Gametic specialization of centromeric histone paralogs in insect species

Lisa Elaine Kursel

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Harmit Malik, Chair

Catherine Peichel

Barbara Wakimoto

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Gametic specialization of centromeric histone paralogs in insect species

Lisa Elaine Kursel

Chair of the Supervisory Committee:
Harmit S. Malik, Member
Basic Sciences Division of the Fred Hutchinson Cancer Research Center
Howard Hughes Medical Institute

The centromere is a specialized chromatin region that is critical for faithful chromosome segregation. Centromere function is conferred epigenetically, by the presence of a histone H3 variant called CenH3 that replaces H3 in nucleosomes at centromeres. Since CenH3 was identified as a centromere-specific histone, countless studies have demonstrated the deleterious consequences of CenH3 perturbation. Paradoxically, numerous studies have also demonstrated that CenH3 evolves rapidly. The centromere drive hypothesis posits that CenH3 evolves rapidly because it is engaged in a co-evolutionary arms race with centromeric DNA. As evidence supporting the molecular mechanisms of centromere drive grows, it seems likely that CenH3's function as a drive suppressor may require rapid evolution. However, rapid evolution may also be at odds with CenH3's essential mitotic role. Therefore, encoding both the mitotic and drive suppressor functions of CenH3 in a single gene creates an intralocus conflict – the evolutionary tension resulting from simultaneous optimization of multiple functions encoded by one gene. CenH3 gene duplication and specialization is one way to resolve this intralocus conflict. Here, I examined the evolution and cytological localization of CenH3 paralogs in *Drosophila* and mosquitoes. Although CenH3 is typically thought of as a single copy gene, I found that most *Drosophila* and mosquito species encode more than one CenH3 paralog.

Furthermore, my analyses suggest that these CenH3 paralogs have acquired specialized germline functions. My characterization of CenH3 paralogs provides an opportunity to dissect how evolution has shaped the fundamental process of chromosome segregation in unanticipated ways.

Table of contents

List of Figures

List of Tables

Acknowledgements

This work would not have been possible without my advisor, Harmit. I cannot express enough gratitude to Harmit for his brilliant ideas, his endless support and for creating a lab environment that fosters scientific creativity and excellence. I am also deeply grateful to Aida de la Cruz whose tireless efforts keep the Malik lab running smoothly. I am indebted to all of my lab mates for the hours they have spent reading manuscripts, giving feedback on talks, planning and troubleshooting experiments and helping me become the scientist I am today. I would also like to thank my dissertation committee, Sue Biggins, Cecelia Moens, Katie Peichel and Barbara Wakimoto, for their guidance and mentorship over years. In particular I would like to thank Katie and Barbara for their feedback on my written dissertation. I must especially thank Barbara for her feedback on Chapter 3 and for helping me learn to present and interpret cytological data in a rigorous and specific way. I am also thankful to my sources of funding including the Genome Sciences Training Grant and the Cellular and Molecular Biology Training Grant.

I would also like to acknowledge my family, especially my mom and dad, Ellen and Peter, and my sister and brother, Jackie and Cameron. I'm not sure that they have ever once doubted my abilities, and for that I am grateful. Special thanks to my Aunt Cynthia and Uncle Jim who provided me with a supportive community from my very first day in Seattle.

Finally, thanks to my friends for helping me find balance, and to Sean, for being my partner in science and beyond.

Chapter 1. Introduction

Adapted from previously published work:

Kursel LE, Malik HS. The cellular mechanisms and consequences of centromere drive. Current opinion in cell biology. 2018 June.

Kursel LE, Malik HS. Centromeres. Current Biology. 2016 June.

1.1 The remarkable diversity of centromeres

Omnis cellula e cellula: all cells come from cells. By the mid-19th century, scientists had rejected the notion that organisms could spontaneously arise from non-living matter, and agreed that cells could divide to form new cells. By the end of the 19th century, the process of mitosis had been described through direct cytological observations and chromosomes ('colored bodies') in the nucleus were identified as the supplier of genetic material[1]. Experiments carried out by Theodor Boveri in sea urchin embryos demonstrated that cells required a full set of chromosomes for development and having too many or too few chromosomes (aneuploidy) caused developmental defects[2].

How do eukaryotic cells orchestrate the exact partitioning of chromosomes to daughter cells? Centromeres are key. Centromeres were first described by Walther Flemming in 1882 as the primary constrictions on chromosomes to which fibers emanating from the spindle made physical connections[1]. By the early 1900s, centromeres acquired a genetic definition: the sites on a chromosome responsible for its inheritance[2].

How are centromeres specified on a chromosome? Louise Clarke and John Carbon were the first to genetically characterize centromeres. They identified centromere-linked genes on both sides of *Saccharomyces cerevisiae* chromosome III, and then 'walked' along the chromosome between the centromere-linked genes, using a technique called overlap-hybridization, to map the entire region. Finally, they were able to identify centromeric

sequences, which were capable of conferring mitotic and meiotic stability to a replicating plasmid upon insertion[3]. They concluded that the centromeric DNA sequences must, therefore, specify the centromeres on budding yeast chromosomes.

A subsequent screen assayed fragmented pieces of the *S. cerevisiae* genome for their ability to confer mitotic stability onto plasmids[4,5]. This screen identified several such sequences, eventually identifying one 125 bp sequence (termed the CEN, for centromeric) per *S. cerevisiae* chromosome (Table 1-1, Figure 1-1). Additional experiments showed that these sequences were both necessary and sufficient to mediate chromosome segregation in *S. cerevisiae*; i.e., it represented a genetically defined 'point' centromere[6]. The CEN sequence was later shown to be capable of mediating assembly of a multi-protein complex termed the kinetochore, which provides the connection between the chromosomal centromere and microtubules to mediate proper chromosome segregation[7,8].

Studies outside of budding yeasts have shown that the short, genetic centromeres of *S. cerevisiae* and its relatives are an exception, not the rule (Table 1-1, Figure 1-1). In other organisms, centromeres are much larger and often made up of repetitive DNA. In the fission yeast *Schizosaccharomyces pombe*, the minimum chromosomal segment that is capable of high-fidelity mitotic and meiotic segregation was found to be 35–50 kilobases long with a 3–5 kb non-repetitive, AT-rich central core flanked by repetitive elements[9,10]. Subsequent studies suggested that only the non-repetitive central core recruits kinetochores while the flanking repeats recruit additional non-centromeric proteins (e.g., heterochromatin and cohesion proteins) that aid in segregation fidelity[11,12] (Figure 1-1, Table 1-1).

In other organisms, centromeric DNA can be entirely comprised of AT-rich repetitive elements called satellite repeats. For instance, human centromeres, ranging in size from several hundred kilobases to several megabases, are made up of arrays of repetitive alpha-satellite DNA[13]. Alpha satellite DNA consists of a 171 bp monomeric unit, which assembles into higher order repeats (HORs). HORs are nearly identical to each other at the center of the satellite

repeat array (likely the result of recombination-driven homogenization) whereas they accumulate both nucleotide changes and transposable element insertions toward the edges[13] (Figure 1-1). Similarly sized repeat units have also been identified in other species, including other mammals and plants. For example, *Arabidopsis thaliana* centromeres have a tandem array of 178 base-pair repeats, and *Oryza sativa* (rice) centromeres contain 155 base-pair *CentO* tandem repeats. Although many centromeric repeat units in plants and animals tend to be ~150 base-pairs, they can also be much shorter; for example, the *Drosophila melanogaster* centromeric region consists of pentameric and other short satellite repetitive DNA units[14] (Figure 1-1).

Not all centromeres in plants and animals contain repetitive satellite arrays. For instance, the centromere of chromosome 8 in rice contains active genes and very few *CentO* repeats[15]. Furthermore, on rare occasion, human centromeres can spontaneously arise *de novo* on chromosomes that have lost their centromere due to a chromosomal rearrangement[16]. These 'neocentromeres' can provide centromere function, i.e., serve as the site of kinetochore assembly and chromosome attachment to the cell division machinery, despite being completely devoid of the centromeric alpha satellite DNA.

Thus, surprisingly, it seems that there may be no rule when it comes to centromeric DNA. The repetitive nature and sheer size of the complex centromeres of humans, plants, and flies have made centromere sequencing and assembly quite challenging, resulting in a literal gap in our knowledge of genome sequences. However, despite these differences, thematic elements of centromeres are nevertheless found in the 'regional' centromeres of humans, plants and flies (as opposed to the 'point centromeres' of budding yeasts): they are large, AT-rich and often contain repetitive DNA. These 'regional centromeres' can be regarded as encoding multiple microtubule-interacting kinetochore units per centromere, each equivalent to the single kinetochore unit assembled on each *S. cerevisiae* CEN sequence.

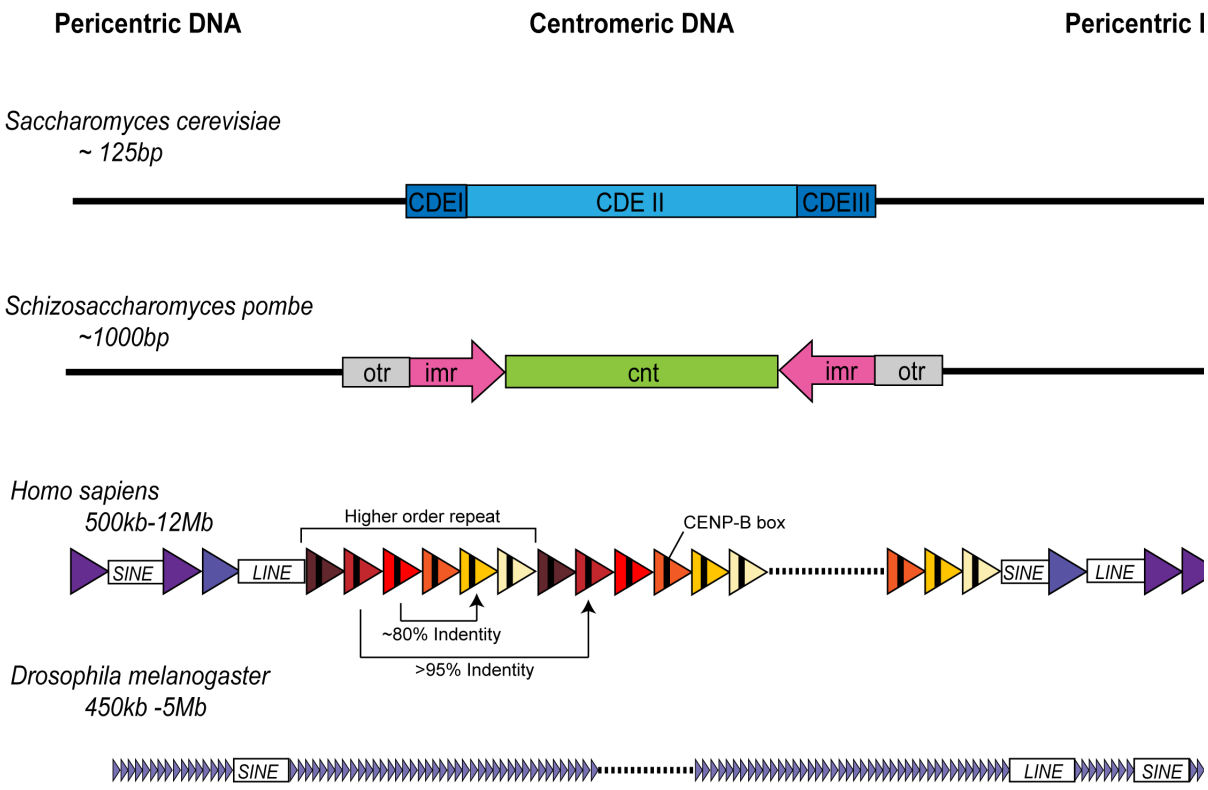


Figure 1-1. Centromere types. Schematic of four eukaryotic centromeres, adapted from Roach *et al.*[17] "Rapid evolution of centromeres and centromeric/kinetochore proteins". *Saccharomyces cerevisiae* has a genetically defined point centromere consisting of three DNA elements, CDEI, CDEII and CDEIII. CDEI is a binding site for the transcription factor Cbfl[18], CDEIII is a binding site for the CBF3 kinetochore complex[7] and CDEII is the site of centromeric nucleosome assembly[19]. *Schizosaccharomyces pombe's* centromere consists of a non-repetitive central core (cnt) that is flanked by innermost repeats (imr) and outer most repeats (otr) which form inverted repeats flanking the central core. *Homo sapiens* have large regional centromeres that consist of tandem arrays of higher order repeats (HORs). Each repeat unit of *H. sapiens* centromeric DNA (indicated by a triangle) is 171bp long and contains a CENP-B box sequence motif. The CENP-B box provides a binding site for the only centromeric protein known to bind a particular DNA sequence, CENP-B. The pericentric DNA of *H. sapiens* chromosomes contains disordered alpha-satellite monomers and transposable element insertions. *Drosophila melanogaster* also has large repetitive centromeres but the individual repeat units are much smaller (5bp) than the 171bp repeat in humans and other organisms. Transposable element insertions are also common in *D. melanogaster* pericentric DNA.

The discovery of CEN sequences in budding yeasts led to the initial idea that all centromeres may be genetically defined by centromeric sequence. The subsequent discoveries of regional centromeres led to the pendulum shifting in the other direction to a model that most active centromeres are defined by the presence of a centromere-specific protein called CenH3[20]. First identified as CENP-A in mammals[21,22], CenH3 is a histone H3-like protein that replaces H3 in nucleosomes at centromeres. CenH3 localizes to functional centromeres (including human and fungal neocentromeres) and is constitutively localized at centromeres throughout the cell cycle. Although sperm-specific proteins (protamines) replace the bulk of canonical histones during animal spermiogenesis, CenH3 persists on sperm centromeres and acts as a heritable centromere marker on paternal chromosomes in both mammals and *Drosophila*[23,24]. CenH3 is such a reliable marker of centromeres that most studies use its presence (by chromatin immunoprecipitation) to identify the centromere, especially in cases where sufficiency tests (by plasmid or minichromosome segregation) for centromere function are impossible. Indeed, CenH3 recruitment to a previously non-centromeric repetitive-DNA array is sufficient to bestow it with centromeric function for several cell divisions[25].

CenH3 is thought to be ubiquitous in eukaryotes and is essential for life in nearly all organisms where it has been identified, including mammals, *Drosophila*, *S. cerevisiae*, plants such as maize and rice, and protists including *Tetrahymena thermophila* and *Giardia lamblia*. CenH3's precise localization to centromeres is maintained by a combination of a CenH3-specific chaperone that deposits CenH3 and ubiquitin ligases that mediate its degradation upon mis-incorporation outside of centromeres[26,27]. CenH3 localization also appears to be self-reinforcing, perhaps suggesting a unique centromeric chromatin environment. In some organisms like *D. melanogaster*, CenH3 recruitment is sufficient to recruit all the proteins necessary to make stable attachments between the DNA and cell division machinery[25]. CenH3 recruitment may also help explain the finding that many centromeric repeat units are ~150 base-pairs long i.e., sufficient to recruit one CenH3-containing nucleosome.

Until recently, CenH3 presence appeared to be the hallmark of functional centromeres in all organisms. However, it is now clear that some organisms completely lack CenH3. For example, *Trypanosoma brucei* (the parasite that causes African sleeping sickness in humans) possesses centromere proteins that share no detectable homology to CenH3 or to any other previously identified centromeric or kinetochore protein[28]. Furthermore, four independent losses of CenH3 occurred in insects, each one coinciding with a transition from monocentric to holocentric chromosomes[29]. It is currently unclear whether CenH3 loss facilitated the transition to holocentricity in these insect lineages, or vice-versa. Although CenH3 is lost in these insects, many other conserved kinetochore proteins have been retained.

Other centromeric proteins also demonstrate evolutionary lability in eukaryotes. In particular, the composition of the inner kinetochore (proteins proximal to centromeric DNA) is much more varied than the outer kinetochore (proteins that interact with microtubules). In fact, very few inner kinetochore proteins are universally conserved in insects. Whereas vertebrate inner kinetochores are comprised of 16 Constitutive Centromere-Associated Network (CCAN) proteins, most of these are missing in insects. Instead, organisms like *Drosophila melanogaster* have a simple repertoire of inner kinetochore proteins including CenH3 (called Cid in *Drosophila*), CENP-C and CAL1, which acts as a CenH3 chaperone in insects.

In sum, there appears to be a remarkable lack of consensus architecture for a chromosomal locus that orchestrates cell division, an essential, conserved process in eukaryotes (Figure 1-2). Nevertheless, the many insights gained from molecular, genetic, biochemical and cytological studies of centromeres in many organisms point to the centromeric histone, CenH3, as the critical marker for functional centromeres in most eukaryotes.

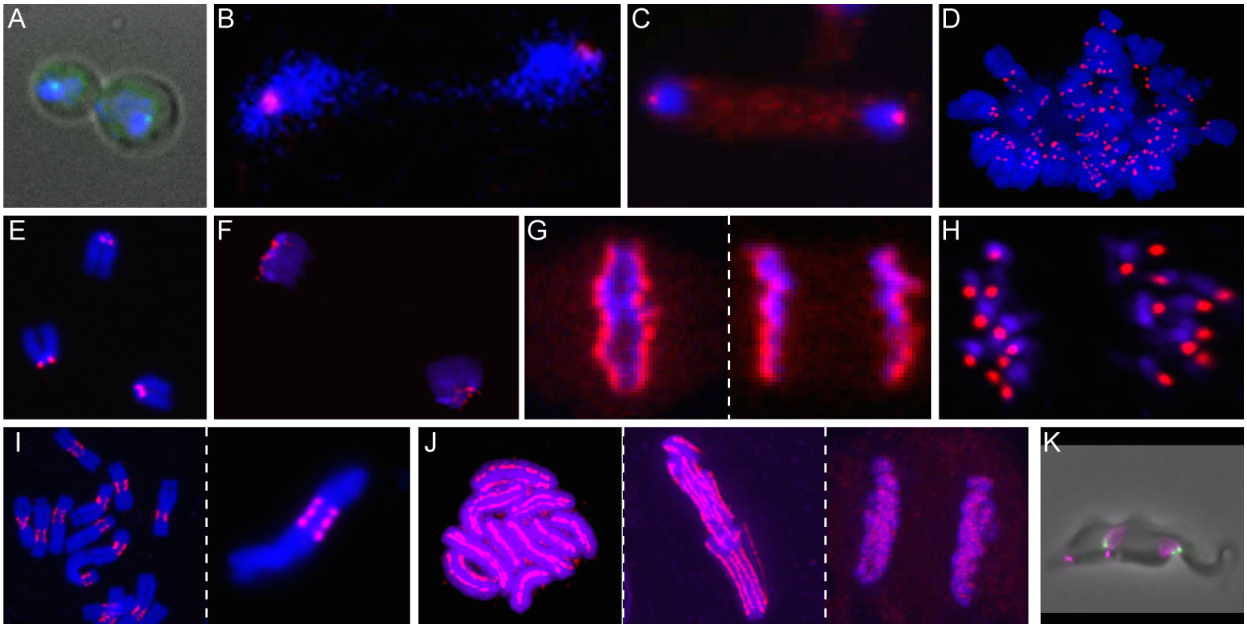


Figure 1-2. The architectural diversity of centromeres. Centromeres from a selection of eukaryotic organisms are represented by a series of immunofluorescence images showing DNA (DAPI, blue) and centromeres marked by centromeric proteins (green or magenta) unless otherwise specified. (A) Clustered point centromeres in anaphase from budding yeast *S. cerevisiae* (Image: Erica Hildebrand/ Sue Biggins lab). Clustered regional centromeres in late anaphase from two fungi: (B) *C. albicans* (Image: Neha Varshney/ Kaustuv Sanyal lab) and (C) *S. pombe* (Image: Sharon White/ Robin Allshire lab). (D) Regional centromeres of *Homo sapiens* metacentric chromosomes during anaphase (Image: Geert Kops lab). (E) Regional centromeres of *M. musculus* acrocentric chromosomes during metaphase (centromeric minor satellite DNA repeats shown in red, Image: Aiko Otsubo/ Michael Lampson lab). (F) Regional centromeres in *D. melanogaster* S2 cells during late anaphase (Image: Leah Rosin/ Barbara Mellone lab). (G) Holocentric chromosomes of the nematode *C. elegans* in early anaphase (left panel) and late anaphase (right panel) of a one-cell embryo (Image: Bram Prevo/ Arshad Desai lab). (H) Regional, metacentric chromosomes of *A. thaliana* during anaphase (Image: Mohan Marimuthu/ Luca Comai, Simon Chan lab). (I) 'Poly-metacentric' metaphase chromosomes of the plant *P. sativus* (left, right panels). (Image: Pavel Neumann). (J) Holocentric chromosomes of the plant *R. pubera* during pre-metaphase (left panel), metaphase (middle panel) and anaphase (right panel) (Image: Andreas Houben lab). (K) Regional centromere-containing chromosomes of the kinetoplastid *T. brucei* cell during anaphase. Centromeric proteins shown in green, DNA in magenta (Image: Bungo Akiyoshi lab).

