Stepwise evolution of the centriole-assembly pathway

Zita Carvalho-Santos, Pedro Machado, Pedro Branco, Filipe Tavares-Cadete, Ana Rodrigues-Martins, José B. Pereira-Leal* and Mónica Bettencourt-Dias*

Instituto Gulbenkian de Ciência, Rua da Quinta Grande, 6, P-2780-156 Oeiras, Portugal

*Authors for correspondence ([jleal@igc.gulbenkian.pt; mdias@igc.gulbenkian.pt])

Accepted 1 March 2010
Journal of Cell Science 123, 1414-1426
© 2010. Published by The Company of Biologists Ltd
doi:10.1242/jcs.064931

Summary
The centriole and basal body (CBB) structure nucleates cilia and flagella, and is an essential component of the centrosome, underlying eukaryotic microtubule-based motility, cell division and polarity. In recent years, components of the CBB-assembly machinery have been identified, but little is known about their regulation and evolution. Given the diversity of cellular contexts encountered in eukaryotes, but the remarkable conservation of CBB morphology, we asked whether general mechanistic principles could explain CBB assembly. We analysed the distribution of each component of the human CBB-assembly machinery across eukaryotes as a strategy to generate testable hypotheses. We found an evolutionarily cohesive and ancestral module, which we term UNIMOD and is defined by three core ancestral components (SAS6, SAS4/CPAP and BLD10/CEP135), that correlates with the occurrence of CBBs. Unexpectedly, other players (SAK/PLK4, SPD2/CEP192 and CP110) emerged in a taxon-specific manner. We report that gene duplication plays an important role in the evolution of CBB components and show that, in the case of BLD10/CEP135, this is a source of tissue specificity in CBB and flagella biogenesis. Moreover, we observe extreme protein divergence amongst CBB components and show experimentally that there is loss of cross-species complementation among SAK/PLK4 family members, suggesting species-specific adaptations in CBB assembly. We propose that the UNIMOD theory explains the conservation of CBB architecture and that taxon- and tissue-specific molecular innovations, gained through emergence, duplication and divergence, play important roles in coordinating CBB biogenesis and function in different cellular contexts.

Key words: Centriole, Basal body, Evolution, Drosophila, Flagella, Comparative genomics

Introduction
The structure of the centriole and the basal body (CBB) is remarkably conserved, comprising microtubule triplets arranged in a ninefold symmetrical configuration (Fig. 1). CBBs are found in all crown eukaryotic groups (Fig. 2A,B; supplementary material Table S1), as a centriole, within the context of a centrosome, and/or as a basal body, tethered to the membrane. This suggests that they were present in the last eukaryotic common ancestor (LECA) (Azimzadeh and Bornens, 2004; Cavalier-Smith, 2002) and that secondary loss occurred in specific branches, such as yeasts and higher plants (Fig. 2A,B; supplementary material Table S1). The conservation of CBB architecture and its structural assembly intermediates (Fig. 1) suggests the existence of common molecular assembly machinery, already present in the LECA. On the other hand, CBBs are assembled in a multiplicity of contexts, such as different cell-cycle phases or cellular locations, suggesting the need for tailored assembly pathways. Moreover, CBBs can have a wide range of functions (Beisson and Wright, 2003; Bettencourt-Dias and Glover, 2007; Delattre and Gonczy, 2004): in humans, they assemble centrosomes, and motile and sensory cilia; in Caenorhabditis elegans, they never form motile cilia; and in green algae, such as Chlamydomonas, they only form motile cilia. The conservation of the structure contrasts with the diversity of assembly contexts and functions, raising an interesting paradox.

To investigate CBB assembly in eukaryotes, we focused on the evolution of the molecular mechanisms that control this process. We used comparative genomics, a strategy that brought major insights into the origin and evolution of the assembly of cellular structures such as the nuclear pore complex (Devos et al., 2004; Mans et al., 2004), the peroxisome (Gabaldon et al., 2006) and cilia (Avidor-Reiss et al., 2004; Li et al., 2004; Wickstead and Gull, 2007). We focused on six proteins shown to be required for CBB biogenesis in humans (Fig. 1): SPD2/CEP192, SAK/PLK4, SAS6, SAS4/CPAP, BLD10/CEP135 and CP110 (Cunha-Ferreira et al., 2009a; Kleylein-Sohn et al., 2007). Orthologs of some of these proteins have been functionally described in other species (Fig. 1).

Results
A molecular toolkit to detect the CBB-assembly machinery
CBB proteins have eluded automatic comparative genomics screens for novel ciliary components (Avidor-Reiss et al., 2004; Baron et al., 2007; Li et al., 2004). They generally contain several coiled-coil domains (Fig. 3; supplementary material Fig. S1), which carry little phylogenetic signal (Rose et al., 2005). Our detailed bioinformatics analysis of each protein family revealed new conserved regions, other than coiled-coil regions (Fig. 3; supplementary material Figs S1-S6), that characterize each protein with previously untapped phylogenetic depth and breath. Our detailed approach also included the characterization of the phylogenetic distribution of known domains within specific taxonomical groups (e.g. the polo boxes of PLKs).

