Evolutionary insights into the role of the essential centromere protein CAL1 in *Drosophila*

Ragini Phansalkar · Pascal Lapierre ·
Barbara G. Mellone

Published online: 21 July 2012
© Springer Science+Business Media B.V. 2012

Abstract Centromeres are essential *cis*-elements on chromosomes that are crucial for the stable transmission of genetic information during mitotic and meiotic cell divisions. Different species employ a variety of centromere configurations, from small genetically defined centromeres in budding yeast to holocentric centromeres that occupy entire chromosomes in *Caenorhabditis*, yet the incorporation of nucleosomes containing the essential centromere-specific histone H3 variant CENP-A is a common feature of centromeres in all eukaryotes. In vertebrates and fungi, CENP-A is specifically deposited at centromeres by a conserved chaperone, called HJURP or Scm3, respectively. Surprisingly, homologs of these proteins have not been identified in *Drosophila*, *Caenorhabditis*, or plants. How CENP-A is targeted to centromeres in these organisms is not known. The *Drosophila* centromeric protein CAL1, found only in the Diptera

genus, is essential for CENP-A localization, is recruited to centromeres at a similar time as CENP-A, and interacts with CENP-A in both chromatin and pre-nucleosomal complexes, making it a strong candidate for a CENP-A chaperone in this lineage. Here, we discuss the conservation and evolution of this essential centromere factor and report the identification of a “Scm3-domain”-like region with similarity to the corresponding region of fungal Scm3 as well as a shared predicted alpha-helical structure. Given the lack of common ancestry between Scm3 and CAL1, we propose that an optimal CENP-A binding region was independently acquired by CAL1, which caused the loss of an ancestral Scm3 protein from the Diptera lineage.

Keywords Centromere · chromosome · mitosis · meiotic drive · kinetochore · CENP-A · CAL1 · CID · histone variants · chromatin

Responsible Editor: Rachel O'Neill and Beth Sullivan

Electronic supplementary material The online version of this article (doi:10.1007/s10577-012-9299-7) contains supplementary material, which is available to authorized users.

R. Phansalkar · B. G. Mellone (✉)
Department of Molecular and Cell Biology,
University of Connecticut,
354 Mansfield Road, U2131, R247,
Storrs, CT 06269, USA
e-mail: barbara.mellone@uconn.edu

P. Lapierre
Bioinformatics Facility, Biotechnology/Bioservices Center,
University of Connecticut,
Storrs, CT, USA

Abbreviations

BLAST	Basic local alignment search tool
CAL1	Chromosome alignment defect 1
CATD	CENP-A targeting domain
CENP	Centromere protein
CID	Centromere identifier
DNMT3	DNA methyl transferase 3
HHPred	HMM-HMM prediction
HJURP	Holliday junction recognizing protein
HMM	Hidden Markov model
KNL2	Kinetochore null 2
LRT	Likelihood ratio test
Mis18	Missegregation 18

M18BP1	Mis18 binding protein 1
PAML	Phylogenetic analysis by maximum likelihood
Scm3	Suppressor of chromosome missegregation 3

Introduction

Cells use specialized regions of chromosomes, called centromeres, to define the location where kinetochores form during mitosis and meiosis. Kinetochores physically connect sister chromatids to spindle microtubules and mediate their transit towards opposite poles of forming daughter cells. Centromeres are stably inherited from one cell generation to the next and from parent to progeny. Within a species, the centromere position is maintained throughout evolutionary times. Centromere inheritance relies on the faithful incorporation of a unique type of chromatin, which contains the essential and universally conserved histone H3 variant, CENP-A (called CID in *Drosophila*) (Allshire and Karpen 2008). CENP-A loss from centromeres results in the failure to assemble kinetochores (Liu et al. 2006) while mislocalization or direct targeting of CENP-A to chromosome arms is sufficient to generate ectopic kinetochores (Heun et al. 2006; Mendiburo et al. 2011). Therefore, the presence of CENP-A within chromatin defines the position of the centromere.

Centromere identification relies on two pathways. First, pre-existing CENP-A nucleosomes must be stably transmitted to the newly forming sister chromosomes during DNA replication (centromere maintenance). Second, newly synthesized CENP-A must be incorporated onto pre-existing CENP-A nucleosome “templates” to replenish the nucleosomes lost during DNA replication (centromere assembly). While little is known about the pathways and molecules involved in centromere maintenance, a lot has been learned about the process of centromere assembly, particularly thanks to recent studies in human cells and *Xenopus* extracts (Barnhart et al. 2011; Bassett et al. 2012; Guse et al. 2011; Moree et al. 2011; Shuaib et al. 2010; Dunleavy et al. 2009; Foltz et al. 2009).

It has long been appreciated that the DNA underlying centromeres is dramatically divergent across

evolutionarily distant organisms. With the exception of yeast, mouse, and some human cell lines (Clarke and Carbon 1980; Hahnenberger et al. 1989; Harrington et al. 1997; Ikeno et al. 1998; Moralli et al. 2006), this DNA is not sufficient to attract CENP-A. Aside from the holocentric organism *C. elegans*, where CENP-A deposition occurs de novo at each cycle (Gassmann et al. 2012), metazoans use pre-existing CENP-A chromatin as a cue for further CENP-A recruitment at each cell cycle. This phenomenon is highlighted by the observation that once CENP-A is incorporated onto DNA, through either direct or indirect targeting strategies to chromosome arms, it is stably transmitted by the endogenous assembly pathway (Barnhart et al. 2011; Mendiburo et al. 2011). CENP-A has virtually no turnover at centromeres in animal cells as it is distributed to sister centromeres semi-conservatively (Hemmerich et al. 2008; Jansen et al. 2007; Mellone et al. 2010) and thus is always available to provide the “signal” needed by the CENP-A assembly machinery. Here, we explore the CENP-A assembly pathway of vertebrates and fungi, which harbor conserved centromere assembly factors, and of insects, which appear to have evolved an independent set of molecules to accomplish this essential function. Using bioinformatics and phylogenetics, we provide new insights into how CENP-A assembly may be mediated in these species and speculate that, despite the divergence at the protein level, the underlying mechanisms may be conserved.