Table 1-1: Representative centromere configurations in eukaryotes.

Species	CenH3 presence	Centromere Features
Saccharomyces cerevisiae (fungi)	yes	Genetically defined, point centromeres (125 bp)
Candida albicans (fungi)	yes	Regional centromeres with 3 kb AT-rich core sequences unique to each chromosome that lack flanking repeats
Schizosaccharomyces pombe (fungi)	yes	Regional centromeres with 4-7 kb AT-rich core sequences flanked by inner and outer heterochromatic repeats
Homo sapiens (metazoa)	yes	Regional centromeres that are hundreds of kilobases long, consisting of higher order arrays of AT-rich alpha satellite repeats. Results from human artificial chromosome experiments and CENP-A mapping suggest preference for dimeric satellite sequences containing CENP-B DNA binding motifs.
Mus musculus (metazoa)	yes	Regional centromeres that are hundreds of kilobases long; arrays of an AT-rich 'minor satellite' flanked by arrays of 'major satellites'
Drosophila melanogaster (metazoa)	yes	Regional centromeres, with different combinations of simple repetitive satellites (e.g. AATAT) interspersed by transposable elements
Bombyx mori (metazoa)	no	Holocentric chromosomes; unknown centromeric DNA sequences; CenH3 replacement proteins also unknown.
Caenorhabditis elegans (metazoa)	yes	Holocentric (polycentric) chromosomes; CenH3 function appears to be dispensable during meiosis
Arabidopsis thaliana (viridiplantae)	yes	Regional centromeres consisting of hundreds of kilobases of AT-rich 178 bp satellite sequences
Oryza sativa (viridiplantae)	yes	Regional centromeres consisting of hundreds of kilobases of AT-rich 155 bp *CentO* repeats, and CRR retrotransposons; some centromeres (e.g. chr 8) appear to be mostly devoid of repeats
Pisum sativum (viridiplantae)	yes	Poly-metacentric chromosomes, consisting of a few dispersed centromeric domains, each comprised of repetitive satellites and retrotransposons; domains separated by several megabases of DNA but act functionally as monocentric (single primary constriction); speculated to represent transition to holocentricity.
Rhynchospora pubera (viridiplantae)	yes	Holocentric chromosomes, consisting of dispersed centromeric repeat arrays (3-16 kb) interspersed with euchromatin domains
Trypanosoma brucei (kinetoplastidae)	no	Regional centromeres based on cytology and topoisomerase-based mapping; novel kinetochore proteins-no homology to known centromere or kinetochore proteins

1.2 The centromere paradox

Although centromere architecture is remarkably diverse[30], in model systems such as *Drosophila*, mouse and human cells, the centromeric histone, CenH3 (CENP-A in mammals[21], Cid in *Drosophila*[31]) is generally accepted as the universal mark of functional centromeres. Despite being essential for chromosome segregation in most eukaryotes[32-34], *CenH3* evolves rapidly[35,36]. Thus, paradoxically, the key protein responsible for centromere designation in eukaryotes is less conserved than one would expect given its critical role in an essential process. This rapid evolution despite the expectation of constraint is referred to as the 'centromere paradox'[37].

1.3 The centromere drive hypothesis

One explanation for the rapid evolution of essential centromeric proteins like CenH3 is that centromeric proteins are engaged in a genetic conflict[37]. In plants and animals, asymmetric female meiosis provides an opportunity for centromeres to act as selfish genetic elements. The centromere drive model proposes that 'driving' centromeres take advantage of this asymmetry and favor their own inclusion in the oocyte rather than the polar body. As a result, driving centromeres are over-represented in the next generation. Henikoff, Ahmad and Malik proposed that centromere alleles with expanded centromeric satellite DNA repeats could bind more centromeric proteins and provide additional microtubule attachment sites creating a stronger centromere that could (via an unknown mechanism) achieve preferential inclusion in the oocyte (Figure 1-3).

Cheating centromeres are thought to be tolerated in female meiosis. However, in males, differences in heterochromatin content of paired chromosomes are predicted to cause non-disjunction and infertility. Reduced fertility in males would drive the evolution of suppressors of centromere drive. Alleles of genes encoding centromeric proteins with altered DNA-binding are

prime candidates for centromere drive suppressors (Figure 1-3). This model predicts that centromeric proteins must evolve rapidly in order to mitigate fitness costs associated with centromere drive[37].

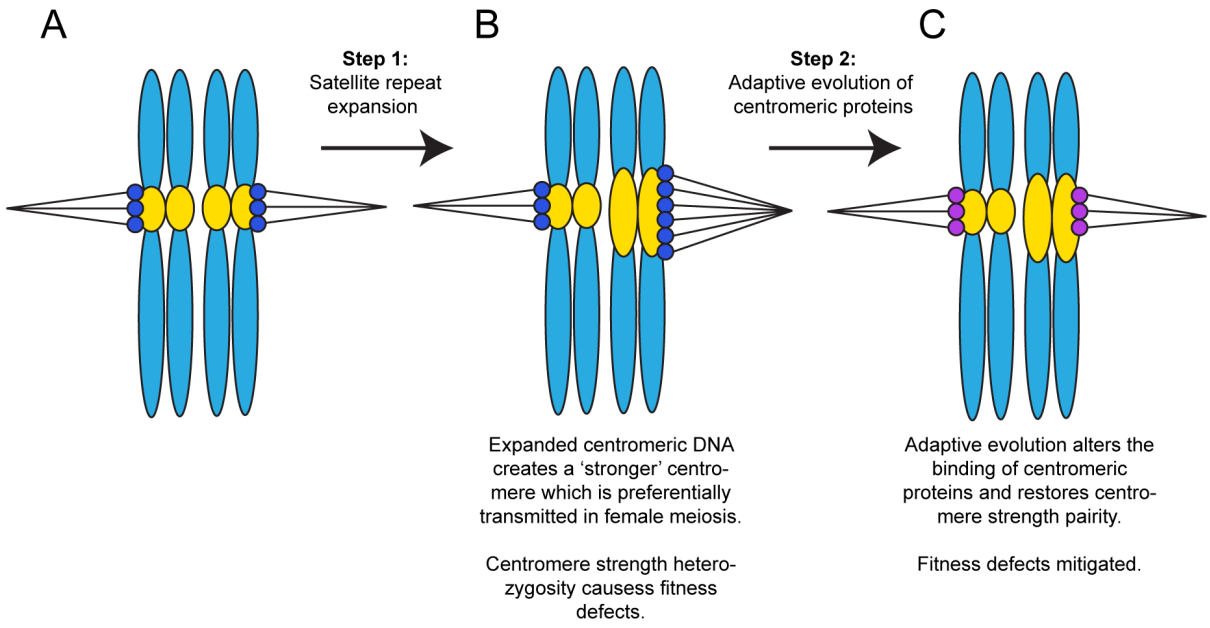


Figure 1-3. The centromere drive hypothesis. Schematic representing paired homologous chromosomes in meiosis I. Centromeric DNA (yellow) is repetitive and prone to repeat copy number variation. If centromeric DNA on one homolog undergoes repeat expansion (A to B), it might be able to recruit more centromeric proteins (dark blue circles) and more microtubules (black lines) than its homologous counterpart (B). The increased recruitment of centromeric proteins creates a 'stronger' centromere that can act as a selfish genetic element and bias its own transmission in female meiosis. However, selfish centromeres cause fitness defects. To mitigate the fitness costs of driving centromeres, centromeric proteins that bind centromeric DNA evolve rapidly to restore parity between homologs in meiosis (C, centromeric proteins are purple, color change represents adaptation).

In its simplest form, there are three primary requirements for centromere drive (Figure 1-4). First, centromere drive relies on asymmetry of the meiosis I spindle. The asymmetric meiosis I spindle enables centromeres to orient themselves relative to each other, creating an opportunity for competition. Second, a preferred centromere position on the asymmetric meiosis I spindle must predictably dictate chromosomal fates *i.e.,* whether chromosomes are retained in the egg or degraded in polar bodies. The third requirement is centromere heterozygosity, in which homologous chromosomes have different propensities to exploit the asymmetry of the female meiotic spindle.

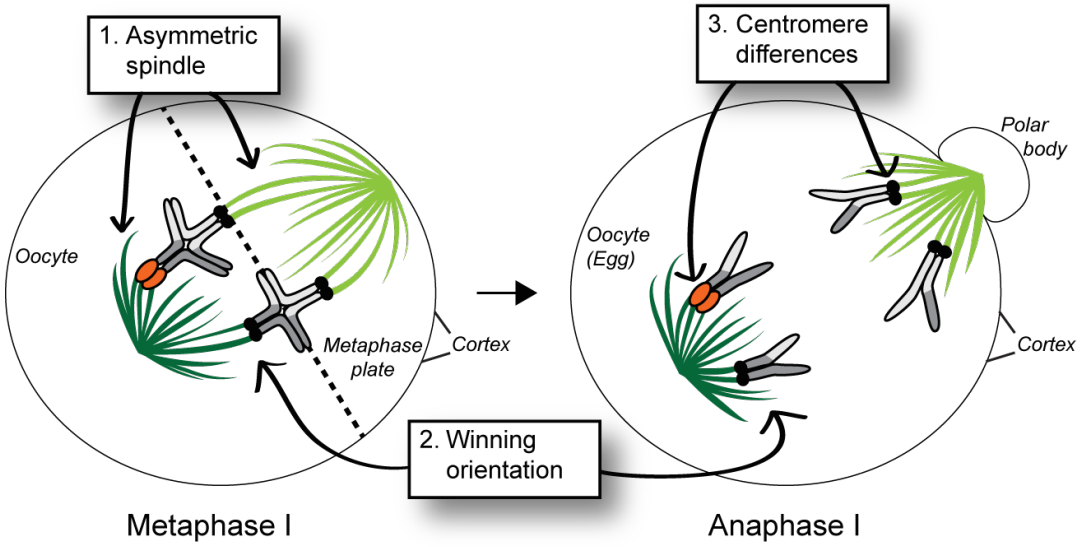


Figure 1-4. The three cellular requirements of centromere drive. Schematic of mouse MI oocyte with two pairs of homologous chromosomes: one pair with unbalanced centromeres (orange and black centromeres) and one pair with balanced centromeres (both black centromeres). In order for centromere drive to occur, three basic requirements must be met. First, the meiotic I spindle must be asymmetric (dark green vs light green microtubules). Second, there must be a preferred (winning) orientation that dictates which chromosomes will segregate to the egg instead of the polar body. In mouse oocytes, chromosomes positioned toward the center of the oocyte nucleus will end up in the egg while chromosomes positioned near the oocyte cortex will segregate to the polar body. Third, there must be centromeric heterozygosity (in the unbalanced pair, one homolog has large, orange, centromeres, while the other homolog has small black centromeres). Chromosomes drawn here are acrocentric (have their centromeres on one end) and are paired with one crossover each in late metaphase I. They segregate to opposite poles in anaphase I. DNA is in shades of grey.

The centromere drive model was proposed nearly 20 years ago, but until recently, there was almost no experimental evidence providing a mechanistic basis for any of the three requirements for centromere drive. There were, however, a handful of studies that linked centromere expansions with meiotic drive. Studies in *Mimulus guttatus* monkeyflower populations found meiotic drive associated with an expansion of the repetitive, pericentric "*D"* locus. In *Drosophila melanogaster*, quantitative measurements of satellite DNA also revealed inheritance patterns consistent with centromere drive[38]. Although these and other systems have contributed much insight into to the evolution and ecology of centromere drive, over the past few years there have been significant gains in our understanding of the mechanistic basis for centromere drive due to cell biological research performed in mouse oocytes.

1.4 Advances in the mechanistic understanding of centromere drive

The original centromere drive model[37] proposed that stronger centromeres (centromeres that are more successful in segregating to the egg during female meiosis) recruit more centromeric proteins, compared to their homologous centromere competitors. This hypothesis was elegantly demonstrated in crosses between mouse strains possessing homologous chromosomes with differential recruitment of centromere proteins[39]. For example, in heterozygous CHPO/CF-1 mice, particular CF-1 chromosomes are preferentially transmitted through female meiosis (meiotic drive). The ability of CF-1 chromosomes to preferentially transmit was directly correlated with the increased amount of inner and outer kinetochore proteins recruited to CF-1 versus CHPO centromeres[39]. Moreover, chromosomes that were subject to meiotic drive were positioned off-center at the meiosis I metaphase plate (Figure 1-4). This study established that increased levels of centromere proteins correlate with increased likelihood of transmission to the egg in female meiosis. These experiments also helped establish a powerful model system that enabled further mechanistic studies of the meiotic

spindle, chromosome behavior on the spindle and the molecular basis for centromere asymmetry.

What is the molecular basis of meiotic spindle asymmetry?

The meiotic spindle in oocytes is different from mitotic spindles in that it lacks microtubule-organizing centers called centrosomes. Instead, meiotic chromosomes organize the microtubules of the MI spindle, which forms first in the center of the oocyte but then moves in an actin-dependent manner perpendicularly towards the oocyte cortex (Figure 1-5). A recent study investigated meiotic spindle asymmetry by examining post-translational modifications of microtubules in mouse oocytes[40]. It found that the meiosis I spindle is preferentially enriched for α-tubulin tyrosination on its cortical side (which would result in polar body inclusion) but depleted on the egg side (Figure 1-5C). α-tubulin tyrosination results in decreased microtubule stability. Thus, the cortical side microtubules are more dynamic than those oriented towards the egg side. Intriguingly, this asymmetry is not evident in early meiosis I when symmetric spindles are formed in the center of mouse oocytes, but is established upon spindle positioning at the cortex (Figure 1-5, compare A and C).

These observations suggested that some cortical signaling likely induces the asymmetry of an otherwise symmetric MI spindle. Further experiments showed that cortical signaling by CDC42, a plasma membrane-associated small GTPase involved in a variety of cell polarization processes, is required for the asymmetric tyrosination[40] (Figure 1-5C). Expression of a dominant-negative or a constitutively active CDC42 decreases or increases α-tubulin tyrosination respectively. The CDC42 signaling is established by a chromatin-based gradient of RAN, a small GTPase with well-established roles in microtubule dynamics. RAN signaling helps to activate CDC42 and polarize the oocyte cortex. Abrogation of this chromosome-directed RAN signaling is sufficient to eliminate the spindle asymmetry and disrupt biased chromosome

orientation required for centromere drive in spite of the proximity of the spindle to the oocyte cortex.

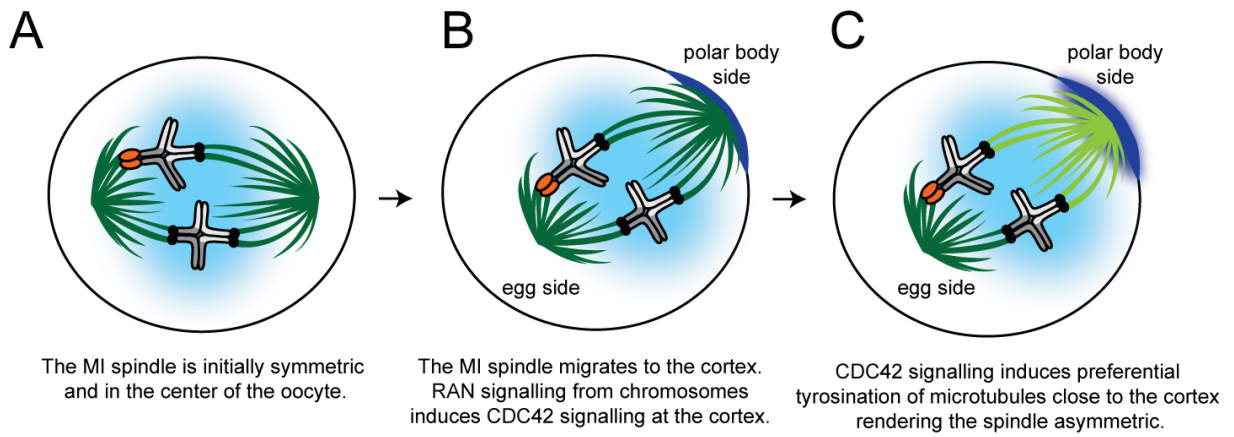


Figure 1-5. Steps to spindle polarization in mouse oocyte meiosis I. Schematic of mouse MI oocyte with two pairs of homologous chromosomes: one pair with centromeres of different 'strength' (orange and black centromeres) and one pair with centromeres of equal 'strength' (both black centromeres). (A) The meiotic I spindle is initially symmetric and is located in the center of the oocyte. (B) The spindle migrates toward the cell cortex where a RAN-GTP gradient (light blue) emanates from the chromosomes and induces CDC42 signaling (dark blue), creating a polarized cortex. (C) CDC42 signaling from the polarized cortex induces tyrosination of α-tubulin (light green microtubules) on the side of the spindle closest to the cortex. The microtubules emanating from the spindle pole oriented toward the center of the oocyte remain primarily de-tyrosinated (dark green microtubules).

How do strong centromeres exploit the asymmetry of the meiotic spindle?

In order to undergo preferential segregation to the egg, chromosomes with stronger centromeres are predicted to preferentially orient on the meiosis spindle by exploiting its inherent asymmetry. Two recent studies showed how strong centromeres achieve this winning orientation. The first of these studies showed that all centromeres (strong or weak) detached from microtubules in proximity to the meiotic spindle poles. This detachment requires the action of Aurora A kinase, which primarily localizes to the spindle poles. A second study showed that

stronger centromeres (those that achieve preferential orientation) are more likely to detach than weaker centromeres, and do so at higher frequency on the cortical ('wrong') side, likely related to the inherently lower stability conferred by α-tubulin tyrosination (Figure 1-6). Although it seems counterintuitive that the 'strong' centromeres are more likely to detach, these properties give 'stronger' centromeres a higher chance of orienting toward the egg pole, thereby achieving preferential transmission into the next generation: centromere drive. This is the key cellular process by which stronger centromeres win.

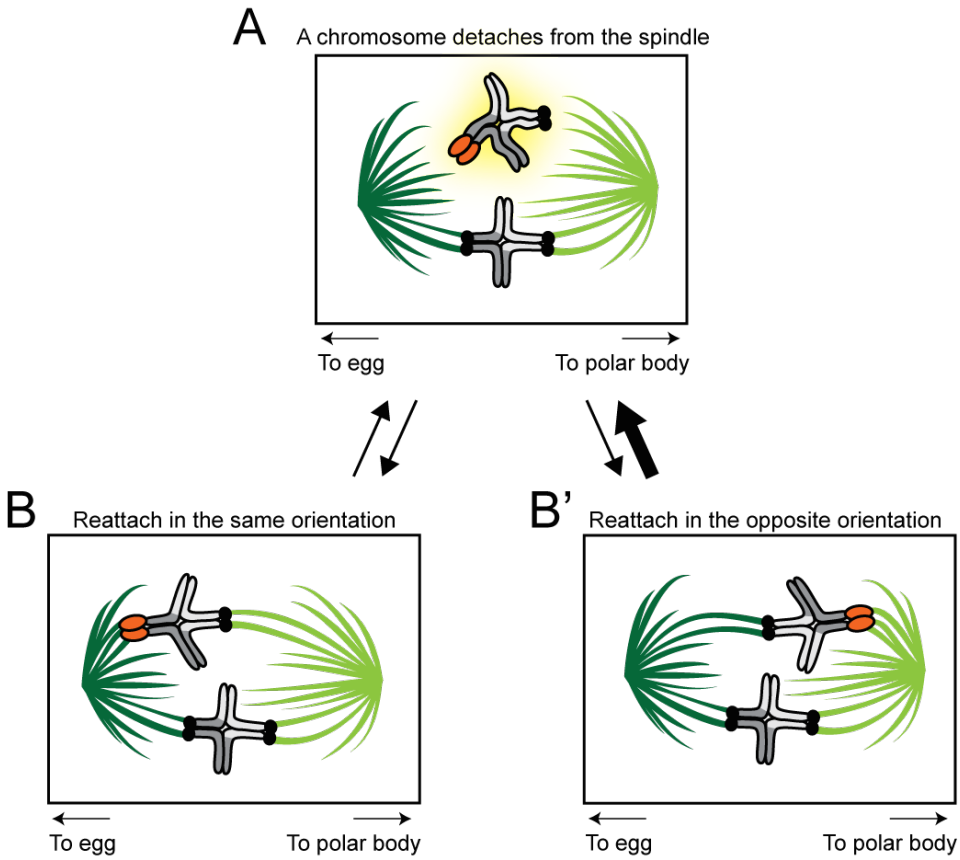


Figure 1-6. How strong centromeres drive. (A) When a chromosome detaches from the meiotic I spindle, it has equal likelihood of reattaching in the same orientation (B) or opposite orientation (B') (equal weight arrows going from (A) to (B) and (A) to (B'). However, a driving centromere is more likely to detach from the cortical spindle (light green microtubules oriented toward polar body, thick arrow from (B) to (A) represents higher likelihood) than from the spindle oriented toward the egg (dark green microtubules, thin arrow from (B') to (A) represents lower likelihood).