A core ancestral module defines the centriole ninefold symmetry
The universality of the CBB structure suggests the existence of an ancestral CBB-assembly mechanism. Recent studies have, in fact, suggested that several components of the flagella apparatus, such as the molecules needed to make the motile axoneme, are likely to be ancestral (Avidor-Reiss et al., 2004; Li et al., 2004; Wickstead and Gull, 2007).
To investigate the existence of such a universal CBB-assembly mechanism, we searched for homologs of known CBB-assembly proteins in a set of 26 representative eukaryotic species, covering the crown eukaryotic groups and representing the diversity of function and architecture (including absence) of CBBs (Fig. 2A,B; see supplementary material Tables S1 and S2). We calculated the correlation between the presence of each molecule and the presence of the CBB, using a normalized Hamming distance (Fig. 2). Given the poor annotation of the proteomes of certain species and the absence of structural information regarding the existence of a CBB in others, we arbitrarily defined that the presence of a molecule and the occurrence of the CBB structure were correlated if this occurred in at least 80% of the species (Fig. 2). To our surprise, given the conservation of the CBB structure, only a subset of CBB-assembly proteins obey the criteria above defined: SAS4/CPAP, SAS6, and BLD10/CEP135 (Fig. 2). This evolutionarily cohesive behavior suggests that these three molecules are part of the same functional ancestral module in CBB assembly, which, for simplicity, we will call UNIVERSAL MODule (UNIMOD). Amongst the six studied families, the UNIMOD components are, in fact, the only ones required to define the CBB architecture: SAS6 and BLD10/CEP135 form the cartwheel, a structure involved in the specification and stabilization of CBB ninefold symmetry (Fig. 1) (Hiraki et al., 2007; Matsuura et al., 2004; Nakazawa et al., 2007; Rodrigues-Martins et al., 2007a), whereas SAS4/CPAP is required for assembling or stabilizing elongating centriolar microtubules (Fig. 1) (Dammermann et al., 2008; Kohlmaier et al., 2009; Schmidt et al., 2009; Tang et al., 2009). Our results suggest that the conservation of the CBB structure in the eukaryotic tree of life is achieved by the preservation of an assembly mechanism based on a set of conserved structural components – the UNIMOD. Similar profiles have been assigned to axonemal proteins that are present in organisms such as green algae, humans and trypanosomes, but missing from the higher land plants (Avidor-Reiss et al., 2004; Li et al., 2004; Wickstead and Gull, 2007).

Predicting extra components of the ancestral assembly pathway: PLKs trigger CBB formation

SAK/PLK4 (a polo-like kinase) is indispensable for centriole biogenesis in human cells and Drosophila melanogaster (Bettencourt-Dias et al., 2005; Habelanck et al., 2005; Kleylein-Sohn et al., 2007; Rodrigues-Martins et al., 2007b). High levels of this protein lead to the appearance of supernumerary centrioles through either canonical (Bettencourt-Dias et al., 2005; Habelanck et al., 2005) or de novo (Peel et al., 2007; Rodrigues-Martins et al., 2007b) biogenesis. Because of its importance, we were surprised to observe that SAK/PLK4 is not part of the UNIMOD and is only found in opisthokonts (purple clades in Figs 2 and 4); we therefore investigated what could be triggering CBB biogenesis in other groups. Gene duplication is believed to play a major role in generating complexity of cellular mechanisms in evolution (Ohno, 1970). We tested whether other PLK family members could play a role in CBB biogenesis in other groups. We found that PLKs are present in all branches of the eukaryotic tree of life (Figs 2 and 4). The PLKs outside the opisthokonts contain a kinase domain that clusters with opisthokont Polo/PLK1 rather than SAK/PLK4 (Fig. 4), and possess two polo boxes, similar to Polo/PLK1 (supplementary material Fig. S5D). This suggests that a Polo/PLK1-like protein is the ancestral member of the family that duplicated,
giving rise to SAK/PLK4 prior to the divergence of fungi and animals (Figs 2 and 4). Our results support the scenario that an ancestral Polo/PLK1 triggered CBB biogenesis in the LECA. This is further supported by two observations: human PLK1 (Liu and Erikson, 2002; Tsou et al., 2009) and human PLK2 (Warnke et al., 2004) play a role in centriole duplication, suggesting the presence