Mechanisms of CENP-A assembly

How the presence of the CENP-A mark triggers the assembly of newly synthesized CENP-A at every cell cycle has been the subject of intense investigation. The molecules responsible for this pathway must accomplish two tasks: recognition of pre-existing CENP-A chromatin and assembly of new CENP-A. A key player in this process is the CENP-A assembly factor HJURP, which has been found in tetrapods as well as a choanoflagellate (Sanchez-Pulido et al. 2009; Foltz et al. 2009; Dunleavy et al. 2009). A homolog, called Scm3, is also present in fungi (Camahort et al. 2007; Mizuguchi et al. 2007; Pidoux et al. 2009; Stoler et al. 2007). HJURP selectively recognizes pre-nucleosomal CENP-A from canonical histone H3 and targets it to centromeres, starting in telophase and continuing

through G1 (Jansen et al. 2007; Lagana et al. 2010). Both HJURP and Scm3 have been shown to possess CENP-A assembly activities in vitro (Dechassa et al. 2011; Barnhart et al. 2011; Shivaraju et al. 2011). In humans, HJURP is recruited to the centromere by M18BP1 (Barnhart et al. 2011), a subunit of the Mis18 complex, which is essential for CENP-A incorporation (Fujita et al. 2007; Hayashi et al. 2004). M18BP1 is recruited to centromeres by direct binding to CENP-C (Moree et al. 2011; Dambacher et al. 2012), a constitutive centromere protein that connects the centromere to the kinetochore (Screpanti et al. 2011; Przewloka et al. 2011) and which binds to the very C-terminal residues of CENP-A (Guse et al. 2011).

Despite the crucial importance of Scm3 and HJURP and despite the fact that these proteins are shared between organisms as divergent as *Homo sapiens* and *Saccharomyces cerevisiae* (which last shared a common ancestor about one billion years ago), Scm3 and HJURP are not universal among metazoans as they appear to be absent from *Drosophila* and *Caenorhabditis* species (Sanchez-Pulido et al. 2009). A functional screen for regulators of the localization of CID also did not identify a HJURP/Scm3 homolog in *Drosophila* (Erhardt et al. 2008). *Drosophila* and *Caenorhabditis* species may have unrecognizable HJURP/Scm3 orthologs due to rapid evolution occurring in these lineages (Sanchez-Pulido et al. 2009); alternatively, novel proteins capable of interacting with CENP-A and assembling it onto chromatin may have emerged in their genomes. Notably, plants also do not harbor HJURP/Scm3 homologs (Dubin et al. 2010). Thus, HJURP and Scm3, unlike CENP-A, are not universal in Eukarya.

The ability to distinguish CENP-A from histone H3 is crucial to ensure centromere integrity. A region spanning part of α -helix 1, loop 1, and α -helix 2 within CENP-A (called the centromere targeting domain (CATD)) has been shown to be critical for CENP-A centromeric localization in humans, *Drosophila* and yeast (Shelby et al. 1997; Vermaak et al. 2002; Collins et al. 2007) and for recognition by HJURP (Bassett et al. 2012; Sekulic et al. 2010; Foltz et al. 2009). In HeLa cells, a chimeric histone H3 containing the CENP-A CATD (H3^{CATD}) was shown to be sufficient for centromeric localization and kinetochore function, suggesting that this region is critical for the assembly of centromeric factors (Black et al.

2007). Furthermore, recombinant co-expressed chimeric H3^{CATD} and histone H4 directly interact with GST-HJURP in vitro (Foltz et al. 2009). In contrast, H3^{CATD} chimeras do not localize to centromeres in *Drosophila* (Moreno-Moreno et al. 2011) or *Arabidopsis* (Ravi et al. 2010), demonstrating that this domain, albeit important, is not sufficient for centromere targeting in these species. It is possible that the alternative CENP-A chaperones used in organisms lacking HJURP/Scm3 have different specificity requirements.

Another unusual feature of *Drosophila* species is that they also lack members of the Mis18 complex (which in humans are Mis18 α , Mis18 β , and M18BP1) whereas an M18BP1 homolog, called KNL2, is present in nematodes (Maddox et al. 2007). Along with HJURP recruitment (Barnhart et al. 2011), functions of the Mis18 complex include transient histone H3 acetylation at the centromere in humans (Ohzeki et al. 2012), as well as centromeric recruitment of DNMT3 and DNA methylation in mice (Kim et al. 2012). This suggests that the Mis18 complex functions by recruiting the CENP-A assembly machinery and by creating the correct epigenetic context for its incorporation.

***Drosophila* CAL1 is an essential regulator of CID assembly**

An intriguing candidate that might fulfill the roles of either HJURP or the Mis18 complex, or both, in *Drosophila* is CAL1 (chromosome alignment defect 1) (Goshima et al. 2007; Erhardt et al. 2008). Depletion of CAL1 completely abolished the localization of CID and CENP-C and resulted in failure to segregate chromosomes (Erhardt et al. 2008). Using yeast two-hybrid assays, it was shown that the N terminus of CAL1 (residues 1–407) interacts with CID, while the C terminus (residues 699–979) interacts with CENP-C. The middle region (aa 392–722), which is less conserved across CAL1 orthologs, is required for the nucleolar localization of CAL1, but its function, if one exists, is unknown. Interestingly, CID and CENP-C do not interact in the absence of CAL1 (Schittenhelm et al. 2010). This is in sharp contrast with observations from reconstituted CENP-A chromatin in *Xenopus* egg extracts, where the C-

terminal CENP-A tail is sufficient to recruit CENP-C (Guse et al. 2011).

The observation that CAL1 is required to bring CID and CENP-C together in a complex has led to the proposal that CAL1 may simply be a bridging factor (Schittenhelm et al. 2010). However, the observation that CAL1 and CID interact in pre-nucleosomal complexes, and that CAL1 is recruited along with CID in mitosis (Mellone et al. 2010), suggests that CAL1 may also play a role in delivering CID to the centromere.

