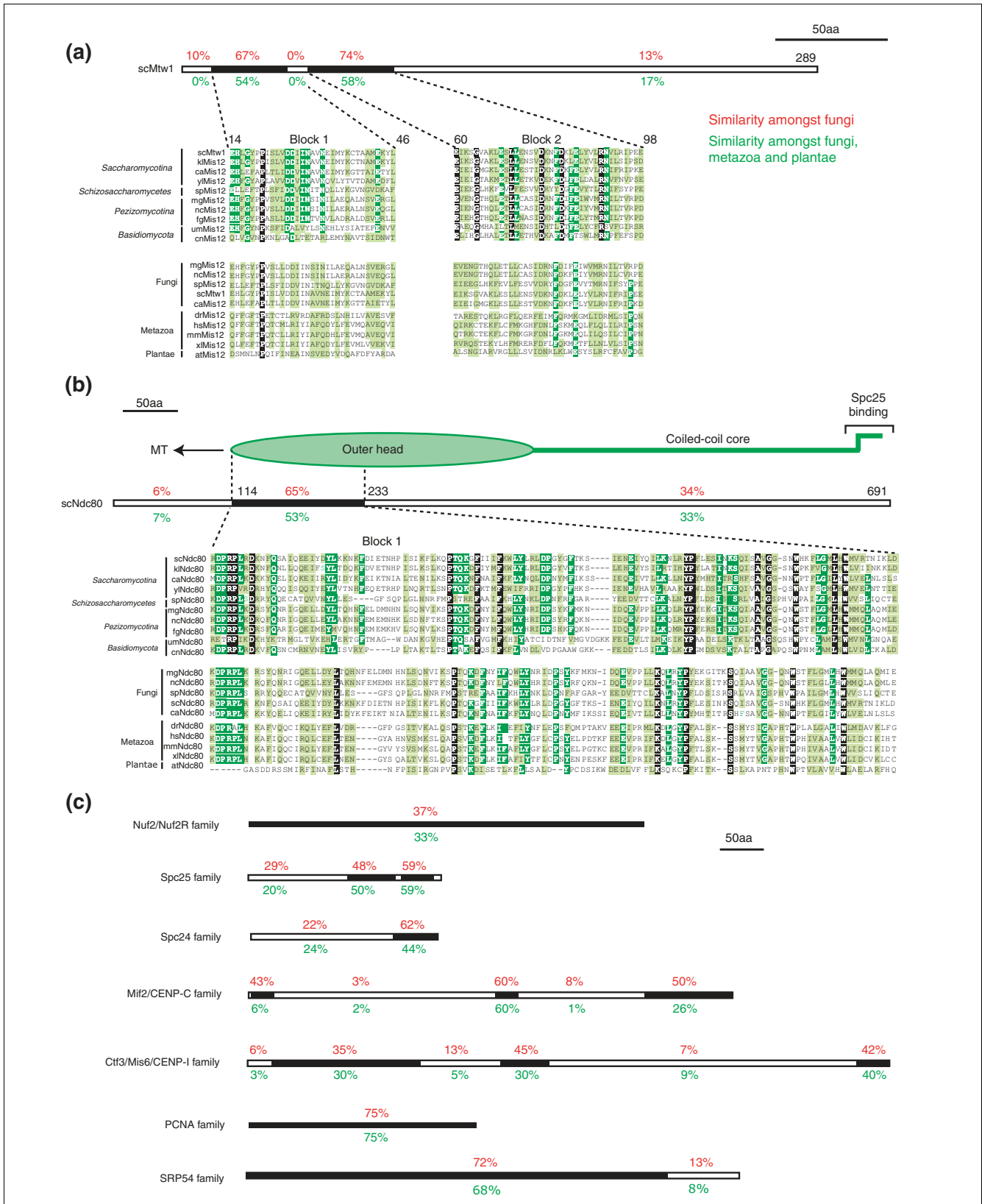


Figure 2 (see legend on previous page)

When 55 *S. cerevisiae* kinetochore proteins (including the CBF3 subunits discussed above) were used in PSI-BLAST queries to search 14 fully annotated fungal genomes (Additional data file 1), 41 were found to have orthologs in organisms with both point and regional *CENs* (Figure 3). These proteins included kinetochore regulators such as the Mad1-3, Bub1, BubR1/Mad3 and Mps1 checkpoint proteins and the Ipl1-AuroraB kinase, as well as many structural components. In addition to the 41 proteins mentioned above, conservation was observed for proteins such as Skp1 [31], Cbf1 [30,36] and some MAPs [37] that function at kinetochores as well as at other locations in the cell. As noted above, these proteins are likely to have been conserved for reasons other than their presence at kinetochores, and they cannot be used to infer overall similarity in kinetochore structure. In this respect, kinesin motor proteins are also difficult to analyze. Eukaryotic cells contain multiple kinesins, which are known to fall into 14 highly conserved protein families based on sequence, structure and function [38]. Typically, each kinesin has more than one cellular function and kinetochores in different organisms recruit different kinesin family members, making it difficult to determine (in the absence of experimentation) which kinesins should be considered kinetochore associated.

Leaving these complications aside, among 55 fungal kinetochore components analyzed, 11 were found in the 7 organisms with point *CENs* and nowhere else, implying that they are specific to a CDEI-II-III *CEN* architecture (Figure 3). These 11 proteins include the CBF3 subunits scCtf13, scCep3 and scNdc10 described above, the non-essential *CNN1* gene product, 1 subunit of the SPC105 complex (Ydr532c), two subunits of the COMA linker complex (scAme1 and scOkp1) and 4 proteins that require COMA for *CEN*-association (scMcm22, scMcm16, scNkp1 and scNkp2). Among organisms in which they are found, the 11 point *CEN*-specific proteins are as well or better conserved than ubiquitous kinetochore proteins, implying that failure to identify orthologs in more distant fungi is a consequence of their actual absence. We therefore propose that approximately 20% of the overall kinetochore in fungi containing CDEI-II-III *CENs* is specialized to their simple *CENs*. As expected, these specialized kinetochore subunits include proteins in direct contact with *CEN* DNA (Figure 3).

### Identification of novel human kinetochore proteins

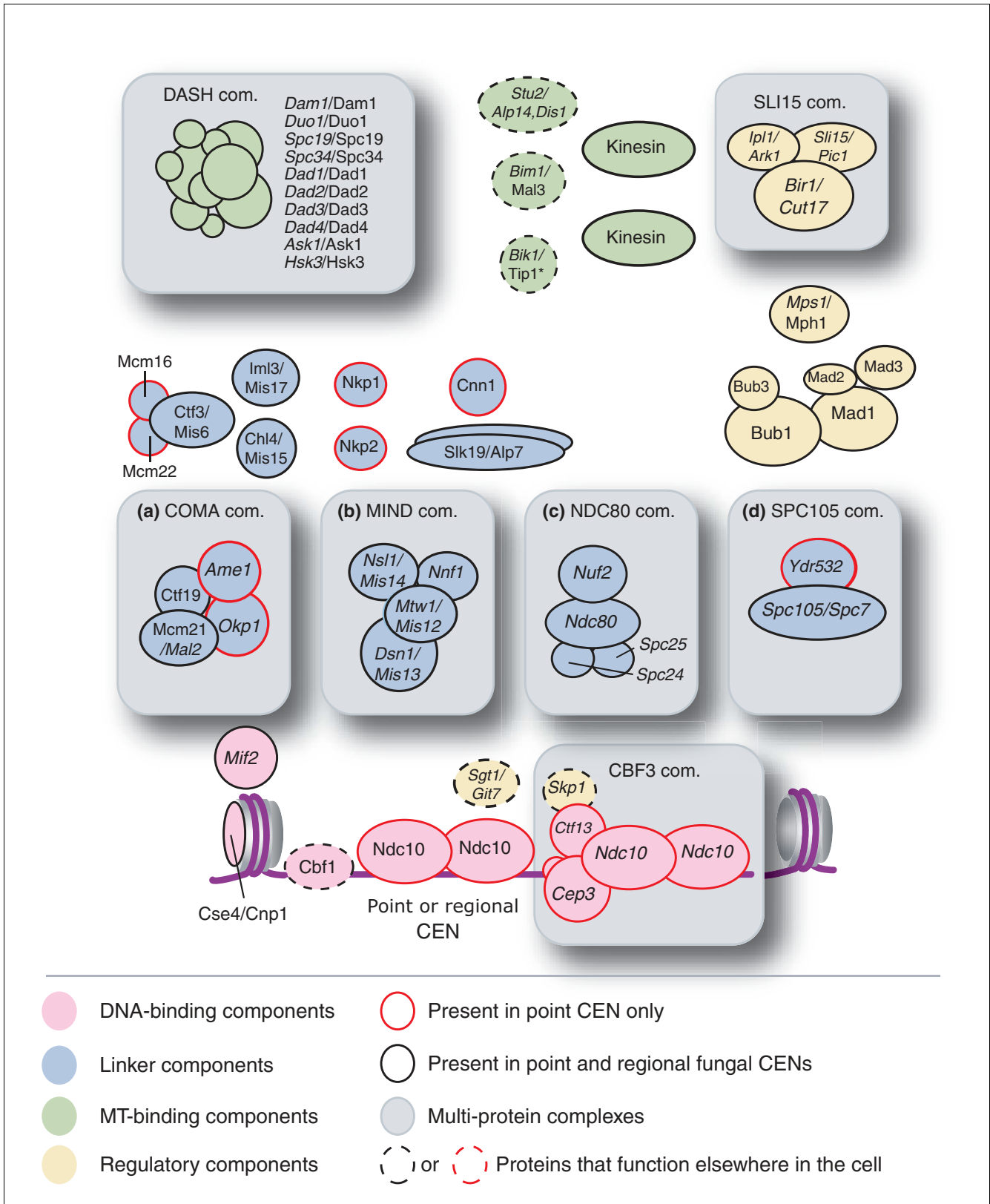
Based on success in identifying fungal orthologs of *S. cerevisiae* kinetochore proteins, we expanded our set of target

organisms to higher eukaryotes (see Figure 4 for a schematic of the approach). Alignments were created for 41 ubiquitous fungal proteins and conserved blocks determined. The non-redundant NCBI protein database was then searched for these conserved blocks using PSI-BLAST or Prosite pattern searching algorithms (see Materials and methods for details). Potential orthologs differing greatly in size from the fungal proteins and candidates with well-established non-kinetochore functions were eliminated from further consideration. The remaining proteins were then aligned to confirm the presence of conserved blocks. This search led to the identification, in a wide variety of organisms, of previously unreported orthologs of many *S. cerevisiae* kinetochore proteins (Additional data file 1), among which were four new human kinetochore proteins (Figure 4). Recent analysis of *S. pombe* kinetochore complexes by mass spectrometry revealed the presence of a set of proteins for which orthologs could not be found in *S. cerevisiae* [39,40]. When conserved sequence blocks from these *S. pombe* proteins were used to search the genomes of higher eukaryotes, two additional human proteins were flagged as likely kinetochore subunits (Figure 4). Regardless of which fungi contributed to the sequence blocks, the most highly conserved kinetochore subunits were invariably regulatory proteins such as the Mad and Bub checkpoint proteins and the Aurora B kinase. Structural proteins such as Ndc80<sup>Hec1</sup>, Nuf2, CENP-C<sup>Mif2</sup> and Mis12<sup>Mtw1</sup> were considerably more diverged.

