Centromere Targeting Element within the Histone Fold Domain of Cid

Danielle Vermaak,† Hillary S. Hayden,‡ and Steven Henikoff*†

Howard Hughes Medical Institute and Basic Sciences Division,† Fred Hutchinson Cancer Research Center, Seattle, Washington 98109

Received 29 May 2002/Returned for modification 18 July 2002/Accepted 30 July 2002

Centromeres require specialized nucleosomes; however, the mechanism of localization is unknown. Drosophila sp. centromeric nucleosomes contain the Cid H3-like protein. We have devised a strategy for identifying elements within Cid responsible for its localization to centromeres. By expressing Cid from divergent Drosophila species fused to green fluorescent protein in Drosophila melanogaster cells, we found that D. bimucronata Cid fails to localize to centromeres. Cid chimeras consisting of the D. bipectinata histone fold domain (HFD) replaced with segments from D. melanogaster identified loop I of the HFD as being critical for targeting to centromeres. Conversely, substitution of D. bipectinata loop I into D. melanogaster abolished centromeric targeting. In either case, loop I was the only segment capable of conferring targeting. Within loop I, we identified residues that are critical for targeting. Most mutations of conserved residues abolished targeting, and length reductions were deleterious. Taken together with the fact that H3 loop I makes numerous contacts with DNA and with the adaptive evolution of Cid, our results point to the importance of DNA specificity for targeting. We suggest that the process of deposition of (Cid.H4)2 tetramers allows for discriminating contacts to be made between loop I and DNA, providing the specificity needed for targeting.

The centromere is the locus responsible for poleward movement of the chromosomes during cell division. All eukaryotic centromeres are characterized by specialized nucleosomes containing an H3-like histone (CenH3) in place of canonical H3 (6, 36). In the absence of CenH3, centromere function and localization of other kinetochore components are abolished (2, 5, 14, 27, 35). Overexpression of human CenH3 (CENP-A) leads to its mislocalization to chromosome arms where it recruits kinetochore components (42). Therefore, CenH3-containing nucleosomes form the basis for organizing the kinetochore.

Canonical nucleosomes consist of 146 bp of DNA wrapped around an octamer of two each of the core histones H2A, H2B, H3, and H4 and are the fundamental packaging unit of chromatin (45). The first step in chromatin assembly is deposition of an (H3.H4)2 tetramer, which by itself organizes 120 bp of DNA (8, 11, 24). Subsequent deposition of two H2A.H2B dimers completes a nucleosome core particle (45). CenH3 replaces canonical H3 in nucleosomes both in vivo and in vitro (1, 2, 5, 18, 25, 29, 33, 34, 35, 47). CenH3 and H3 do not heterodimerize in the same nucleosome (1, 3, 33) and are therefore probably found as separate tetramers with histone H4 before incorporation into chromatin. Maintenance of CenH3 chromatin depends upon targeting of (CenH3.H4)2 tetramers to centromeres, a process that is poorly understood.

CenH3s differ from canonical H3 in that they have diverged histone fold domains (HFDs), noncanonical N-terminal tails, and slightly longer loop I (LI) regions (12, 36). Differences between CenH3s and canonical H3 appear to specify centromeric targeting of CenH3s. Deletion analysis showed that the HFD and not the N-terminal tail of CENP-A targeted centromeres in a human cell line (33, 37). However, no single targeting element could be mapped by swapping amino acids between CENP-A and canonical H3 (33, 37). For Saccharomyces cerevisiae CenH3 (Cse4), targeting was not specifically assayed, although residues throughout the HFD were similarly required for centromere function (9, 16). Since these regions were predicted to contact the DNA, it was proposed that DNA specificity was important for centromeric targeting of CenH3 (9, 16, 33, 37). However, this proposal appears to be contradicted by the lack of DNA sequence conservation at centromeres (6, 12, 36) and by the existence of human neocentromeres, known to be devoid of the α-satellite DNA found at canonical human centromeres (7, 18, 19, 30, 43).