Exactly how stronger centromeres preferentially detach remains unknown. It is possible that centromeres recruit factors like Aurora kinases that can destabilize microtubule-kinetochore attachments in a stoichiometric manner, such that stronger centromeres recruit more kinase and their attachments to spindle microtubules are, therefore, more likely to be destabilized. If this were the case, proteins that enhance the recruitment or retention of such factors would be expected to directly enhance success in female meiosis, much like kinetochore protein recruitment is predicted to do under the original centromere drive model.

What is the molecular basis of centromere strength?

In the original centromere drive model, Henikoff, Ahmad and Malik proposed that increased recruitment of centromeric proteins is linked to the quantity of underlying centromeric DNA[37]. A recent study compared stronger and weaker centromeres in mouse strains that have identical sequences of both repetitive centromeric satellite DNA (called the minor satellite) and the centromeric histone CENP-A[41]. Closer investigation revealed that the stronger centromeres possessed 6-10 times more copies of the minor satellite repeats than the weaker centromeres. As predicted in 2001[37], these results establish that more centromeric DNA can recruit more centromeric proteins, which can explain the differences in centromere strength that underlie centromere drive.

1.5 Cellular consequences of centromere drive.

Accelerated evolution of centromeric satellite DNA (in size and sequence) is only the first prediction of the centromere drive model. The second prediction is that centromere drive must have deleterious fitness consequences, either directly as a result of expanded or mismatched centromeric strengths or indirectly due to the hitchhiking of deleterious alleles linked to driving centromeres. Finally, the third prediction of centromere drive is that centromeric proteins must

co-evolve with centromeric DNA in order to suppress centromere drive's deleterious consequences.

Recent studies have attempted to test both of these predictions, starting with the question of whether centromere drive has deleterious consequences. Early work based on human carriers of Robertsonian fusions (in which two acrocentric chromosomes fuse their centromeres to create one metacentric chromosome) suggested that these expanded centromeres cause no somatic (mitotic) defects but result in lower male fertility in the heterozygous state[42,43]. This led to the model that the primary deleterious effects of centromere drive must be in male meiosis or gametogenesis. Supporting this model, taxa lacking male meiosis appear to have not undergone rapid evolution of their centromeric histones suggesting they are not subject to deleterious effects despite apparently undergoing centromere drive[44,45]. For instance, the centromeric satellites of haplodiploid fire ants cover nearly a third of the length of their chromosomes[46]. Since males are haploid and do not undergo meiosis, it has been hypothesized that these animals do not experience the direct deleterious consequences of centromere drive. As a result, there is not strong selection for drive suppressors, and driving centromeres can expand unchecked.

However, other recent findings have challenged the hypothesis that the primary deleterious consequence of centromere drive is in male fertility. Recent work in the *Mimulus* system has shown that driving centromeres cause a myriad of deleterious effects including reduced male and female viability[47] in individuals homozygous for driving centromeres. However, it is possible that the effects of deleterious hitchhiking alleles, which may have accumulated in centromere-linked heterochromatin blocks, have obscured the true deleterious effects of centromere drive in this system. Thus, there is still no robust evidence for the deleterious effects of centromere drive.

Do centromeric proteins such as CenH3 co-evolve with rapidly evolving centromeric DNA? If so, they should be specifically adapted to their own genome. Recently, researchers

addressed this question by investigating whether divergent CenH3 orthologs can complement loss of the endogenous CenH3 allele in *Arabidopsis thaliana*[48]. They found that untagged CenH3 orthologs from *Lepidium oleraceum* and *Zea mays* are surprisingly capable of supporting both mitotic and meiotic function in *A. thaliana*. Heterologous CenH3 bearing plants are fully fertile and yield viable seeds when selfed, comparable to wild-type crosses (Figure 1-7, A and B). This would suggest that *A. thaliana* CenH3 has not specifically adapted to *A. thaliana* centromeres, even though *A. thaliana* CenH3 has been shown to evolve rapidly[35]

How do we reconcile the full functionality of CenH3 orthologs, with the evolutionary signature of their rapid evolution? A partial answer is revealed in crosses using pollen from *A. thaliana* plants encoding *A. thaliana* CenH3 and ovules from *A. thaliana* plants encoding a divergent *L. oleraceum* CenH3 ortholog (Figure 1-7C). In this cross, although fertilization proceeds normally, the maternal chromosomes, whose centromeres are packaged in *L. oleraceum* CenH3, undergo dramatic chromosome segregation defects[48]. In contrast, the paternal chromosomes, whose centromeres are packaged in *A. thaliana* CenH3, undergo proper chromosome segregation. As a result, many progeny plants are aneuploid or haploid, containing solely paternal chromosomes.

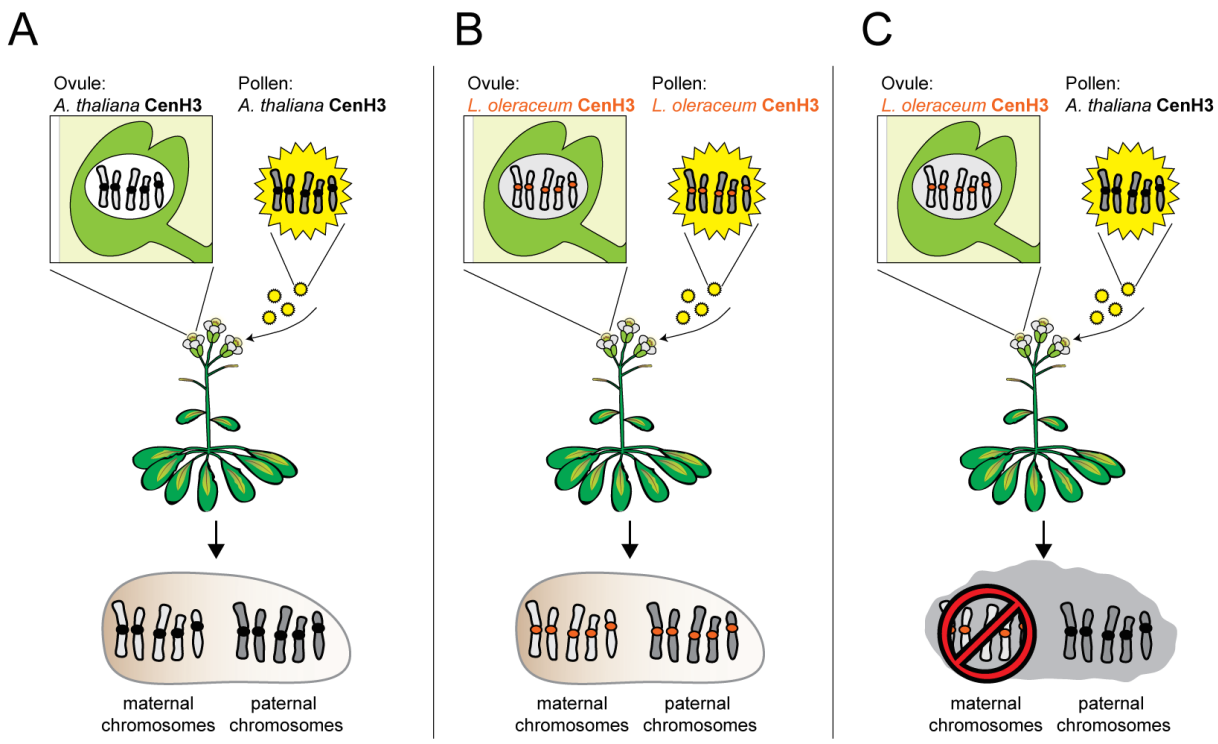


Figure 1-7. CenH3 incompatibility induces aneuploidy in *Arabidopsis*. A *CenH3* null mutant *Arabidopsis thaliana* plant can be fully rescued by an *A. thaliana CenH3* transgene (black centromeres). (A) A self-cross between pollen and ovules from an *A. thaliana CenH3* transgene plant results in healthy seeds that develop into phenotypically wild-type, fertile plants. (B) A *CenH3* null mutant *A. thaliana* plant can also be rescued by an orthologous *CenH3* transgene from *L. oleraceum* (orange centromeres). A self-cross between pollen and ovules from an *L. oleraceum CenH3* transgene plant results in healthy seeds that develop into phenotypically wild-type, fertile plants. (C) However, when pollen from an *A. thaliana CenH3* transgenic plant is crossed to ovules from a *L. oleraceum CenH3* transgenic plant, the resulting progeny have high rates of aneuploidy which is entirely attributable to defects in the maternal, *L. oleraceum*-CenH3-packaged, genome. Note: all plants have an *A. thaliana* genetic background.

What is the nature of the developmental defect induced when genomes packaged by heterologous CenH3s are outcrossed to wild-type *A. thaliana?* One possibility is that *L. oleraceum* CenH3 recognizes a different set of satellites than *A. thaliana* CenH3. However, ChIP-seq experiments with *A. thaliana* and *L. oleraceum* CenH3 proteins revealed no significant differences in satellite sequence binding in *A. thaliana* genomes[49]. Although these experiments cannot rule out the possibility that orthologous CenH3 proteins have lower stability than *A. thaliana* CenH3 despite correct localization, they do weaken the possibility that DNA sequence-binding preferences can fully explain functional differences between CenH3 orthologs[49,50].

In Drosophila species, rapid, divergent evolution of centromeric proteins can also lead to incompatibilities in protein-protein interactions. For instance, the *D. melanogaster* CenH3 chaperone CAL1 is incompatible with the *D. bipectinata* CenH3 protein, leading to the mislocalization of exogenous *D. bipectinata* CenH3 in *D. melanogaster* tissue culture cells[51]. Co-expression of the *D. bipectinata* CAL1 protein can alleviate this mislocalization. However, the finding that plant CenH3 orthologs are intrinsically capable of performing meiotic and mitotic functions in *A. thaliana* weakens the likelihood that protein-protein incompatibilities can explain the *Arabidopsis* aneuploidy-induction phenotype[48]. Thus, although these studies do find evidence in support of the coevolution of centromeric proteins, they leave the mechanistic basis of centromeric incompatibilities unresolved[50].

1.6 Specialized CenH3 paralogs might resolve intralocus conflict

Centromere drive might explain why centromeric proteins, such as CenH3, evolve rapidly. However, it leaves unresolved the tension between the requirement of functional constraint (CenH3 mutation is highly deleterious[32-34,52]) and the observation of rapid evolution[36,53,54]. Essentially, CenH3 performs multiple functions that might have different functional optima depending on the cellular context. In mitotic cells, CenH3 faithfully provides the site of kinetochore assembly. In the male germline, CenH3 might act as a drive suppressor. This type of evolutionary scenario involving functions with divergent fitness optima is called intralocus conflict – the evolutionary tension created by simultaneous optimization of multiple functions in one gene. Intralocus conflicts occur most often in genes that have different functions in males and females or different functions in different tissues[55,56]. The result is less than ideal; selection cannot achieve optimal function for either role.

Intralocus conflict can be resolved via gene duplication and specialization. Duplication and specialization of mitochondrial genes in the *Drosophila* male germline provides one example of the resolution of intralocus conflict. In the testis, the strongest selective force shaping mitochondrial function might be increased production of fast swimming sperm. However, mutations that allow for increased mitochondrial energy production and faster swimming sperm might also increase the mitochondrial mutation rate due to increased production of reactive oxygen species. Since sperm mitochondria are not transmitted to offspring, the high mutation rate is not selected against in the male germline. However, a high mutation rate would be problematic in the female germline and in somatic tissues. Gallach *et al.* proposed that testis-specific paralogs of mitochondrial genes might allow organisms to achieve optimal mitochondrial function simultaneously in somatic tissues and testes[55]. More broadly, Gallach and Betran proposed that resolution of intralocus conflict might explain the high rate of retention of duplicate genes[57]

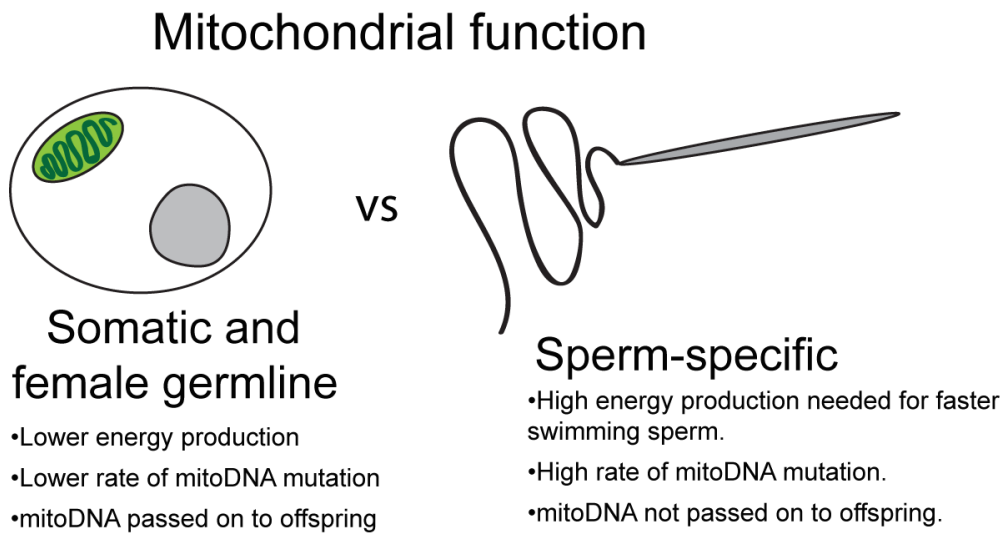


Figure 1-8. Intralocus conflict – the mitochondrial gene example. Somatic and female germline mitochondrial function might face different selective pressures than male germline mitochondrial function. In the male germline, selection will favor mutations in nuclear-encoded mitochondrial genes that allows for increased energy production, which is necessary for sperm swimming and nuclear remodeling. However, this higher rate of energy production causes an increased rate of mitochondrial DNA mutation due to higher production of reactive oxygen species. A higher mitochondrial DNA mutation rate is tolerated in sperm because sperm mitochondria are not passed on to the next generation. In the female germline and in somatic cells, selection will favor nuclear-encoded mitochondrial gene variants that result in lower rates of mitochondrial DNA mutation and therefore have lower rates of energy production, insufficient for sperm swimming. The evolutionary tension caused by simultaneous optimization of both functions in one gene is called intralocus conflict.

Does CenH3 have different functions in males and females or in different tissues or cell types? Several lines of evidence suggest that it might. First, as described above, centromere drive suppression in the male germline might require that CenH3 evolves rapidly whereas mitotic CenH3 function may benefit from evolutionary constraint. In *Drosophila,* differences in the timing of CenH3 chromatin assembly hint at different functions for CenH3 in somatic versus germline tissues. In mitotic cells, CenH3 is loaded after mitosis in G1[58]. However, in the *Drosophila* male germline, CenH3 is loaded in two phases: once during prophase I and again after exit from meiosis[59]. CenH3 might also have specialized requirements in the male versus female germline in the production of male and female gametes. During the production of male

gametes, the sperm nucleus undergoes a dramatic transition from histone-based chromatin to chromatin that is packaged by protamines. In short, nearly all of the histones are removed and are replaced by highly basic proteins called protamines[60-62]. Even though CenH3 is a histone protein, it is not removed from sperm chromatin during this process. Studies in mammals find the presence of CenH3 in mature sperm[24] and studies in *Drosophila melanogaster* found that loss of paternal CenH3 on sperm chromatin results in early embryonic lethality[23]. Therefore, the retention of CenH3 in the highly compact chromatin environment of male gametes may favor a different amino acid composition than CenH3 in cycling mitotic or meiotic cells. The female germline poses another specific challenge for the centromeric histone. In female meiosis in humans and mice, oocyte nuclei arrest in prophase I for extended periods of time (years in humans, months in mice)[63,64]. In mouse, oocyte centromere function does not seem to depend on the loading of new CenH3 because mice with a conditional knockout of CenH3 in prophase I are fully fertile[63]. This means that the individual CenH3 molecules are capable of stably persisting in a mouse oocyte for a year. Thus, the male and female germline pose specific challenges for centromere inheritance through the gametes that may put different selective pressures on CenH3 in males and females.

CenH3 might be subject to intralocus conflict if a single copy of CenH3 is incapable of optimally performing mitotic, meiotic and germ cell function. Like the mitochondrial gene example, gene duplication and specialization could allow CenH3s paralog to achieve optimality for different functions. The potential for functional interrogation of intralocus conflict within *CenH3* makes the identification and study of *CenH3* duplications intriguing.

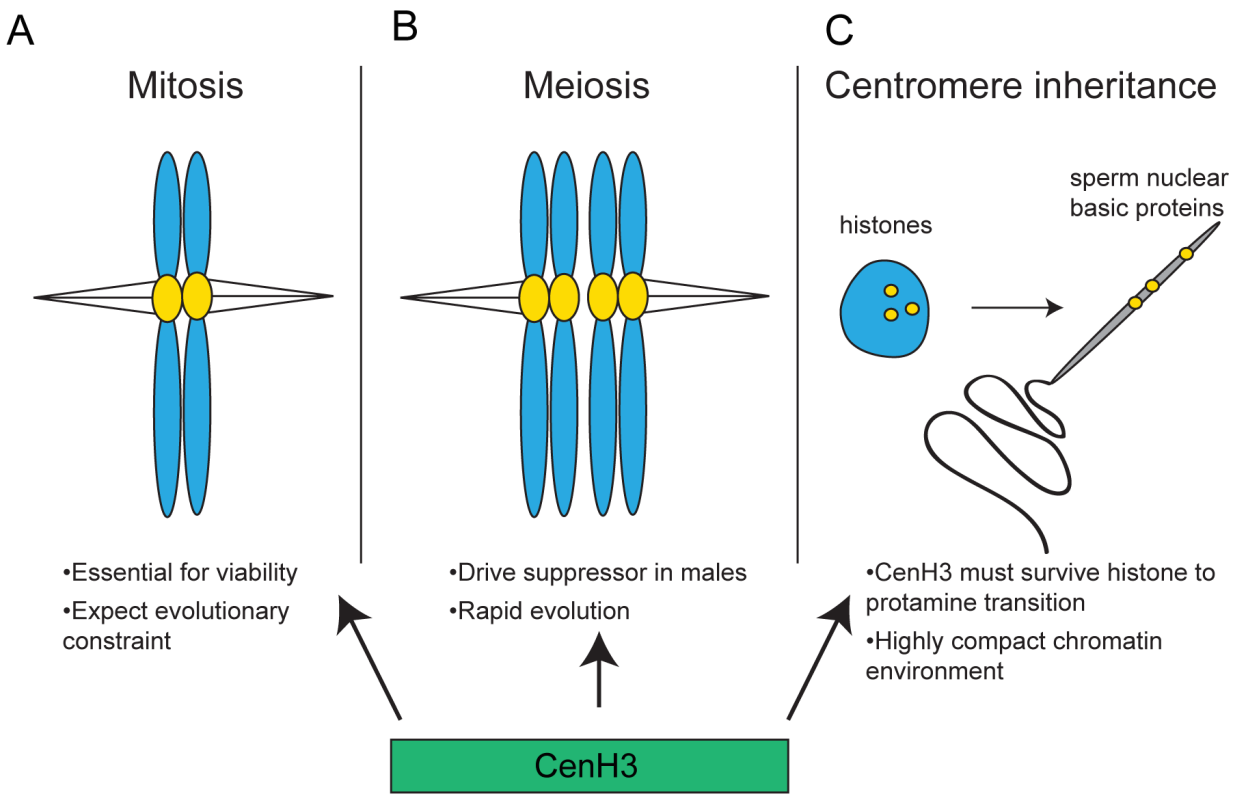


Figure 1-9. The multiple functions of CenH3. Schematic demonstrating the different functional requirements of CenH3 in different cellular contexts. (A) CenH3 is essential for mitotic cell divisions and therefore essential for viability. Given CenH3's essentiality, we would expect that it evolves under evolutionary constraint. (B) CenH3 may function as a suppressor of centromere drive in male meiosis. CenH3's drive-suppressor function may require rapid evolution. (C) CenH3 is essential for centromere inheritance as the trans-generational epigenetic mark of centromeres in sperm. Nearly all histones are removed during sperm development and are replaced by protamines, allowing the sperm nucleus to become highly compact. Because CenH3 is essential for centromere inheritance, it must not be removed during the histone to protamine transition.

1.7 Previously described centromeric histone duplications

Previous evolutionary studies of *CenH3* have suggested that *CenH3* is typically a single copy gene. This conclusion primarily comes from studies in polyploid plant species that found that one copy of *CenH3* is typically lost after whole genome duplication events[65,66]. Furthermore, maintaining two copies of *CenH3* might be risky because of the possibility of acquiring a deleterious or dominant negative mutation in an essential gene. Therefore, most duplicate *CenH3* genes were thought to quickly become pseudogenized.