Identification of novel CAL1 orthologs outside of the *Drosophila* genus

Determining whether or not CAL1 has homologs beyond *Drosophila* could shed light on its origin and determine if CAL1 and HJURP/Scm3 homologs coexist in non-*Drosophilid* species. There are currently 12 orthologs of CAL1 listed on FlyBase, all from within *Drosophila*: *Drosophila melanogaster*, *Drosophila simulans*, *Drosophila sechellia*, *Drosophila yakuba*, *Drosophila erecta*, *Drosophila ananassae*, *Drosophila pseudoobscura*, *Drosophila persimilis*, *Drosophila grimshawi*, *Drosophila mojavensis*, *Drosophila virilis*, and *Drosophila willistoni*. Our searches using the Basic Local Alignment Search Tool (BLAST) revealed no additional significant hits aside from the known CAL1 orthologs. Since position-specific iterated BLAST (PSI-BLAST) is a more sensitive BLAST, we used it to identify more distant homologs. A second iteration identified an unnamed sequence (AGAP004338-PA) from *Anopheles gambiae* that showed some conservation with CAL1 and a third iteration identified an unnamed sequence (EFR29609) from *Anopheles darlingi*, which also showed similarity with CAL1. Next, the query sequence was divided according to the three functional domains of *D. melanogaster* CAL1 (Schittenhelm et al. 2010) and separate BLASTp (protein BLAST) searches were performed for each. A BLASTp of the N terminus, in addition to all the known CAL1 homologs within *Drosophila*, identified hypothetical proteins from *Culex quinquefasciatus* (XP_001843218.1) and *Aedes aegypti* (AaeL_AAEL009578). We conclude from these findings that CAL1 is conserved outside of the *Drosophila* genus. Interestingly, a

BLASTp of the C terminus identified only the known *Drosophila* CAL1 orthologs, suggesting that the N terminus is more conserved than the C terminus across distant dipteran species. BLASTp of the middle region of CAL1 only identified CAL1 orthologs from *D. yakuba* and *D. simulans*, consistent with the previously reported divergence of this part of CAL1 (Erhardt et al. 2008; Schittenhelm et al. 2010).

Additional BLASTp, BLASTn, and BLASTx (identification and comparison of protein coding sequences in genomic DNA) were carried out using the CAL1 sequence of *D. melanogaster* alongside the putative CAL1 ortholog from *D. grimshawi*, which, being a more distant relative of the *D. melanogaster* CAL1, could help to identify more distant homologs in the species tree. However, no additional sequences with significant similarity were found, leading us to speculate that CAL1 is not present outside of Diptera. Importantly, BLAST searches within Diptera species using HJURP and Scm3 sequences and alignments did not yield any significant hits, consistent with the idea that HJURP/Scm3 are not conserved in Diptera and that CAL1 and HJURP/Scm3 are mutually exclusive proteins.

Our BLAST searches also indicated that there are no obvious paralogs of CAL1 in *Drosophila*, suggesting that it is unlikely that CAL1 originated through a gene duplication event. Alternatively, paralogs could be unrecognizable due to high levels of diversification. A maximum likelihood gene tree of all the known CAL1 homologs, including the new orthologs we identified, was identical to the fly species tree, consistent with high conservation of CAL1 and with the presence of CAL1 in the common ancestor of *Drosophila* species (Fig. 1).

The N terminus of CAL1 contains a “Scm3 domain”-like region

After BLAST searches failed to identify any relatives of CAL1 outside Diptera genomes, we used HHPred to identify homologous regions within CAL1 in other species (Soding et al. 2005). HHPred identifies hits by comparing a profile hidden Markov model (HMM) of the input alignment to HMMs of sequences in a variety of databases. Because it searches alignment as opposed to sequence databases, and scores based on

Fig. 1 Maximum likelihood tree of all the known CAL1 orthologs. Nucleotide and amino acid alignments for the newly identified Diptera sequences from *A. gambiae* (E value of $8e^{-43}$), *A. darlingi* (E value of $1e^{-12}$), *C. quinquefasciatus* (E value of $6e^{-43}$), and *A. aegypti* (E value of $1e^{-30}$), and the 12 *Drosophila* CAL1 orthologs from Flybase were generated based on the translated amino acid sequences using Muscle in TranslatorX (Abascal et al. 2010). The Maximum likelihood tree for the Diptera CAL1 orthologs was calculated using the protein alignments in FastTree (Price et al. 2010) with the sequence replicate for the bootstrap values generated in Seqboot (Felsenstein 2005). The scale represents the average number of amino acid substitutions per site

both sequence similarity and secondary structure similarity, HHPred is a far more sensitive homolog detection algorithm than BLAST. For example, HHPred was recently used to demonstrate that HJURP and Scm3 shared common ancestry (Sanchez-Pulido et al. 2009).

Surprisingly, an initial HHPred query of the full-length *D. melanogaster* CAL1 protein sequence yielded a *Kluyveromyces lactis* Scm3 sequence from the UniProt database matching a short segment (amino acids, 5–40) from the conserved N terminus of CAL1 with low confidence (E value of 13). The region of *K. lactis* Scm3 most similar to CAL1 (residues 58–97) is part of the conserved “Scm3 domain”, a 52-amino acid long region conserved in Scm3 and HJURP, which mediates interaction with CENP-A in yeast and humans (Barnhart et al. 2011; Shuaib et al. 2010; Aravind et al. 2007; Mizuguchi et al. 2007; Sanchez-Pulido et al. 2009; Bassett et al. 2012). CAL1 orthologs from *D. willingstoni*, *D. virilis*, *D. mojavensis*, and *D. grimshawi* all identified the same

region of *K. lactis* Scm3 in HHPred, albeit with low confidence. Interestingly, an HHPred of the full-length CAL1 from *A. gambiae*, which last shared a common ancestor with *D. melanogaster* over 60 million years ago, identified the same region (residues 58–93) of *K. lactis* Scm3 as the top hit (E value of 7.1). Remarkably, *A. gambiae* CAL1 contains the residues Lys-Tyr (positions 112–113 in *S. cerevisiae* Scm3) which are part of the Scm3 domain and are highly conserved in all Scm3 and HJURP orthologs (Sanchez-Pulido et al. 2009). An alignment showing the region identified by HHPred from *K. lactis* Scm3 and representative CAL1 orthologs is shown in Fig. 2a. Because *K. lactis* Scm3 was only a marginally significant hit in the initial HHPred search, its appearance was subject to changes in parameters and database updates. However, the presence of multiple Scm3 hits for several of the queries and the appearance of *K. lactis* Scm3 in multiple searches increased the overall probability of sequence similarity between this and CAL1.