---

**Fig. 2. Phylogenetic distribution of families of molecular players in CBB biogenesis.** (A) Simplified taxonomic tree representing crown eukaryotic groups in different colors (these groups contain a common ancestor and all its descendants) [modified from Baldauf (Baldauf, 2003) and Hedges (Hedges, 2002)]. Unikonts include eukaryotic cells that, for the most part, have a single emergent flagellum. Unikonts can be divided into opisthokonts (which propel themselves with a single posterior flagellum; animals, fungi and choanoflagellates) and amebozoa (Cavalier-Smith, 2002). Bikonts include eukaryotic organisms with two emergent flagella (Cavalier-Smith, 2002). Branch color code: purple – opisthokonts, blue – amebozoa, green – plants, yellow – alveolates, orange – heterokonts, brown – excavates and discicristates. Organisms with incomplete genomes are indicated in grey. Detailed taxonomical and genome-sequencing status information is found in supplementary material Table S2. (B) Distribution of the structures present in each organism: internal architecture of the centrioles (within the centrosome), basal bodies and axonemes of cilia and/or flagella present in each species (see supplementary material Table S1 for details). In some cases, because of lack of information, we used structural information from other species in the same group (see supplementary material Table S1). Note that the presence of CBBs correlates with the occurrence of a flagellated stage. The basal body was lost in species that do not assemble flagella, such as yeasts (*S. pombe*, *S. cerevisiae*). (C) Phylogenetic distribution of molecular players in CBB biogenesis. Protein families and their role in CBB biogenesis, as indicated in Fig. 1, are schematized at top. Black boxes represent the presence of homologous sequences that are bidirectional best hits to family members in humans; grey boxes represent homologous sequences that are bidirectional best hits to family members in organisms other than human; white boxes indicate no detectable ortholog.
of a residual function in this process; and in Trypanosoma brucei, the depletion of PLK1 leads to defects in basal-body duplication and cytokinesis (Hammarton et al., 2007).

What could be the consequences of this duplication event? In humans and D. melanogaster, Polo/PLK1 is known to have important roles in the cell cycle, such as entry and progression in mitosis and cytokinesis, and γ-tubulin recruitment to the centrosome (Archambault and Glover, 2009). This explains the presence of PLKs in species that do not assemble CBBs. On the other hand, since SAK/PLK4 emerged, it became strictly correlated with CBBs.

Fig. 3. Sequence features of the CBB-assembly proteins. For each protein family, we show the architecture of the human sequence, but it can vary slightly among different organisms (e.g. SAS4/CPAP in supplementary material Fig. S2). We further show a conservation plot of the multiple sequence alignment. Note that gaps of non-aligned residues spanning ≥25% of the total sequence were removed for clarity. All the regions identified in this study as being conserved among orthologs are referred to as conserved regions (CRs) and are represented in the schematic of each protein. HMMs of each region are available from www.evocell.org. (A) SAS6 contains a long coiled-coil region spanning residues 160-480. A previously identified domain named PISA (Present In SAS6) (Leidel et al., 2005) is the first highlighted conserved region (SAS6 CR1: residues 44-87). This domain was further used as an additional criterion to classify SAS6 orthologs. We identified a second conserved region, SAS6 CR2 (residues 123-157), adjacent to the PISA domain. It contains a recently identified phosphorylation residue (Kitagawa et al., 2009). (B) SAS4/CPAP orthologs show conservation in a single region corresponding to the G-box (residues 1159-1264). This C-terminal conserved region is characterized by several nonamer motifs with a glycine or a glutamine in the first position, and is part of a larger region called the TCP10 domain (T complex protein 10), which was initially identified in the TCP10 protein (Hung et al., 2000). (C) BLD10/CEP135 contains long coiled-coil domains that mask their sequence similarity. Two conserved regions were identified (BLD10/CEP135 CR1: residues 6-147 and BLD10/CEP135 CR2: residues 678-1021). In vertebrates, we found a protein, TSGA10, that shares a high degree of similarity (65%) with BLD10/CEP135. Both domains, CR1 and CR2, were used to identify homologous sequences. (D,E) PLKs have an N-terminal kinase domain and one or two C-terminal ‘polo boxes’. Human PLK1-3 and Drosophila Polo have two Polo boxes, whereas SAK/PLK4 has a single Polo box. SAK/PLK4 has distinct conserved regions: the previously named cryptic polo box (Swallow et al., 2005) (residues 696-836) and a second conserved region between the kinase domain and the cryptic polo box (CR1: residues 283-569) that includes a regulatory phosphodegron (Cunha-Ferreira et al., 2009b; Holland et al., 2010; Rogers et al., 2009). (F) SPD2/CEP192 has a conserved region previously named the SPD2 domain (SPD2/CEP192 CR2: residues 1169-1417). We identified two other conserved regions: CR1 (residues 841-1089) and CR3 (residues 1466-1918). (G) A conserved region of CP110 was identified as CP110 CR1 (residues 794-910), which localizes within a Ca2+-independent calmodulin-binding region.
Fig. 4. The kinase domain is sufficient to distinguish all the PLK subfamilies. Maximum likelihood phylogenetic tree of the kinase domain of members of the different PLK subfamilies (a NiMA kinase was used as root). Branch labels represent bootstrap support in percentage. Using this strategy, we were able to distinguish the lineage of SAK/PLK4 from the other members of the PLK family, albeit with low bootstrap support. Please note that PLKs found outside opisthokonts cluster with budding yeast PLK (Cdc5), which is shown to be a functional homolog of human PLK1 (Lee and Erikson, 1997). The fact that Drosophila (as well as other insects) contains a single PLK that clusters with human PLK1 supports Polo/PLK1 being the ancestral kinase that underwent several duplication rounds in animals, to give rise to PLK1, 2, 3 and 5.
as shown by its disappearance in the yeasts, in which the CBB was lost concomitant with spindle pole body (SPB) emergence (Fig. 2; supplementary material Table S1). The evidence presented above strongly suggests that an ancestral Polo/PLK1 had both mitotic and CBB biogenesis functions. Upon duplication followed by subfunctionalization, this ancestral Polo/PLK1 generated Polo/PLK1 and SAC/PLK4, allowing the uncoupling of more general cell-cycle functions from CBB biogenesis.