More recently, there is growing evidence that retention of *CenH3* gene duplicates is somewhat common, especially in plants. Wheat, barley, monkey flowers, peas, *Arabidopsis* and *Luzula* all have independent duplications of *CenH3*[66-74]. Moreover, the wheat and barley *CenH3* paralogs show signs of functional specialization. In wheat, *αCenH3* and *βCenH3* have different expression patterns. *αCenH3* is expressed at all stages of the mitotic cell cycle but *βCenH3* is only expressed during interphase in somatic cells. However, *βCenH3* is expressed during all stages of meiosis. *αCenH3* and *βCenH3* also have different knockdown phenotypes. Knockdown of *αCenH3* impacts plant growth and development whereas knockdown of *βCenH3* impacts reproductive fitness[67]. In barley, one of the two *CenH3* paralogs is primarily expressed in embryonic and reproductive tissues while the other paralog is widely expressed[68]. In both wheat and barley, the reproductive and somatic functions of CenH3 may be encoded by different *CenH3* paralogs.

In peas, monkey flowers and *Arabidopsis*, retention of *CenH3* duplicates seems to coincide with major changes in centromere architecture. Pea chromosomes have acquired a polymetacentric centromere configuration in which chromosomes have an elongated primary constriction with multiple discrete sites of kinetochore assembly. Even though the two pea *CenH3* paralogs evolve under different evolutionary constraints[71], they completely co-localize at all centromeres[70]. Therefore, the retention of pea *CenH3* paralogs is likely due to the

requirement for increased dosage of *CenH3* to accommodate expanded polymetacentric chromosome configuration. In monkey flowers, *CenH3* gene duplication coincides with widespread chromosomal fission events resulting in a dramatic increase in chromosome number. While it is possible that monkey flower *CenH3* paralogs have also been maintained for 'dosage' purposes, the detection of different selective forces acting on *CenH3* paralogs and the fact that centromere-associated meiotic drive has been observed in monkey flower species led Finseth *et al.* to hypothesize that, in this instance, *CenH3* paralogs were retained for suppression of centromere drive[75]. Finally, in *Arabidopsis*, there is no evidence that *CenH3* paralogs have acquired tissue specific functions, but *Arabidopsis* centromeres have undergone recent changes in satellite sequence content. Therefore, *CenH3* duplicates may be retained to bind different satellite sequences[72]. Taken together, these studies indicate that *CenH3* duplication and retention events in plants are actually somewhat common. Some studies estimate that ~10% of diploid plant species contain two *CenH3* genes[48].

Unlike in plants, only three cases of *CenH3* duplications have been previously described in animals. These duplication events have occured in cows[76] and in two species of *Caenorhabditis, C. elegans* and *C. remanei*[77,78]. Recent *CenH3* gene family expansion gave rise to the ten new copies of *CenH3* in cows[76]. However, most of the cow paralogs have pseudogenized and only two are expressed. There has been no functional characterization of either expressed paralog. Interestingly, the *C. elegans* and *C. remanei* paralogs arose from independent duplication events[77,78] but functional studies of the *C. elegans CenH3* duplicate have so far indicated that the paralog is not expressed and has no clear function[78]. Therefore, even though plant *CenH3* duplications are now considered to be somewhat common, the majority of animal species examined so far have a single *CenH3* gene.

1.8 Layout of dissertation

In my dissertation, I comprehensively studied the evolution and localization of duplicate *Cid* genes in *Drosophila.* I hypothesize that duplicate *Drosophila Cid* genes have acquired specialized germline functions.

In Chapter 2, I used an in-depth phylogenetic approach to characterize *Cid* gene duplication events in *Drosophila*. Surprisingly, based on my analysis of over 100 Cid paralogs and phylogenetic inference, my results suggest that most *Drosophila* species likely encode two or more *Cid* genes. Some *Cid* paralogs have testis-restricted expression patterns and encode distinct N-terminal tails that may represent specialize protein-protein interaction sites. This suggests that *Drosophila Cid* paralogs may have acquired specialized function following duplication.

In Chapter 3, I took a cytological approach to determine the localization of the two Cid paralogs, Cid1 and Cid5, in *D. virilis.* I found that Cid1 and Cid5 have distinct germline localization patterns and are alternately retained in gametes. Cid1 is the only paralog detectable in somatic cells and the mature oocyte nucleus. In contrast, Cid5 is only found in the male germline and is the only detectable Cid paralog in mature sperm. I hypothesize that the distinct localization patterns of Cid1 and Cid5 may reflect specialized functions. Moreover, I propose that the functional specialization of Cid1 in females and of Cid5 in males could be indicative of sexually antagonistic functions – optimal female and male Cid function may be incompatible.

In Chapter 4, I extend my evolutionary analysis of duplicate *Cid* genes to mosquitoes in order to test whether similar evolutionary forces may have driven retention of *Cid* paralogs in other Dipteran species. I found that, like in *Drosophila*, *Cid* duplications are also common and long-lived in mosquitoes. However, whereas some *Drosophila Cid* paralogs have testis-biased expression patterns, some mosquito *Cid* paralogs show ovary-biased and early embryo –

biased expression patterns. Interestingly, the centromeric histone chaperone, *CAL1* has also duplicated in mosquitoes.

In Chapter 5, I discuss future prospects and avenues of research for the centromeric histone duplication field. I propose that the rapid coevolution between centromeric DNA and the centromeric histone selects for the retention of specialized *Cid* genes. This is the first suggestion that the centromeric histone is subject to intralocus conflict and that intralocus conflict prevents the simultaneous optimization of multiple centromeric histone functions. I propose that gene duplication and specialization resolves intralocus conflict, adding a new insight into the consequences of rapidly evolving centromeric proteins.

Chapter 2. Recurrent gene duplication leads to diverse repertoires of centromeric histones in *Drosophila* species

Adapted from previously published work:

Kursel LE, Malik HS. Recurrent gene duplication leads to diverse repertoires of centromeric histones in *Drosophila* species. Molecular biology and evolution. 2017 June.

2.1 Abstract

Despite their essential role in the process of chromosome segregation in most eukaryotes, centromeric histones show remarkable evolutionary lability. Not only have they been lost in multiple insect lineages, but they have also undergone gene duplication in multiple plant lineages. Based on detailed study of a handful of model organisms including *Drosophila melanogaster*, centromeric histone duplication is considered to be rare in animals. Using a detailed phylogenomic study, we find that *Cid*, the centromeric histone gene, has undergone at least four independent gene duplications during *Drosophila* evolution. We find duplicate *Cid* genes in *D. eugracilis* (*Cid2*), in the *montium* species subgroup (*Cid3, Cid4*) and in the *Drosophila* subgenus (*Cid5)*. We show that Cid3, Cid4, and Cid5 all localize to centromeres in their respective species. Some *Cid* duplicates are primarily expressed in the male germline. With rare exceptions, *Cid* duplicates have been retained after birth, suggesting that they perform non-redundant centromeric functions, independent from the ancestral *Cid*. Indeed, each duplicate encodes a distinct N-terminal tail, which may provide the basis for distinct protein-protein interactions. Finally, we show some *Cid* duplicates evolve under positive selection whereas others do not. Taken together, our results support the hypothesis that *Drosophila* Cid duplicates have acquired specialized functions. Thus, these gene duplications provide an unprecedented opportunity to dissect the multiple roles of centromeric histones.

2.2 Introduction

Centromeres are the chromosomal regions that link DNA to the spindle during cell division, thus ensuring faithful segregation of genetic material. Proper centromere function is critical for eukaryotic life. Centromeric defects can result in aneuploidy and cycles of chromosome breakage[79,80] with catastrophic consequences for genome stability and fertility. Despite the fact that centromeres are essential for life, centromere architecture is remarkably diverse[30]. Centromeric DNA sequences[13,81,82] and centromeric proteins[36,54,83] also evolve rapidly in diverse organisms. This diversity and rapid evolution make it nearly impossible to name a single defining feature of all centromeres. However, the hallmark of many centromeres is the presence of a specialized centromeric H3 variant called CenH3 (CENP-A in mammals[21], Cid in *Drosophila*[31]). Despite being essential for chromosome segregation in most eukaryotes[32-34], *CenH3* evolves rapidly[35,36] Thus, paradoxically, proteins and DNA that mediate chromosome segregation in eukaryotes are less conserved than one would expect given their participation in an essential process. This rapid evolution despite the expectation of constraint is referred to as the 'centromere paradox'[37].

Genetic conflicts provide one potential explanation for the rapid evolution of centromeric DNA and proteins. In both animals and plants, the asymmetry of female meiosis provides an opportunity for centromere alleles to act selfishly to favor their own inclusion in the oocyte and subsequent passage into offspring rather than the polar body. In female meiosis, centromeric expansions[84] and differential recruitment of centromeric proteins resulting in centromere strength variation between homologs[39] may provide the molecular basis of segregation distortion. In males, however, expanded centromeres and centromere strength variation are thought to result in reduced fertility[42,84]. This lower fertility is predicted to drive the evolution of genetic suppressors of centromere drive, including alleles of centromeric proteins with altered

DNA-binding affinity. Under this model, centromeric proteins evolve rapidly in order to mitigate fitness costs associated with centromere drive[37].

Centromere drive and its suppression provide an explanation for the rapid evolution of both centromeric DNA and centromeric proteins. However, it invokes the relentless, rapid evolution of essential proteins such as CenH3, whose mutation could be highly deleterious[32-34,52]. A simpler way to allow for the rapid evolution of centromeric proteins without compromising their essential function would be via gene duplication. Duplication and specialization of centromeric proteins would allow one paralog to function as a drive suppressor in the male germline, while allowing the other to carry out its canonical centromeric role. Gene duplication as a way of separating functions with divergent fitness optimums has been previously invoked to explain the high frequency of duplicate gene retention, including retention of testis-expressed gene duplicates that carry out mitochondrial functions[57] (see Chapter 1, Figure 1-8). Gene duplications allow organisms to achieve optimal mitochondrial function simultaneously in somatic tissues and testes. By the same reasoning, if a single-copy gene is incapable of achieving the multiple fitness optima that are required for multiple centromeric functions (*e.g.*, mitosis versus meiosis), gene duplication could allow each duplicate to achieve optimality for different functions, thereby resolving intralocus conflict[57] (Figure 1-9).

At least five independent gene duplications of *CenH3* have been described in plants[66,68,70-73,75]. In most cases, both protein variants are widely expressed and co-localize at centromeres during cell divisions[70,71]. However, in barley, one *CenH3* paralog is widely expressed while the other is only expressed in embryonic and reproductive tissues[68]. In cases that have been examined closely, *CenH3* duplicates are subject to divergent selective pressures (*i.e.* one paralog evolves under positive selection but the other does not)[71,75]. Indeed, *CenH3* duplications in *Mimulus guttatus* have been hypothesized to result from centromere drive suppression[75].

In animals, *CenH3* is thought to have independently duplicated in the holocentric nematodes *Caenorhabditis elegans* and *C. remanei*[77,78]. Detailed studies have only been performed on the *CenH3* duplicate in *C. elegans,* and these have yet to elucidate a clear function[78]. *CenH3* duplications have also been described in Bovidae (including cows) where recent gene family expansion has resulted in ten copies of *CenH3*[76]*.* However, only two of the ten cow *CenH3* duplicates have retained open reading frames and all cow *CenH3* duplicates remain poorly characterized[76]. Furthermore, many systems in which *CenH3* has been extensively studied (predominant mammalian systems, such as mice and humans, and model organisms like *D. melanogaster)* have only one copy of *CenH3*.

To comprehensively study the incidence of *CenH3* duplication in a well-studied animal lineage, we took advantage of the recent sequencing of high-quality genomes from multiple *Drosophila* species. These genomes are at a close enough evolutionary distance to allow inferences of gains, losses and selective constraints. Despite there being only one copy of *CenH3* in *D. melanogaster,* we were surprised to find that some *Drosophila* species had two or more copies of *CenH3.* This motivated our broader analysis of *CenH3* duplication and evolution throughout *Drosophila.* In total, we find at least four independent *Cid* duplications over *Drosophila* evolution. Cytological analyses confirm that these *Cid* duplicates encode *bona fide* centromeric proteins, two of which are expressed primarily in the male germline. Based on their retention without loss over long periods of *Drosophila* evolution, and analysis of their selective constraints, we infer that these duplicates now perform non-redundant centromeric roles, possibly as a result of specialization. Overall, this suggests that *Drosophila* species encoding a single *CenH3* gene may be in the minority.

2.3 Results

Four Cid duplications in the Drosophila genus: ancient retention and recent recombination

Although their N-terminal tails are highly divergent, CenH3 histone fold domains (HFD, ~100 aa) are highly conserved and recognizably related to canonical H3[22,85]. Thus, sequence similarity searches based on either CenH3 or even canonical H3 HFDs are sufficient to identify putative CenH3 homologs in fully sequenced genomes; inability to find homologous genes can be indicative of true absence[29]. To identify all CenH3 homologs in Drosophila, we performed a tBLASTn search using both the canonical H3 and the *D. melanogaster* CenH3 (Cid) HFD as a query against 22 sequenced *Drosophila* genomes, as well as genomes from two additional Dipteran species. We recorded each *Cid* gene "hit" as well as its syntenic locus in each species (Figure 2-1). Consistent with previous studies, we found no additional *Cid* genes in the *D. melanogaster* genome or in closely related species of the *melanogaster* species subgroup[31,86]. In addition, we found that orthologs of the *Cid* gene in *D. melanogaster* have been preserved in their shared syntenic location in each of the *Drosophila* species we examined, except in *D. eugracilis* where it has clearly pseudogenized. We also found *Cid* orthologs in the shared syntenic context in a basal *Drosophila* species, *D. busckii,* as well as *Phortica variegata,* which belongs to an outgroup sister clade of *Drosophila*. Based on these findings, we conclude that an ortholog of *D. melanogaster Cid1* was present in the common ancestor of *Drosophila* in the shared syntenic location. We denote this orthologous set of genes in this shared syntenic location as *Cid1*.

Our analysis also identified four previously undescribed *Cid* duplications in *Drosophila* (Figure 2-1). The first of these was in *D. eugracilis*, which has a pseudogene at the ancestral *Cid1* shared syntenic location but also encodes a full-length *Cid* gene in a new genomic location (Figure 2-1). We refer to this gene as *Cid2*. We sequenced an additional 8 strains of *D.*

eugracilis to see if there were any cases of dual retention of both *Cid1* and *Cid2* in this species. In all cases, we found that *Cid1* orthologs were pseudogenized; they all contained a two base pair deletion leading to a frame shift after the first nine amino acids and a stop codon after 12 amino acids. *D. eugracilis* represents a unique case wherein the ancestral *Cid1* was lost and replaced by a recent duplicate, *Cid2*. Based on additional sequencing (below) it remains the only case of *Cid1* loss described in *Drosophila*.

In addition to the *Cid* duplicate in *D. eugracilis,* we found two new *Cid* paralogs in *D. kikkawai*, which belongs to the *montium* subgroup of *Drosophila*. Thus, *D. kikkawai* encodes three *CenH3* genes: the ancestral *Cid1,* as well as *Cid3* and *Cid4* (Figure 2-1). *Cid3* is located in close proximity to the original *Cid1* gene, whereas *Cid4* is present at a distinct genomic location. *Cid1, Cid3* and *Cid4* are quite different from one another at the sequence level. Their N-terminal tails only share ~25% amino acid identity, whereas pairwise amino acid identity of their HFD ranges from 80% (Cid1 and Cid3*)* to 55% (Cid3 and Cid4) to 45% (Cid1 and Cid4*)*. To study the age and evolutionary retention of these *Cid* paralogs, we sequenced these three syntenic loci from 16 additional species of the *montium* subgroup, for which no genomic sequences are publically available. We found that *Cid1*, *Cid3* and *Cid4* have been almost completely preserved in the *montium* subgroup (Figure 2-2) with one exception: the *Cid3* ortholog is pseudogenized in *D. mayri* (Figure 2-2A). Due to the lack of a complete genome sequence, we cannot rule out the possibility that *D. mayri* encodes a *Cid3*-like gene elsewhere in its genome. Based on these findings, we conclude that *Cid3* and *Cid4* were born from duplication events in the common ancestor of the *montium* subgroup at least 15 million years ago[87].

The fourth Cid duplication was found in the three species of the *Drosophila* subgenus: *D. virilis, D. mojavensis* and *D. grimshawi* (Figure 2-1, 'Additional *Cid* genes' column). Each of these species encodes *Cid1* and *Cid5,* which have an average pairwise amino acid identity of 60% in the HFD but only 15% in the N-terminal tail. To investigate the age and evolutionary retention of *Cid1* and *Cid5*, we sequenced both genes from an additional 11 species from the

virilis species group. We found that both *Cid1* and *Cid5* have been completely preserved (Figure 2-2B). Thus, we conclude that *Cid5* was born in the common ancestor of *Drosophila* subgenus at least 40 million years ago[87].

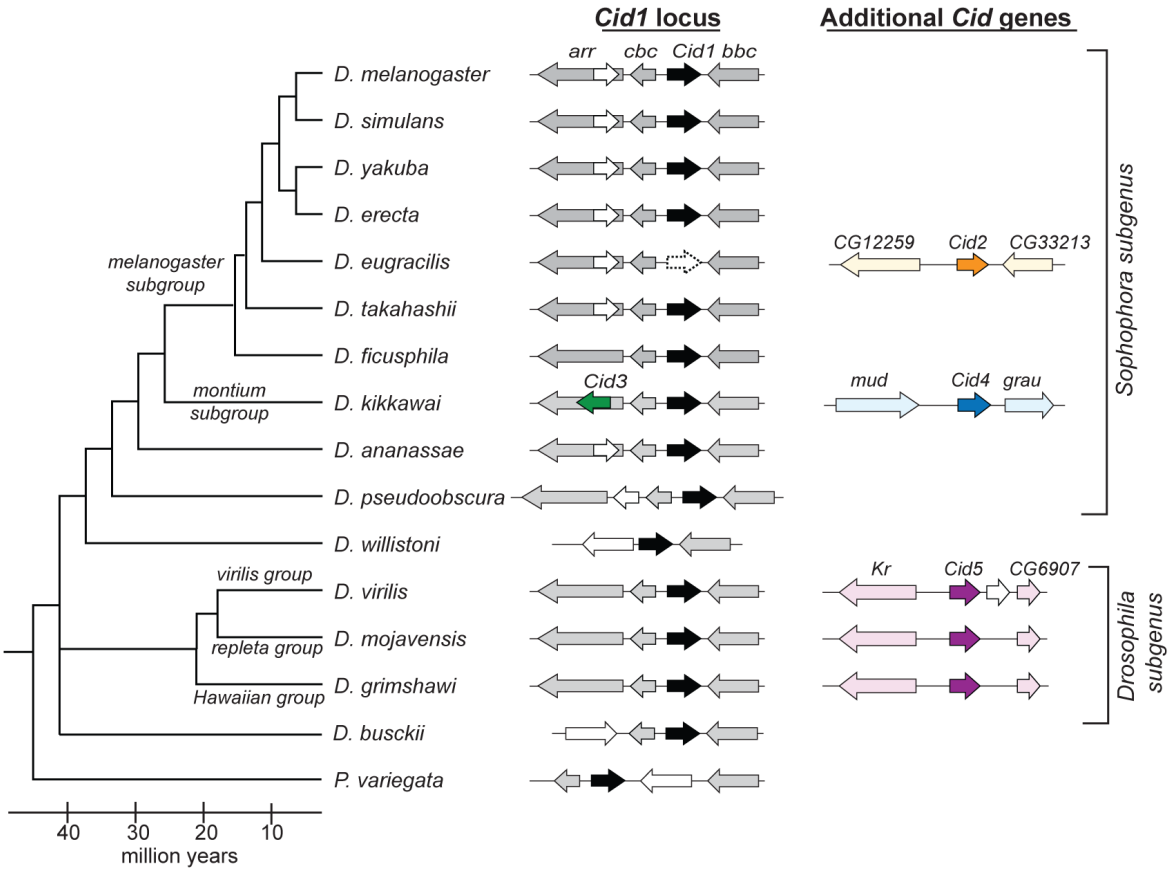


Figure 2-1. Identification of *Cid* duplication events across *Drosophila* evolution. A *Drosophila* species cladogram is presented with *Phortica variegata* as an outgroup. The genomic context of representative *Cid* paralogs identified by tBLASTn using previously published genome sequences is schematized to the right of each species. Within a species, each locus depicted is contained on a unique genomic scaffold (see Table 2-4 in Methods for detailed scaffold information). *Cid1* is the ancestral locus based on its presence in almost all species, including the outgroup species *P. variegata* (black arrow, see column labeled '*Cid1* locus'). In total, we found four *Cid* duplication events resulting in the birth of the genes *Cid2, Cid3, Cid4* and *Cid5* (see '*Cid1* locus' and 'Additional *Cid* genes' columns, dark orange, dark green, dark blue and dark purple arrows). We also found one *Cid1* pseudogene ('*Cid1* locus' column, empty arrow, dashed outline) in *D. eugracilis*. Arrows colored in a lighter version of the corresponding *Cid* gene color represent genes that define the shared syntenic locus of each paralog. White arrows represent genes that are present in a locus, but do not define the locus since they are present in fewer than 50% of the represented species. We do not provide gene names for these 'white arrow' genes. Genes that define each syntenic locus are named based on the *D. melanogaster* gene name.