The four human proteins representing hitherto unrecognized orthologs of *S. cerevisiae* kinetochore subunits were provisionally named hsNnf1-Related (hsNnf1R; also known as PMF1 [41]; Figures 4 and 5), hsNsl1R (also known as DC8 or DC31), hsMcm21R and hsChl4-R. hsNnf1R shares with its fungal counterpart 2 conserved blocks of 30 to 35 residues with 47% and 67% similarity, hsNsl1R shares 1 conserved block of 35 residues with 43% similarity, hsMcm21R shares 3 conserved blocks of 15 to 30 residues with 46%, 87% and 33% similarity and hsChl4R shares 2 conserved blocks of 20 and 50 amino acids with 45% and 40% similarity (Figure 5). The potential human orthologs of *S. pombe* Fta1 and Sim4 were provisionally named hsFta1R and hsSim4R (also known as Solt [42]). hsFta1R shares with its fungal counterpart three conserved sequence blocks of 40, 25 and 30 residues with 48%, 49% and 58% similarity and hsSim4R one block of 27 residues with 65% similarity (Figure 6). Elsewhere we will describe experimental data showing that hsChl4R, hsNsl1R,

**Figure 3** (see following page)

Fungal kinetochores contain a set of point centromere specific components. Schematic model of kinetochore subunit organization based on the architecture of the *S. cerevisiae* kinetochore. Kinetochore proteins can be roughly divided into DNA-binding (pink), linker (blue), MT-binding (green) and regulatory layers (yellow). Within each layer many proteins are organized into multi-protein complexes, for example, the linker layer is composed of at least four complexes (gray boxes (a) to (d)): COMA, NDC80, MIND and SPC105. Protein names are given for *S. cerevisiae* first and *S. pombe* second, while essential genes (italic letters) and non-essential (normal letters) is indicated. Protein names followed by an asterisk indicate that this specific ortholog is known not to localize to kinetochores. The kinesins present at kinetochores in *S. cerevisiae* are Kip3 (Kinesin-8), Cin8 (Kinesin-5), Kip1 (Kinesin-5) and Kar3 (Kinesin-14), while in *S. pombe* they are Klp5 (Kinesin-8), Klp6 (Kinesin-8) and Klp2 (Kinesin-14) (for nomenclature see [38]).



**Figure 3** (see legend on previous page)

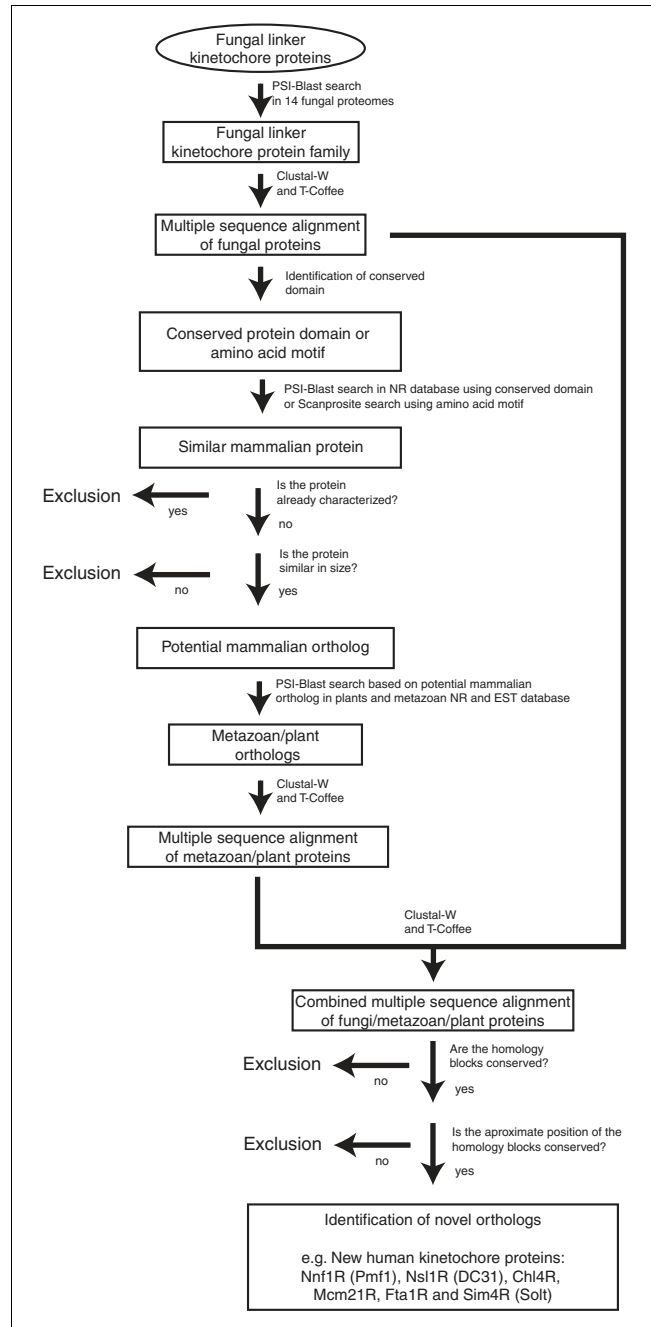


hsMcm21R, hsNnf1R, hsFta1R and hSim4R localize to kinetochores in human cells and are required for accurate chromosome segregation (AD McAinsh *et al.*, submitted). Importantly, for the purposes of the current analysis, the identification of new human kinetochore proteins means that one or more subunits are present in metazoans for each of the four multi-protein linker complexes forming the core of the *S. cerevisiae* kinetochore. Thus, it appears that simple point *CENs* in budding yeast and complex regional *CENs* in human cells probably share fundamental architectural similarities.

*S. cerevisiae* DASH is a 10-protein MT-binding complex that has attracted considerable recent interest because it forms rings encircling MTs [43,44]. DASH subunits are conserved among fungi but we have found few if any potential orthologs in higher eukaryotes. The closest match to a DASH protein in humans, NYD-SP28 [45], has an amino-terminal domain of about 30 amino acids 40% similar to *S. cerevisiae* Spc34 (Additional data file 2). The *Chlamydomonas reinhardtii* ortholog of NYD-SP28 localizes to the flagellum [46], implying that NYD-SP28 might be involved in interactions with MTs. Our preliminary conclusion is that higher eukaryotes do not contain a protein complex closely related to fungal DASH, although further investigation of NYD-SP28 is warranted.

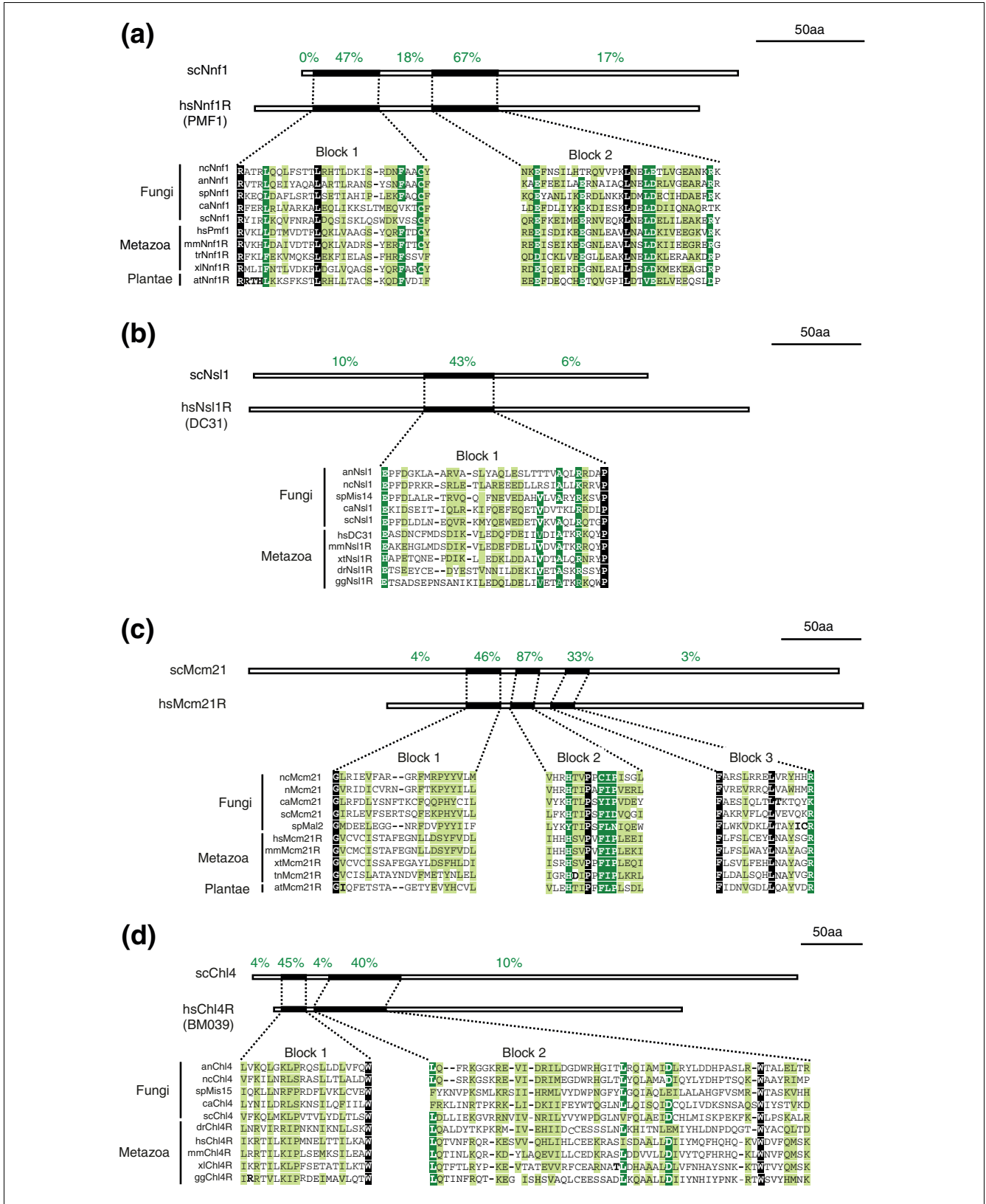
**Correspondence between human kinetochore proteins and their yeast counterparts**

Several kinetochore proteins first identified in human cells have previously been shown to have fungal orthologs, including CENP-C (orthologous to scMif2p [47]) and CenH3<sup>CENP-A</sup> (orthologous to scCse4 [48]). We therefore wondered whether additional orthologs might be found in fungi for kinetochore proteins hitherto characterized only in higher eukaryotes, such as CENP-E, CENP-H, Rod, Zwint and Zwilch [49-53]. We found that, among fungal proteins, hsCENP-H is most similar to *S. pombe* spFta3 (Figure 7a), which was shown recently to be a fission yeast kinetochore protein [39]. It has been suggested previously that *S. cerevisiae* scNnf1 is the budding yeast CENP-H ortholog [54] (Figure 7b) but we find that scNnf1 is actually much more similar to hsNnf1R<sup>Pmf1</sup> and spNnf1 than to CENP-H (Figure 7c). We therefore propose that CENP-H is orthologous to the fungal Fta3 family of proteins. Searches using PSI-BLAST revealed that the Fta3 protein, like the Sim4 and Fta1 proteins with which it interacts in *S. pombe* [39], has apparent orthologs only in organisms with regional *CENs* (Additional data file 1). The presence of Sim4 and Fta1 in the budding yeast *Yarrowia lipolytica*, which has regional *CENs*, but not in yeasts with point *CENs*, is striking, since *Y. lipolytica* is significantly closer in overall sequence to *S. cerevisiae* than to *S. pombe*. We therefore conclude that Fta3, Sim4 and Fta1 are members of a class of kinetochore proteins found specifically in fungi and metazoa with regional *CENs* and not in fungi with point *CENs*.