An HHPred of full-length *K. lactis* Scm3 did not yield any *Drosophila* sequences, but, predictably, included HJURP (residues 14–55), which matched residues 58–99 in *K. lactis* Scm3 (E value of 0.18). An alignment spanning a region of 36 residues in length, including the region in *D. melanogaster* CAL1 that matched *K. lactis* Scm3 from a subset of CAL1 sequences (*D. melanogaster*; *D. mojavensis*, *D. grimshawi*, *D. persimilis*, *D. erecta*, *D. ananassae*, *D. yakuba*, and *D. willingstoni*), was run in HHPred to determine if including a broader phylogenetic representation of Drosophilid CAL1 sequences improved the probability of similarity. This analysis identified the *K. lactis* Scm3 sequence (residues 76–97) as the top hit (E value of 0.55), and the *S. cerevisiae* Scm3 (residues 106–126) from a Cse4+Scm3+H4 single-chain fusion database entry (E value of 4.8) as another significant hit. Thus, the inclusion of more distant *Drosophila* species increased the significance of these HHPred searches in identifying similarity between CAL1 and Scm3.

The profile of the fungal Scm3 protein alignment was then compared with that of *Drosophila* CAL1, and vice versa, specifically for these 36 residues using HHPred. In each comparison, sequence similarity between these two families was statistically significant (E values of 5.1×10^{-5} and 6.3×10^{-5} , respectively; Fig. 2b). A similar analysis of the profiles of fungal Scm3 and metazoan HJURP alignments in HHPred

Fig. 2 A region within the N terminus of CAL1 shows homology to part of the “Scm3 domain” of Scm3. **a** Multiple alignment of the region identified by HHpred (<http://toolkit.tuebingen.mpg.de/hhpred>; Soding et al. 2005) of representative sequences from two (*S. cerevisiae* and *K. lactis*) Scm3 orthologs and four (*D. melanogaster*, *D. persimilis*, *D. mojavensis*, and *D. grimshawi*) CAL1 orthologs. Residues 87–125 of *S. cerevisiae* Scm3 are part of the so-called Scm3 domain. The magenta boxes highlight the Lys-Tyr and the Gly residues, which are highly conserved between tetrapod

HJURP and fungal Scm3 (Sanchez-Pulido et al. 2009). Residues in Scm3 found at the Scm3/Cse4 interface are marked by a blue “C” while residues at the histone H4 interface are marked by a red “H” (Cho and Harrison 2011). Asterisks indicate residues in this region of CAL1 that are under purifying selection (see text). **b** E values corresponding to profile-to-profile comparisons performed with HHalign (<http://toolkit.tuebingen.mpg.de/hhalgn>) between the fungal Scm3 and the fly CAL1 Scm3-like domain alignments. Arrows indicate the direction of the profile search

yielded highly significant similarity (E value of $<10^{-5}$) further supporting the previously identified common ancestry between HJURP and Scm3 (Sanchez-Pulido et al. 2009). Interestingly, PSIPRED, a secondary structure prediction algorithm, predicted the presence of an α -helix in this region of CAL1; similarly, the corresponding region in *K. lactis* Scm3 forms a slightly longer α -helix (residues 44–103) in the region that makes extensive contacts with both Cse4 and histone H4 (Sanchez-Pulido et al. 2009; Cho and Harrison 2011). Thus the similarity, as far as this region of CAL1 is concerned, extends beyond the sequence to the predicted protein structure.

Despite the apparent similarity of this CAL1 region to the Scm3 domain, our analyses do not support the existence of common ancestry between CAL1 and Scm3, consistent with a previous study which failed to identify Scm3 homologs in *Drosophila* (Sanchez-Pulido et al. 2009). CAL1 is a much larger protein than either HJURP or *K. lactis* Scm3 and the sequence similarity appears limited to a region within the Scm3 domain, with no discernible conservation across other parts of the protein sequence. However, given the functional similarities between CAL1 and Scm3/HJURP in the process of CENP-A loading, and the fact that

the N terminus of CAL1 interacts with CID (Schittenhelm et al. 2010), we speculate that the CAL1 N terminus may have independently evolved a region that resembles part of the Scm3 domain of Scm3 to carry out CENP-A binding through convergence. Future studies aimed at elucidating the structure of the CAL1 N terminus in complex with CID, as well as studies addressing the ability of CAL1 to recruit CID, are needed to determine whether CAL1 and HJURP/Scm3 are functionally and, at least in part given the findings above, structurally analogous proteins.

Is CAL1 evolving under positive selection?

Previous studies have shown that CENP-A is evolving adaptively in *Drosophila*, *Arabidopsis*, and primates (Malik and Henikoff 2001; Talbert et al. 2002; Schuler et al. 2010). It has been proposed that adaptive evolution of centromere-binding proteins occurs to counterbalance the harmful effects of centromere drive, an expansion of centromeric satellites that leads to the preferential segregation of the “stronger” centromere into the oocyte of organisms with asymmetric female meiosis (Henikoff et al. 2001; Malik and Henikoff 2002). The difficulty in identifying orthologs

for CAL1, coupled with the adaptive evolution of CID and the functional association between these two proteins, raises the question of whether the evolutionary history of CAL1 may also be under positive selection.