**SPD2 and CP110 emerged in a taxon-specific manner**

A surprising observation is that SPD2/CEP192 and CP110, two proteins crucial for centriole biogenesis and function in humans, emerged in a taxon-specific manner (Fig. 2). SPD2/CEP192 is present in *Dictyostelium discoideum* (Fig. 2), having been lost in *Entamoeba histolytica* and at the base of fungi. *D. discoideum* is a well-characterized amoeba that does not assemble CBBs. Instead, it has a microtubule-organizing center (MTOC) called the nucleus-associated body (NAB), where SPD2/CEP192 was recently shown to localize (Schulz et al., 2009). This suggests that the ancestral function of SPD2/CEP192 was pericentriolar material (PCM) recruitment to the MTOC, independent of the presence of CBBs. PCM proteins, such as SPD2, might have acquired a role in recruiting CBB-assembly proteins to the centrosome (Dammermann et al., 2004; Loncarek et al., 2008). In animals, SPD2/CEP192 is essential for CBB biogenesis in contexts in which less PCM is available. In agreement, *C. elegans* and *D. melanogaster* SPD2/CEP192 are essential for the recruitment of PCM to the PCM-naked sperm CBB and its duplication upon fertilization (Dix and Raif, 2007; Kemp et al., 2004; Pelletier et al., 2004). By contrast, *D. melanogaster* SPD2/CEP192 is dispensable for both PCM recruitment and CBB duplication in somatic cells (Dix and Raif, 2007; Giansanti et al., 2008).

CP110 only appeared in animals (Fig. 2). It localizes to a distal centriole compartment, and is needed for centriole reduplication in S-phase-arrested human cells and to define centriole length (Chen et al., 2002; Kleylein-Sohn et al., 2007; Kohlmaier et al., 2009; Schmidt et al., 2009). We hypothesize that CP110 was added to the centrosome-assembly pathway in animals as an innovation. We found that a binding partner of CP110, CEP97, has a very similar phylogenetic distribution to CP110 (supplementary material Fig. S7). These results both suggest that the two proteins might work in a complex in all animals and validate the use of phylogenetic distributions as a screening strategy to find potential binding partners. *Drosophila* CP110 and CEP97 localize to centrioles and are necessary for centriole duplication in S2 cells (supplementary material Fig. S8A,B,D,E) (Dobbelaraet al., 2008). CP110 in humans participates in other processes, such as preventing centrioles from nucleating cilia (Kleylein-Sohn et al., 2007; Spektor et al., 2007) and cytokinesis (Tsang et al., 2006). It has been proposed that centrioles might play an important role in signaling the event of cellular abscission in cytokinesis (Piel et al., 2001). It is possible that CP110 emerged in animals to allow further coordination of centriole duplication with ciliogenesis and/or cytokinesis.

**Extreme sequence divergence**

Our expectation was that, considering the extreme structural conservation of CBBs, we were facing a highly conserved set of components. To our surprise, in the process of defining conserved regions in CBB-assembly components (Fig. 3; supplementary material Figs S1 and S5), we found their sequences to be highly divergent. We explored whether this divergence could underlie the evolution of CBBs, using conservation scores, an estimate of the divergence of a pair of proteins or conserved protein regions (Lopez-Bigas and Ouzounis, 2004) (Fig. 5). A baseline for conserved molecules are the cell-cycle kinases, whose conservation is evident from the rescue of a *cdc2* fission yeast mutant and a *cdc5* budding yeast mutant by their human CDK1 and PLK1 counterparts, respectively (Lee and Erikson, 1997; Lee and Nurse, 1987). Their conservation scores (CS), calculated between the human sequence and either the *Drosophila* or zebrafish sequences, are *CS* _Drosophila_ = 0.75; *CS* _Zebrafish_ = 0.86 for CDK1, and *CS* _Drosophila_ = 0.51; *CS* _Zebrafish_ = 0.76 for PLK1. By contrast, SAC/PLK4 is much more divergent (*CS* _Drosophila_ = 0.18; *CS* _Zebrafish_ = 0.25; Fig. 5A,B). This divergence is more pronounced outside the kinase domain (Fig. 5C).
5A,C), which leads us to hypothesize that there was a fast change in the regulation of this enzyme on the evolutionary timescale.

We tested this hypothesis experimentally, taking advantage of the fact that overexpression of both *D. melanogaster* and human SAK/PLK4 leads to overduplication of centrioles (Bettencourt-Dias et al., 2005; Habedanck et al., 2005; Kleylein-Sohn et al., 2007; Rodrigues-Martins et al., 2007b). Whereas human SAK/PLK4 induced centriole amplification in human osteosarcoma cells (U2OS), the *D. melanogaster* counterpart did not, despite being able to localize to centrioles (Fig. 6A,B) and being expressed at similar or higher levels (supplementary material Fig. S9A). The reverse was also true, human SAK/PLK4 did not induce centriole amplification in *Drosophila* S2 cells (Fig. 6C,D; supplementary material Fig. S9B). It is thus possible that the divergence of these sequences has...
functional implications, leading to changes in protein regulation in a taxon-specific manner.