**Figure 4** Schematic describing the sequence-search based approach used to identify fungal, metazoan, and plant orthologs of the kinetochore proteins scNnf1, scNsl1, scChl4, scMcm21, spSim4 and spFta1. Since such sequence-based searches can yield a significant number of false positives, strict exclusion criteria were applied to ensure the identification of orthologs.

In contrast to CenH3<sup>CENP-A</sup>, CENP-C and CENP-H, potential orthologs of the human CENP-E, Rod, Zwint and Zwilch proteins were not found in any of the fungi examined. The apparent absence of a fungal Rod or Zwilch is particularly interesting, since their binding partner at human kinetochores, Zw10, has a potential ortholog in *S. cerevisiae*, Dsl1



**Figure 5** (see legend on next page)

**Figure 5** (see previous page)

Identification of potential orthologs of scNnf1, scNsl1, scMcm21 and scChl4 in humans. *S. cerevisiae* (a) Nnf1, (b) Nsl1, (c) Mcm21 and (d) Chl4 were aligned with five fungal, four metazoan and one plant sequence. White letters on black denote identical residues, white letters on green, identical residues in  $\leq 80\%$  of the organisms and black letters on green, similar residues in  $\leq 80\%$  of the organisms. Schematic drawings above the alignments indicate the length of the *S. cerevisiae* proteins and the percentages denote the degree of similarity of successive sequence blocks (black boxes).

[55]. Both hsZw10 and scDsl1 play a role in membrane trafficking during interphase [56], but scDsl1 is not known to localize to kinetochores. Thus, whereas human hsZw10 functions in vesicle-MT and chromosome-MT interaction, scDsl1 appears to have only the former function, presumably because Rod and Zwi10 are not present. The absence of Zw10 from fungal kinetochores is also sufficient to explain the absence of Dynein: the Rod/Zw10/Zwi10 (RZZ) complex is needed for the association of Dynein with human and *Drosophila* kinetochores [57]. Considering these data together, we conclude that animal cell kinetochores contain proteins, currently comprising perhaps 25% of the total (and likely to increase), that are absent in fungi with either regional or point *CENs*.

**Evolutionary relationships among kinetochores**

Thus far, we have distinguished only between point and regional *CENs* but a more nuanced view can be obtained from phylogenetic analysis of kinetochore structural proteins. As a reference for these comparisons, a tree was constructed by combining data on three well-conserved eukaryotic proteins:  $\alpha$ -tubulin, PCNA and SRP54 (Figure 8a; this reference tree closely matches reference trees constructed by others [58,59]). The reference tree exhibited prototypical clustering of fungi in one branch and metazoa in another so that *Drosophila* and *C. elegans* were much closer to humans than *S. pombe* or *S. cerevisiae*. However, the phylogenetic trees for Ndc80<sup>Hec1</sup> and Nuf2 were remarkably different: overall sequence divergence was much greater and *Drosophila* and *C. elegans* Ndc80<sup>Hec1</sup> (or Nuf2) proteins were not significantly more similar to their human than their fungal counterparts (Figure 8b, c). *Drosophila* Ndc80 and Nuf2 were particularly striking in occupying a branch of the phylogenetic tree distant from all other animals. This great divergence in *Drosophila* kinetochore protein sequence is also illustrated by the fact that, apart from regulatory components, such as the Mad-Bub proteins and a few MAPs, only a limited number of structural kinetochore proteins have been identified in flies (for example, CENP-C [60], CenH3<sup>CID</sup> [61], the RZZ complex [62], Ndc80<sup>Hec1</sup>, Nuf2, and Mis12<sup>Mtw1</sup>; Figure 9).

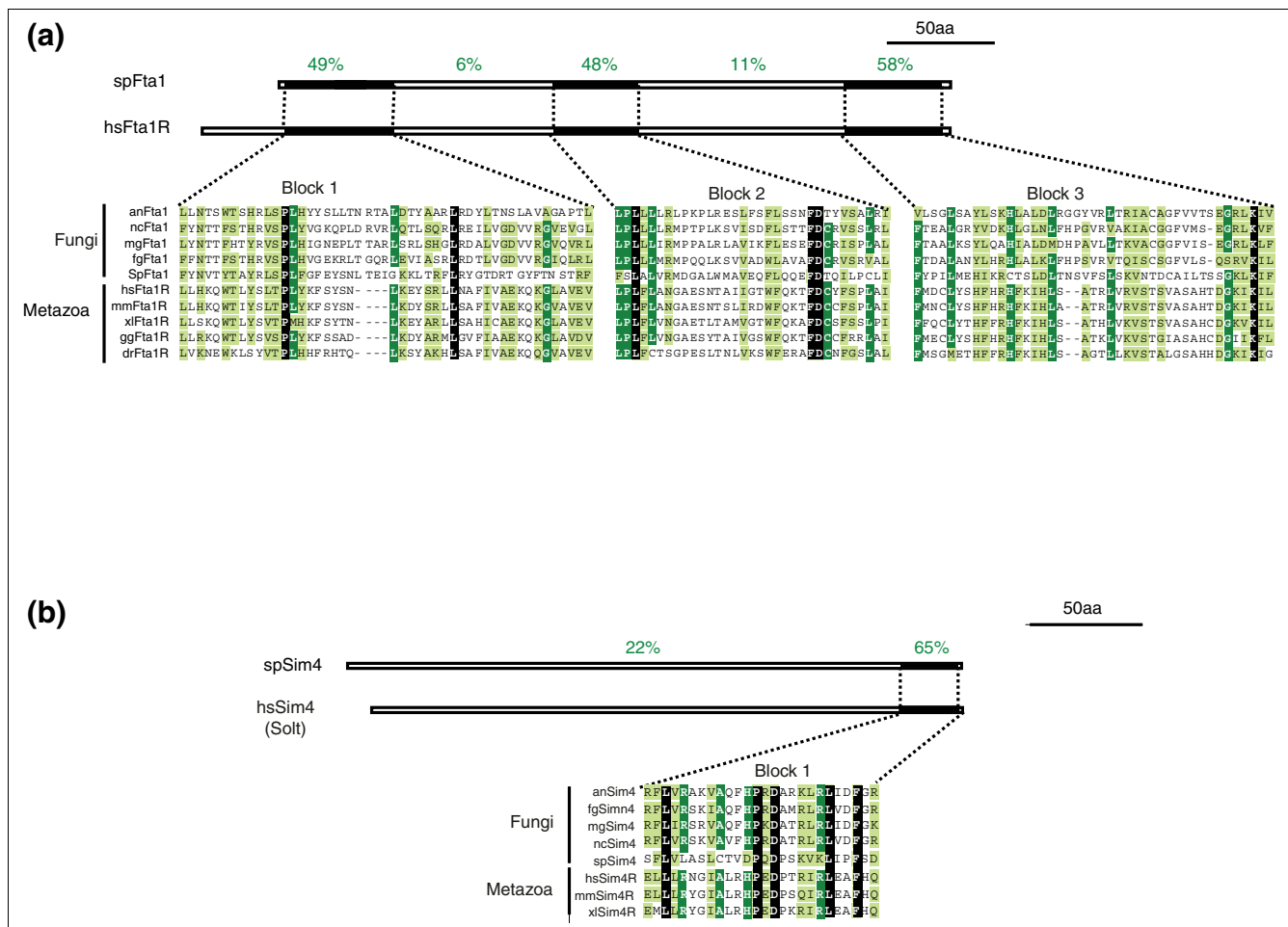
**Organization of the simplest kinetochore**

*Encephalitozoon cuniculi* is a microsporidium and intracellular parasite that has been subjected to considerable evolutionary pressure to reduce its genome to the smallest possible size. As a consequence, *E. cuniculi* and related microsporidia have the smallest known eukaryotic proteome (1,997 potential open reading frames) and many cellular structures

in *E. cuniculi* lack redundant and non-essential genes [63]. Using our HMM for CDEI-II-III, no sequences similar to point *CENs* were found on any of the 11 *E. cuniculi* chromosomes, nor were CBF3 proteins found by PSI-BLAST (Figure 10a). We therefore speculate that *E. cuniculi* contains a regional *CEN* of some sort. Orthologs of CenH3 and CENP-C<sup>Mif2</sup> are present in *E. cuniculi*, as are all four components of the NDC80 linker complex, three components of MIND and SPC105 (Figure 10b, Additional data file 3). No subunits of COMA, the fourth *S. cerevisiae* linker, were found. Among regulatory proteins, *E. cuniculi* Ipl1/Aurora B and Survivin<sup>Bir1</sup> orthologs were present as were Mps1 and Bub3, but not other proteins required for the spindle assembly checkpoint in yeast or human cells (Figure 10b). When Cdc20, an essential activator of the anaphase promoting complex (APC/C) was examined for sequence motifs, further evidence was obtained that *E. cuniculi* lacks a spindle checkpoint. APC/C is an E3 ligase required for the ubiquitination of proteins whose destruction is necessary at the metaphase-anaphase transition [64]. In all eukaryotes examined to date, an activated form of the Mad2 checkpoint protein binds to Cdc20 via a short conserved peptide so as to block Cdc20 from activating APC/C, thereby arresting cells at the metaphase-to-anaphase transition [65,66] (Figure 10c). *E. cuniculi* Cdc20 contains the WD-domain implicated in APC/C interaction but lacks any sequence similar to a Mad2 binding domain (Figure 10c), implying that it is not subject to checkpoint control. From these data we conclude that *E. cuniculi* probably contains a very simple kinetochore, based on a regional *CEN* that contains about one-half the proteins found in *S. cerevisiae*. In contrast, other large multi-protein structures in *E. cuniculi* are only slightly less complex than their higher eukaryotic counterparts. For example, *E. cuniculi* ribosomes are composed of 77 subunits as compared to 84 subunits in *S. cerevisiae*. Symptomatic of the simplicity of the *E. cuniculi* kinetochore is the absence of the vast majority of potential MAPs. Nonetheless, it is significant that the *E. cuniculi* kinetochore contains three of the four linker complexes that appear to form the core of budding yeast and human kinetochores.