Evolutionary forces acting on proteins can be measured by comparing the rates of nonsynonymous (dN) and synonymous (dS) nucleotide substitutions between coding sequences from closely related species. The dN and dS are expected to be equal for proteins under neutral selection (dN/dS or $\omega=1$). Negative or purifying selection results in dN/dS of <1 and indicates that changes in the protein are deleterious and thus eliminated from the population. Positive selection (dN>dS) can indicate that a sequence is evolving adaptively, as is often observed as a consequence of genetic conflict, such as that existing between virus and host in which co-evolving host–parasite systems are perpetually changing in an unceasing arms race (Valen 1973).

Analysis of full-length CAL1 using the M0 model in PAML (Yang 2007) resulted in a global dN/dS <1 (0.18400), which suggests that the gene overall is not subject to positive selection. Analysis of the N terminus of CAL1 resulted in the lowest dN/dS ratio (dN/dS=0.094), compared with the C terminus (dN/dS=0.182) and the middle region (dN/dS=0.625), suggesting these individual regions are also not under positive selection. It is worth noting that the middle region of CAL1 cannot be aligned correctly due to the presence of many gaps and high sequence variability, thus, the higher dN/dS ratio for this region is most likely due to alignment artifacts.

Analyses of the dN/dS ratios for sites under positive selection were conducted in PAML for the N and C termini of CAL1. No sites were found to be under positive selection with significant confidence. A likelihood ratio test (LRT) for the whole sequence, and for the N terminus, C terminus, and middle regions comparing the M7 and M8 models in PAML showed that the model allowing for positive selection (M8) was statistically better than the neutral model (M7) for the whole sequence and middle region, but not for the N and C termini (Table 1). However, the lack of sites found under significant positive selection in either the N or C termini suggests that this is a result of the presence of a few sites with dN/dS ratios of >1 in the middle region of the protein, which again is likely due to alignment artifacts.

To determine whether selective pressure varied along the lineages at any point in time, we carried out a branch-site model analysis in PAML to calculate the dN/dS ratios (ω) of each branch of the reference tree using the full-length CAL1 sequence, as well as the N- and C-terminal regions (Fig. 3). Five branches were identified with $\omega>1$ for at least one of the three sequences. For each branch with $\omega>1$, an LRT was performed comparing a model under neutral evolution (fixed ω to 1), versus a model allowing the branch of interest to be under positive selection (Table 2). Two branches were found to have adaptively evolved (Fig. 3): the branch separating *D. grimshawi*, *D. virilis*, and *D. mojavensis* from the other species and the branch leading to *D. persimilis*. The corresponding codons under positive selection are within the N terminus of CAL1 (Table 3), suggesting that part of the CID binding region experienced positive selection in these lineages. It will be interesting to determine whether these residues are involved in forming the interface between the N terminus of CAL1 and CID and whether the corresponding residues in CID also evolved adaptively in any of the tree branches.

Given the lack of detectable diversifying selection within CAL1, we sought to determine whether the conserved N and C termini are evolving under neutral or purifying selection (Pond and Frost 2005). A total of 123 sites in the N terminus (~30 % of N-terminal sites) and 50 sites in the C terminus (~20 % of C-terminal sites) were found to be under purifying selection with significant *p* values, while only one such site was identified in the middle region (codon 696; see Supplemental Table 1). Several residues from within the Scm3 domain-like region identified by HHPred were found to be under purifying selection, consistent with high conservation in this region (Fig. 2a). Collectively, our analyses show that CAL1 is not evolving adaptively and that, in fact, several sites within the N and C termini of CAL1 are under purifying selection.

Thus, it appears that CAL1 does not participate in the meiotic drive observed for CID in this species group. We speculate that, while CID is rapidly evolving at sites where it makes contact with the DNA (Henikoff et al. 2001; Malik and Henikoff 2001; Malik et al. 2002), the interface between CAL1 and CID remains unchanged to ensure robust CID-CAL1 interaction and reliable CID recruitment at centromeres.

Table 1 LRT of positive selection over the full sequence, N terminus, C terminus, and middle region of CAL1

CAL1 regions	2 Δ l (M7 vs. M8)	<i>p</i> values	ω	Percentage of sites
Whole protein	21.822	1.83E⁻⁰⁵	3.663	0.057
N terminus	1.517	0.468	1.000	0.019
Middle region	12.925	0.002	3.395	0.198
C terminus	-0.001	1.000	2.402	1E⁻⁰⁵

PAML was used to search for positive selection in CAL1 under a global model. dN/dS ratios were calculated for the full-length sequence, N terminus (Codon 1–399 based on the *D. melanogaster* sequence), C terminus (729–979), and middle region (400–728). Two models were used to test for residues under positive selection (M7 vs. M8). M7 (beta) determines the number of categories of ω between 0 and 1 (10 in this analysis) estimated by a beta continuous distribution. M8 (beta and ω) is similar to M7 with an additional category of $\omega > 1$. A Chi-square test of the parameter 2 Δ l (twice the difference in log likelihoods between models M7 and M8 in PAML) was used to determine the *p* value for evidence of positive selection for sites under the M7 vs. M8 model. Significant *p* values of <0.05 (**bold typeface**) correspond to regions of CAL1 where a model allowing for sites with $\omega > 1$ (M8) is better than the neutral (M7) model. The ω values correspond to $\omega > 1$ estimated under the M8 model and the percentage of sites column is the number of sites falling under that ω for the sequence analyzed

Fig. 3 Detection of positive selection on CAL1 in *Drosophila*. Unrooted maximum likelihood tree of 12 *Drosophila* species with the ω (dN/dS) values as calculated with using the branch-site model M1 in PAML. The three values on each branch correspond to the ω calculated using the full-length, N and C terminus of the CAL1 sequences. $\omega < 1$ values indicate branches under purifying selection, while $\omega > 1$ values identify branches identified under positive selection (**blue lines**). Branches 2 and 3 were determined to be statistically significant for positive selection (**red lines**) using an LRT which compared each branch of interest under a null model (H0) of neutral or purifying selection, against a model (H1) that allowed the tested branch to evolve under positive selection in PAML (see Table 2). The *scale* represents the average number of amino acid substitutions per site