Recent evolutionary studies of Drosophila and Arabidopsis CenH3s support a role for DNA specificity in localizing CenH3-containing tetramers (12, 21, 39). In Drosophila, both the N-terminal tail and LI of the Cid HFD were found to be evolving adaptively (21). LI is known to form a DNA-binding domain together with LII of H4 (20), suggesting that LI mediates DNA specificity in Cid-containing chromatin, presumably at the level of targeting of (Cid.H4)2 tetramers. Given the lack of mechanistic insight into the process of CenH3 targeting to centromeres, we have investigated the targeting determinants of Cid.

We reasoned that amino acid swaps between H3 and CenH3 did not reveal a targeting element because, although they are expected to have similar structures (20, 29, 47), they perform different functions and are therefore evolving under different selective constraints. H3-containing nucleosomes mediate essential processes that impact chromatin, such as transcription and gene silencing (15), whereas CenH3-containing nucleosomes are required for centromere function and maintenance.

* Corresponding author. Mailing address: Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N., Seattle, WA 98109. Phone: (206) 667-4515. Fax: (206) 667-5889. E-mail: steveh@fhcrc.org.
† Present address: LifeSpan BioSciences, Inc., Seattle, WA 98121.

7553
In addition, (H3.H4)$_2$ tetramers are assembled onto the bulk of genomic DNA, whereas (CenH3.H4)$_2$ tetramers only need to be incorporated into a small subset of the genomic DNA sequences found at their respective centromeres. These differences in DNA-binding requirements may be driving the rapid, adaptive evolution of CenH3s (12, 21, 39). In contrast, H3 is highly conserved. Furthermore, assembly of canonical H3 and CenH3s use different processes, with most H3 being deposited in a replication-coupled manner and CenH3 deposition occurring in a replication-independent manner throughout the cell cycle (1, 32). The interpretation of swap experiments between canonical H3 and CenH3 is further complicated by the possible dimerization of chimeric proteins with endogenous partners.

Here we examine CenH3 targeting by making swaps within a single lineage, Drosophila. Given the rapid evolution of Cid (12, 21) we predict that Cid from a diverged Drosophila species will not have the ability to target D. melanogaster centromeres. Such a diverged Cid would be a suitable recipient in swap experiments aimed at mapping targeting determinants. By using this approach, we have delineated a single targeting element, and it corresponds to LI. We show that LI is both necessary and sufficient for centromeric targeting and identify single amino acids within LI that are essential for targeting. (Cid.H4)$_2$ tetramers may be targeted to centromeres by a LI mediated discrimination step during their assembly.

**MATERIALS AND METHODS**

**Constructs.** All coding regions were amplified by PCR and cloned into two separate plasmid vectors for expression from either an hsp70 heat shock (HS) or Cid promoter as C-terminal green fluorescent protein (GFP) fusion proteins with a 6-amino-acid linker (ORFVPAT) between the last amino acid and GFP. Proteins expressed from the vector with a Cid promoter (13) also contained an N-terminal myc tag (MEOKLISEEDLSR) flanked by a ByII restriction enzyme site encoding amino acids RS. The vector with an HS promoter (13) was modified to contain bases TAGAACA immediately before the ATG start site. This allowed higher expression levels from the HS promoter than had been observed before (1, 13). PCR primers contained 5¢/H11032 and 3¢/H11032, and coding sequence, starting from the first codon after the ATG for full-length Cid proteins (22). To delete the N-terminal tail, restriction enzyme sites were 5¢ of R119 of D. melanogaster Cid (CidMel) or the equivalent amino acid of other H3-like proteins. Swap proteins and mutants were constructed by two consecutive PCR steps. First, two products were produced, by using either the 5¢ or 3¢ primer for the entire coding region and either a downstream or an upstream primer spanning the region of the swap junction or mutation. These two PCR products were combined and used as a template for the second PCR step with 5¢ and 3¢ primers for the entire coding region followed by cloning using restriction sites outlined above. All constructs were verified by sequencing.