**Discussion**

Extensive genetic and biochemical experimentation has made *S. cerevisiae* kinetochores the best characterized structures involved in chromosome-MT attachment [5]. *S. cerevisiae* kinetochores contain upwards of 70 protein subunits assembled into 14 or more multi-protein complexes. In this study we used similarity-based sequence searching to ascertain

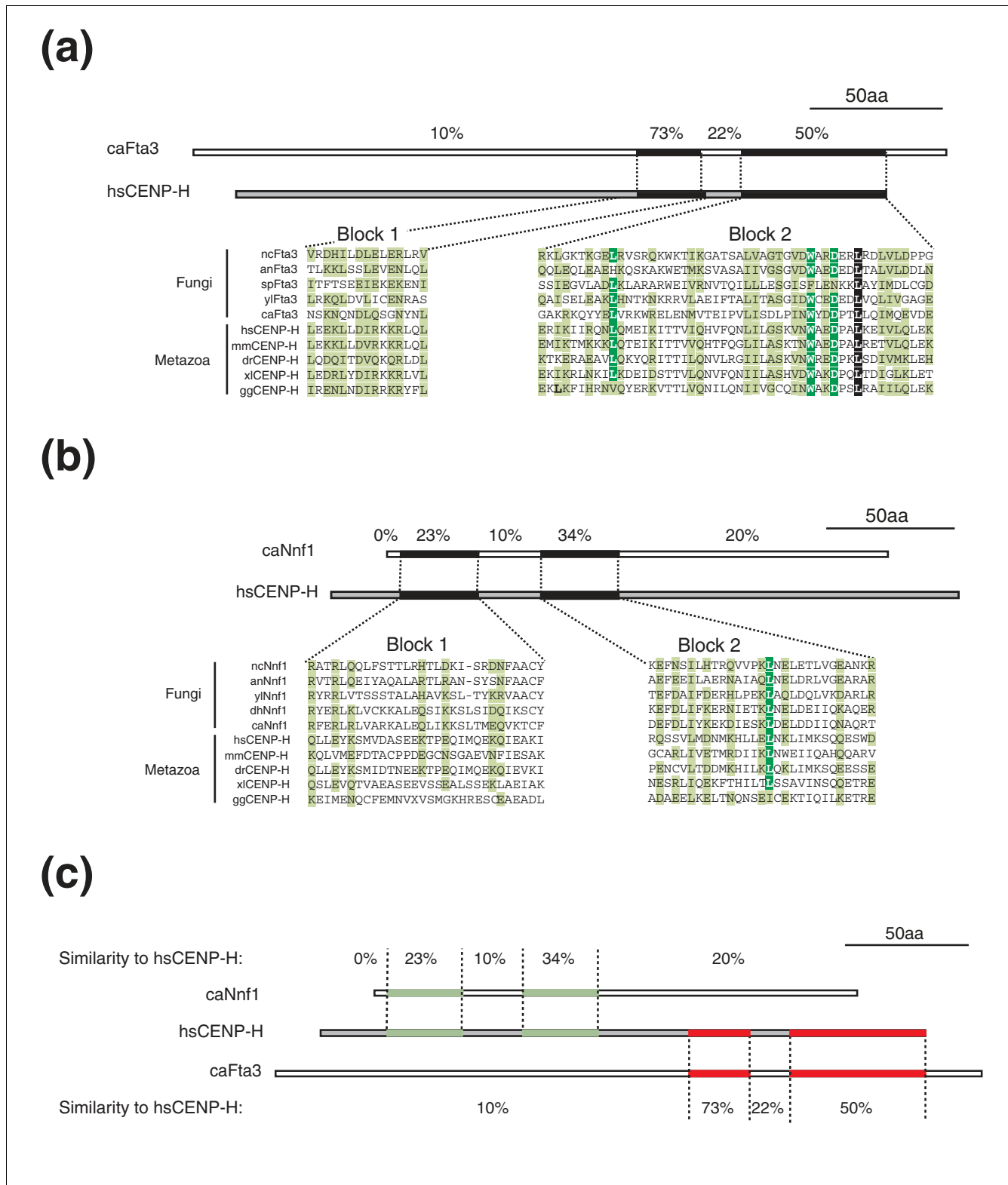
**Figure 6**

Identification of potential orthologs of spFta1 and spSim4 in humans. *S. pombe* (a) Fta1 and (b) Sim4 were aligned with five fungal, and three to five metazoan sequences. White letters on black denote identical residues, white letters on green, identical residues in  $\geq 80\%$  of the organisms and black letters on green, similar residues in  $\geq 80\%$  of the organisms. Schematic drawings above the alignments indicate the length of the *S. cerevisiae* proteins and the percentages denote the degree of similarity of successive sequence blocks (black boxes).

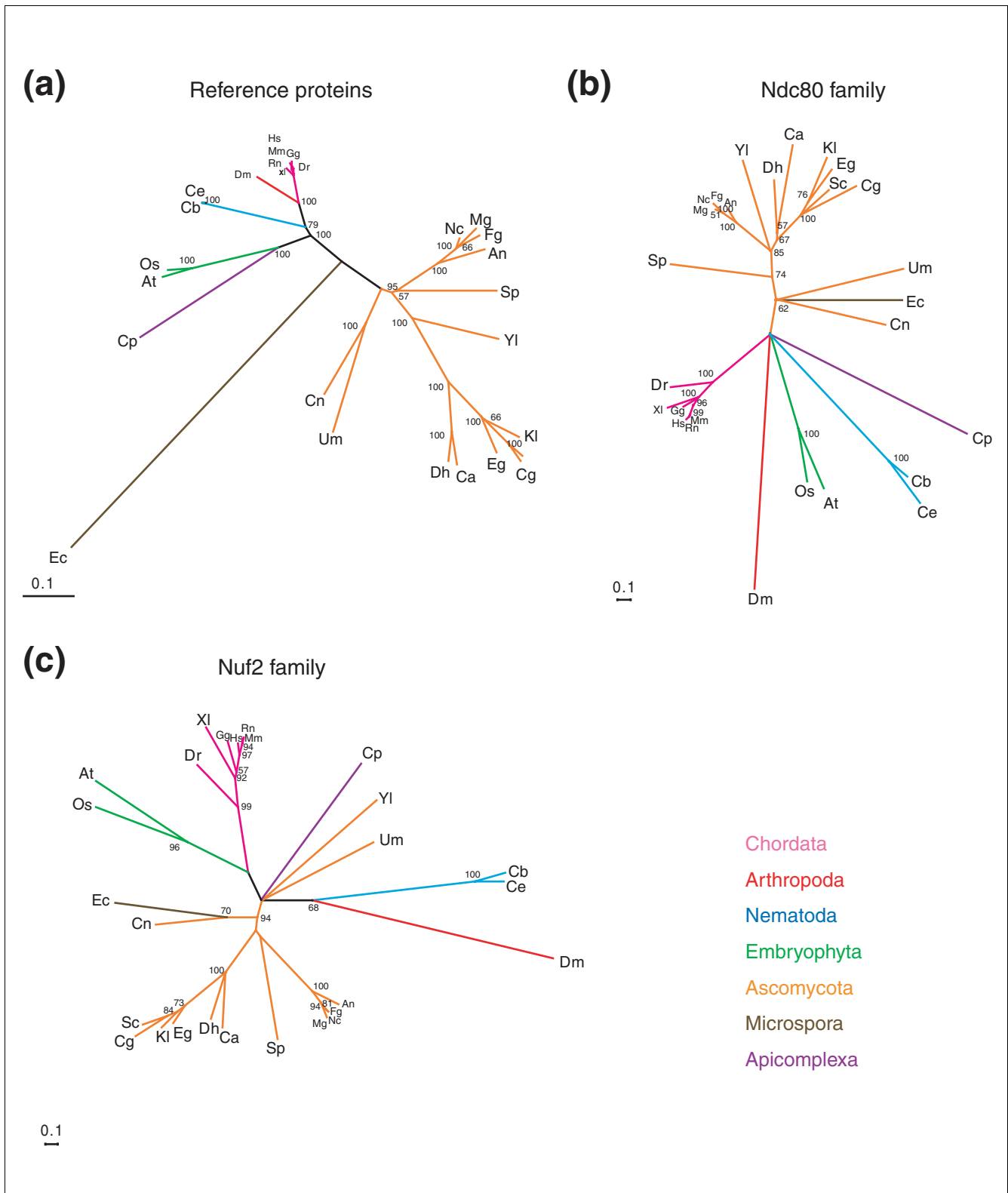
which *S. cerevisiae* kinetochore proteins have orthologs in 15 fungi, 11 metazoa and 2 plants (Additional data file 1) with the overall aim of determining which structural features of *S. cerevisiae* kinetochores have been conserved throughout evolution. The analysis is not as straightforward as might be assumed, because kinetochore proteins are among the most rapidly evolving proteins in the genome [67]. In addition, the structure and sequence of *CEN* DNA has diverged widely from organism to organism. Whereas fungi closely related to *S. cerevisiae* contain 125 to 225 bp *CENs* with a CDEI-CDEII-CDEIII structure, most other organisms contain much longer regional *CENs* with few if any conserved sequence elements.