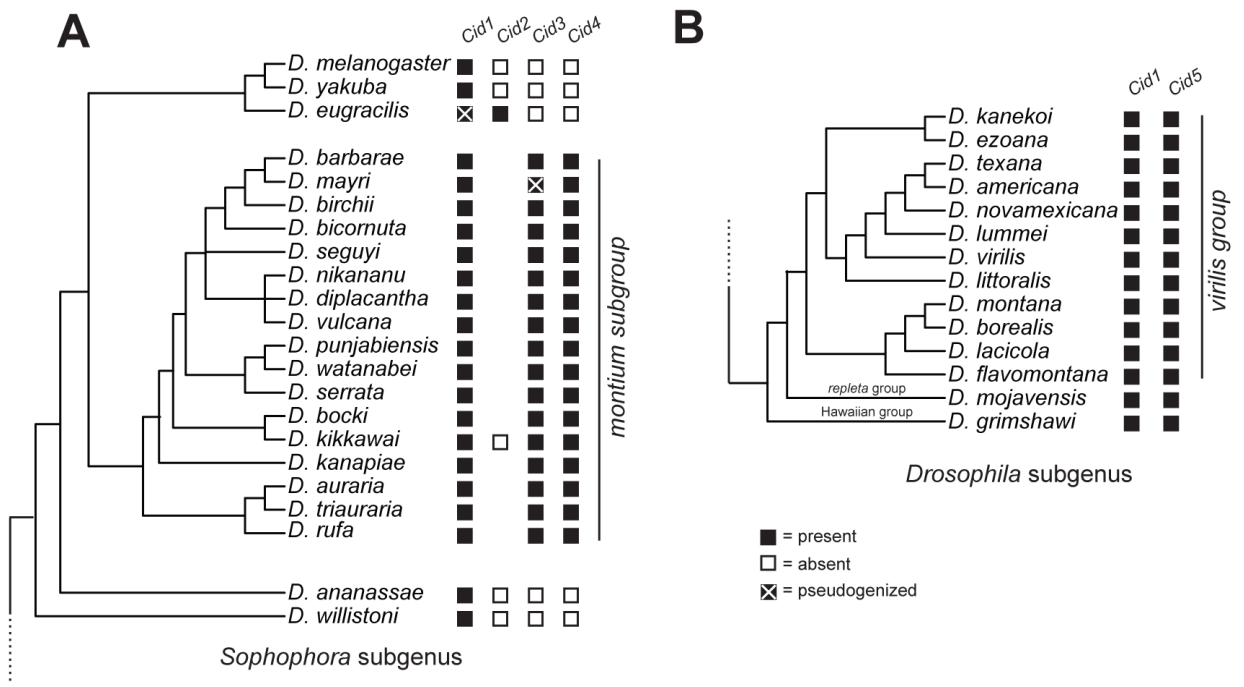


Figure 2-2. *Cid* paralogs are retained following duplication. (A) Summary of *Cid* paralog presence across the *Sophophora* subgenus with an expanded *montium* subgroup. The presence (black box) or absence (white box) of each *Cid* paralog as determined by PCR and Sanger sequencing is displayed next to each species. The lack of a box means that we did not attempt to amplify the locus. *Cid1, Cid3* and *Cid4* were preserved in almost all *montium* subgroup species with the exception of a *Cid3* pseudogene in *Drosophila mayri* (black box with a white X). This analysis indicated that *Cid3* and *Cid4* were born 20 – 30 million years ago. (B) Summary of *Cid* paralog presence across the *Drosophila* subgenus with an expanded *virilis* group. *Cid1* and *Cid5* were completely preserved in all examined *virilis* group species. We conclude that *Cid5* was born 40 – 50 million years ago in the common ancestor of the *Drosophila* subgenus.

To more rigorously test the paralogy and age of the *Cid* duplicates, we performed phylogenetic analyses (Figure 2-3). The N-terminal tails of all the Cid proteins were too divergent to be aligned, so we built a codon-based DNA alignment of the HFD of all *Drosophila Cid* genes, including *Cid1* orthologs sequenced in a previous survey[86]. We then used maximum likelihood (Figure 2-3) and neighbor-joining (Figure 2-4) analyses to construct a phylogenetic tree based on this alignment. We were able to draw the same conclusions from both trees except for one major difference, which we discuss below. Both phylogenetic analyses were in agreement with expected branching topology of *Drosophila* species[87] and concurred with our analyses of shared synteny (Figure 2-1). For instance, *D. eugracilis Cid2* (Figure 2-3 clade A, orange branch) grouped with *Cid1* genes of the *melanogaster* group with high confidence. Its closest phylogenetic neighbor was the *Cid1* pseudogene from *D. eugracilis,* supporting *Cid2*'s species-specific origin in a recent ancestor of *D. eugracilis*. We also found that the *Cid1* and *Cid5* genes of the *Drosophila* subgenus form monophyletic sister clades (clade D is sister to clade E, Figure 2-3 and Figure 2-4). We found that *D. busckii* and *D. albomicans* encode *Cid1* genes (clade E), based on phylogeny and shared synteny. However, whereas *D. albomicans* also encodes *Cid5*, *D. busckii* does not (clade D). The phylogenetic resolution between *Cid1* and *Cid5* clades is strong enough to suggest that the *Cid5* duplication may have predated the split between *D. busckii* and other members of the *Drosophila* subgenus, but that *Cid5* was subsequently lost in *D. busckii*.

We also found that the *Cid4* genes from the *montium* subgroup form a monophyletic clade (Figure 2-3, clade B) that forms sister clade to the *montium* subgroup *Cid1* and *Cid3* genes (clade C). The *melanogaster* subgroup *Cid1* genes (clade A) formed an outgroup to *montium* subgroup genes *Cid1, Cid3* and *Cid4* (clade A is an outgroup to clade B and C). This was the only major difference in branching topology between the maximum likelihood and neighbor-joining analyses; the latter (Figure 2-4) placed the *Cid4* genes from the *montium* subgroup (clade B) as a sister lineage to the *melanogaster* subgroup *Cid1* clade (clade A).

Since Cid1 is expected to be the ancestral gene in both subgroups, we favor the tree topology suggested by the maximum likelihood analysis. Both analyses reveal an unexpected intermingling of the *montium* subgroup *Cid1/Cid3* genes into a single clade (Figure 2-3 and Figure 2-4, clade C). This intermingled phylogenetic pattern could be the result of multiple, independent duplications of *Cid3* from *Cid1* in the *montium* subgroup. Alternatively, this pattern could reflect the effects of recurrent gene conversion, in which at least the HFD regions of *Cid1* and *Cid3* were homogenized by recombination

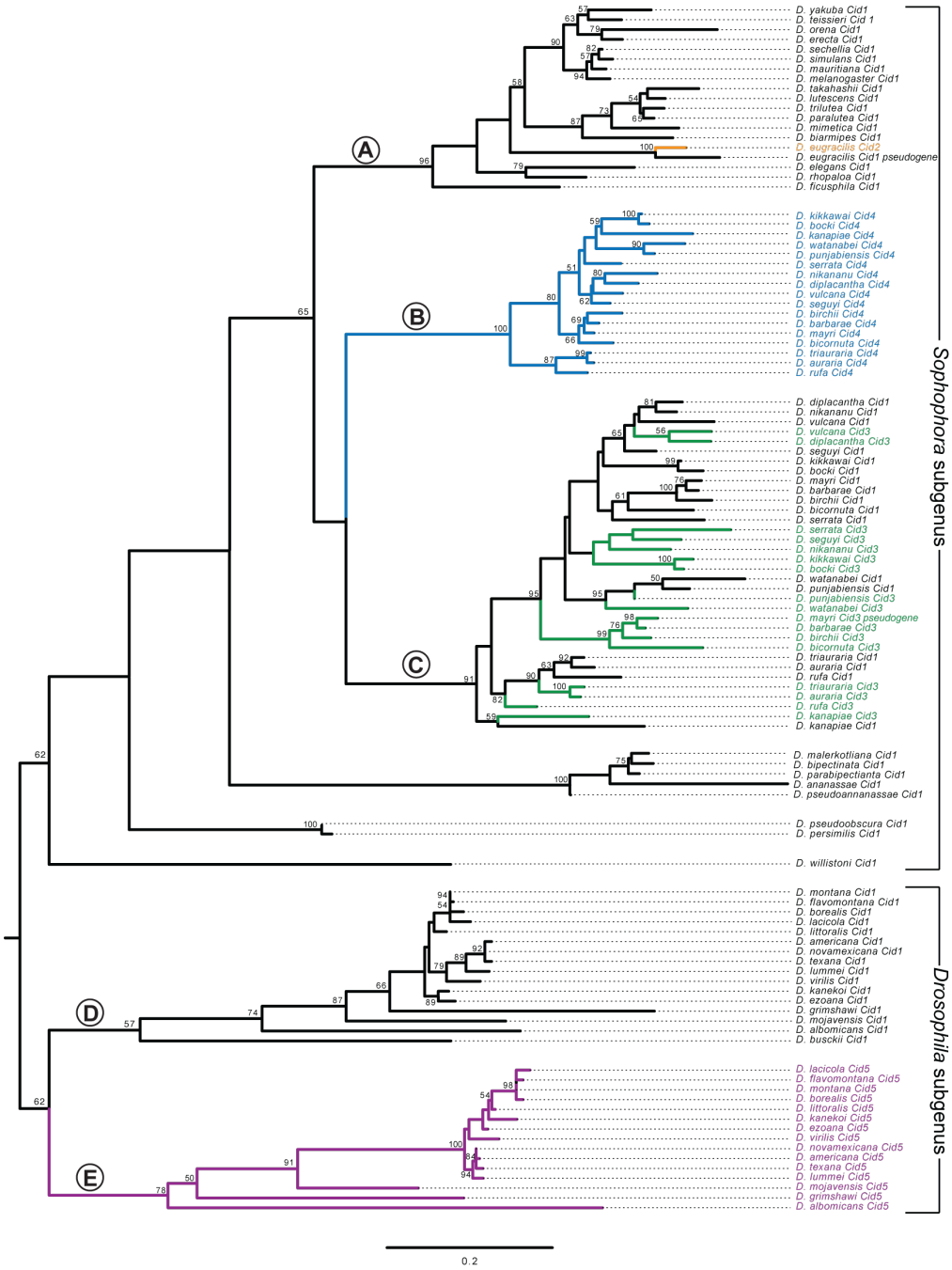


Figure 2-3. Maximum likelihood tree of all *Drosophila Cid* paralogs. We performed maximum likelihood phylogenetic analyses using PhyML with a nucleotide alignment of the histone fold domain of all *Cid* paralogs. We found that *Drosophila* subgenus *Cid1* (clade E), *Drosophila* subgenus *Cid5* (clade D) and *montium* subgroup *Cid4* (clade B) all formed well-supported monophyletic clades suggesting a single origin for these *Cid* paralogs. In contrast, *montium* subgroup *Cid1* and *Cid3* grouped together (clade C), consistent with our finding that they may be undergoing recurrent recombination (Figure 2-5). Selected clades (labeled with letters A – E) are further discussed in the main text. Bootstrap values greater than 50 are shown. The tree is arbitrarily rooted to separate the *Sophophora* and *Drosophila* subgenera. Scale bar represents number of substitutions per site.

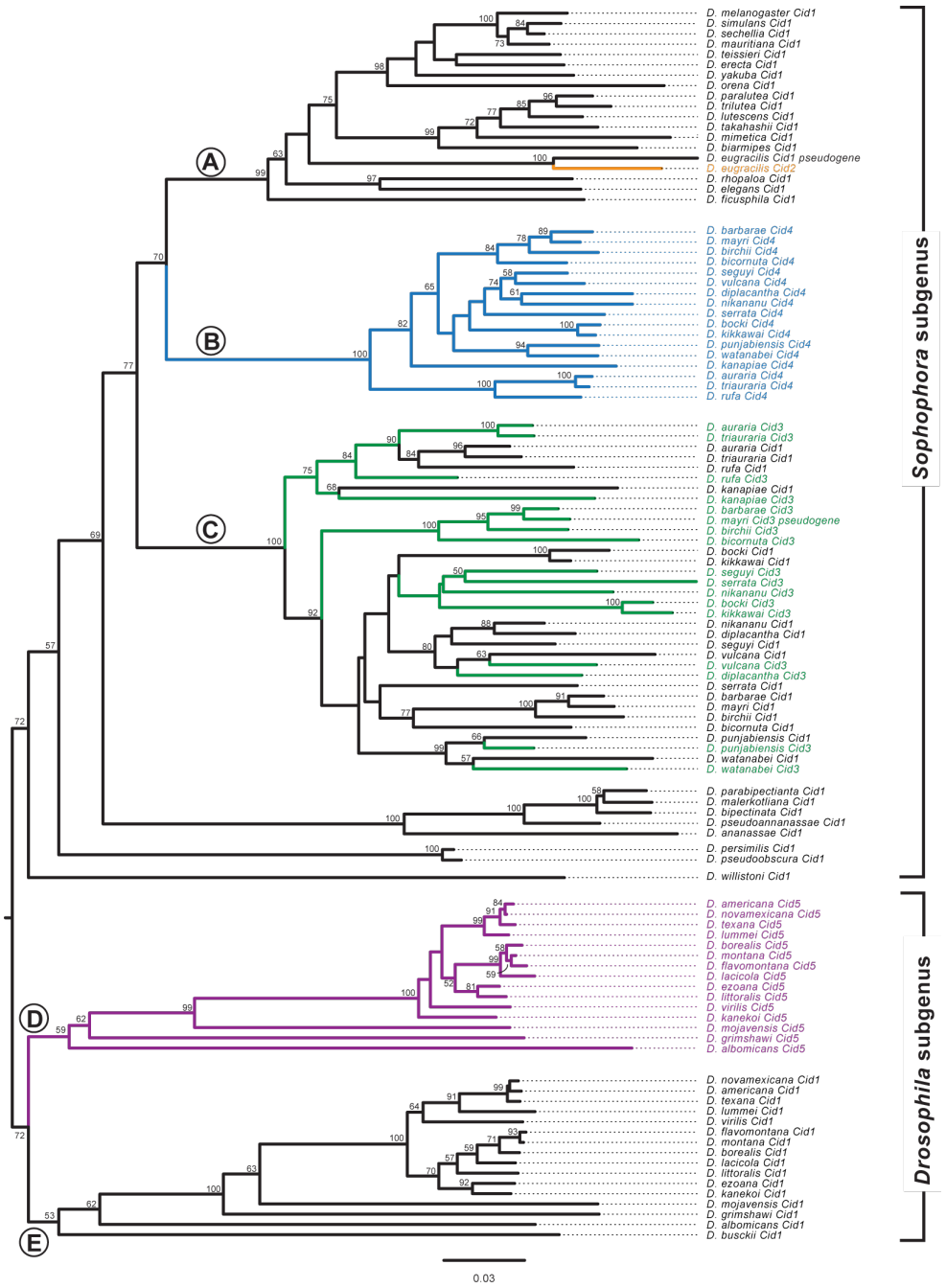


Figure 2-4. Neighbor-joining tree of all *Drosophila Cid* paralogs. We performed neighbor joining phylogenetic analyses with a nucleotide alignment of the histone fold domain of all *Cid* paralogs. We found that *Drosophila* subgenus *Cid1* (clade E), *Drosophila* subgenus *Cid5* (clade D) and *montium* subgroup *Cid4* (clade B) all formed well-supported monophyletic clades suggesting a single origin for these *Cid* paralogs. In contrast, *montium* subgroup *Cid1* and *Cid3* grouped together (clade C), consistent with our finding that they may be undergoing recurrent recombination (Figure 2-5). Selected clades (labeled with letters A – E) are further discussed in the main text. Bootstrap values greater than 50 are shown. The tree is arbitrarily rooted to separate the *Sophophora* and *Drosophila* subgenera. Scale bar represents number of substitutions per site.

Gene conversion between *Cid1* and *Cid3* could be facilitated by the close proximity of their genomic locations (see Figure 2-1, '*Cid1* locus' column), since frequency of gene conversion is inversely proportional to the distance between recombining sequences[88]. We used GARD (Genetic Analysis for Recombination Detection) analyses[89] to formally test for recombination between *Cid1* and *Cid3* from the *montium* subgroup. Consistent with our hypothesis of gene conversion, we found strong evidence for recombination between *Cid1* and *Cid3* ($p = 0.0002$) but not between *Cid1* and *Cid4.* The predicted recombination breakpoint is at the transition between the N-terminal tail and HFD (Figure 2-5A). Indeed, when we made a maximum likelihood tree from segment 1 alone (consisting primarily of the N-terminal tail), *Cid1* and *Cid3* formed the expected monophyletic clades distinct from each other (Figure 2-5B). However, when we made a maximum likelihood tree of the HFD, we found evidence for at least three specific instances of gene conversion (Figure 2-5C, recombination highlighted by asterisks). The HFD is important for Cid's interaction with other nucleosome proteins as well as for centromere targeting [51,90-93]. We speculate that such a recombination pattern allows Cid1 and Cid3 to perform distinct functions due to their divergent N-terminal tails whereas the homogenization of the HFD ensures that both proteins retain localization to the centromeric nucleosome. This pattern of ancient divergence followed by recurrent gene conversion may also partially explain the discrepant phylogenetic position of the *Cid1/Cid3* clade from the *montium* subgroup relative to the *Cid4* clade from the same subgroup (compare Figure 2-3 to Figure 2-4).

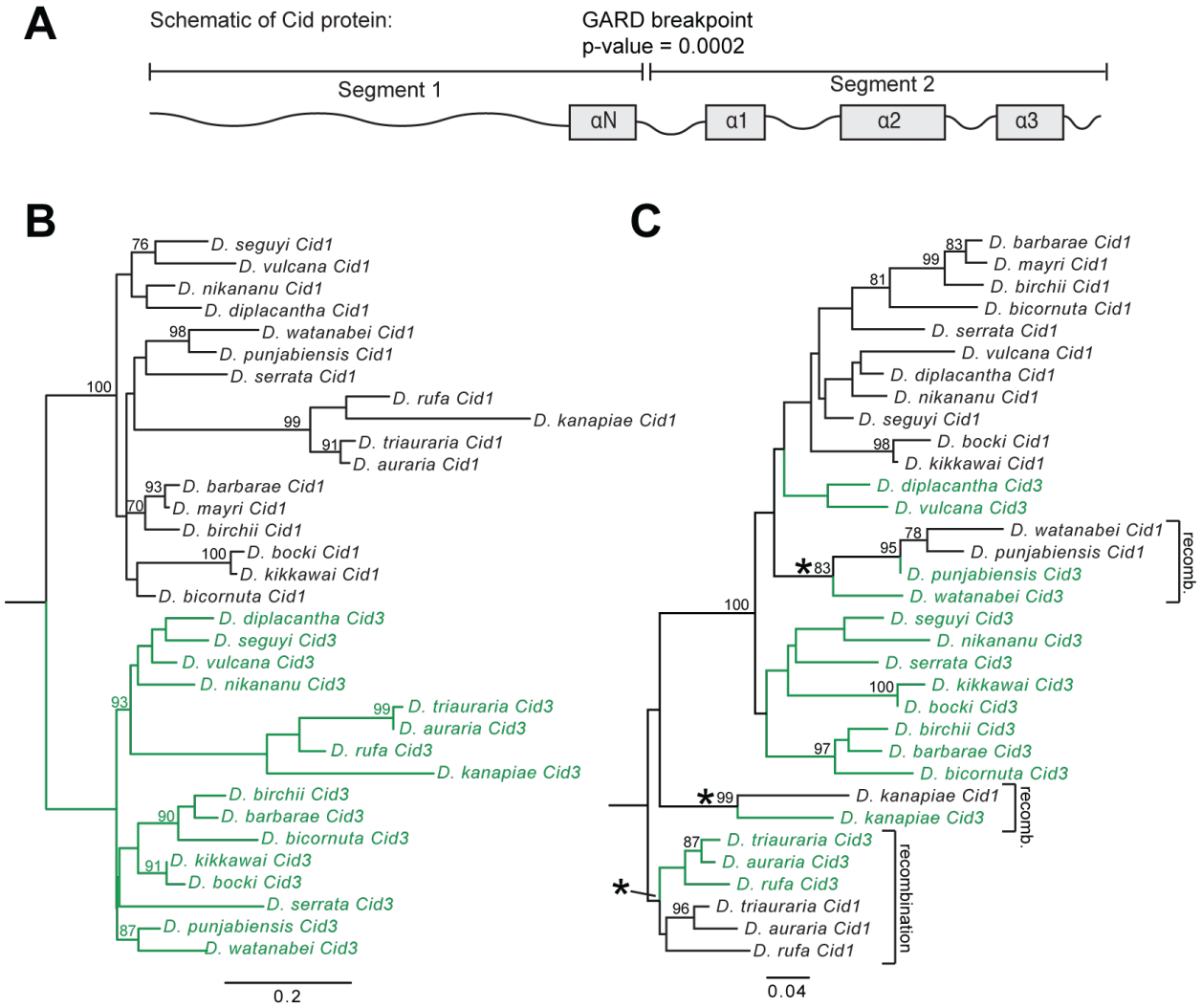


Figure 2-5. Recurrent gene conversion in *mountium* subgroup *Cid1* and *Cid3*. (A) We used the Genetic Algorithm for Recombination Detection (GARD[89]) to test for recombination in the *montium* subgroup *Cid1* and *Cid3.* GARD identified one significant (p=0.0002) breakpoint between the N-terminal tail and the histone fold domain. (B, C) Maximum likelihood phylogenetic trees from an alignment of GARD segment 1 (B) and GARD segment 2 (C) were subsequently generated using PhyML. Bootstrap values above 75 are displayed. Asterisks indicate branches along which gene conversion likely occurred. Scale bar represents nucleotide substitutions per site.