Table 2 LRT for branches under positive selection

CAL1 branch	2 Δ l (H0 vs. H1)	<i>p</i> values
Complete sequence		
Branch 1	3.789	0.052
Branch 2	22.270	0.000
Branch 3	10.522	0.001
N terminus		
Branch 1	0.917	0.338
Branch 2	32.240	1.4E⁻⁰⁸
Branch 3	12.381	4.0E⁻⁰⁴
C terminus		
Branch 1	1.110	0.292
Branch 4	1.248	0.264
Branch 5	0.000	1.000

Five branches were identified to be potentially under positive selection using the branch-site model (M1) in PAML (see Fig. 3). An LRT was performed for these five branches to determine the *p* values associated with the $\omega > 1$ at these branches. Each branch was tested under a null model (H0) of neutral or purifying evolution against an alternate model (H1) that allowed the branch of interest to evolve under positive selection. *p* values of 0.05 or below (*bold typeface*) correspond to branches where the model with positive selection is favored over the neutral model. According to this analysis, branches 2 and 3 experienced positive selection with most of the signal coming from the N terminus of CAL1 (see Fig. 3)

Could CAL1 have replaced an ancestral Scm3-like chaperone in flies?

The presence of HJURP/Scm3 proteins in such divergent species as fungi and vertebrates, and their

absence in plants, insects, and nematodes, is an elusive puzzle of centromere biology. The absence of universal CENP-A chaperones could be indicative of the existence of alternative molecular mechanisms mediating CENP-A assembly in diverse species. On the other

Table 3 Residues under positive selection identified under the branch-site model LRT

Full sequence			N terminus		
Codon	AA	<i>p</i> values	Codon	AA	<i>p</i> values
Branch 2					
97	S	0.952	97	S	0.974
206	A	0.993	122	V	0.972
218	C	0.998	206	A	0.997
262	D	0.974	218	C	0.999
284	A	0.976	262	D	0.981
307	P	0.963	284	A	0.990
311	S	0.979	307	P	0.974
			311	S	0.991
			399	K	0.952
Branch 3					
358	Q	0.986	358	Q	0.990

Bayes Empirical Bayes analysis identified nine codons for branch 2 and one codon for branch 3 with probability values of being under positive selection greater than 95 %

Fig. 4 Models for centromere assembly in vertebrates and *Drosophila*. **a** In vertebrates, CENP-C binds pre-existing CENP-A nucleosomes (red and blue circles) via the small C-terminal tail of CENP-A (curvy gray line). During telophase, M18BP1 localizes to the centromere through its interaction with CENP-C. From late telophase through G1, M18BP1 recruits HJURP in complex with new CENP-A/H4 dimers (yellow half circle). HJURP assembles new CENP-A onto centromeric nucleosomes (place holder H3.3 (Dunleavy et al. 2011) not shown). **b** In *Drosophila*, CENP-C marks the centromere for

CID incorporation through its interaction with the pre-existing CID (red and blue circles) in complex with CAL1 and CENP-C (dashed arrow). During mitosis, newly synthesized CID is delivered to the centromere by soluble CAL1. New CID/H4 assembly (yellow half circle) could be mediated by CAL1 directly or through the recruitment of a different CID assembly factor (orange). A possible loading factor is the p55 subunit of CAF1, which has been shown to stimulate CID nucleosome assembly in vitro (Furuyama et al. 2006), although the functional relevance of this activity has not been demonstrated in vivo

hand, the observed lack of homologs could be misleading, and in fact evolutionarily distinct chaperones could use the same molecular strategies of HJURP/Scm3 to recognize and assemble CENP-A.

The evolutionary divergence spanned by the *Drosophila* genus exceeds that of mammals, when generation time is taken into account; therefore, these species provide an excellent model to study how conserved functions are maintained despite sequence divergence (Clark et al. 2007). In this lineage, HJURP/Scm3 and Mis18 complex homologs are absent while the new protein CAL1 shows high conservation in the regions known to interact with CENP-A and CENP-C. The surprising identification of a Scm3 domain-like region within CAL1, at both the sequence and predicted structure levels, raises the possibility that this particular region is optimal for CENP-A recruitment and has been acquired independently by Scm3 and CAL1. This region contains a high number of residues under purifying selection, which suggests that upon the establishment of CAL1 as a robust CID loading partner, and perhaps upon loss of a Scm3-like ancestral chaperone, the CAL1/CID interaction interface was unable to change without an effect on its binding capacity. Similarly, the presence of residues under purifying selection in the C terminus of CAL1 underscores the crucial importance of

the interaction between CAL1 and CENP-C. The C terminus of CAL1 could function by directing the CID pre-nucleosomal complex to the centromere, in a manner similar to human M18BP1, which recruits HJURP in complex with CENP-A (Barnhart et al. 2011). We propose that CAL1 is functionally analogous to HJURP in delivering CID to the centromere and M18BP1 in mediating centromere recognition through CENP-C binding in *Drosophila* (Fig. 4).

The discovery of HJURP/Scm3 and of its mechanisms of action has provided much sought-after insights into the pathways mediating accurate CENP-A recognition and assembly. If CAL1 indeed functions as a CID chaperone in Diptera, it will be interesting to investigate what advantages this protein offered over HJURP/Scm3 in this lineage, and whether worm and plant-specific chaperones with distinct homology signatures exist. Alternatively, CAL1 (and equivalent CENP-A binding factors specific to worms and plants) could function as an adaptor bringing together soluble CID and a general histone-assembly factor thereby mediating CID assembly indirectly. While this glimpse into the evolutionary history of CAL1 has afforded model predictions for refining the role of CAL1 in centromere assembly, future cell biological and structural studies will be needed to test this and other models.