Guided by experimental data on established orthologies in yeast, humans and other organisms, we base most of the conclusions in this paper on the characterization of proteins that share blocks of homologous sequence. In several cases, we also draw inferences from a failure to identify homologous proteins. We recognize that this failure represents a negative

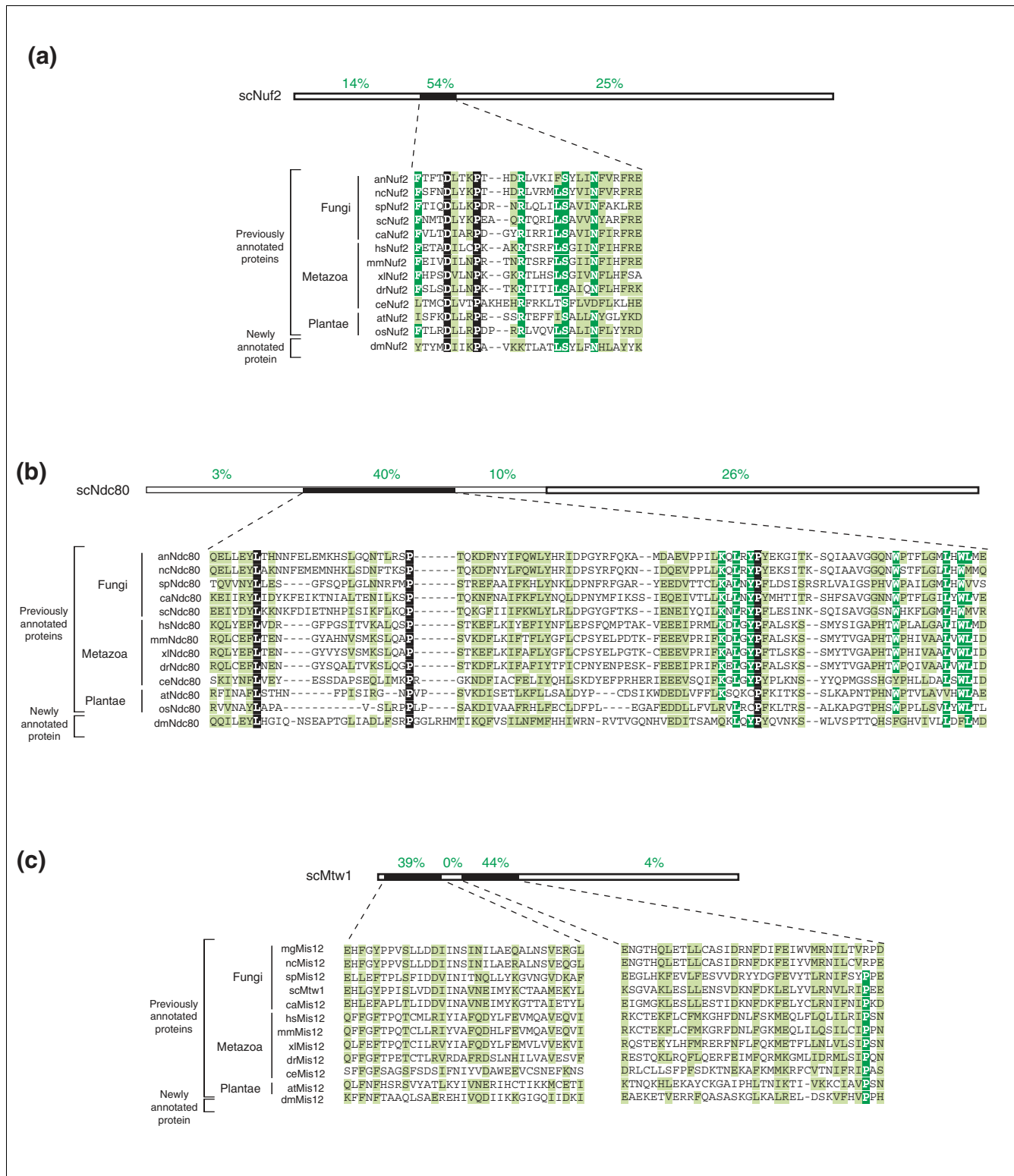
result with many potential causes. However, in cases in which a kinetochore protein is conserved among organisms A, B and C whereas a second kinetochore protein is well-conserved only in species A and B and undetectable in C (and multiple related species), a tentative conclusion can be drawn that the second protein is actually absent from C. For example, we find that CBF3, an essential *CEN*-binding protein in *S. cerevisiae*, has orthologs in seven budding yeasts containing *CEN* DNA conforming to a CDEI-CDEII-CDEIII organization but not in organisms with regional *CENs*. In contrast, other kinetochore proteins similar in their degree of sequence conservation to CBF3 subunits among point *CEN*-containing yeast (approximately 45% to 50% similarity) are found throughout fungi. Thus, we provisionally conclude that CBF3 is present in only fungi with CDEI-CDEII-CDEIII centromeres. Despite the potential for occasional error, our use of both positive and negative findings makes it possible to draw broad conclusions about the organization and possible origins of simple and



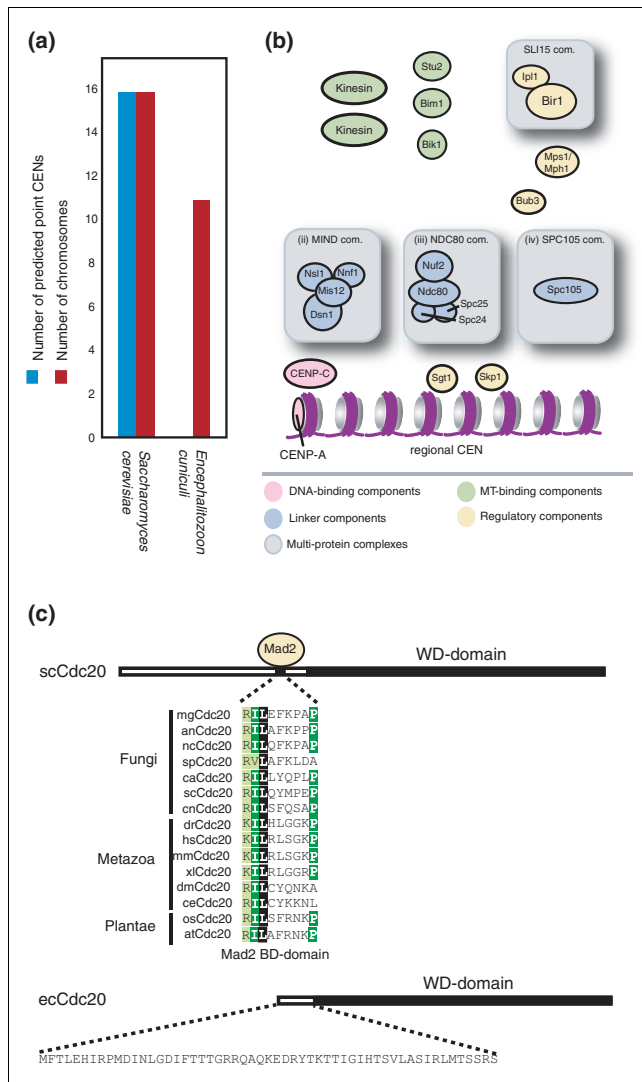
**Figure 7**  
 The human kinetochore protein CENP-H is more closely related to a novel family of fungal proteins than the Nnf1 family. Multiple sequence alignments of metazoan CENP-H proteins and either (a) fungal Fta3 family proteins or (b) fungal Nnf1 family of proteins. Sequences were annotated as in Figure 5. (c) Comparison of sequence similarity between human conserved domains of CENP-H and *C. albicans* Nnf1 and *C. albicans* Fta3.



**Figure 8**  
 Phylogenetic analysis of kinetochore protein conserved domains. Radial phylogenetic trees were assembled for **(a)** reference proteins ( $\alpha$ -tubulin, the signal recognition protein SRP54 and the DNA replication factor PCNA), **(b)** the Ndc80 family and **(c)** the Nuf2 family. For bootstrap analysis, sample size equals 100. Nodes with support less than 50% were collapsed. The accession number for each protein is described in Additional data file 1.



**Figure 9** Identification and annotation of (a) Nuf2, (b) Ndc80 and (c) Mis12 orthologs in *D. melanogaster*. Schematic drawing above the alignment indicate the length of the *S. cerevisiae* proteins and the percentages denote the degree of similarity of successive sequence blocks (black boxes). White letters on black denote identical residues, white letters on green, identical residues in  $\geq 80\%$  of the organisms and black letters on green, similar residues in  $\geq 80\%$  of the organisms. Accession numbers are described in Additional data file 1.



**Figure 10**  
 Identification of a minimal kinetochore in *E. cucullii*. (a) HMM described in Figure 1a, b failed to find a CDEI-II-III structure in the genome of *E. cucullii*. The Green bar indicates point *CENs* identified and black bars the number of chromosomes. (b) Speculative model of *E. cucullii* kinetochore subunit organization. Proteins colored in pink, blue, green or yellow represent components of the DNA binding, linker, regulatory or microtubule-binding layers, respectively, based on kinetochore organization in *S. cerevisiae*. Potential multi-protein complexes are highlighted with a grey box. (c) Sequence alignment of fungal, metazoan and plantae *Cdc20* showing the conserved Mad2 binding site. Note: *E. cucullii* lacks both the conserved Mad2 binding site and an ortholog of the Mad2 protein. Schematic drawings indicate the length of the *S. cerevisiae* and *E. cucullii* proteins, the position of the WD-domain (black box) and the position where Mad2 binds.

complex kinetochores that would not be possible based on a more conservative approach.

### Origins of point centromeres

Based on the simple structure of their *CENs*, it is widely assumed that *S. cerevisiae* kinetochores represent an ancestral structure from which complex regional kinetochores

evolved. Several findings in the current work suggest, however, that CDEI-II-III *CENs* arose in combination with a set of 11 proteins as a specialization of a regional *CEN*. First, all annotated organisms containing point *CENs* (*S. cerevisiae*, *C. glabrata*, *K. lactis*, and *E. gossypii*) have a common origin in one relatively shallow branch of the fungal phylogenetic tree. Were CDEI-II-III sequences an ancestral *CEN*, the current distribution of regional *CENs* would require loss of point *CENs* from multiple independent evolutionary branches. Second, we could obtain no evidence for CDEI-II-III *CEN* DNA or CBF3 proteins in the microsporidium *E. cucullii*, which is thought to have arisen through an ancient divergence in the fungal kingdom [68].

If the speculation that CDEI-II-III point-*CENs* evolved from regional *CENs* is correct, we must consider the possible existence of other short *CENs* that are also based on sequence-specific DNA binding interactions just not CBF3. By way of precedent, the emergence of CDEI-II-III *CENs* is coincident with large-scale chromosomal changes that gave rise to the *HMR*, *HML* and *MAT* loci, thereby changing the sexual potential of *S. cerevisiae* and related yeasts [29]. *S. pombe* and its close relatives undergo mating type switching analogous to that in *S. cerevisiae*, but the molecular mechanisms of switching are completely different [69]. Functional analysis of fungi with short uncharacterized *CENs* will be needed to test the speculation that just as different forms of mating-type switching have developed based on distinct biochemistry, point *CENs* with structures other than CDEI-II-III might exist.