Cell culture and immunostaining. D. melanogaster Kc167 cells were cultured and transfected with 10 μg of DNA as described previously (13). For inducible constructs, HS was for 1 h at 37°C, 16 to 24 h after transfection. Recovery was for 2 to 6 h as indicated. For the Cid promoter, cells were processed 24 h after transfection. Immunostaining was carried out as described with an anti-Cid antibody and Texas red secondary antibody (13). The anti-Cid polyclonal antibody recognizes an 18-amino-acid peptide present in the N-terminal tail of D. melanogaster, D. simulans, and D. erecta Cid proteins (13). Images were analyzed using Delta Vision software (Applied Precision).

All transfections were done at least three times. Transfection efficiency varied from 50 to 90%. Data from transfections were only used if CidMel-GFP (full length or HFD) showed high transfection levels in a parallel experiment, thus controlling for transfection efficiency of different populations of Kc cells. Constructs that were scored as displaying no targeting had no detectable GFP at the centromere for at least 1,000 cells examined. For constructs scored as targeting centromeres, Cid-GFP expressed from either the Cid or HS promoters could typically be detected at the centromere in 250 to 800 cells out of every 1,000, with intensities up to 100-fold above background for the HS promoter. It should be noted that expression from the Cid promoter is cell cycle specific (13) and at low levels compared to the high expression level from the HS promoter. Although some proteins were targeted when expressed from the HS promoter but not from the Cid promoter, we never observed the converse, i.e., targeting when expressed from the Cid promoter but not from the HS promoter. The reproducible absence of green dots at the centromeres in all cells examined when a fusion protein was expressed from the Cid promoter is likely to reflect reduced targeting efficiency rather than variation in expression levels, because all open reading frames could be produced at high levels from the HS promoter irrespective of targeting, as evidenced by green fluorescence throughout the entire nucleus.

**RESULTS**

When CidMel is transiently expressed as a C-terminal GFP fusion protein in Drosophila cultured cells, it appears as green dots that overlap with endogenous Cid (immunostained in red) at centromeres (13) (Fig. 1, D. melanogaster). Given the rapid, adaptive evolution of Cid (21, 22), we expected that more diverged Cids may have lost the ability to target D. melanogaster centromeres. To find such a Drosophila Cid that does not target D. melanogaster centromeres, we expressed Cid-GFP from D. simulans (CidSim), D. erecta (CidEre), D. lutescens (CidLut), and D. pseudoobscura (CidPse) in D. melanogaster cells. These species are 2.5 to 36 million years diverged from D. melanogaster, and their Cids range from 97 to 58% identical to CidMel at the amino acid level in the HFD (Table 1) (22). CidSim, CidEre, CidLut, and CidPse-GFP fusion proteins targeted D. melanogaster centromeres such that they could be visualized as green dots overlapping with endogenous CidMel (red dots) (Fig. 1). However, one of the more distantly related Cids, CidBip, had lost the ability to target D. melanogaster centromeres, even when it was present at high levels in the nucleus (Fig. 1, D. bipeckata).

We further analyzed the behavior of CidBip-GFP compared to CidMel-GFP in D. melanogaster cells. CidMel-GFP targeting of centromeres was observed in both interphase and mitotic cells after expression at low levels (Fig. 2A and C). At high expression levels, CidMel-GFP localized both to centromeres and euchromatin (Fig. 2B and D). In contrast, CidBip-GFP was never observed at D. melanogaster centromeres, but there were two kinds of interphase patterns. One is a euchromatic pattern (Fig. 2E), where no localization is seen at the heterochromatic chromocenter, which contains the centromeres (1). The other pattern shows CidBip-GFP distributed uniformly throughout the nucleus (Fig. 2F).