Taxon-specific divergence might be extreme in *C. elegans*, for which we did not find a SAK/PLK4 ortholog (Figs 2 and 4). The kinase ZYG1 in worms plays an important role upstream of SAS6 and SAS4, similar to human SAK/PLK4 (Bettencourt-Dias et al., 2005; Delatte et al., 2006; Habedanck et al., 2005; Kleylein-Sohn et al., 2007; Pelletier et al., 2006), and has been speculated to be its ortholog (Bettencourt-Dias et al., 2005; Song et al., 2008). When expressed in human and *Drosophila* cells, ZYG1 localized to centrosomes (Fig. 6A,C), although it did not induce centriole amplification (Fig. 6A-D). We further investigated the relationship of these kinases. We analyzed the phylogeny of their kinase domains and compared the structures of the C termini of ZYG1 and SAK/PLK4. We found a strongly supported monophyletic group of PLKs that included the known *C. elegans* PLKs 1-3, but not ZYG1, which is more similar to the centrosome kinases NIMA and MPS1 (Fig. 6E-G). Using fold recognition (3D-PSSM) (Kelley et al., 1999), we detected polo boxes in the C termini of both Polo/PLK1 and SAK/PLK4 kinases, but not in ZYG1 (data not shown). Moreover, we generated hidden Markov models (HMMs) of the so-called ‘cryptic polo box’ domain of animal SAK/PLK4, which targets it to the centrosome (Habedanck et al., 2005). This model was able to detect the distantly related SAK/PLK4 of the fungi *Batrachochytrium dendrobatidis*, but no *C. elegans* protein. The lack of both sequence similarity and supportive phylogenetic models (Fig. 6E-G) strongly supports the hypothesis that these molecules are not orthologs, that is, they do not share the same ancestry. Instead, the fact that ZYG1 can localize to centrosomes in *Drosophila* and human cells, and that it also plays a role upstream of SAS6 and SAS4 in *C. elegans* suggests a scenario of convergent evolution of ZYG1 and SAK/PLK4.

We were surprised to observe that the structural components of the UNIMOD were also very divergent, contrary to other structural proteins, such as tubulins, actins and myosins (Fig. 5B and data not shown). We wondered whether the presence of coiled coils could contribute to UNIMOD divergence. Coiled-coil conservation varies substantially, according to their function: protein-protein interaction motifs diverge very little, whereas protein domains that work as spacers and rods are more divergent [e.g. skeletal muscle myosin and nuclear mitotic apparatus protein (NuMA) diverge 2.1% and 18% between rat and human, respectively] (White and Erickson, 2006). We observed medium (8-12%) to high divergence (22%) of the UNIMOD coiled coils, suggesting that these sequences function as spacers or rods (White and Erickson, 2006) and thus contribute to UNIMOD divergence. Supporting this hypothesis for coiled-coil function as rods and spacers is the fact that *Chlamydomonas reinhardtii* BLD10 coiled-coil truncations lead to the assembly of smaller cartwheel spokes (Hiraki et al., 2007; Matsuura et al., 2004) (supplementary material Fig. S6).

In principle, high protein divergence could potentially mask the ancient origin of the non-UNIMOD proteins. However, we think that this is not the case for two main reasons. First, we found proteins with regions showing some degree of similarity but different protein architecture in all eukaryotic branches (Fig. 2; supplementary material Fig. S4). Second, when comparing conserved domains that define the UNIMOD, such as PISA and G-box domains, flagellated fungi and *Chlamydomonas* are less divergent from human than *Drosophila* proteins; however, SPD2, SAK/PLK4 and CP110 were found in *Drosophila* but in none of these other branches.

### Tissue specificity through subfunctionalization

We found two paralogs of SAS4/CPAP and BLD10/CEP135 in vertebrates, TCP10 and TSGA10, respectively (Figs 2 and 3). These vertebrate paralogs display the conserved G box and BLD10/CEP135 conserved region 2 (CR2), respectively. These duplicates are, in general, shorter than the ancestor family member present in organisms such as *Chlamydomonas* and *Drosophila*; in the case of TSGA10, it lacks BDL10/CEP135 CR1 (Figs 2, 3, Fig. 7A). What could be the role of these vertebrate paralogs in CBB assembly? *Chlamydomonas* and human BDL10/CEP135 have been shown to be important for early steps in CBB assembly (Hiraki et al., 2007; Kleylein-Sohn et al., 2007; Matsuura et al., 2004). TSGA10 is mainly expressed in testes and its absence is also associated with male sterility in humans (Modarressi et al., 2000). This protein localizes to the flagellum of mouse and bovine sperm (Behnam et al., 2006; Modarressi et al., 2004), suggesting a role in the assembly of sperm flagella. We propose two scenarios to explain this function of TSGA10 in the assembly of sperm flagella: subfunctionalization (partition of ancestral functions into the two duplicates) or neofunctionalization (acquisition of a new function by one duplicate).