### Evolution of kinetochore proteins

Sequence comparison reveals that conservation among orthologous kinetochore proteins is invariably restricted to relatively short sequence blocks embedded in longer regions of low sequence similarity. The restriction of sequence similarity to small blocks explains the relative difficulty in finding orthologs and the widespread assumption that yeast and human kinetochores are very different. Henikoff and colleagues [67] have studied the evolutionary divergence of CenH3 and CENP-C<sup>Mif2</sup> in some detail and propose that kinetochore proteins are under positive selection in plants and animals as a consequence of meiotic drive by *CEN* DNA during female meiosis. Rapid evolution in protein sequence is most apparent in worms and flies, and in this study we have added only dmNdc80, dmNuf2 and dmMis12 to the list of likely structural *Drosophila* kinetochore proteins. Why the rate of kinetochore protein evolution is so much greater in flies and worms as compared to mammals, plants and fungi remains a mystery but it is reminiscent of data on other key regulators of chromosome segregation. Securin and its protease separase are also highly diverged in *D. melanogaster*: *Drosophila* securin, unlike the human and yeast proteins, consists of two separate gene products, called *three rows* and *pimples*, that interact with an unusually short separase [70]. Moreover, unlike the majority of eukaryotes that utilize an

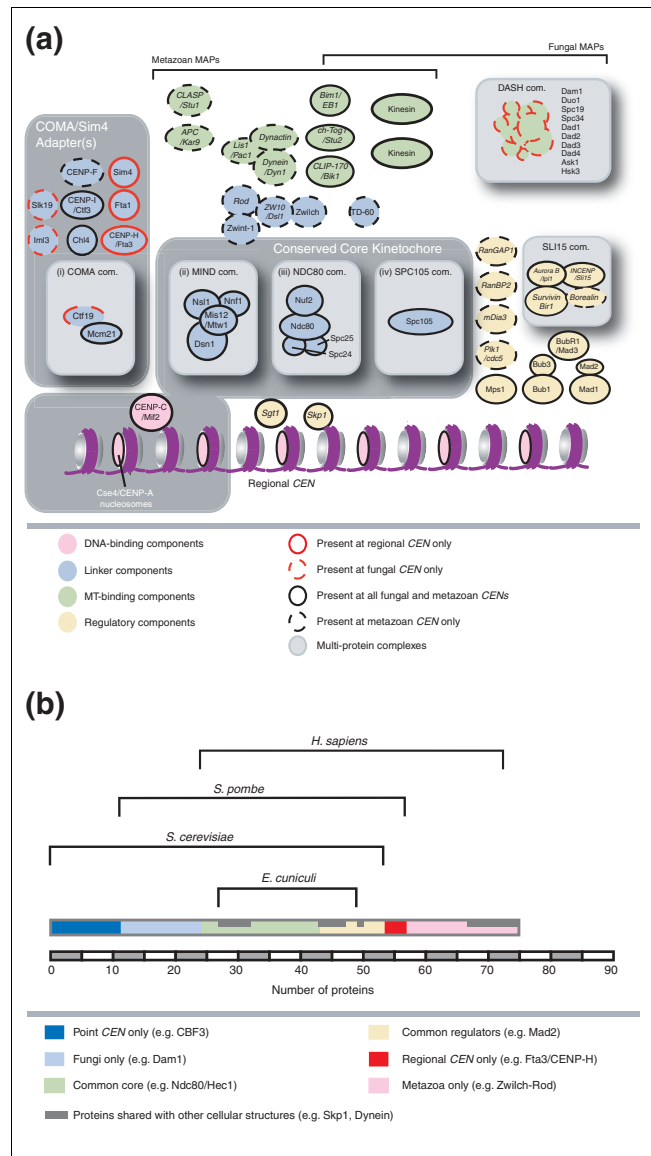


RNA-templated reverse transcriptase to replicate telomeres, *D. melanogaster* uses an alternative mechanism based on transposition of the HeT-A and TART retrotransposable elements [71]. It seems very likely that several distinct classes of kinetochore arose early in evolution. Perhaps surprisingly, fungal kinetochores appear to be as good a model for their human counterparts as kinetochores in organisms such as worms and flies.

For the majority of kinetochore proteins we have little knowledge of their biochemical functions or their structure. It is tempting to speculate that conserved sequence blocks represent protein-protein interaction domains or interaction surfaces under tight evolutionary pressure. However, with very few exceptions (for example, the kinase domains of checkpoint and regulatory proteins and motor domains of kinesins), blocks of conserved sequence do not correspond to recognizable functional domains. This stands in contrast to the situation in nuclear pore complexes, in which highly conserved and recognizable domains correspond to key functional units [72]. The most abundant structural elements in kinetochore proteins are coiled coils, which are known to function in protein-protein association [73] and act as springs and levers [74]. Coiled coils in the budding yeast spindle pole body protein scSpc42 also create a crystalline core involved in spindle pole body duplication [75]. Biochemical and electron microscopy experiments have shown that the heptad repeat domains in all four subunits of the *S. cerevisiae* NDC80 complex associate to form an extended stalk that is linked to two globular heads [32,33]. Whether the stalk is simply a spacer or some sort of mechanical element remains unexplored. Only detailed structural and biochemical experiments, backed up by analysis *in vivo*, will reveal the logic of sequence conservation among kinetochore subunits.

**A conserved molecular core of the kinetochore**

A key conclusion in this paper is that four multi-protein linkers that form the core of the *S. cerevisiae* kinetochore, MIND, the SPC105 complex, the NDC80 complex and COMA are also likely to be present in a wide variety of species (Figure 11). Along with CenH3 and CENP-C, SPC105, MIND and NDC80 complexes are ubiquitous. In budding yeast, linker complexes are thought to form a bridge between proteins in direct contact with DNA and those that bind MTs [5,14,15], and it will be important to show that this is also true in other organisms. Prior to the current work, biochemical experiments had led to the identification of SPC105, MIND and NDC80 complexes in *S. pombe*, *C. elegans* and human cells [18,19] but our systematic sequence analysis extends these observations to a greater variety of organisms, including *E. cucullii*, a microsporidium with a remarkably small proteome. The presence of the structural kinetochore proteins listed above appears to be more fundamental for chromosome segregation than a Mad2-dependent spindle assembly checkpoint, which does not seem to exist in *E. cucullii*. Thus, ascertaining the precise molecular functions of the MIND complex, NDC80 complex,



**Figure 11**  
 Evolutionary development of kinetochores from yeast to mammals. **(a)** Model of the kinetochore using protein subunit positions derived from the organization of the *S. cerevisiae* kinetochore. Proteins present in all fungal and mammalian CENs are outlined in black while proteins present only in fungi and mammals with regional CENs are outlined in red. Red dotted outlines indicate proteins that are only present in fungi. Black dotted outlines indicate that either this protein only exists in metazoans or that only the metazoan ortholog is present at kinetochore. Proteins colored in pink, blue, green or yellow represent components of the DNA binding, linker, regulatory or microtubule-binding layers, respectively, based on kinetochore organization in *S. cerevisiae*. Potential multi-protein complexes are highlighted with a light gray box and the conserved kinetochore core, or COMA/Sim4 adaptor with a dark gray box. Protein names are given for *H. sapiens* first and then *S. cerevisiae* when different. Italic lettering indicates that the protein has additional functions in the cell. The kinesins present at kinetochores in *H. sapiens* are CENP-E (Kinesin-7) and MCAK (Kinesin-13), and in *S. cerevisiae* Kip3 (Kinesin-8), Cin8 (Kinesin-5), Kip1 (Kinesin-5) and Kar3 (Kinesin-14) (for nomenclature see [38]). **(b)** Quantification of the number of kinetochore proteins, and their respective evolutionary class, in *S. cerevisiae*, *S. pombe*, *E. cucullii* and *H. sapiens*.

SpC105, CenH3 and CENP-C and of the macromolecular assemblies in which they participate is a key task in the study of kinetochore biology.

### Diverged kinetochore components

Budding yeast with CDEI-II-III point *CENs* contain a set of 11 proteins that are not present in fungi such as *S. pombe* or *C. albicans* (Figure 11). Three of the eleven point-*CEN* specific proteins are involved in sequence-specific binding to CDEIII while six are part of the COMA complex or of a COMA-dependent assembly pathway. Only three of the eleven components of the COMA pathway in *S. cerevisiae* (Mcm21<sup>Mal2</sup>, Chl4<sup>Mis15</sup> and Ctf3<sup>Mis6/CENP-1</sup>) are conserved among fungi and mammalian kinetochores (Figure 11). In *S. pombe*, an alternative set of eight proteins, including spSim4 and spFta1-7, are bound to the COMA components spMcm21<sup>Mal2</sup>, spChl4<sup>Mis15</sup> and spMis6<sup>Ctf3</sup>. At least three of these proteins (CENP-H<sup>Fta3</sup>, Fta1 and Sim4) are members of a class of proteins found in fungal and metazoan organisms with regional *CENs* whereas the other four proteins have no obvious orthologs (Figure 11a). Overall, these data point to COMA and COMA-associated proteins as kinetochore components with a particularly high degree of sequence divergence through evolution. It seems reasonable to speculate that COMA helps to accommodate kinetochore subunits that are highly conserved among regional and point *CENs*, such as the NDC80 complex, to diverged components, such as CBF3. By analogy, it seems likely that specialized proteins have evolved to meet the special structural demands of holocentric *CENs*; ceKNL-3, a kinetochore protein bound to the *C. elegans* MIND and NDC80 complexes [18] but absent from other kinetochores, may be an early example of a holocentric adaptor.

### The logic of kinetochore assembly

The MT binding components of kinetochores are unlike kinetochore structural components in that almost all are involved in multiple MT-based processes (Figure 11). In humans for example, EB1<sup>Bim1</sup> and APC<sup>Kar9</sup> are found not only at kinetochores, but also at sites of MT association with the cell cortex; CLIP-170<sup>Bik1</sup> and Dynein play important roles in vesicle trafficking and ch-Tog1<sup>Stu2</sup> is required for spindle assembly. From yeast to humans, only one or two of the six to ten kinetochore MAPs and motors are specific to kinetochores. CENP-A functions in most organisms to determine *CEN* location without recognizing *CEN*-specific sequences; similarly, the NDC80-MIND-SPC105-COMA complexes must determine the specialized biochemistry of MT-kinetochore linkages without resort to many kinetochore-specific MAPs.

### Conclusion

We conclude that critical structural features of kinetochores are conserved from yeast to man, despite highly divergent *CEN* sequences. It appears that both short *S. cerevisiae* point centromeres and complex metazoan regional centromeres arose from a common ancestor that probably had regional

centromeres. Both simple and complex kinetochores contain conserved SPC105, MIND and NDC80 complexes along with more variable COMA complexes. This core assembly is supplemented by adaptor proteins specific to organisms with point, regional or holocentric *CENs*. The key to understanding kinetochore biology is now to determine how specialized adaptors and conserved core complexes interact with inner centromere components such as CenH3 and CENP-C to assemble structures capable of binding to and regulating microtubules through the recruitment of MAPs and motors.