Time course analysis revealed that the two different patterns are found in populations of cells at different stages of the cell cycle. GFP fusion proteins were induced for 1 h by an HS pulse. Kc cells show a very regular cell cycle, including a G2 phase of ~5 h (1). Therefore, mitotic chromosomes seen 2 h after HS must have been in G2 phase at the time of the induction, whereas mitotic chromosomes seen 6 h after HS must have been from cells that were induced during S phase. A pulse of CidMel-GFP targeted it to centromeres both in S-phase and G2 cells (Fig. 2C and D). CidBip-GFP was not incorporated into chromatin in G2 phase cells (Fig. 2G) but could be visualized in euchromatin where the protein had been induced during S phase (Fig. 2H). We conclude that CidBip-GFP is competent for assembly into D. melanogaster chromosomes in S phase, but not into centromeres. CidMel must therefore contain a centromeric targeting element that CidBip does not.
The HFD of CENP-A has been shown to target it to human centromeres (33, 37). We found that the HFD of Cid Mel similarly targets it to *D. melanogaster* centromeres (Fig. 3A). Just like the full-length Cid Bip, the HFD of Cid Mel did not localize to *D. melanogaster* centromeres (Fig. 3A). Furthermore, a chimera of the Cid Mel N-terminal tail and the Cid Bip HFD also did not localize to *D. melanogaster* centromeres, and conversely, a chimera of the Cid Bip N-terminal tail and Cid Mel HFD did localize (not shown). Therefore, targeting determinants map to the HFD. The HFD of Cid Bip has 59% amino acid identity to the Cid Mel HFD (Table 1) (22) with differences spread throughout the domain. To select candidate targeting determinants, we compared the HFDs of Cids that are localized to *D. melanogaster* centromeres (Cid Mel and Cid Lut) to the Cid Bip HFD that is not localized (Fig. 3B). We concentrated on amino acids that are unique to the Cid Bip HFD compared to both Cid Mel and Cid Lut.

We substituted the NR, LI, or CR regions of Cid Bip with those of Cid Mel and assayed centromeric targeting in *D. melanogaster* cells. Strikingly, we found that a single segment of Cid Mel, LI, was sufficient to target Cid to *D. melanogaster* centromeres when introduced into Cid Bip (Fig. 3C). Neither the Cid Mel NR nor the CR targeted Cid to centromeres when introduced into the Cid Bip HFD (Fig. 3C). Furthermore, replacing LI of Cid Mel with that of Cid Bip disrupted Cid Mel HFD centromeric targeting (Fig. 3D), whereas the NR or CR regions of Cid Bip HFD had no effect on centromeric targeting of Cid Mel HFD. These data show that LI is both necessary and sufficient for centromeric targeting of Cid.

LI of Cid Mel is 1 amino acid longer than that of Cid Bip and differs from it in 12 of the remaining 14 amino acids (Fig. 3A). To identify critical amino acids for centromere targeting of Cid Mel LI, we performed a mutational analysis. Mutations were introduced into the Cid Bip HFD with LI of Cid Mel (HFD-Bip LI Mel), because this construct localizes to centromeres yet is not expected to dimerize with endogenous Cid Mel. Shorter segments of Cid Mel LI did not target Cid to *D. melanogaster* centromeres, suggesting that multiple contacts within LI are required for targeting (Fig. 4A; M1, M2, and M3 mutants). To
identify key targeting residues, we introduced single-amino-acid scanning mutations to alanine or glycine throughout *D. melanogaster* LI (Fig. 4A, M4 to M18). The results are displayed together with a sequence Logos representation of *Drosophila* LI in Fig. 4B.

Three point mutations (M5, M14, and M17 [Fig. 4A]) each abolished centromeric targeting. Five other point mutations (M4, M6, M8, M13, and M15) resulted in weak targeting, because Cid HFD-GFP could only be detected at the centromere when expressed at high levels from the HS promoter. Interestingly, these detrimental mutations cluster towards either end of LI (Fig. 4B, red lines), whereas amino acids in the center of LI were insensitive to mutation (Fig. 4A, M9, M10, M11, and M12; Fig. 4B, green line). In fact, targeting was unaffected when three central acidic amino acids were mutated to neutral residues (Fig. 4A, M21), or even to basic amino acids (Fig. 4A, M22). An alignment of Cid LI from different *Drosophila* species (22) supports this functional characterization of central, hypervariable amino acids flanked by more conserved regions.