We proceeded to test these scenarios in a model organism, *D. melanogaster*, which contains a single BDL10/CEP135 family member. These scenarios can be distinguished by the presence (subfunctionalization) or absence (neofunctionalization) of a *Drosophila* BLD10 (DmBLD10) function in flagella biogenesis, besides the expected role in centriole biogenesis. To test this, we used two approaches, RNAi in tissue culture cells and a mutant fly-fruit stock for BLD10/CEP135 (supplementary material Fig. S10A-C,D; Fig. S10A). We confirmed that DmBLD10 protein is absent from hemizygous mutant spermatocytes, whereas it localizes along centrioles in wild-type flies (supplementary material Fig. S10B). In line with its putative described ancestral function, we and others found that the protein localizes in the centrosomes of Dmel cells and RNAi leads to a decrease in centrosome number (supplementary material Fig. S10A-C,E) (Bettencourt-Dias et al., 2005; Dobbelaere et al., 2008; Rodrigues-Martins et al., 2007a). A role in centriole biogenesis is further supported by the observation that DmBLD10 mutant spermatocytes show shorter centrioles and premature centriole disengagement associated with defects in meiosis I of spermatogenesis (Fig. 7B-D; supplementary material Fig. S10D-F), similar to other mutants in which centriole biogenesis is impaired (Rodrigues-Martins et al., 2007a). We thus conclude that DmBLD10 is involved in centriole biogenesis, although the consequences of its absence are not as severe compared with SAS6 mutants (supplementary material Fig. S10G-I) (Bettencourt-Dias et al., 2005; Blachon et al., 2009; Peel et al., 2007; Rodrigues-Martins et al., 2007a).

We investigated a possible role for DmBLD10 in sperm formation. As in humans lacking TSGA10, DmBLD10 mutant males were sterile, suggestive of sperm malfunction (supplementary material Fig. S10C). The male infertility phenotype was not due to the inability of short centrioles to build axonemes, because the number of axonemes in 64 spermatid cysts of DmBLD10 mutants was similar to the one observed in the wild type (supplementary material Fig. S10D; Fig. S11A). However, we observed that the central microtubule pair, a structure essential for flagellum motility, was absent in mutant axonemes (Fig. 7E,F). The central pair is nucleated from a distal area of the basal body called the transition zone (McKean et al., 2003). Accordingly, we observed DmBLD10 to localize in a more distal region of the basal body (supplementary material Fig. S11B).
Our results and those from a recent report (Mottier-Pavie and Megraw, 2009) suggest that DmBLD10 mutant males are infertile because this molecule is needed for the assembly of the central microtubule pair of the axoneme. These data clearly support the subfunctionalization scenario, whereby two distinct ancestral functions of BLD10/CEP135 were present in a single protein in animals and were split between duplicates in vertebrates (Fig. 2). In this respect, it is interesting that TCP10, the dupeicate of SAS4/CPAP, is mainly expressed in testes and was originally identified as a member of the t-complex locus linked to male sterility (Cebra-Thomas et al., 1991; Schimenti et al., 1988). It will be important to investigate whether this molecule is also involved in flagella biogenesis.

The origin of the CBB-assembly machinery
Our detailed bioinformatics analysis of each protein family revealed the conserved regions (Fig. 3; supplementary material Figs S1-S6) that characterize each protein. These regions, considered together with the UNIMOD, represent a genomic identifier of the CBB. A long-standing debate revolves around the origin of these structures, with suggestions that the flagellum and its basal body have a bacterial origin, resulting from endosymbiosis (Dolan et al., 2002). We can now use these conserved regions to investigate whether the CBB ancestral core has bacterial counterparts. We generated profile HMMs of the conserved regions identified in this study and used them to search a database of 586 bacterial and 50 archaeal genomes. With the exception of the kinase domain of Polo, which is related to many protein kinase domains in bacteria and archaea (Kannan et al., 2007), we could not detect any positive hits suggestive of putative homologous sequences. This result indicates a eukaryotic origin of the CBB.

Discussion
The conservation of the morphology of the CBB structure contrasts with the diversity of contexts in which it assembles and operates in eukaryotic life. Focusing on the phylogenetic distribution of six proteins essential for centriole assembly in humans, we found that, in contrast to the previously observed conservation of ciliary and flagella components (Avior-Reiss et al., 2004; Li et al., 2004), CBB-assembly mechanisms evolved in a stepwise fashion (Figs 2 and 8). We propose that a subset of these proteins, which belong to what we call the universal module (UNIMOD), are necessary to both centriole and flagella biogenesis. These proteins have a similar phylogenetic distribution to that previously observed for ciliary and flagella components, and it is likely that new centriole components, such as POC1 (Keller et al., 2009; Pearson et al., 2009), will also fall into this subset. Furthermore, the set of proteins needed to form a centriole is likely to be larger than the UNIMOD, including proteins that also have non-centriolar functions and are present in organisms that do not have CBBs, such as α- and γ-tubulins and centrin. Mechanisms such as duplication with subfunctionalization of ancestral components (e.g. PLK and the BLD10/CEP135 families, Figs 6 and 7), divergence (e.g. SAK/PLK4, Figs 4, 5 and 6) and the emergence of new genes (e.g. SPG2/CEP192 and CP110; Fig. 2) play important roles in the evolution of CBB biogenesis. We have shown experimentally that subfunctionalization might have played a role in CBB evolution at least twice. In the case of BLD10/CEP135, duplication and subfunctionalization with the generation of TSGA10 is likely to be important in the development of tissue-specific mechanisms of CBB assembly and flagella formation (Fig. 7). In the case of the PLK family, the appearance of SAK/PLK4 with subfunctionalization (Fig. 4) is likely to play a role in uncoupling the regulation of CBB biogenesis from other cell-cycle events performed by PLKs. We have also shown experimentally that divergence in the PLK4 family leads to loss of cross-species complementation (Figs 5 and 6), which might create conditions for further development of species-specific regulation of CBB-assembly mechanisms. Finally, the emergence of novel molecules might have allowed adaptation.
Evolution of centriole assembly