## Materials and methods

### Sequence-similarity searches

Database searches were performed on NCBI non-redundant and EST databases using PSI-BLAST and BLAST (protein-protein BLAST (blastp) and genomic BLAST (tblastn)) [76]. Pattern searches were performed using ScanProsite [77]. Multiple sequence alignments were built with ClustalW, MUSCLE and T-Coffee and edited by hand [78-80]. Coiled coil predictions were based on the COILS program using a window size of 28 [81]. Human Nnf1R and Fta1R were identified in PSI-BLAST searches using the full-length *S. pombe* Fta1 or *S. cerevisiae* Nnf1 protein sequences as the queries. To identify human Mcm21R, the *S. cerevisiae* Mcm21 protein sequence was first used as a query in PSI-BLAST searches that yielded fungal Mcm21 related proteins. These proteins were assembled in a multiple sequence alignment from which the motif [HYF]-[KRHDENQ]-[VLI]-x-[HYF]-[ST]-[IVL]-[P]-x-x-[IL]-x-[ILV] was derived, and then used in a pattern search to identify metazoan orthologs. To identify orthologs of *S. cerevisiae* Nsl1 and Chl4, *S. pombe* Sim4 or *H. sapiens* CENP-H conserved blocks were first identified. PSI-BLAST searches were carried out using *S. pombe* Sim4, *S. cerevisiae* Nsl1, *S. cerevisiae* Chl4 or *H. sapiens* CENP-H as query sequences. This approach identified a set of fungal Sim4, Nsl1 or Chl4 related proteins and a set of metazoan CENP-H related proteins. Each set of proteins was then assembled into multiple sequence alignments and conserved blocks identified (amino acids 1 to 143 for *U. maydis* Nsl1, 1 to 114 for *S. cerevisiae* Chl4, 341 to 373 for *S. pombe* Sim4 and 224 to 269 for *H. sapiens* CENP-H). The sequences present in these conserved blocks were then used in PSI-BLAST searches to identify new fungal (for CENP-H) or metazoan (for Sim4, Chl4 or Nsl1) proteins.

### Phylogenetic analysis

Phylogenetic alignments were generated with MUSCLE using GBLOCKS to identify conserved blocks [82]. Conserved blocks were selected only if single positions were conserved in at least 50% of the sequences, with higher stringency at flanking positions (80%). A maximum of eight contiguous non-conserved positions were allowed. The minimum block length was five amino acids. Positions with gaps were allowed only if their number did not exceed 50%. Conserved blocks and the number of positions used for each protein family are

described in Additional data file 5. To calculate the distances between sequences we took a maximum likelihood approach using TREE-PUZZLE [83] with the 'Pairwise distance calculation only' option, the Jones-Taylor-Thornton substitution matrix [84] and gamma-distributed rates (eight categories) to account for rate heterogeneity (parameters were estimated from the dataset). A neighbor-joining tree was constructed from the distance matrix with the NEIGHBOR program from the PHYLIP package [85]. Reliability of the dataset was assessed by bootstrap. We generated 100 permutation datasets using the SEQBOOT program from the PHYLIP package. From these 100 datasets we calculated distance matrices and constructed neighbor-joining trees using the parameters described above. TREE-PUZZLE was then used with the 'Consensus of user defined trees' option to generate a consensus tree from all neighbor-joining trees (nodes with support less than 50% were collapsed) [86]. Trees were visualized using the SPLITS TREE tool [87]. Amino acid similarity percentages used in multiple sequence alignments are given in Additional data file 4.

#### Hidden Markov model based-modeling

The point *CEN* model was constructed from three different sub-models based on the known structure of point *CENs*. The first sub-model searched for CDEI-like regions in the query sequence using the [T|G]CA[C|G|T][A|C|G]TG motif. The second sub-model then searched for adjacent CDEII-like AT rich regions. The CDEII region was modeled with a HMM using CDEII from *S. cerevisiae* [88]. For the negative model, *S. cerevisiae* genomic DNA was used (the effect of including *CENs* in the genomic DNA was disregarded). For both datasets, base transition frequencies were determined and the transition matrix for the HMM was calculated. The quality of the HMM was evaluated by screening annotated budding yeast genomes and assessment with a bit score:

$$S(x) = \log \frac{P(x | model+)}{P(x | model-)}$$

Given the identification of CDEI and CDEII sequence elements, a third sub-model searched for an adjacent CDEIII motif using an expression based on the highly conserved CCGGAA motif. Positive hits were evaluated with the bit score calculated from the CDEII HMM, length distribution, AT length, AT runs and synteny.

#### Additional data files

The following additional data are available with the online version of this paper. Additional data file 1 contains accession numbers of all proteins that are used in this study. Additional data file 2 shows the multiple sequence alignment of *S. cerevisiae* Spc34, a subunit of the multi-protein DASH complex, with a set of fungal orthologs and a set of related metazoan proteins (NYD-Sp28 family). Additional data file 3 contains multiple sequence alignments of the *E. cuniculi* kinetochore

proteins Ndc80, Nuf2R, Mis12/Mtw1, Nnf1, Spc105 and CENP-C amongst five fungi. Additional data files 4 and 5 list amino acid similarities used in all multiple sequence alignments and homology blocks used in phylogenetic analysis, respectively.

#### Acknowledgements

We thank Daniel Huson (University of Tübingen) for help with the computational centromere screen and members of the Sorger lab for helpful discussions. ADM was supported by a fellowship from the Jane Coffin Childs Fund for Medical Research and PM by an EMBO long-term fellowship. This work was supported by NIH grants CA84179 and GM51464.

#### References

1. Koshland DE, Mitchison TJ, Kirschner MW: **Polewards chromosome movement driven by microtubule depolymerization in vitro.** *Nature* 1988, **331**:499-504.
2. Cleveland DW, Mao Y, Sullivan KF: **Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling.** *Cell* 2003, **112**:407-421.
3. Choo KH: **Domain organization at the centromere and neocentromere.** *Dev Cell* 2001, **1**:165-177.
4. Fitzgerald-Hayes M, Clarke L, Carbon J: **Nucleotide sequence comparisons and functional analysis of yeast centromere DNAs.** *Cell* 1982, **29**:235-244.
5. McAinsh AD, Tytell JD, Sorger PK: **Structure, function, and regulation of budding yeast kinetochores.** *Annu Rev Cell Dev Biol* 2003, **19**:519-539.
6. Sanyal K, Baum M, Carbon J: **Centromeric DNA sequences in the pathogenic yeast *Candida albicans* are all different and unique.** *Proc Natl Acad Sci USA* 2004, **101**:11374-11379.
7. Houben A, Schubert I: **DNA and proteins of plant centromeres.** *Curr Opin Plant Biol* 2003, **6**:554-560.
8. Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF: **Genomic and genetic definition of a functional human centromere.** *Science* 2001, **294**:109-115.
9. Sun X, Wahlstrom J, Karpen G: **Molecular structure of a functional *Drosophila* centromere.** *Cell* 1997, **91**:1007-1019.
10. Clarke L, Amstutz H, Fishel B, Carbon J: **Analysis of centromeric DNA in the fission yeast *Schizosaccharomyces pombe*.** *Proc Natl Acad Sci USA* 1986, **83**:8253-8257.
11. Albertson DG, Thomson JN: **The kinetochores of *Caenorhabditis elegans*.** *Chromosoma* 1982, **86**:409-428.
12. Wiens GR, Sorger PK: **Centromeric chromatin and epigenetic effects in kinetochore assembly.** *Cell* 1998, **93**:313-316.
13. Mellone BG, Allshire RC: **Stretching it: putting the CEN(P-A) in centromere.** *Curr Opin Genet Dev* 2003, **13**:191-198.
14. De Wulf P, McAinsh AD, Sorger PK: **Hierarchical assembly of the budding yeast kinetochore from multiple subcomplexes.** *Genes Dev* 2003, **17**:2902-2921.
15. Nekrasov VS, Smith MA, Peak-Chew S, Kilmartin JV: **Interactions between centromere complexes in *Saccharomyces cerevisiae*.** *Mol Biol Cell* 2003, **14**:4931-4946.
16. Scharfenberger M, Ortiz J, Grau N, Janke C, Schiebel E, Lechner J: **Nsl1p is essential for the establishment of bipolarity and the localization of the Dam-Duo complex.** *EMBO J* 2003, **22**:6584-6597.
17. Bharadwaj R, Qi W, Yu H: **Identification of two novel components of the human NDC80 kinetochore complex.** *J Biol Chem* 2004, **279**:13076-13085.
18. Cheeseman IM, Niessen S, Anderson S, Hyndman F, Yates JR 3rd, Oegema K, Desai A: **A conserved protein network controls assembly of the outer kinetochore and its ability to sustain tension.** *Genes Dev* 2004, **18**:2255-2268.
19. Obuse C, Iwasaki O, Kiyomitsu T, Goshima G, Toyoda Y, Yanagida M: **A conserved Mis12 centromere complex is linked to heterochromatic HPI and outer kinetochore protein Zwint-1.** *Nat Cell Biol* 2004, **6**:1135-1141.
20. Goshima G, Kiyomitsu T, Yoda K, Yanagida M: **Human centromere chromatin protein hMis12, essential for equal segregation, is independent of CENP-A loading pathway.** *J*