Drosophila Cid paralogs localize to centromeres

There are three possible outcomes following a functional gene duplication event: subfunctionalization, neofunctionalization and redundancy, which often leads to the loss of one paralog. Because we observe the co-retention of most *Cid* duplicates for millions of years (with the exception of *Cid1* loss in *D. eugracilis* and *Cid3* loss in *D. mayri)*, it is unlikely that duplicate *Cid* genes have been retained for redundant functions. We therefore wanted to distinguish between the possibilities of subfunctionalization and neofunctionalization for duplicate *Cid* genes.

It is not unprecedented that a histone variant paralog might develop a new function. For example, in mammals, the H2B variant SubH2Bv acquired a non-nuclear role in acrosome development in sperm[94]. To assess the possibility that the *Cid* paralogs may have acquired a non-centromeric role (*i.e.,* have become neofunctionalized), we turned to cell biological analyses to determine their localization. Previous studies showed that Cid1 orthologs (including those from *D. bipectinata* and *D. virilis*) can fail to localize to *D. melanogaster* centromeres, due to changes at the interface between Cid1 and its chaperone protein CAL1[51,93]. We therefore decided to test the localization of selected Cid paralogs in tissue culture cells from the same species.

Among all *montium* subgroup species that contain *Cid1, Cid3* and *Cid4,* cell lines were available only from *D. auraria* (cell line ML83-68, DGRC). We cloned the *Cid1, Cid3* and *Cid4* genes from *D. auraria* and tagged each with an N-terminal Venus tag to aid in visualization. We then transfected these constructs individually into *D. auraria* cells. We found that each Venus-Cid paralog localized in a similar manner, in punctate foci in a DAPI-intense region of the cells (Figure 2-7A). This pattern is highly characteristic of centromere localization[95]. To confirm this, we co-stained the cells with an antibody against CENP-C, a constitutively centromeric protein. Since no *D. auraria*-specific CENP-C antibodies were available, we first confirmed that the *D.*

melanogaster CENP-C antibody appropriately marked centromeres in *D. auraria.* Indeed, the *D. melanogaster* CENP-C antibody recognized foci at the primary constriction of *D. auraria* metaphase chromosomes (Figure 2-6). Moreover, we found that Venus-Cid1, Venus-Cid3 and Venus-Cid4 all co-localized with CENP-C in this cell line (Figure 2-7A). Based on this, we conclude that all the *D. auraria* Cid paralogs localize to centromeres.

We similarly tested the localization of *D. virilis* Cid1 and Cid5 in a *D. virilis* cell line (WR Dv-1). Unfortunately, the antibody raised against *D. melanogaster* CENP-C did not recognize *D. virilis* centromeres likely due to the high divergence between the CENP-C orthologs from the two species. We therefore co-transfected Venus-Cid1 and FLAG-Cid5. We found that Cid1 and Cid5 co-localize at nuclear foci, in a staining pattern that is typical of centromeric localization (Figure 2-7B). This suggests that despite their divergence, all Cid duplicates retain the ability to be recognized and deposited at centromeres by the existing machinery including CAL1, the chaperone that deposits *Drosophila* centromeric histones[51]. Alternatively, Cid paralog proteins might achieve centromeric co-localization by forming heterodimers with Cid1. Together, these results support the hypothesis that *Cid* duplicates have been retained to perform a centromeric function. Our cytological findings do not formally rule out the possibility of neofunctionalization; *Cid* duplicates might have been retained to perform a new centromeric function.

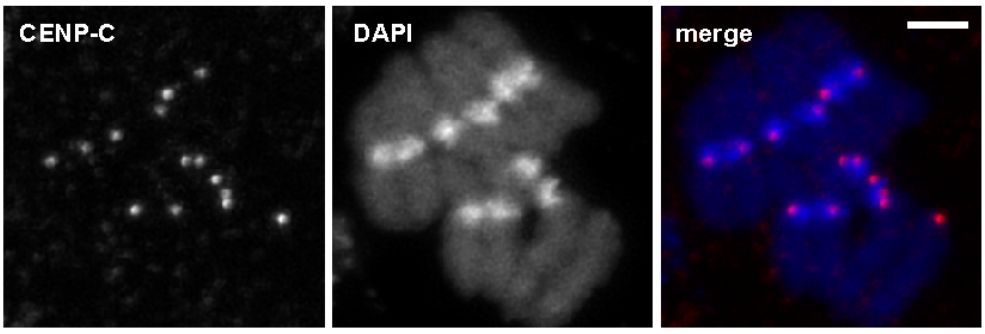


Figure 2-6. *D. melanogaster* CENP-C antibody localizes to *D. auraria* centromeres. *D. auraria* cell line ML83-68 was fixed and stained with *D. melanogaster* anti-CENP-C (red in merged image) and imaged using a confocal microscope. DNA (blue in merged image) shows metaphase chromosomes. Scale bar = 2 μm.

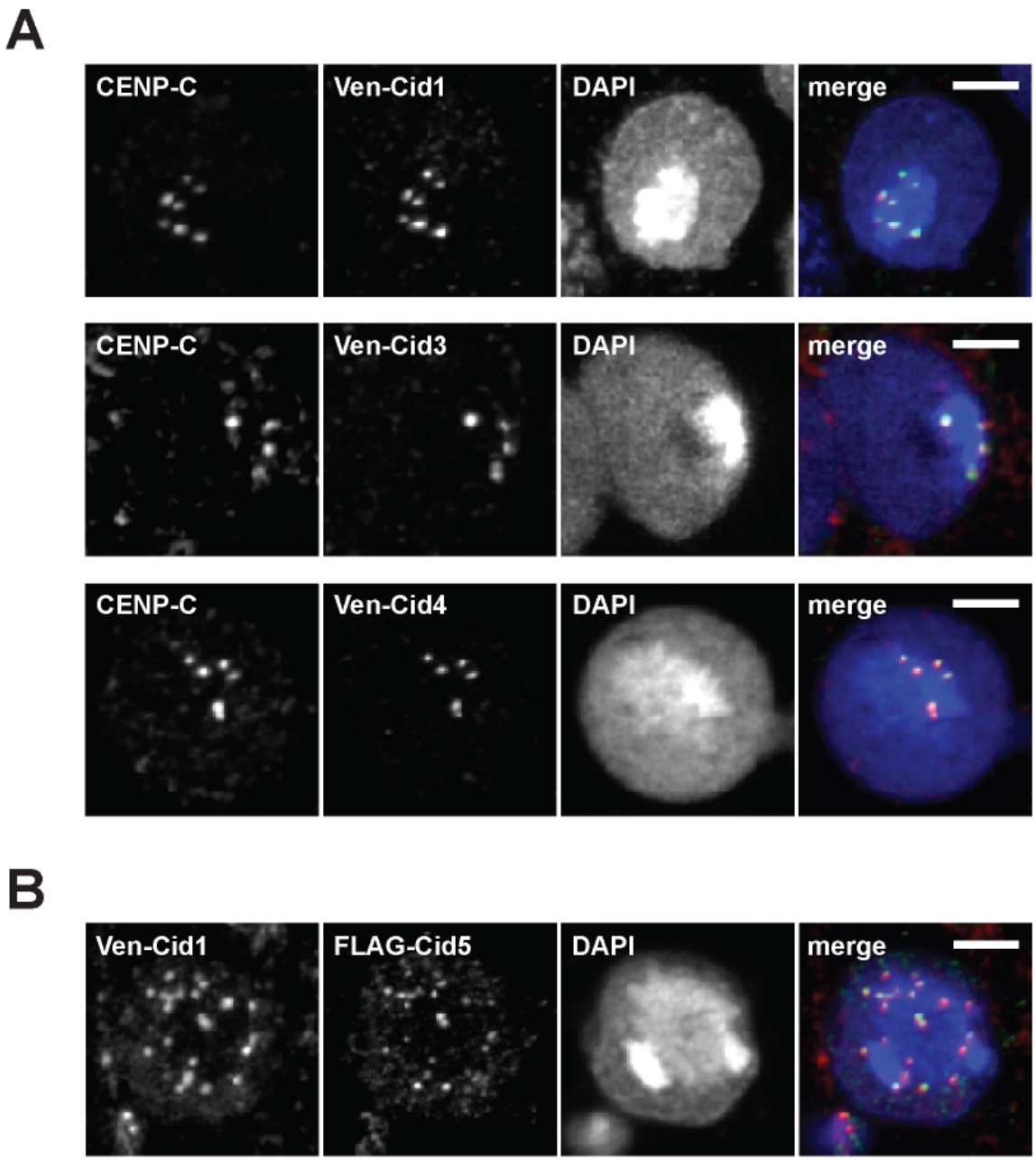


Figure 2-7. Proteins encoded by *Cid* paralogs localize to centromeres in cell culture. (A) Venus-tagged *D. auraria* Cid1, Cid3 *and* Cid4 were transiently transfected in a *D. auraria* cell line (top, middle and bottom panels, respectively). Cells were fixed and co-stained with a *D. melanogaster* CENP-C antibody (red in merged image) and anti-GFP (green in merged image). These data show co-localization of all three *montium* subgroup Cid proteins with CENP-C. (B) We co-transfected Venus-tagged Cid1 and FLAG-tagged Cid5 from *D. virilis* into a *D. virilis* cell line. Venus-Cid1 (red in merged image) and FLAG-Cid5 (green in merged image) both formed co-localized foci in the nucleus. All scale bars = 2 μm.

Testis restricted expression of Cid3 and Cid5

One means by which subfunctionalization can occur is by tissue-specific expression[96,97]. Duplicate genes could retain different subsets of promoter and enhancer elements from their parent gene, requiring both genes' expression to fully recapitulate parental gene expression[98]. We therefore wondered whether any of the *Cid* duplicates showed tissue-specific expression. We expected that at least one *Cid* paralog in each species must have maintained mitotic function and would therefore be widely expressed in somatic tissues. To test this, we first looked for expression of *Cid* paralogs in *D. auraria* and *D. virilis* tissue culture cell lines, which are derived from embryonic and larval tissues, respectively. We extracted RNA from both cell lines and performed RT-PCR. After 30 cycles of PCR, we detected a faint *Cid1* band in addition to a robust *Cid4* band in the *D. auraria* cell line (Figure 2-8A). In the *D. virilis* cell line, we detected expression of *Cid1* but not *Cid5* after 30 cycles of PCRs (Figure 2-8B). We did not detect *Cid3* (*D. auraria*) or *Cid5* (*D. virilis*) in this assay, which suggests that both genes are either not expressed or are expressed at low levels in tissue culture cells. From this analysis, we predict that Cid4 *(*and possibly Cid1*)* performs somatic Cid function in *D. auraria (i.e.,* mitotic cell divisions for growth) and that Cid1 performs somatic Cid function in *D. virilis.*

To further explore tissue specific expression, we performed RT-qPCR on dissected male and female *D. virilis* and *D. auraria* flies (whole fly, head, testes/ovaries, carcass). We performed the same analysis for *D. melanogaster,* which only encodes a single *Cid1* gene, for comparison. In *D. melanogaster,* we found that *Cid1* expression is highest in testes and ovaries and is relatively low in head and carcass (Figure 2-9). This is not surprising since testes and ovaries contain higher numbers of actively dividing cells than the head and the carcass. Similarly, in *D. auraria* and *D. virilis,* we found low expression of *Cid* paralogs in the head and the carcass of male and female flies (Figure 2-9). Interestingly, we found that the expression of *Cid3* in *D. auraria* and *Cid5* in *D. virilis* was primarily restricted to the male germline (Figure

2-8C, Figure 2-8D). We also found that *Cid1* and *Cid4* in *D. auraria* as well as *Cid1* in *D. virilis* are expressed in both testes and ovaries.

We wanted to extend our expression analyses of the *Cid* paralogs to other species containing duplicate *Cid* genes. We performed RT-qPCR on two additional *montium* subgroup species (*D. kikkawai* and *D. rufa)* and on two additional *Drosophila* subgenus species (*D. montana* and *D. mojavensis).* In all cases, *Cid3* or *Cid5* expression was detected in testes but not in ovaries. *Cid1* and *Cid4* expression patterns were similar across species too, with the exception of *Cid1* in *D. rufa,* which expressed at very low levels in ovaries (Figure 2-8C, Figure 2-8D, Figure 2-9).

Our findings are consistent with the hypothesis of tissue-specific specialization of the *Cid* paralogs in both the *montium* subgroup and the *virilis* group. These results also suggest that *Cid3* and *Cid5* were retained to perform a testis-specific function. In contrast, the other *Cid* paralogs are expressed in both somatic and germline tissues. However, these analyses lack the cellular resolution necessary to conclude whether the expression patterns are mutually exclusive or overlapping in tissues where multiple *Cids* are expressed. Moreover, in the *montium* subgroup, *Cid4* is expressed broadly in a pattern similar to *D. melanogaster Cid1,* and it is the primary *Cid* duplicate expressed in somatic cells. This suggests that *Cid4,* and not *Cid1,* performs canonical *Cid* function in *montium* subgroup species.

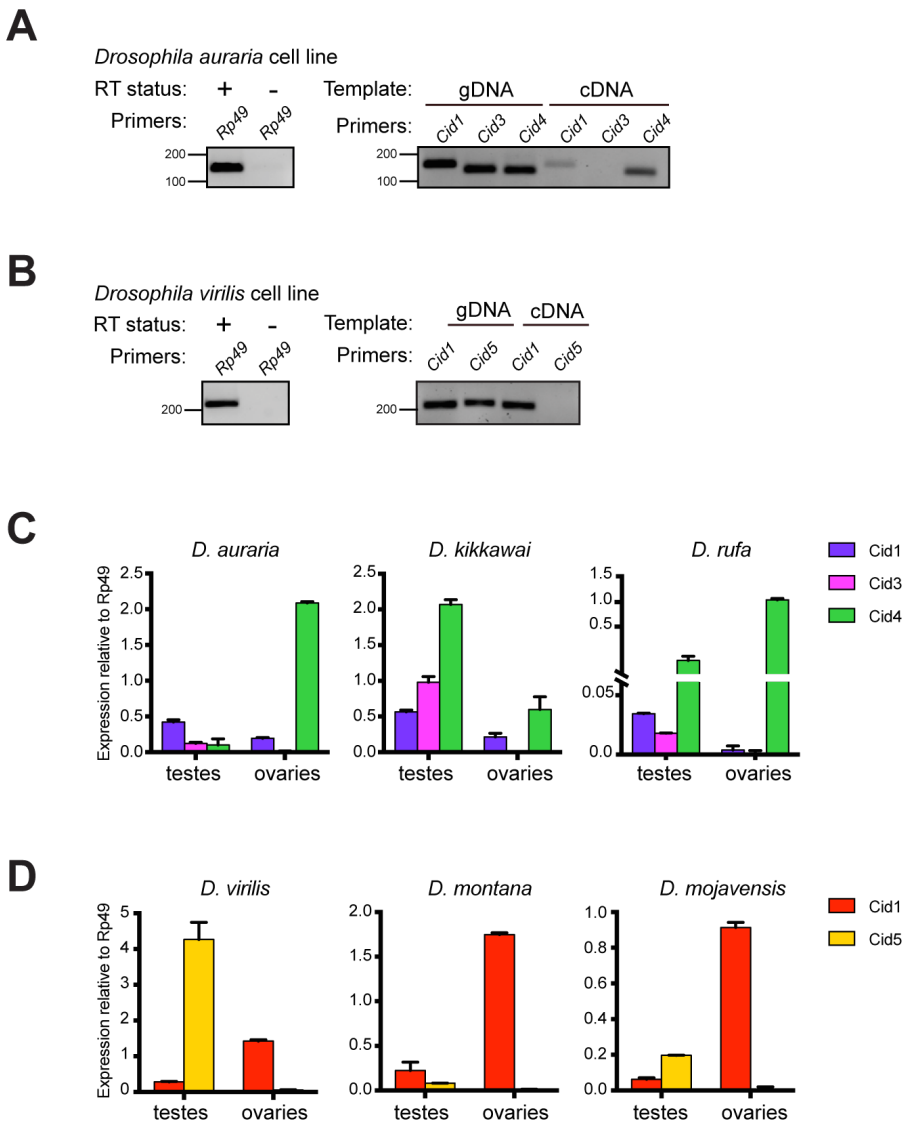


Figure 2-8. Male germline-restricted expression of some *Cid* paralogs. (A) Left gel: RNA samples used for *D. auraria* RT-PCR were free of DNA contamination as indicated by performing 35-cycle PCR for *Rp49* on cDNA samples generated with (+) and without (-) reverse transcriptase. Right gel: 30-cycle PCR performed with either genomic DNA (gDNA) or cDNA for *Cid1, Cid3* and *Cid4* from a *D. auraria* cell line. We detected both *Cid1* and *Cid4* expression but the *Cid4* expression band was more robust than the *Cid1* band. We did not detect expression of *Cid3* in this cell line. (B) Left gel: as in (A), RNA samples used for *D. virilis* RT-PCR were free of DNA contamination. Right gel: RT-PCR analyses of *Cid1* and *Cid5* from a *D. virilis* cell line at 30 cycles revealed only the expression of *Cid1*. We did not detect *Cid5* by RT-PCR. (C) RT-qPCR for *Cid1, Cid3* and *Cid4* from dissected tissues from three *montium* subgroup species revealed that *Cid1* and *Cid4* are expressed in both the testes and the ovaries whereas *Cid3* expression is testis restricted. (D) RT-qPCR from dissected tissues from three species from the *Drosophila* subgenus revealed that *Cid1* is expressed in the testes and ovaries of all three species whereas *Cid5* is only expressed in the testes. All RT-qPCR was normalized using *Rp49* as a control. Error bars represent standard deviation calculated from three technical replicates.

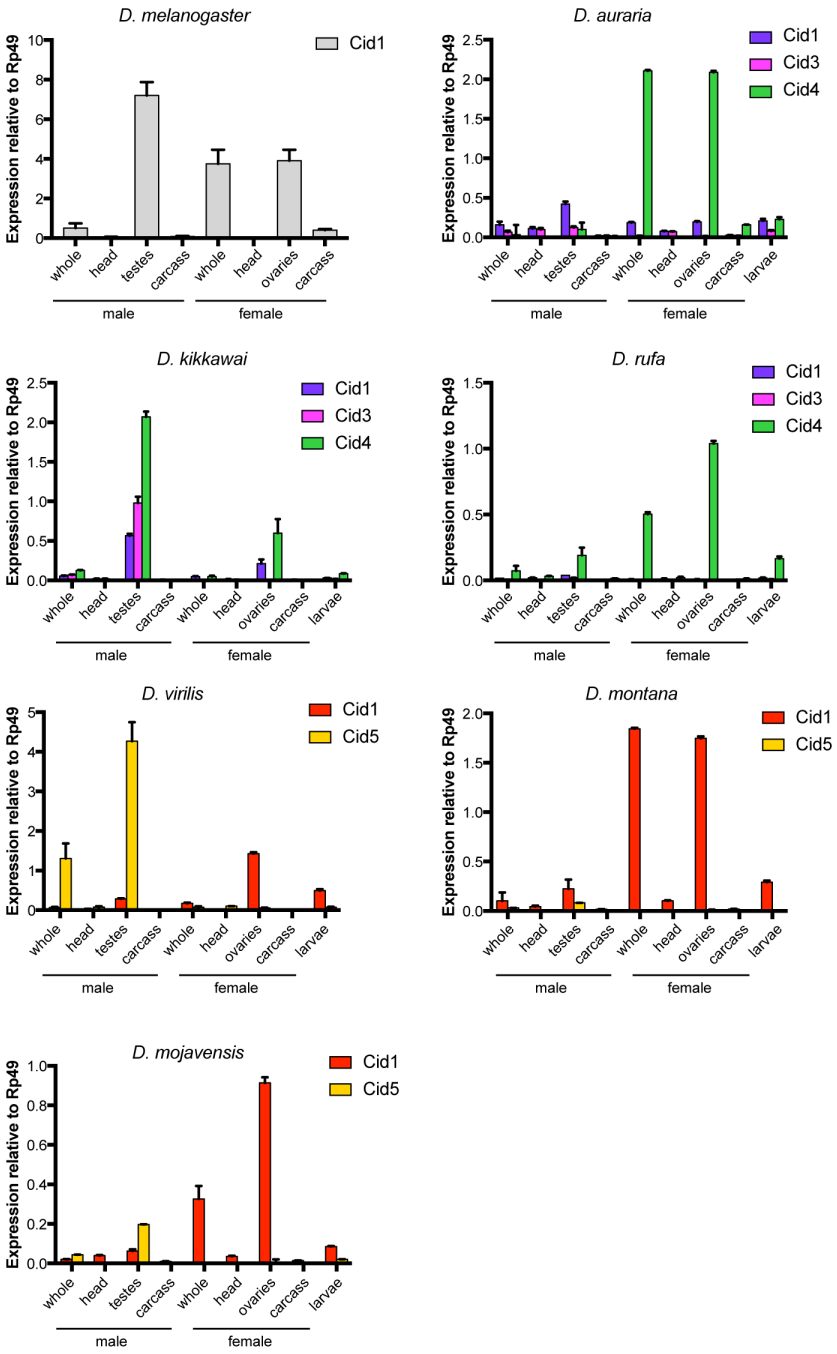


Figure 2-9. Cid paralog expression in multiple tissue types. RT-qPCR from dissected tissues from *D. melanogaster Cid1* (gray bars), *Cid1, Cid3 and Cid4* from three montium subgroup species (purple, pink and greens bars, respectively) and *Cid1* and *Cid5* from three Drosophila subgenus species (red and yellow bars, respectively). All RT-qPCR was normalized using Rp49 as a control. Error bars represent standard error calculated from three technical replicates.