Materials and Methods

Sequence analysis

We used the following approaches for the identification and classification of homologous proteins. (1) We searched for putative orthologs using BLASTP and iterative BLASTP in non-redundant protein databases (Altschul et al., 1990; Altschul et al., 1997; Schaffer et al., 2001) using the full human sequence of each family in eukaryotic species with complete, draft assembly or ongoing genome sequencing (supplementary material Table S2). We considered proteins to be orthologs as reciprocal best hits in BLASTP to the full human sequence (Overbeek et al., 1999). Top-scoring hits were further characterized and specific conserved regions were mapped for each family in multiple sequence alignments (Fig. 3). (2) To further query genome databases, we used regions of high conservation, either previously defined by others or identified in this study, in multiple sequence alignments of the bona fide members of each family. (3) We further investigated the negative results by querying the databases using family members of closely related species or using profile HMMs created with bona fide members of the family or specific conserved regions (using HMMER 2.3.2) (Eddy, 1998). (4) We used TBLASTN (Altschul et al., 1997) whenever sequences were too divergent or much shorter than other members of the family to search for the full protein sequence. (5) We further considered as orthologs those sequences that, although not obeying the first criterion for orthology (see above), were bidirectional best hits to members of the family in closely related species or to the most conserved regions in the human sequence (shown in Fig. 2 as grey boxes). (6) When possible, our orthology assignments were aided by phylogenetic analysis.

Fig. 8. Stepwise evolution of CBB assembly in eukaryotes. Our work suggests that an ancestral and universal module composed of SAS6, BLD10/CEP135 and SAS4/CPAP was present in the LECA and dictates or enforces CBB ninefold symmetry. We propose that CBB biogenesis is controlled by a trigger: an ancestral Polo/PLK1 kinase in bikonts that has both cell-cycle and CBB biogenesis functions, and its SAK/PLK4 duplicate in unikonts, which is involved only in CBB biogenesis. Note that, because of the lack of completely sequenced genomes of amebozoans containing CBBs, such as Physarum polycephalum, we cannot pinpoint whether SAK/PLK4 emergence occurred at the base of opisthokonts (purple clade) or before their divergence from amebozoa. Other proteins, such as SPD2/CEP192 and CP110, emerged in a taxon-specific manner to add new functions or regulatory steps to the conserved CBB structure, as discussed in the text. SPD2/CEP192 emerged in unikonts and its ancestral function is likely to be in PCM recruitment. SPD2/CEP192 might play an important role in CBB formation in contexts in which there is little PCM, such as duplication of the basal body upon fertilization. CP110 might play an important role in coordinating CBB assembly with cilia biogenesis and cytokinesis (see references in main text).
Correlation Molecule: CBB was calculated using the formula: 100\% \times [number of species showing correlation \times (n)\% of total number of species], where n is the total of species containing both CBB and the molecule, and n’ is the total of species containing both CBB and the molecule. Only sequenced species and species for which ultrastructure information exists were considered in this correlation (supplementary material Tables S1 and S2). Putative homologs that do not strictly satisfy our orthology criteria (grey squares in Fig. 2) were considered as negative hits. Multiple sequence alignments were performed using Muscle 3.6 with the default settings (Edgar, 2004a; Edgar, 2004b). The alignments were represented using Jalview v.2.3 (Waterhouse et al., 2009) with the BLOSUM62 color settings. The species used in the alignments are underlined in supplementary material Table S2. Organism-specific sequences larger than five residues were excluded, the alignment and are highlighted in supplementary material Fig. S5. Protein conservation values (Fig. 3) were obtained from these alignments using Jalview v.2.3 (Waterhouse et al., 2009) – each residue of the alignment is classified from 0 to 11 according to the percentage of aligned residues (these values are shown as a percentage). This information was shown graphically for each subset of protein orthologs. Regions in the alignment with more than 25% gaps are not scored and hence not included. HMMS were built using HMMer (http://hmmer.wustl.edu/) (Eddy, 1998) and these models were used to query specific genomes. A hit was considered significant if the e-value was lower than 0.1 and the bit-score was positive. We used this strategy for BDL10/CEP135, SPD2/CEP192, CP110 and the cryptic polo-box domain of known SAK/PLK4 orthologs, but still we were unable to find further orthologs. Phylogenies were inferred using: (i) neighbor joining (Saitou and Nei, 1987) as implemented in ClustalW 2.0 (Thompson et al., 1994) (1000 bootstraps); (ii) maximum likelihood (Felsenstein, 1981) in the Phylip 3.5 package (Felsenstein, 1981) and (iii) the Bayesian method implemented in MrBayes v.3.1.2 (with Blossum62, fixed amino acid rate mode and the program running until the error standard deviation was lower than 0.01). Trees were drawn using FigTree v.1.0.4 (http://tree.bio.ed.ac.uk/software/).