- Cell Biol* 2003, **160**:25-39.
21. Liu ST, Hittler JC, Jablonski SA, Campbell MS, Yoda K, Yen TJ: **Human CENP-I specifies localization of CENP-F, MAD1 and MAD2 to kinetochores and is essential for mitosis.** *Nat Cell Biol* 2003, **5**:341-345.
  22. McClelland ML, Gardner RD, Kallio MJ, Daum JR, Gorbosky GJ, Burke DJ, Stukenberg PT: **The highly conserved Ndc80 complex is required for kinetochore assembly, chromosome congression, and spindle checkpoint activity.** *Genes Dev* 2003, **17**:101-114.
  23. Tirnauer JS, Canman JC, Salmon ED, Mitchison TJ: **EB1 targets to kinetochores with attached, polymerizing microtubules.** *Mol Biol Cell* 2002, **13**:4308-4316.
  24. Kitagawa K, Hieter P: **Evolutionary conservation between budding yeast and human kinetochores.** *Nat Rev Mol Cell Biol* 2001, **2**:678-687.
  25. Wigge PA, Kilmartin JV: **The Ndc80p complex from *Saccharomyces cerevisiae* contains conserved centromere components and has a function in chromosome segregation.** *J Cell Biol* 2001, **152**:349-360.
  26. Dujardin D, Wacker UL, Moreau A, Schroer TA, Rickard JE, De Mey JR: **Evidence for a role of CLIP-170 in the establishment of metaphase chromosome alignment.** *J Cell Biol* 1998, **141**:849-862.
  27. Henikoff S, Dalal Y: **Centromeric chromatin: what makes it unique?** *Curr Opin Genet Dev* 2005, **15**:177-184.
  28. Dujon B, Sherman D, Fischer G, Durrrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neugevele C, Talla E, et al.: **Genome evolution in yeasts.** *Nature* 2004, **430**:35-44.
  29. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature* 2003, **423**:241-254.
  30. Kuras L, Thomas D: **Identification of the yeast methionine biosynthetic genes that require the centromere binding factor I for their transcriptional activation.** *FEBS Lett* 1995, **367**:15-18.
  31. Cardozo T, Pagano M: **The SCF ubiquitin ligase: insights into a molecular machine.** *Nat Rev Mol Cell Biol* 2004, **5**:739-751.
  32. Ciferri C, De Luca J, Monzani S, Ferrari KJ, Ristic D, Wyman C, Stark H, Kilmartin J, Salmon ED, Musacchio A: **Architecture of the human ndc80-hecl1 complex, a critical constituent of the outer kinetochore.** *J Biol Chem* 2005, **280**:29088-29095.
  33. Wei RR, Sorger PK, Harrison SC: **Molecular organization of the Ndc80 complex, an essential kinetochore component.** *Proc Natl Acad Sci USA* 2005, **102**:5363-5367.
  34. Kline-Smith SL, Sandall S, Desai A: **Kinetochore-spindle microtubule interactions during mitosis.** *Curr Opin Cell Biol* 2005, **17**:35-46.
  35. Malik HS, Henikoff S: **Conflict begets complexity: the evolution of centromeres.** *Curr Opin Genet Dev* 2002, **12**:711-718.
  36. Biswas K, Rieger KJ, Morschhauser J: **Functional characterization of CaCBFI, the *Candida albicans* homolog of centromere binding factor 1.** *Gene* 2003, **323**:43-55.
  37. Browning H, Hackney DD, Nurse P: **Targeted movement of cell end factors in fission yeast.** *Nat Cell Biol* 2003, **5**:812-818.
  38. Lawrence CJ, Dawe RK, Christie KR, Cleveland DW, Dawson SC, Endow SA, Goldstein LS, Goodson HV, Hirokawa N, Howard J, et al.: **A standardized kinesin nomenclature.** *J Cell Biol* 2004, **167**:19-22.
  39. Liu X, McLeod I, Anderson S, Yates JR 3rd, He X: **Molecular analysis of kinetochore architecture in fission yeast.** *EMBO J* 2005, **24**:2919-2930.
  40. Pidoux AL, Richardson W, Allshire RC: **Sim4: a novel fission yeast kinetochore protein required for centromeric silencing and chromosome segregation.** *J Cell Biol* 2003, **161**:295-307.
  41. Wang Y, Devereux W, Stewart TM, Casero RA Jr: **Cloning and characterization of human polyamine-modulated factor-1, a transcriptional cofactor that regulates the transcription of the spermidine/spermine N(1)-acetyltransferase gene.** *J Biol Chem* 1999, **274**:22095-22101.
  42. Yamashita A, Ito M, Takamatsu N, Shiba T: **Characterization of Solt, a novel SoxLZ/Sox6 binding protein expressed in adult mouse testis.** *FEBS Lett* 2000, **481**:147-151.
  43. Miranda JJ, De Wulf P, Sorger PK, Harrison SC: **The yeast DASH complex forms closed rings on microtubules.** *Nat Struct Mol Biol* 2005, **12**:138-143.
  44. Westermann S, Avila-Sakar A, Wang HW, Niederstrasser H, Wong J, Drubin DG, Nogales E, Barnes G: **Formation of a dynamic kinetochore-microtubule interface through assembly of the Dam1 ring complex.** *Mol Cell* 2005, **17**:277-290.
  45. Jishage M, Fujino T, Yamazaki Y, Kuroda H, Nakamura T: **Identification of target genes for EWS/ATF-1 chimeric transcription factor.** *Oncogene* 2003, **22**:41-49.
  46. Pazour GJ, Agrin N, Leszyk J, Witman GB: **Proteomic analysis of a eukaryotic cilium.** *J Cell Biol* 2005, **170**:103-113.
  47. Meluh PB, Koshland D: **Evidence that the MIF2 gene of *Saccharomyces cerevisiae* encodes a centromere protein with homology to the mammalian centromere protein CENP-C.** *Mol Biol Cell* 1995, **6**:793-807.
  48. Stoler S, Keith KC, Curnick KE, Fitzgerald-Hayes M: **A mutation in CSE4, an essential gene encoding a novel chromatin-associated protein in yeast, causes chromosome nondisjunction and cell cycle arrest at mitosis.** *Genes Dev* 1995, **9**:573-586.
  49. Williams BC, Li Z, Liu S, Williams EV, Leung G, Yen TJ, Goldberg ML: **Zwilch, a new component of the ZW10/ROD complex required for kinetochore functions.** *Mol Biol Cell* 2003, **14**:1379-1391.
  50. Sugata N, Munekata E, Todokoro K: **Characterization of a novel kinetochore protein, CENP-H.** *J Biol Chem* 1999, **274**:27343-27346.
  51. Starr DA, Williams BC, Li Z, Etemad-Moghadam B, Dawe RK, Goldberg ML: **Conservation of the centromere/kinetochore protein ZW10.** *J Cell Biol* 1997, **138**:1289-1301.
  52. Rattner JB, Rao A, Fritzler MJ, Valencia DW, Yen TJ: **CENP-F is a ca 400 kDa kinetochore protein that exhibits a cell-cycle dependent localization.** *Cell Motil Cytoskeleton* 1993, **26**:214-226.
  53. Yen TJ, Compton DA, Wise D, Zinkowski RP, Brinkley BR, Earnshaw WC, Cleveland DW: **CENP-E, a novel human centromere-associated protein required for progression from metaphase to anaphase.** *EMBO J* 1991, **10**:1245-1254.
  54. Westermann S, Cheeseman IM, Anderson S, Yates JR 3rd, Drubin DG, Barnes G: **Architecture of the budding yeast kinetochore reveals a conserved molecular core.** *J Cell Biol* 2003, **163**:215-222.
  55. Andag U, Schmitt HD: **Dsl1p, an essential component of the Golgi-endoplasmic reticulum retrieval system in yeast, uses the same sequence motif to interact with different subunits of the COPI vesicle coat.** *J Biol Chem* 2003, **278**:51722-51734.
  56. Hirose H, Arasaki K, Dohmae N, Takio K, Hatsuzawa K, Nagahama M, Tani K, Yamamoto A, Tohyama M, Tagaya M: **Implication of ZW10 in membrane trafficking between the endoplasmic reticulum and Golgi.** *EMBO J* 2004, **23**:1267-1278.
  57. Starr DA, Williams BC, Hays TS, Goldberg ML: **ZW10 helps recruit dynactin and dynein to the kinetochore.** *J Cell Biol* 1998, **142**:763-774.
  58. Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF: **A kingdom-level phylogeny of eukaryotes based on combined protein data.** *Science* 2000, **290**:972-977.
  59. Gribaldo S, Philippe H: **Ancient phylogenetic relationships.** *Theor Popul Biol* 2002, **61**:391-408.
  60. Heeger S, Leismann O, Schittenhelm R, Schraidt O, Heidmann S, Lehner CF: **Genetic interactions of separase regulatory subunits reveal the diverged *Drosophila* Cenp-C homolog.** *Genes Dev* 2005, **19**:2041-2053.
  61. Blower MD, Karpen GH: **The role of *Drosophila* CID in kinetochore formation, cell-cycle progression and heterochromatin interactions.** *Nat Cell Biol* 2001, **3**:730-739.
  62. Karess R: **Rod-Zw10-Zwilch: a key player in the spindle checkpoint.** *Trends Cell Biol* 2005, **15**:386-392.
  63. Vivares CP, Gouy M, Thomarat F, Metenier G: **Functional and evolutionary analysis of a eukaryotic parasitic genome.** *Curr Opin Microbiol* 2002, **5**:499-505.
  64. Peters JM: **The anaphase-promoting complex: proteolysis in mitosis and beyond.** *Mol Cell* 2002, **9**:931-943.
  65. Luo X, Tang Z, Rizo J, Yu H: **The Mad2 spindle checkpoint protein undergoes similar major conformational changes upon binding to either Mad1 or Cdc20.** *Mol Cell* 2002, **9**:59-71.
  66. Sironi L, Mapelli M, Knapp S, De Antoni A, Jeang KT, Musacchio A: **Crystal structure of the tetrameric Mad1-Mad2 core complex: implications of a 'safety belt' binding mechanism for the spindle checkpoint.** *EMBO J* 2002, **21**:2496-2506.
  67. Talbert PB, Bryson TD, Henikoff S: **Adaptive evolution of centromere proteins in plants and animals.** *J Biol* 2004, **3**:18.
  68. Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, Prensier G, Barbe V, Peyretailade E, Brottier P, Wincker P, et al.: **Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*.** *Nature* 2001, **414**:450-453.

69. Dalgaard JZ, Vengrova S: **Selective gene expression in multigene families from yeast to mammals.** *Sci STKE* 2004, **2004**:re17.
70. Jager H, Herzig A, Lehner CF, Heidmann S: **Drosophila separase is required for sister chromatid separation and binds to PIM and THR.** *Genes Dev* 2001, **15**:2572-2584.
71. Louis EJ: **Are Drosophila telomeres an exception or the rule?** *Genome Biol* 2002, **3**:REVIEWS0007.
72. Baptiste E, Charlebois RL, MacLeod D, Brochier C: **The two tempos of nuclear pore complex evolution: highly adapting proteins in an ancient frozen structure.** *Genome Biol* 2005, **6**:R85.
73. Newman JR, Wolf E, Kim PS: **A computationally directed screen identifying interacting coiled coils from Saccharomyces cerevisiae.** *Proc Natl Acad Sci USA* 2000, **97**:13203-13208.
74. Rose A, Meier I: **Scaffolds, levers, rods and springs: diverse cellular functions of long coiled-coil proteins.** *Cell Mol Life Sci* 2004, **61**:1996-2009.
75. Bullitt E, Rout MP, Kilmartin JV, Akey CW: **The yeast spindle pole body is assembled around a central crystal of Spc42p.** *Cell* 1997, **89**:1077-1086.
76. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
77. Gattiker A, Gasteiger E, Bairoch A: **ScanProsite: a reference implementation of a PROSITE scanning tool.** *Appl Bioinformatics* 2002, **1**:107-108.
78. Higgins DG: **CLUSTAL V: multiple alignment of DNA and protein sequences.** *Methods Mol Biol* 1994, **25**:307-318.
79. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302**:205-217.
80. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**:1792-1797.
81. Lupas A, Van Dyke M, Stock J: **Predicting coiled coils from protein sequences.** *Science* 1991, **252**:1162-1164.
82. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.** *Mol Biol Evol* 2000, **17**:540-552.
83. Schmidt HA, Strimmer K, Vingron M, von Haeseler A: **TREE-PUZ-ZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics* 2002, **18**:502-504.
84. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences.** *Comput Appl Biosci* 1992, **8**:275-282.
85. Felsenstein J: **PHYLIP - Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5**:164-166.
86. Felsenstein KM, Lewis-Higgins L: **Processing of the beta-amyloid precursor protein carrying the familial, Dutch-type, and a novel recombinant C-terminal mutation.** *Neurosci Lett* 1993, **152**:185-189.
87. Huson DH, Bryant D: **Application of phylogenetic networks in evolutionary studies.** *Mol Biol Evol* 2006, **23**:254-267.
88. Durbin R, Eddy SR, Krogh A, Mitchison G: *Probabilistic Models of Proteins and Nucleic Acids* Cambridge: Cambridge University Press; 1998.