Acknowledgments We gratefully acknowledge Seth Kasowitz, Asav Dharia, Paul Talbert, Peter Gogarten, and Craig Nelson for help with the evolutionary analyses of CAL1 and Karolin Luger for suggesting HHPred. This work was funded by the National Science Foundation award number 1024973 to BGM.

References

- Abascal F, Zardoya R, Telford MJ (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res* 38:W7–W13
- Allshire RC, Karpen GH (2008) Epigenetic regulation of centromeric chromatin: old dogs, new tricks? *Nat Rev Genet* 9:923–937
- Aravind L, Iyer LM, Wu C (2007) Domain architectures of the Scm3p protein provide insights into centromere function and evolution. *Cell Cycle* 6:2511–2515
- Barnhart MC, Kuich PH, Stellfox ME et al (2011) HJURP is a CENP-A chromatin assembly factor sufficient to form a functional de novo kinetochore. *J Cell Biol* 194:229–243
- Bassett EA, Denizio J, Barnhart-Dailey MC et al (2012) HJURP uses distinct CENP-A surfaces to recognize and to stabilize CENP-A/histone H4 for centromere assembly. *Dev Cell* 22:749–762
- Black BE, Jansen LE, Maddox PS et al (2007) Centromere identity maintained by nucleosomes assembled with histone H3 containing the CENP-A targeting domain. *Mol Cell* 25:309–322
- Camahort R, Li B, Florens L et al (2007) Scm3 is essential to recruit the histone H3 variant Cse4 to centromeres and to maintain a functional kinetochore. *Mol Cell* 26:853–865
- Cho US, Harrison SC (2011) Recognition of the centromere-specific histone Cse4 by the chaperone Scm3. *Proc Natl Acad Sci U S A* 108:9367–9371
- Clark AG, Eisen MB, Smith DR et al (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450:203–218
- Clarke L, Carbon J (1980) Isolation of a yeast centromere and construction of functional small circular chromosomes. *Nature* 287:504–509
- Collins KA, Camahort R, Seidel C, Gerton JL, Biggins S (2007) The overexpression of a *Saccharomyces cerevisiae* centromeric histone H3 variant mutant protein leads to a defect in kinetochore biorientation. *Genetics* 175:513–525
- Dambacher S, Deng W, Hahn M et al (2012) CENP-C facilitates the recruitment of M18BP1 to centromeric chromatin. *Nucleus* 3:101–110
- Dechassa ML, Wyns K, Li M et al (2011) Structure and Scm3-mediated assembly of budding yeast centromeric nucleosomes. *Nat Comm* 2:313
- Dubin M, Fuchs J, Graf R, Schubert I, Nellen W (2010) Dynamics of a novel centromeric histone variant CenH3 reveals the evolutionary ancestral timing of centromere biogenesis. *Nucleic Acids Res* 38:7526–7537
- Dunleavy EM, Roche D, Tagami H et al (2009) HJURP is a cell-cycle-dependent maintenance and deposition factor of CENP-A at centromeres. *Cell* 137:485–497
- Dunleavy EM, Almouzni G, Karpen GH (2011) H3.3 is deposited at centromeres in S phase as a placeholder for newly assembled CENP-A in G phase. *Nucleus* 2:146–157
- Erhardt S, Mellone BG, Betts CM et al (2008) Genome-wide analysis reveals a cell cycle-dependent mechanism controlling centromere propagation. *J Cell Biol* 183:805–818
- Foltz DR, Jansen LE, Bailey AO et al (2009) Centromere-specific assembly of CENP-a nucleosomes is mediated by HJURP. *Cell* 137:472–484
- Fujita Y, Hayashi T, Kiyomitsu T et al (2007) Priming of centromere for CENP-A recruitment by human hMis18alpha, hMis18beta, and M18BP1. *Dev Cell* 12:17–30
- Furuyama T, Dalal Y, Henikoff S (2006) Chaperone-mediated assembly of centromeric chromatin in vitro. *Proc Natl Acad Sci U S A* 103:6172–6177
- Gassmann R, Rechtsteiner A, Yuen KW (2012) An inverse relationship to germline transcription defines centromeric chromatin in *C. elegans*. *Nature* 484:534–537
- Goshima G, Wollman R, Goodwin SS et al (2007) Genes required for mitotic spindle assembly in *Drosophila* S2 cells. *Science* 316:417–421
- Guse A, Carroll CW, Moree B, Fuller CJ, Straight AF (2011) In vitro centromere and kinetochore assembly on defined chromatin templates. *Nature* 477:354–358
- Hahnenberger KM, Baum MP, Polizzi CM, Carbon J, Clarke L (1989) Construction of functional artificial minichromosomes in the fission yeast *Schizosaccharomyces pombe*. *Proc Natl Acad Sci U S A* 86:577–581
- Harrington JJ, Van Bokkelen G, Mays RW, Gustashaw K, Willard HF (1997) Formation of de novo centromeres and construction of first-generation human artificial microchromosomes. *Nat Genet* 15:345–355
- Hayashi T, Fujita Y, Iwasaki O et al (2004) Mis16 and Mis18 are required for CENP-A loading and histone deacetylation at centromeres. *Cell* 118:715–729
- Hemmerich P, Weidtkamp-Peters S, Hoischen C et al (2008) Dynamics of inner kinetochore assembly and maintenance in living cells. *J Cell Biol* 180:1101–1114
- Henikoff S, Ahmad K, Malik HS (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* 293:1098–1102
- Heun P, Erhardt S, Blower MD et al (2006) Mislocalization of the *Drosophila* centromere-specific histone CID promotes formation of functional ectopic kinetochores. *Dev Cell* 10:303–315
- Ikeno M, Grimes B, Okazaki T et al (1998) Construction of YAC-based mammalian artificial chromosomes. *Nat Biotechnol* 16:431–439
- Jansen LE, Black BE, Foltz DR, Cleveland DW et al (2007) Propagation of centromeric chromatin requires exit from mitosis. *J Cell Biol* 176:795–805
- Kim IS, Lee M, Park KC et al (2012) Roles of Mis18alpha in epigenetic regulation of centromeric chromatin and CENP-A loading. *Mol Cell* 46:260–273
- Lagana A, Dorn JF, De Rop V et al (2010) A small GTPase molecular switch regulates epigenetic centromere maintenance by stabilizing newly incorporated CENP-A. *Nat Cell Biol* 12:1186–1193
- Liu ST, Rattner JB, Jablonski SA, Yen TJ (2006) Mapping the assembly pathways that specify formation of the trilaminar kinetochore plates in human cells. *J Cell Biol* 175:41–53