A hypervariable region flanked by more-conserved amino acids may be a hallmark of LI in CenH3 histones in general. For example, CENP-A targeting is unaffected by replacement of the central four hypervariable amino acids (T79, R80, G81, and V82) by only two (KT) (33) (Fig. 5A), whereas deletion of three central hypervariable amino acids of Cse4 (K172, D173, and Q174) similarly does not affect viability in a null background (9, 16) (Fig. 5A). In contrast, changing conserved residues causes severe effects: point mutation of W86 of CENP-A revealed a requirement for an aromatic amino acid for targeting, and mutation of the respective W178 in Cse4 to either Y, D, or E was lethal (16, 33) (Fig. 5A).
We note that position 173 of all Cids contains only threonine or serine (22). Mutation of this threonine of Cid Mel LI to alanine abolished targeting (Fig. 4B, M17), consistent with the possibility that phosphorylation may play a role in LI function. Targeting was also abolished by mutation of this threonine to a glutamic acid (Fig. 4A, M19), which mimics constitutive phosphorylation (Fig. 4A).

All CenH3s have a longer LI region than that of canonical H3 (Fig. 5B). We found that shortening D. melanogaster LI by one amino acid compromised centromeric targeting of HFDBip LIMel (Fig. 4A, M23 and M25). Shortening LI by two or more amino acids abolished all targeting (Fig. 4A, M24 and M26), even when the deletion was within the hypervariable region where mutation of the same amino acids had no effect on targeting (Fig. 4A, compare M26 with M21 and M22). The added length of the LI peptide backbone compared to that of H3 appears to be necessary to ensure targeting, rather than any specific interactions of the acidic side chains of D166, D167, and E168. In contrast, the Cse4 LI was found to tolerate a three-amino-acid deletion (16). In this case, LI is not the primary targeting determinant, but rather Cse4 is targeted by CBF3 binding to its cognate centromeric DNA element.

FIG. 3. LI is necessary and sufficient for targeting of Cid to D. melanogaster centromeres. (A) Cid Mel HFD-GFP localizes to D. melanogaster centromeres, whereas Cid Bip HFD-GFP does not. Representative fields of transfected cells are shown. The secondary structure of the Cid HFD is represented by an orange (Cid Mel) or blue (Cid Bip) line, with α-helices shown as boxes. (B) Alignment of the HFD of Drosophila H3 with the HFD of Cid Mel (M) (orange), Cid Lut (L), and Cid Bip (B) (blue). Hyphens denote gaps in the alignment. The secondary structure of H3 HFD is represented by a line above the sequence, with α-helices shown as boxes. Regions that were swapped between Cid Mel HFD and Cid Bip HFD are underlined in the Cid Bip HFD sequence and marked NR (NR of the HFD), LI, and CR (CR of the HFD). Note that the alignment cannot accurately identify the beginning and the end of LI in the Cid HFD (22). (C) Typical cells transfected with CidMel HFD-GFP with the NR (HFDBip NR Mel), LI (HFDBip LI Mel), or CR (HFDBip CR Mel) segments of Cid Mel substituted for their native counterparts. The secondary structure for each construct is shown schematically underneath the image with segments from D. bipectinata and D. melanogaster colored blue and orange, respectively. (D) Typical cells transfected with Cid Mel HFD-GFP with substitutions of the NR (HFDMel NR Bip), LI (HFDMel LI Bip) or CR (HFDMel CR Bip) regions of Cid Bip. The secondary structure for the HFD of each construct is shown schematically as in panel C.
CDEIII (28), and so LI may be evolving under relaxed constraints. We conclude that the crucial targeting function of CidMel LI is mediated by several amino acids located at both ends of the loop and that a minimum length of LI appears to be required for targeting.