Differential retention of N-terminal tail motifs and the evolution of new motifs following Cid duplication

Given their sequence divergence and different expression patterns, it seems likely that *Cid* paralogs may have been retained to perform distinct functions. Unlike the structural constraints that shape the HFD, the N-terminal tail of Cid is highly variable in length and sequence. We speculated that analyses of selective constraint in the N-terminal tail might present an additional opportunity to determine if specialization had occurred among the Cid paralogs. Although the specific function of the N-terminal tail has yet to be elucidated for *Drosophila* Cid, studies in humans and fission yeast have shown that the N-terminal tail is important for recruitment and stabilization of inner kinetochore proteins[52,99,100]. Furthermore, post-translational modifications of the N-terminal tail have been shown to be important for CENP-A mitotic function[101] and for facilitating interaction between two CENP-A molecules[102].

Conserved motifs provide an avenue to evaluate differential selective constraint in the N-terminal tail of different *Drosophila Cid* paralogs[48]. Motifs are regions of high similarity among protein sequences. They represent putative sites of protein-protein interaction and post-translational modification. We reasoned that we might be able to use the presence of certain N-terminal tail motifs as a proxy for various functional domains. We therefore used the motif generator algorithm, MEME[103], to identify conserved motifs in the N-terminal tail from six different groups of *Drosophila* Cid proteins: *melanogaster* group Cid1 (single copy genes only), *montium* subgroup Cid1, *montium* subgroup Cid3, *montium* subgroup Cid4, *virilis* group Cid1, and *virilis* group Cid5. We then used the motif search algorithm, MAST[104], to search for each motif in all Cid proteins. In total we found 10 unique motifs. Finally, we overlaid our motif analysis with the *Drosophila* species tree to gain insight into the evolution of N-terminal tail motifs (Figure 2-10A).

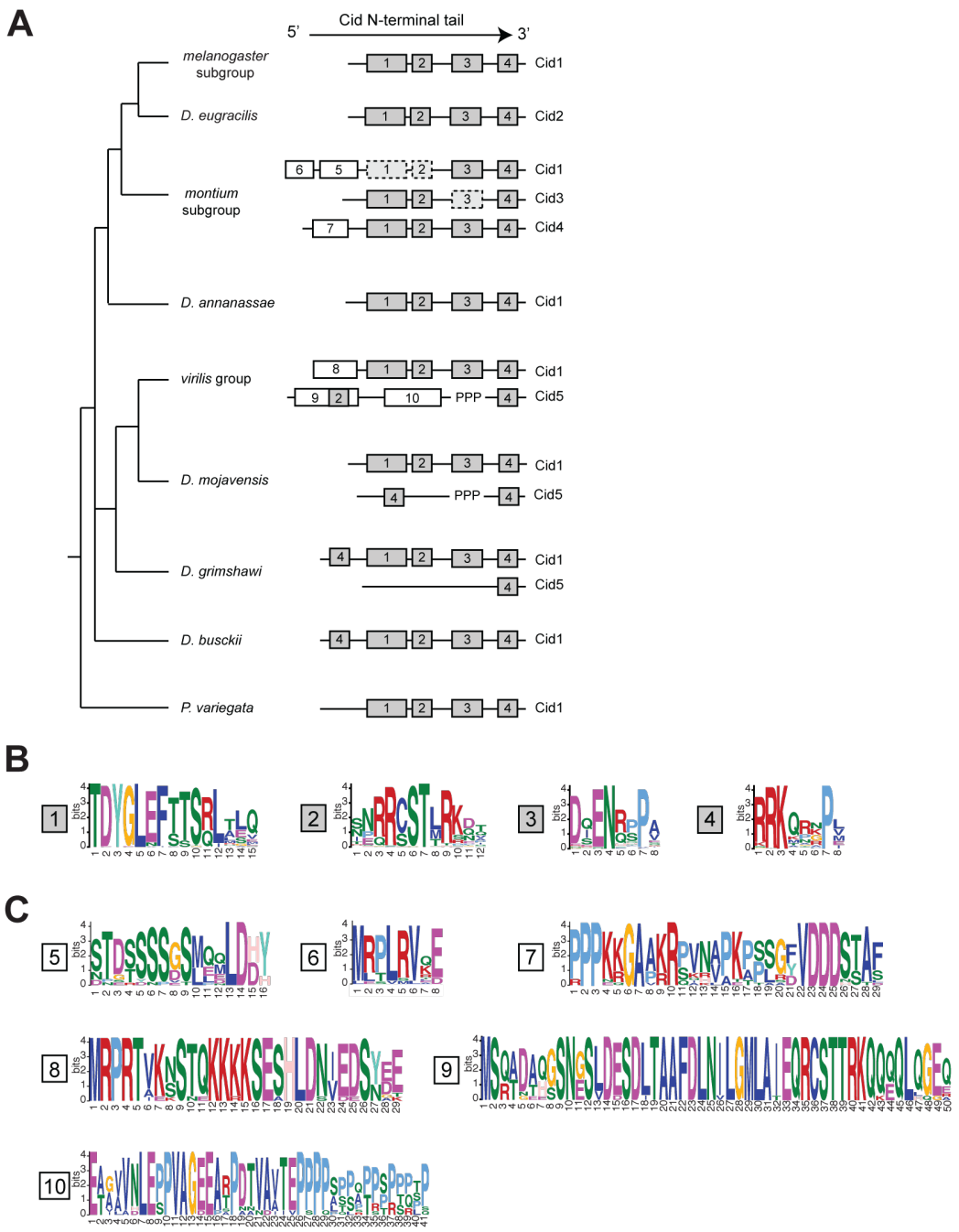


Figure 2-10. Evolution of N-terminal motifs among all Cid proteins. (A) A *Drosophila* species tree with a schematic of N-terminal tail motifs identified by MEME and MAST displayed to right of each species or species group. Each number represents a unique motif that does not statistically match any other motif in the figure with the exception of motif 2 and 9 (see Materials and methods). Gray boxes indicate 'core' motifs 1 – 4, which are present in all single copy *Cid* genes. White boxes indicate lineage specific motifs. 'PPP' indicates the position of the variable proline-rich region in Cid5. Dashed boxes indicate cases in which a given motif was present in ~50% of species. (B) Logos generated by MEME for consensus motifs 1 – 4. (C) Logos generated by MEME for consensus motifs 5 – 10.

From this analysis we can make several interesting conclusions. First, motifs 1-4 (Figure 2-10B) are conserved in every Cid1 protein when it is the only copy encoded in the genome. These motifs correspond nicely to the motifs we previously identified in the *melanogaster* group using Block Maker[86]. Although their function remains largely uncharacterized, motif 4 has been shown to be involved in recruitment of mitotic checkpoint protein, BubR1[105]. Motif 4 could also play a role in histone-DNA interaction because it is located in the region where the N-terminal tail exits the nucleosome and passes between the two strands of DNA[91]. Motif 4 is the only motif present in all Cid paralogs, which suggests that it performs a general function among all Cids. Given their retention in all single copy Cid-containing *Drosophila* species, we consider motifs 1 – 4 to be the "core" Cid1 motifs (Figure 2-10B) and speculate that all are required for Cid1 function when it is the only centromeric histone protein. Indeed, all *Drosophila* species contain all of these motifs amongst their various Cid paralogs.

Next, we observed that some Cid paralogs had evolved and retained 'new' N-terminal tail motifs (Figure 2-10C). We identified three motifs that evolved in Cid paralogs from the *montium* subgroup; motifs 5 and 6 are found in Cid1 whereas motif 7 is found in Cid4. One might interpret the invention of additional N-terminal tail motifs as evidence of neofunctionalization. Indeed, invention of novel protein-protein interactions to perform new centromeric functions is expected for neofunctionalized paralogs. However, new motifs could also arise in paralogs that have become specialized, to more optimally perform a subset of the pre-existing functions *e.g.,* in the male germline. Thus, specialization could lead to the retention of novel motifs, especially if these motifs would be incompatible with all ancestral functions.

More direct evidence of subfunctionalization emerged from our observation of frequent loss of 'ancestral' motifs 1-3 from Cid1 and Cid3, despite their complete preservation in Cid4 (Figure 2-10A, dotted lines indicate motif is absent from ~50% of queried species). Intriguingly, some Cid1 and Cid3 orthologs in the *montium* subgroup appear to have differentially retained motifs 1-3; Cid1 has motif 3 and Cid3 has motifs 1 and 2. This differential retention of an

ancestrally conserved subset of core motifs is highly suggestive of subfunctionalization[48]. Furthermore, our findings support the hypothesis that in the *montium* subgroup, it is the Cid4 paralog rather than the ancestral Cid1, which performs the canonical functions of centromeric histones carried out by Cid1 in other species, because Cid4 contains all core motifs but *montium* subgroup Cid1 does not. This would also be consistent with our expression analyses, in which Cid4 expresses more robustly than Cid1 in somatic cells (Figure 2-8A).

This pattern of new motif evolution and ancient motif degeneration is also evident in the Cid paralogs from the *virilis* group. In this group of species, the Cid1 paralog has retained the core set of motifs 1-4 but added motif 8. In contrast, Cid5 paralogs have added motifs 9 and10 but lost core motifs 1 and 3. We therefore conclude that the tissue-specific pattern of expression and the differential retention of N-terminal motifs support a general model of subfunctionalization, but that some paralogs may have acquired novel protein-protein interaction motifs perhaps to optimize for new or specialized centromeric functions.

Different evolutionary forces act on different Cid duplicates

Tissue specific expression of some *Cid* paralogs and differential retention of N-terminal tail motifs supports the hypothesis that Cid paralogs may have subfunctionalized. We next considered the possibility that duplicate *Cid* genes were retained to allow optimization for divergent functions. In the *melanogaster* group, *Cid1* (a single copy *Cid* gene) has been shown to evolve rapidly[36], perhaps due to its interaction with rapidly evolving centromeric DNA and the need for drive suppressors in male meiosis[37]. While this rapid evolution might be required for the 'drive suppressor' function, it may be disadvantageous for canonical Cid function (*e.g.,* mitosis). As a result, selection may act differently on *Cid* in the male germline than on somatic or ovary-expressed *Cid*. For instance, some *Cid* paralogs (*e.g.,* those that are expressed primarily in the male germline and may suppress centromere-drive) might evolve under positive selection while others would not.

We used maximum likelihood methods using the PAML suite to test for positive selection on each of the Cid paralogs. For montium subgroup Cid1 and Cid3, we performed each analysis separately on GARD segment 1 and 2 (Figure 2-5). For all other Cid genes we performed PAML analyses on full-length alignments. Consistent with our prediction, we found that some, but not all, Cid paralogs likely evolve under positive selection (Table 2-1). For example, PAML analyses reveal that Cid3 segment 1 evolved under positive selection (Table 2-1, M1 vs M2 p = 0.02 and M8a vs M8 p = 0.01). However, we did not find evidence that Cid5, another male germline-restricted paralog, evolves under positive selection. We note, however, that we were unable to unambiguously align a highly variable proline-rich segment in Cid5's N-terminal tail and excluded this segment from our analyses. If positive selection was occurring in this region, we would be unable to detect it. We also found that *Cid4* evolved under positive selection but *montium* subgroup *Cid1* and *Cid3* segment 2, and *virilis* group *Cid1,* did not (Table 2-1). To ensure that recombination in *Cid1* and *Cid3* segment 2 was not obscuring our ability to detect positive selection in these segments, we re-ran the PAML analyses excluding the species for which we could detect apparent gene coversion events (*D. watanabei, D. punjabiensis, D. kanapiae, D. triauraria, D. auraria and D. rufa).* Exclusion of these species did not affect the conclusions from the PAML analyses; we did not detect positive selection in either *Cid1* or *Cid3* segment 2.

For those genes that PAML identified as having evolved under positive selection (*Cid3* segment 1 and *Cid4),* Bayes Empirical Bayes analyses identified one amino acid in Cid3 and one amino acid in Cid4 as having evolved under positive selection with a high posterior probability (>0.95). In Cid3*,* the positively selected site is adjacent to the αN-helix. In Cid4, the positively selected site is in loop 1 of the HFD (Table 2-1). Interestingly, these are both places where Cid is predicted to contact centromeric DNA[91]. Interestingly, Loop 1 is also the domain that interacts and co-evolves with the centromeric histone chaperone, CAL1[51,93]. These results

are consistent with the hypothesis that both *Cid3* and *Cid4* are engaged in a genetic conflict involving centromeric DNA.

Table 2-1: PAML tests for positive selection on *Drosophila Cid* paralogs. Summary table of M1 vs. M2, M7 vs. M8 and M8a vs. M8 PAML results for each *Cid* gene or gene segment. P-values less than 0.05 are indicated in bold text. PP=posterior probability.

Gene	Species group	# species	Alignment region	M1 vs M2 p-value	M2 % sites with ω > 1 (Avg ω)	M2 BEB (PP > %95)	M7 vs M8 p-value	M8 % sites with ω > 1 (Avg ω)	M8a vs M8 p-value	M8 BEB (PP > %95)
Cid1	Montium	17	GARD seg1	1.000	12.6 (1.0)	--	**0.040**	35.0 (1.0)	1.000	--
			GARD seg2	1.000	2.5 (1.0)	--	0.020	10 (1.5)	0.128	--
Cid3	Montium	16	GARD seg1	**0.023**	2.5 (4.2)	50T	**0.006**	3.9 (3.0)	**0.008**	50T
			GARD seg2	1.000	4.7 (1.0)	--	0.667	7.1 (1.1)	0.955	--
Cid4	Montium	17	Full length	0.059	0.5 (5.4)	135T	**0.052**	1 (4.0)	**0.024**	135T
Cid1	Virilis	12	Full length	0.310	6.1 (2.2)	--	0.168	12.0 (1.9)	0.119	--
Cid5	Virilis	12	Full length	0.628	5.2 (2.1)	--	0.588	9.8 (1.8)	0.322	--

We next used the McDonald-Kreitman (MK) test to look for positive selection in each of the *Cid* paralogs. While PAML detects positive selection occurring recurrently at selected amino acid residues across deep evolutionary time, the MK test detects more recent positive selection distributed over entire genes or protein domains. The MK test assumes that if protein constraints have not dramatically altered over evolution, the ratio of non-synonymous to synonymous fixed differences between species (D_N/D_S) should approximately equal the ratio of non-synonymous to synonymous polymorphisms with a species (P_N/P_S). However, if a higher than expected number of non-synonymous fixed changes are observed (*i.e.,* $D_N/D_S > P_N/P_S$), this would be indicative of positive selection after the divergence of the species.

In order to test for positive selection in the *montium* subgroup using the MK test, we sequenced and compared *Cid1, Cid3* and *Cid4* paralogs from 26 strains of *D. auraria* and 10 strains of *D. rufa.* For *virilis* group *Cids,* we sequenced *Cid1* and *Cid5* paralogs from 10 strains of *D. virilis* and 21 strains of *D. montana.* We found an excess of non-synonymous fixed

differences between *D. auraria* and *D. rufa Cid1* and *Cid3*, suggesting that both genes evolve under positive selection (Table 2-2). Parsing the signal by performing the MK test on just the N-terminal tail or just the HFD domain revealed that *Cid1* and *Cid3* HFD domains evolve under positive selection (Table 2-2). However, we did not find evidence for positive selection in the N-terminal tails. Most of the non-synonymous fixed differences occur in Loop1, which is predicted to contact centromeric DNA[91]. Interestingly, even though PAML analyses detected ancient recurrent positive selection in *montium* group *Cid4*, we did not find strong evidence for recent positive selection since the *D. auraria-D. rufa* divergence using the MK test (p=0.08). We also found no evidence of positive selection having acted on *virilis* group *Cid1* or *Cid5* using the MK test (Table 2-2).

Table 2-2: McDonald-Kreitman tests for positive selection on *Drosophila Cid* genes. Summary of results from McDonald-Kreitman tests for each *Cid* gene or gene segment. P-values less than 0.05 are indicated in bold text. Neutrality index > 1 is indicative of an excess of non-synonymous polymorphisms, which suggests negative selection. Neutrality index < 1 is indicative of an excess of non-synonymous fixed changes, which suggests positive selection.

Gene	Species group	Regoin	# codons	Non-syn fixed	Syn fixed	Non-syn poly	Syn poly	Ratio	N.I.	P-value
Cid1	montium	1-188	188	25	20	31	62	25:20:31:62	0.4	**0.02**
		1-108	108	11	7	26	36	11:7:26:36	0.46	0.18
		109-188	80	14	13	5	26	14:13:5:26	0.18	**<0.01**
Cid3	montium	1-196	196	24	21	24	48	24:21:24:48	0.44	**0.04**
		1-115	115	17	14	22	29	17:14:22:29	0.63	0.36
		116-196	81	7	7	2	19	7:7:2:19	0.11	**0.02**
Cid4	montium	1-224	224	16	31	14	10	16:31:14:10	2.71	0.08
		1-148	124	13	18	13	8	13:18:13:8	2.25	0.25
		149-224	76	3	13	1	2	3:13:1:2	2.17	1
Cid1	virilis	1-228	228	31	34	4	6	31:34:4:6	0.73	0.74
		1-123	123	24	23	4	4	24:23:4:4	0.96	1
		124-228	105	7	11	0	2	7:11:0:2	x	0.52
Cid5	virilis	1-208	208	29	23	12	15	29:23:12:15	0.63	0.36
		1-103	103	23	7	10	8	23:7:10:8	0.38	0.2
		104-208	105	6	16	2	7	6:16:2:7	0.76	1

To summarize our positive selection analyses, we found that *Cid3* has experienced both ancient and recent positive selection in protein domains predicted to contact centromeric DNA. *Cid4* has also experienced ancient, recurrent positive selection at putative DNA-contacting sites, but we found no evidence of recent positive selection in a MK test comparison. This could suggest that *Cid4* was either relieved of its role in such conflict or that the MK test lacks the power to detect selection acting on only a few residues. Similarly, although PAML analyses failed to identify a pattern of ancient, recurrent positive selection, the MK test did reveal positive selection for *montium* subgroup *Cid1* while comparing the entire HFD. In contrast, we did not find evidence for positive selection having acted on *Cid1* and *Cid5* in the *virilis* group by either test.

2.4 Discussion

The availability of many high-quality sequenced genomes as well as the comprehensive understanding of phylogenetic relatedness between species make *Drosophila* an ideal system to study gene duplication and evolution. This facilitated our discovery of four ancient *Cid* duplications in *Drosophila*. We found that while *Cid1* (previously known as just '*Cid')* is preserved in its shared syntenic location in all species examined except one, many species encode one or two additional *Cid* genes. The species of the *montium* subgroup, including *D. kikkawai,* have three *Cid* genes (*Cid1, Cid3 and Cid4),* which were born from a duplication event ~15 million years ago. The species of the *virilis* group, as well as *D. mojavensis* and *D. grimshawi* (*repleta* and *Hawaiian* groups, respectively), have two *Cid* genes (*Cid1* and *Cid5),* which were born from a duplication event ~40 million years ago. These *Cid* duplications have been almost completely preserved in extant species. Despite the fact *Cid* paralogs are divergent from one another at the sequence level, all paralogs have the ability to localize to centromeres when expressed in tissue culture cells. Based on our detailed analysis of two subgenera

(*Drosophila* and *Sophophora*), we predict that over one thousand *Drosophila* species encode two or more *CenH3* (*Cid*) genes[106]. We further conclude that *D. melanogaster* and other *Drosophila* species that have only one *Cid* are the minority; most *Drosophila* species have multiple *Cid* paralogs.