- Maddox PS, Hyndman F, Monen J, Oegema K, Desai A (2007) Functional genomics identifies a Myb domain-containing protein family required for assembly of CENP-A chromatin. *J Cell Biol* 176:757–763
- Malik HS, Henikoff S (2001) Adaptive evolution of CID, a centromere-specific histone in *Drosophila*. *Genetics* 157:1293–1298
- Malik HS, Henikoff S (2002) Conflict begets complexity: the evolution of centromeres. *Curr Opin Genet Dev* 12:711–718
- Malik HS, Vermaak D, Henikoff S (2002) Recurrent evolution of DNA-binding motifs in the *Drosophila* centromeric histone. *Proc Natl Acad Sci U S A* 99:1449–1454
- Mellone BG, Grive KJ, Shteyn V et al (2010) Assembly of *Drosophila* centromeric chromatin proteins during mitosis. *PLoS Genet* 7:e1002068
- Mendiburo MJ, Padeken J, Fulop S, Schepers A, Heun P (2011) *Drosophila* CENH3 is sufficient for centromere formation. *Science* 334:686–690
- Mizuguchi G, Xiao H, Wisniewski J, Smith MM, Wu C (2007) Nonhistone Scm3 and histones CenH3-H4 assemble the core of centromere-specific nucleosomes. *Cell* 129:1153–1164
- Moralli D, Simpson KM, Wade-Martins R, Monaco ZL (2006) A novel human artificial chromosome gene expression system using herpes simplex virus type 1 vectors. *EMBO Rep* 7:911–918
- Moree B, Meyer CB, Fuller CJ, Straight AF (2011) CENP-C recruits M18BP1 to centromeres to promote CENP-A chromatin assembly. *J Cell Biol* 194:855–871
- Moreno-Moreno O, Medina-Giro S, Torras-Llort M, Azorin F (2011) The F box protein partner of paired regulates stability of *Drosophila* centromeric histone H3, CenH3 (CID). *Curr Biol* 21:1488–1493
- Ohzeki JJ, Bergmann JH, Kouprina N (2012) Breaking the HAC barrier: histone H3K9 acetyl/methyl balance regulates CENP-A assembly. *EMBO J* 31:2391–2402
- Pidoux AL, Choi ES, Abbott JK et al (2009) Fission yeast Scm3: A CENP-A receptor required for integrity of sub-kinetochore chromatin. *Mol Cell* 33:299–311
- Pond SL, Frost SD (2005) Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics* 21:2531–2533
- Price MN, Dehal PS, Arkin AP (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490
- Przewlaka MR, Venkei Z, Bolanos-Garcia VM et al (2011) CENP-C is a structural platform for kinetochore assembly. *Curr Biol* 21:399–405
- Ravi M, Kwong PN, Menorca RM et al (2010) The rapidly evolving centromere-specific histone has stringent functional requirements in *Arabidopsis thaliana*. *Genetics* 186:461–471
- Sanchez-Pulido L, Pidoux AL, Ponting CP, Allshire RC (2009) Common ancestry of the CENP-A chaperones Scm3 and HJURP. *Cell* 137:1173–1174
- Schittenhelm RB, Althoff F, Heidmann S, Lehner CF (2010) Detrimental incorporation of excess Cenp-A/CID and Cenp-C into *Drosophila* centromeres is prevented by limiting amounts of the bridging factor Cal1. *J Cell Sci* 123:3768–3779
- Schueler MG, Swanson W, Thomas PJ, Green ED (2010) Adaptive evolution of foundation kinetochore proteins in primates. *Mol Biol Evol* 27:1585–1597
- Screpanti E, De Antoni A, Alushin GM et al (2011) Direct binding of Cenp-C to the Mis12 complex joins the inner and outer kinetochore. *Curr Biol* 21:391–398
- Sekulic N, Bassett EA, Rogers DJ, Black BE (2010) The structure of (CENP-A-H4)(2) reveals physical features that mark centromeres. *Nature* 467:347–351
- Shelby RD, Vafa O, Sullivan KF (1997) Assembly of CENP-A into centromeric chromatin requires a cooperative array of nucleosomal DNA contact sites. *J Cell Biol* 136:501–513
- Shivaraju M, Camahort R, Mattingly M, Gerton JL (2011) Scm3 is a centromeric nucleosome assembly factor. *J Biol Chem* 286:12016–12023
- Shuaib M, Ouararhni K, Dimitrov S, Hamiche A (2010) HJURP binds CENP-A via a highly conserved N-terminal domain and mediates its deposition at centromeres. *Proc Natl Acad Sci U S A* 107:1349–1354
- Soding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* 33:W244–W248
- Stoler S, Rogers K, Weitze S et al (2007) Scm3, an essential *Saccharomyces cerevisiae* centromere protein required for G2/M progression and Cse4 localization. *Proc Natl Acad Sci U S A* 104:10571–10576
- Talbert PB, Masuelli R, Tyagi AP, Comai L, Henikoff S (2002) Centromeric localization and adaptive evolution of an *Arabidopsis* histone H3 variant. *Plant Cell* 14:1053–1066
- Valen V (1973) A new evolutionary law. *Evol Theory* 1:1–30
- Vermaak D, Hayden HS, Henikoff S (2002) Centromere targeting element within the histone fold domain of CID. *Mol Cell Biol* 22:7553–7561
- Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591
- Felsenstein, J (2005) PHYLIP version 3.6 (distributed by the author). Department of Genome Sciences, University of Washington, Seattle