Transmission electron microscopy analysis of testses
Testses from pharate adults were dissected in 183 mM KCl, 47 mM NaCl, 1 mM EDTA and 1 mM Tris-HCl (pH 6.8), transferred to poly-L-lysine coated slides, frozen in liquid nitrogen and then frozen in acetone. DNA was stained with TOTO-3-iodide. Testses were mounted using Vectashield mounting media for fluorescence (Vector Laboratories). Testses were observed using a Leica DMRB microscope and high-resolution fluorescence spectra confocal microscope. Images are presented as maximum-intensity projections. For phase contrast microscopy analysis, testses were dissected in 0.7% NaCl solution and analyzed on an Olympus IMT-2 inverted microscope equipped with a Leica DC 200.

Antibodies
Mouse GT335 anti-polyglutamylated tubulin antibody was kindly provided by Carsten Janke (CNRS, France). The origin of the other antibodies was as follows: chicken anti-D-PLP (Rodrigues-Martins et al., 2007b); rabbit anti-pan-Pol (Sigma; 1:500). Secondary antibodies were purchased from Jackson ImmunoResearch Laboratories, USA, and used at 1:100 for immunostaining and 1:10,000 for western blot. The DmBNL10 antibody was generated in chicken against the peptide C-LADDRYNQARTREVS (Drosophila SAK/PLK4) including an N-terminal hexaHis tag (GenBank isolate # DQ873926).

Western blotting and reverse transcriptase (RT)-PCR
Drosophila SAK/PLK4 entry vector has been described elsewhere (GenBank isolate # DQ873926).

RT-PCR
Standard procedures were used for western blotting. Extracts of U2OS cells were prepared, resuspended the cells in 150 μl lysis buffer (50 mM HEPES pH 8, 200 mM NaCl, 5 mM EDTA, 1% NP-40 and protease inhibitors); all procedures were carried out on ice. Protein concentration was quantified using the Bradford reagent (BioRad) and the same amount of protein applied in the gel. Total RNA was extracted from cells using the RNaseasy mini kit (QUIAGEN) and RNase-free DNase set kit (QUIAGEN), according to the manufacturer's instructions. cDNA synthesis was carried out using the Transcripter First Strand cDNA synthesis Kit (ROCHE). PCR of the gene of interest was carried out using the same primers used for dsRNA synthesis. Amplification products of e14-kb cDNA were used as loading control.

Immunostaining and imaging
U2OS cells were fixed for 3 minutes in dry ice-cold methanol, permeabiled and washed in PBSTB (PBS containing 0.1% Triton X-100 and 1% BSA), and stained for polyglutamylated tubulin. Dmel cells were plated on glass coverslips and fixed in 1% paraformaldehyde in PHEM buffer (60 mM PIPES, 25 mM HEPES, 10 mM EGTA, 4 mM MgCl2). Cells were permeabiled and washed in PBSTB, and stained for Drosophila pericentrin-like protein (D-PLP). Dmel cells were plated on glass coverslips and fixed in dry ice-cold methanol followed by 10 minutes in acetone. DNA was stained with TOTO-3-iodide. Testses were mounted using Vectashield mounting media for fluorescence (Vector Laboratories). Testses were observed using a Leica DMRB microscope and high-resolution fluorescence spectra confocal microscope. Images are presented as maximum-intensity projections. For phase contrast microscopy analysis, testses were dissected in 0.7% NaCl solution and analyzed on an Olympus IMT-2 inverted microscope equipped with a Leica DC 200.

Transmission electron microscopy analysis of testses
Testses from 3- to 5-day-old adults were dissected and fixed in 2.5% glutaraldehyde in PBS (pH 7.2) for 2 hours at 4°C. Testses were post-fixed in OsO4 1% for 1 hour and treated with 1% uranyl acetate for 30 minutes. Samples were then dehydrated in a graded series of alcohols (70%, 90%, 100% for 1 minute each and three times in 100% for 10 minutes). Testses were incubated in propylene oxide three times for 10 minutes, followed by 1:1 propylene-oxide and resin twice for 15 minutes (Glaucet, 1984). Samples were embedded and solidified for 16-48 hours at 60°C. Thin sections (60-80 nm) were cut in a Leica Reichert Ultracut S ultramicrotome, collected on copper grids, and stained with uranyl acetate and lead citrate (Hatay, 1989). Samples were examined and photographed at 80 kV using either a Philips CM10 or a Morgagni 268 transmission electron microscope.