Consistent with our results for the full-length Cid proteins, the LI regions of CidSim and CidLut also targeted Cid to D. melanogaster centromeres when introduced into CidBip HFD (Fig. 4A, M28 and M29). The CidBip HFD with CidPse LI localized to D. melanogaster centromeres when it was expressed at levels high enough to appear in euchromatin too (Fig. 4B). Centromeric targeting correlates with the degree of conservation of the amino acids that are critical for centromeric targeting of CidMel LI (Fig. 4B [number of critical amino acids that are conserved is shown in parentheses]). CidPse LI contains four of the eight critical amino acids and weakly targets Cid to D. melanogaster centromeres. CidBip LI contains only one of the eight amino acids and fails to target Cid to D. melanogaster centromeres. It is interesting that the CidBip N-terminal tail has undergone a nine-amino-acid repeat expansion (22). We have shown that this expansion is significantly similar to DNA minor groove-binding motifs and have suggested that these and other minor groove binding repeats in the Cid N-terminal tail facilitate centromere-specific binding (22). CidBip may rely on its N-terminal tail for targeting to D. bipectinata centromeres. In this way, there would be a relaxed constraint on LI in the D. bipectinata lineage. We conclude that the LI region is a common targeting element of CidMel, CidSim, CidLut, and CidPse for D. melanogaster centromeres, whereas CidBip may have followed a different evolutionary trajectory and lost this function.

**DISCUSSION**

We used Cid from different species of the Drosophila lineage to delineate a centromeric targeting element, LI. It is striking that adaptive evolution of Cid has also occurred in LI (21) and that the amino acids within this region are changing rapidly in arbitrarily defined in histone H3) is indicated (length). Targeting is indicated by a plus sign, and the lack thereof is indicated by a minus sign for both the Cid and HS promoter constructs. (B) Single point mutations to alanine or glycine within D. melanogaster LI that affect targeting are aligned with the logos representation for Drosophila LI (22). The logos representation was generated from an alignment of Cid from a large collection of Drosophila species (22). In this graphical representation, the height of each amino acid indicates the level of conservation across Drosophila species, and amino acids with similar properties are similarly colored. Point mutations that abolish targeting are shown in red, whereas those that allow targeting are shown in green. The lack of targeting at low levels only (when expressed from the Cid promoter) is indicated by lowercase red letters. The red lines indicate more conserved amino acids that are required for targeting, whereas the green line denotes the hypervariable region. The sequence of LI from different species of Drosophila as well as H3 is shown. Hyphens denote gaps in the alignment. Amino acids that are involved in targeting by D. melanogaster LI are shown in red, whereas those that are conserved in other LI regions. The LI length relative to 13 amino acids in H3 is shown (Length), with the number of amino acids conserved with those that are critical for targeting in D. melanogaster (red) in parentheses. Targeting of different LI regions when inserted into the D. bipectinata HFD is indicated by a plus sign (targeting) or minus sign (no targeting) for the Cid and HS promoters.

FIG. 4. Targeting of LI mutants of HFD<sup>Bip</sup> LI<sup>Mel</sup>-GFP. (A) The sequence of LI for each construct is shown, with hyphens indicating gaps in the alignment (22). Amino acids unique to D. melanogaster or D. bipectinata LI are shown in orange and blue, respectively. Equal signs indicate amino acids identical to D. melanogaster LI. For point mutants the amino acid changes in D. melanogaster LI are shown to the right. Numbering is relative to amino acids in D. melanogaster full-length Cid. The number of amino acids in LI (compared to 13 of LI as
FIG. 5. LI regions of CenH3s differ in sequence and length from LI of canonical H3. (A) Alignment of LI and part of the adjacent α1- and α2-helices from a selection of known CenH3 proteins (5, 13, 35, 37, 38, 39). The secondary structure for H3 is shown schematically above the sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study. The length of LI relative to 13 amino acids in H3, is shown next to the sequence. Hyphens denote gaps. Numbers of amino acids are relative to each starting sequence (20). Dashed lines mark the extent of the LI region for CenH3, found to be critical for targeting of Cid in this study.
suggests that LI does not act as a simple recognition site for a centromere-specific Drosophila assembly factor. In addition, such a protein-protein interaction should evolve to an optimal conformation, leading to the conservation of LI rather than the adaptive evolution observed in multiple Drosophila lineages (21, 22).

Instead of targeting the centromere by binding an assembly factor that localizes to centromeres, Cid LI may target centromeres by mediating an interaction between existing centromeric nucleosomes and unassembled Cid tetramers (Fig. 6, step II). Cid LI likely protrudes from both tetramers and nucleosomes, consistent with such an interaction. However, this protein-protein interaction also should evolve to an optimum and so does not explain the adaptive evolution of Drosophila LI (21, 22). Rather than mediating an interaction between an unassembled tetramer and an existing centromeric nucleosome, LI may mediate the correct fitting (32, 36) of new centromeric nucleosomes in between existing ones (Fig. 6, step IV). Such an LI-discriminating assembly mechanism would rely on contacts of LI to both proteins and DNA, consistent with coevolution of LI and centromeric satellites. LI of CidBip may thus prevent targeting of CidMel HFD by posing a structural hindrance to its inclusion within centromeric chromatin.

LI may also specify targeting by preferential assembly of (Cid.H4) tetramers on certain DNA sequences (Fig. 6, step III) (21, 33). Different DNA sequences may compete for binding of (Cid.H4) tetramers during chromatin assembly. Those DNA sequences that are capable of adopting a favorable conformation for interaction could then be strongly preferred during the assembly step. In an equilibrium assembly process, complexes with the lowest free energy will prevail and be available for binding of H2A.H2B dimers and subsequent chromatin maturation. A recent energetic study suggests that the histone-DNA contacts at the central one or two turns of nucleosomal DNA provide a nucleation point for wrapping DNA around the rest of the octamer (4). If assembly indeed starts at the dyad, a noncanonical LI could present a critical energetic barrier 25 bp in either direction from the central contact point. The location of LI at superhelical positions +2.5 and −2.5 of the nucleosome (20) ensures that a DNA molecule will encounter LI at an interval of 50 bp during assembly. Different DNA sequences can have a substantial free-energy difference for canonical nucleosomes (40). In the case of CenH3 nucleosomes, the free-energy differences for different DNAs may be even larger due to constraints imposed by the increased length of LI. This could result in substantial preference for one DNA over another.

Despite differences at the primary sequence level, centromeric satellites contain patterns consistent with a bias in nucleosome assembly: nucleosomal unit repeat lengths are a common occurrence (12), and there is evidence for positioning of CENP-A nucleosomes on α-satellite DNA (41). Although some sequences appear to be preferred at centromeres (30, 44), optimal conformations for assembly of (CenH3.H4) tetramers need not be specified by just one primary sequence (17, 44). Indeed, a periodic secondary structural feature may be required for optimal binding to LI. It is not surprising that such DNA features have escaped detection: predicting positioning strength based on DNA sequence is notoriously difficult even for canonical nucleosomes (40). Neocentromeres may represent DNA sequences with free energy of nucleosome formation higher than those of bona fide centromeres but lower than those of other sequences in the genome. Although neocentromeres do not contain α-satellite DNA, they do show a sequence bias (18, 19), consistent with DNA preference in targeting of (CENP-A.H4) tetramers.

The present study underlines the primary importance of the
Cid HFD and LI for centromeric targeting. We have suggested that the N-terminal tails of Cid may also bind DNA, specifically in the minor groove (22). Such an interaction may result in a unique higher-order structure of centromeric core chromatin (22). Minor groove binding of the Cid N-terminal tail is likely to occur at a step subsequent to wrapping of the DNA around the octamer (Fig. 6) and may contribute to the stability of Cid nucleosomes.

ACKNOWLEDGMENTS

We thank the Drosophila Species Center for the various fly strains used, Harmit Malik and Kami Ahmad for stimulating discussions and comments on the manuscript, and Judy O’Brien for technical assistance. We were supported by the Cancer Research Fund of the Damon Runyon Cancer Research Foundation, DRG-1554 (D.V.), and by the Howard Hughes Medical Institute (H.S.H. and S.H.).

REFERENCES