Our phylogenetic analyses support our synteny-based conclusions, and reveal recurrent recombination between *Cid1* and *Cid3* in *montium* subgroup species. This is the first reported case of recombination between *CenH3* paralogs. Our results suggest that this recombination results in evolutionary homogenization of the histone fold domain between *Cid1* and *Cid3*, while the N-terminal tails of *Cid1* and *Cid3* appear to be evolving independently, perhaps maintaining divergent functions. This recombination could be the genetic mechanism by which *Cid1* and *Cid3* maintain function in the centromeric nucleosome via near-identical HFDs despite having divergent N-terminal tails, which facilitates distinct interactions. This pattern of gene conversion is akin to patterns of recombination seen for paralogous mammalian antiviral proteins, IFIT1 and IFIT1B, in which gene conversion homogenizes the N-terminal oligomerization domain but not the divergent C-terminus, which allows IFIT1 and IFIT1B proteins to have distinct anti-viral specificities[107].

What is the evidence that *Cid* paralogs have distinct functions? The strongest evidence is that they have been co-retained in both the *montium* subgroup and the virilis/repleta/Hawaiian radiation for tens of millions of years. If they performed redundant functions, we predict that one of the paralogs would be lost over this time frame considering the high rate of DNA deletion in *Drosophila*[108]. Indeed, we observed only two instances of Cid duplication followed by pseudogenization (*Cid3* pseudogene in *D. mayri* and *Cid1* pseudogene in *D. eugracilis*) and inferred the possible loss of *Cid5* (in *D. busckii*). Our findings that *Cid3* and *Cid5* are expressed primarily in the male germline, that N-terminal tail motifs have been differentially retained and that different selective pressures have shaped different *Cid* paralogs further supports the idea that these *Cid* paralogs perform non-redundant functions.

Interestingly, our expression and motif analyses strongly suggest that *Cid4* has taken over the primary function of somatic centromeric histone function in *montium* subgroup species. *Cid4* is the primary *Cid* gene expressed in *D. auraria* tissue culture cells and is the only Cid paralog in this species that contains all four of the 'core' N-terminal tail motifs. In contrast, the 'ancestral' *Cid1* is expressed at lower levels than *Cid4*, *Cid3* is primarily expressed in the male germline, and neither *Cid1* nor *Cid3* contain all four 'core' motifs. This finding has implications for future experiments taking an evolutionary approach to study Cid function. The correct *Cid* paralog for such studies must be chosen carefully. Further functional experimentation, such as creating genetic knockouts, will be required to determine the specific function of each *Cid* paralog.

We propose that in species with a single-copy Cid gene, the same protein must perform multiple functions including mitotic cell division in somatic tissues and drive suppression in the male germline. These functions might require different selective pressures to achieve functional optimality. For example, we have previously proposed that drive suppression results in rapid evolution of *Cid* to co-evolve with rapidly evolving centromeric DNA[37] whereas mitotic function might impose purifying selection on *Cid,* minimizing changes in amino acid sequence. Therefore, it could be advantageous to have two copies of *Cid* such that each encodes a separate function. Our results suggest that *Cid3* and *Cid5* are candidate drive suppressors given their male germline-restricted expression. Consistent with this prediction, we detected evidence for positive selection in *Cid3*. In contrast, we did not find evidence that *Cid5* evolves under positive selection. This leaves open the possibility that *Cid5* performs an alternative, centromeric, male germline function independent of potential centromere-drive suppression in meiosis.

If it is advantageous to have multiple *Cid* paralogs*,* why don't more animal species possess more than one gene encoding centromeric histones? We hypothesize that retention of duplicate *Cid* genes requires a defined series of evolutionary events and that the cadence of the

mutations determines the ultimate fate of the duplicated genes[109]. First, the duplication must not be instantaneously harmful; gene expression must be carefully controlled, as Cid overexpression or expression at the wrong time during the cell cycle can be catastrophic[110,111]. Even though other kinetochore proteins might limit Cid incorporation into ectopic sites[112], a duplicate *Cid* gene that acquired a strong or constitutive promoter would almost certainly be detrimental. Furthermore, in order for a duplicate *Cid* gene to be retained, a series of subfunctionalizing mutations must occur (before pseudogenization of either paralog) such that both paralogs are required for complete Cid function. This model, known as duplication-degeneration-complementation[97], most often refers to mutations in the promoters of duplicate genes. However, the same principle could be applied to mutations in coding regions. Since it is easier to introduce a mutation that results in a non-functional *Cid* gene than a subfunctionalized *Cid*, most *Cid* duplicates probably succumb to pseudogenization early in their evolutionary history and, in *Drosophila,* are quickly lost from the genome[108].

The existence of *Cid* duplications in genetically tractable organisms provides an opportunity to study the multiple functions of a gene that is essential when present in a single copy. While we know a lot about the role of *Cid* in mitosis, its roles in meiosis[59] and inheritance of centromere identity through the germline[23] are less well-characterized. Studying *Cid* paralogs that may have specialized for different functions (*e.g.,* meiosis) may allow for detailed analysis of these underappreciated *Cid* functions without the risk of disrupting essential mitotic functions. Future functional studies can now leverage the insight provided by duplicate *Cid* genes, where evolution and natural selection may have already carried out a 'separation of function' experiment.

2.5 Materials and Methods

Drosophila species and strains

Flies were obtained from the *Drosophila* Species Stock Center at UC-San Diego (https://stockcenter.ucsd.edu) and from the *Drosophila* Stocks of Ehime University in Kyoto, Japan (https://kyotofly.kit.jp/cgi-bin/ehime/index.cgi). For a complete list of species and strains used in this study, see Table 2-3.

Identification of Cid orthologs and paralogs in sequenced genomes

Drosophila Cid genes were identified in previously sequenced genomes using both *D. melanogaster Cid1* and *H3* histone fold domain to query the non-redundant database using tBLASTn[113] implemented in Flybase[114] or NCBI genome databases. Since *Cid* is encoded by a single exon in *Drosophila*, we took the entire open reading frame for each *Cid* gene hit. For annotated genomes, we recorded the syntenic locus (3' and 5' flanking genes) of each *Cid* gene hit as indicated by the Flybase genome browser track. For genomes that were sequenced but not annotated (*D. eugracilis, D. takahashii, D. ficusphila, D. kikkawai* and *P. variegata*), we used the 3' and 5' nucleotide sequences flanking the putative *Cid* open reading frame as a query to the *D. melanogaster* genome using BLASTn. We annotated the syntenic locus according to these *D. melanogaster* matches. Each *Cid* gene was named according to its shared syntenic location. It is worth noting that the Flybase gene prediction for *D. virilis Cid5* (GJ21033) includes a predicted intron but we found no evidence that *Cid5* was spliced in any tissue. The results of all BLAST searches are summarized in Table 2-4.

Table 2-3: List of species and strains used in *Drosophila Cid* evolution analyses.

Stock Center	Strain number	Species	Stock Center	Strain number	Species
SanDiego	15010-0951.00	D. americana	SanDiego	15010-0991.13	D. lacicola
SanDiego	14028-0471.00	D. auraria	SanDiego	15010-1001.11	D. littoralis
EHIME	E-11201	D. auraria	SanDiego	15010-1011.08	D. lummei
EHIME	E-11202	D. auraria	SanDiego	14028-0591.01	D. mayri
EHIME	E-11206	D. auraria	SanDiego	15010-1021.00	D. montana
EHIME	E-11207	D. auraria	SanDiego	15010-1021.06	D. montana
EHIME	E-11208	D. auraria	SanDiego	15010-1021.09	D. montana
EHIME	E-11209	D. auraria	SanDiego	15010-1021.11	D. montana
EHIME	E-11211	D. auraria	SanDiego	15010-1021.13	D. montana
EHIME	E-11212	D. auraria	SanDiego	15010-1021.14	D. montana
EHIME	E-11213	D. auraria	SanDiego	15010-1021.15	D. montana
EHIME	E-11214	D. auraria	SanDiego	15010-1021.16	D. montana
EHIME	E-11215	D. auraria	SanDiego	15010-1021.17	D. montana
EHIME	E-11217	D. auraria	SanDiego	15010-1021.18	D. montana
EHIME	E-11220	D. auraria	SanDiego	15010-1021.19	D. montana
EHIME	E-11221	D. auraria	SanDiego	15010-1021.21	D. montana
EHIME	E-11222	D. auraria	SanDiego	15010-1021.22	D. montana
EHIME	E-11223	D. auraria	SanDiego	15010-1021.23	D. montana
EHIME	E-11224	D. auraria	SanDiego	15010-1021.24	D. montana
EHIME	E-11225	D. auraria	SanDiego	15010-1021.25	D. montana
EHIME	E-11226	D. auraria	SanDiego	15010-1021.26	D. montana
EHIME	E-11229	D. auraria	SanDiego	15010-1021.27	D. montana
EHIME	E-11230	D. auraria	SanDiego	15010-1021.28	D. montana
EHIME	E-11231	D. auraria	SanDiego	15010-1021.29	D. montana
EHIME	E-11233	D. auraria	SanDiego	15010-1021.09	D. montana
EHIME	E-11234	D. auraria	SanDiego	14028-0601.01	D. nikananu
EHIME	NGN-27	D. auraria	SanDiego	15010-1031.00	D. novamexicana
SanDiego	14028-0491.01	D. barbarae	SanDiego	14028-0641.00	D. punjabiensis
SanDiego	14028-0511.00	D. bicornuta	SanDiego	14028-0661.02	D. rufa
SanDiego	14028.0521.00	D. birchii	EHIME	E-14801	D. rufa
SanDiego	14028-0751.00	D. bocki	EHIME	E-14802	D. rufa
SanDiego	15010-0961.00	D. borealis	EHIME	E-14803	D. rufa
SanDiego	14028-0586.00	D. diplacantha	EHIME	E-14805	D. rufa
SanDiego	14026-0451.02	D. eugracilis	EHIME	E-14807	D. rufa
SanDiego	14026-0451.03	D. eugracilis	EHIME	E-14809	D. rufa
SanDiego	14026-0451.04	D. eugracilis	EHIME	E-14810	D. rufa
SanDiego	14026-0451.05	D. eugracilis	EHIME	E-14811	D. rufa
SanDiego	14026-0451.07	D. eugracilis	EHIME	E-14812	D. rufa
SanDiego	14026-0451.08	D. eugracilis	SanDiego	14028-0671.02	D. seguyi
SanDiego	14026-0451.09	D. eugracilis	SanDiego	14028-0681.00	D. serrata
SanDiego	14026-0451.10	D. eugracilis	SanDiego	15010-1041.00	D. texana
SanDiego	15010-0971.00	D. ezoana	SanDiego	14028-0651.00	D. triauraria
SanDiego	15010-0981.04	D. flavomontana	SanDiego	15010-1051.00	D. virilis
SanDiego	14028-0541.00	D. kanapiae	SanDiego	15010-1051.08	D. virilis
SanDiego	15010-1061.00	D. kanekoi	SanDiego	15010-1051.09	D. virilis
SanDiego	14028-0561.00	D. kikkawai	SanDiego	15010-1051.118	D. virilis
SanDiego	15010-1051.48	D. virilis	SanDiego	15010-1051.86	D. virilis
SanDiego	15010-1051.49	D. virilis	Malik Lab	A2	D. virilis
SanDiego	15010-1051.51	D. virilis	SanDiego	14028-0711.00	D. vulcana
SanDiego	15010-1051.52	D. virilis	SanDiego	14028-0531.02	D. watanabei

Table 2-4: Summary of Cid BLAST searches in *Drosophila*.

Database	Species|Chr	Location	Gene name	Expect	Identities
Flybase	dmel|2R	2R:1..25286936	CG13329 (cid)	6.00E-58	107/107 (100%)
Flybase	dsec|scaffold_1	scaffold_1:1..14215200	GM21471 (cid)	2.00E-55	103/106 (97%)
Flybase	dsim|Scf_2R	Scf_2R:1..21544594	GD10968 (cid)	3.00E-55	103/106 (97%)
Flybase	dere|scaffold_4845	scaffold_4845:1..22589142	GG20383 (cid)	3.00E-49	92/105 (87%)
Flybase	dyak|2R	2R:1..21139217	GE12545 (cid)	1.00E-45	87/105 (82%)
Flybase	459201496|gb|KB457491.1| Drosophila ficusphila	unplaced genomic scaffold scf7180000454039	N.A.	4.00E-44	84/104 (80%)
Flybase	452187604|gb|KB452656.1| Drosophila rhopaloa	unplaced genomic scaffold scf7180000780288	N.A.	1.00E-43	85/104 (81%)
Flybase	459204446|gb|KB461257.1| Drosophila takahashii	unplaced genomic scaffold scf7180000415379	N.A.	3.00E-43	85/104 (81%)
Flybase	459197794|gb|KB462718.1| Drosophila biarmipes	unplaced genomic scaffold scf7180000302291	N.A.	2.00E-42	81/102 (79%)
Flybase	459205818|gb|KB465323.1| Drosophila eugracilis	unplaced genomic scaffold scf7180000409787	N.A.	5.00E-41	80/104 (76%)
Flybase	459205844|gb|KB465297.1| Drosophila eugracilis	unplaced genomic scaffold scf7180000409758	N.A.	7.00E-37	75/103 (72%)
Flybase	459200343|gb|KB458640.1| Drosophila elegans	unplaced genomic scaffold scf7180000491282	N.A.	3.00E-40	79/104 (75%)
Flybase	459202931|gb|KB459691.1| Drosophila kikkawai (Cid4)	unplaced genomic scaffold scf7180000302476	N.A.	2.00E-28	65/106 (61%)
Flybase	459203095|gb|KB459527.1| Drosophila kikkawai (Cid1)	unplaced genomic scaffold scf7180000302277	N.A.	4.00E-25	60/102 (58%)
Flybase	459203095|gb|KB459527.1| Drosophila kikkawai (Cid3)	unplaced genomic scaffold scf7180000302277	N.A.	7.00E-25	60/102 (58%)
Flybase	dana|scaffold_13266	scaffold_13266:1..19884421	GF13696(cid)	4.00E-28	60/101 (59%)
Flybase	dper|scaffold_4	scaffold_4:1..7162766	GL17403	3.00E-27	62/108 (57%)
Flybase	459199090|gb|KB464107.1| Drosophila bipectinata	unplaced genomic scaffold scf7180000396384	N.A.	4.00E-27	62/102 (60%)
Flybase	dpse|3	3:1..19787792	GA12208 (cid)	9.00E-27	62/108 (57%)
Flybase	480995219|gb|CM001519.2| Drosophila miranda strain MSH22	chromosome 3	N.A.	9.00E-27	61/101 (60%)
Flybase	dmoj|scaffold_6496 (Cid5)	scaffold_6496:1..26866924	GI21176	2.00E-25	57/102 (55%)
Flybase	dmoj|scaffold_6496 (Cid1)	scaffold_6496:1..26866924	GI18331	1.00E-24	57/100 (57%)
Flybase	dwil|scf2_1100000004382	scf2_1100000004382:1..1405142	GK10722	3.00E-24	59/102 (57%)
Flybase	dvir|scaffold_12875 (Cid5)	scaffold_12875:1..20611582	GJ21033	5.00E-24	54/100 (54%)
Flybase	dvir|scaffold_13324 (Cid1)	scaffold_13324:1..2960039	GJ19757	8.00E-24	56/100 (56%)
Flybase	dgri|scaffold_15112 (Cid5)	scaffold_15112:1..5172618	GH22666	6.00E-24	56/101 (55%)
Flybase	dgri|scaffold_15112 (Cid1)	scaffold_15112:1..5172618	GH23161	2.00E-22	54/100 (54%)
Flybase	gi|405989128|gb|JH859027.1| Drosophila albomicans	Dalb_scaffold_62607	N.A.	6.00E-23	57/102 (55%)
Flybase	gi|405984592|gb|JH863563.1| Drosophila albomicans	Dalb_scaffold_102010	N.A.	1.00E-18	47/105 (44%)
NBCI	Drosophila busckii chromosome chr2R, ASM127793v1	NC_030803.1	LOC108596391	6.00E-21	45/88(51%)
NBCI	Phortica variegata	scaffold1864	N.A.	1.00E-22	51/103 (50%)

Identification of Cid orthologs and paralogs in non-sequenced genomes

Approximately 10 whole (5 male, 5 female) flies were ground in DNA extraction buffer (10mM Tris pH 7.5, 10mM EDTA, 100mM NaCl, 0.5% SDS) with Proteinase K (New England Biolabs). Ground flies were incubated for 2 hrs at 55°C. DNA was extracted using phenol-chloroform (Thermo Fisher Scientific) according to the manufacturers instructions. Primers were designed to amplify each *Cid* paralog based on regions of homology in neighboring genes or intergenic regions. Only *Cid* paralogs that were predicted to be present in the species based on related species sequenced genomes were amplified. All PCRs were performed using Phusion DNA Polymerase (New England Biolabs). Appropriately sized amplicons were gel isolated and cloned into the cloning/sequencing vector pCR-Blunt (Thermo Fisher Scientific) and Sanger sequenced with M13F and M13R primers plus additional primers as needed to obtain sufficient coverage of the locus. A complete list of primers used in this study can be found in Table 2-5. A list of primer pairs used to amplify *Cid* paralogs in non-sequenced genomes can be found in Table 2-6. Sequences obtained in this study have been deposited in Genbank with the following accession numbers: KY212539-KY212710, KY124384-KY124460.

Table 2-5: List of primer sequences used in *Drosophila Cid* evolution analyses.

Name	Sequence	Name	Sequence
LEK044	caccATGCGTCCACGCACTGTAA	LEK310	TCAACTCCAACGCAAACATG
LEK045	TCAAAGATTACCATAGGTTTTGCAG	LEK311	TACCACATGCAGGTCCACAT
LEK047	caccATGAGTCAAGCTAATGCACAGAG	LEK317	GCGGGAAGCCTTTAAAAACAAATAAA
LEK048	TCAGAGTTGTCCGTGCAATTT	LEK318	GCGGGAAGCCTTTAAAAACAAACAAA
LEK055	AAACACCGGATCGTGAAGAC	LEK330	TCTCGCCTAGCTCTTGACAGTT
LEK056	TTCGGGGTGTTTGTCGTACT	LEK336	GGGTAATCAGCTGGCCAACC
LEK057	AACAATTGCAAGGCGAAGAG	LEK337	CCATGCCTTGTCGACAGGAA
LEK058	CTGCAAAGGATACGGCTGTT	LEK352	AAGAAGCGCACCAAGCACT
LEK065	ACGCACATTGTGTACGAGGA	LEK353	ATACCCTTGGGCTTGCGC
LEK066	TGACCATTCGTCCAGCATAC	LEK354	ACGGCTTGGAGTTTACCACC
LEK075	CATGTGGGAATTTCCCAACATA	LEK355	GACGGCGGTCTTTTGGATTG
LEK076	TGCCATTTAAATTGAGTAATCA	LEK356	GTTGGCCATTGAACAACGCT
LEK078	CAATTGGGAATTTATTTGTGA	LEK357	CTGGCGACTCCAAATTTGCC
LEK085	GCTCTGCAAAACCTGTGACG	LEK358	CCATGCAGATGCTCGACCAT
LEK087	GCGCTTAATTTCTTACTCACTGAGT	LEK359	GTGTTCCGATTCGAACGTCG
LEK093	CTGCACTCATATAATGCGCGT	LEK360	CAAGCGTCCGGTAAATGCAC
LEK094	GCACATAAACCTCGCTCCCA	LEK361	CTGGCGTCCTCCAGTGTTAG
LEK101	GCAGTGCGGACAAGAACGG	LEK364	caccATGCGAACACTGAGGGTCAA
LEK102	AGCTCCGTGCCGAACAGCTC	LEK365	CTAGCTGCGGTAATGACTAACG
LEK107	CCAACCGGATGAATCGGGAG	LEK366	caccATTATGGCCGCCCGGC
LEK113	GCTCCTGTTCGGTTGGGTTA	LEK367	CTAGGAACGATGATGACTAACGTCG
LEK114	CAAAAGCTGGGCACATTGCT	LEK368	caccATGAGACCACCACCAAAAAGAG
LEK115	TGCTCATGGTCCTTCTGCTG	LEK369	CTAGTGACCCTGATTGCACAG
LEK116	GATGATGGTGGGTGGCAGAA	LEK370	ATCGACAACAGAGTGCGTCG
LEK123	GCGAAATTAATTATGGAGCGTGG	LEK371	ATCAGCAGCACCTCCAGC
LEK124	TGCTCACATTTTGCATGCG	LEK372	GCCCAAGGGTATCGACAACA
LEK125	GTGTAGTCCTTGCCATGGT	LEK373	ATTCGCGCACATTGTGTACG
LEK126	ATTTGTAGCTCGCTCGGCTC	LEK375	GCATTTCGAACACCGGATCG
LEK127	TTCTAGGCCCGAAAGCGTTC	LEK376	CTCGACGACGGTCTTTTGGA
LEK128	ACTCCAACGCAAACATGCAC	LEK377	TCGAACTGATGCACAAGGCT
LEK129	GTTGCACGGAACCATTGCTT	LEK378	CAGCGTTGTTCAATGGCCAA
LEK130	TCAGCAACATGTTCAAGCGC	LEK379	CAGGGTCAAGGAATCGCAGA
LEK134	AAAGTGCATTCCAAACGCCG	LEK380	GCGTCGTAGAGCTGCAATTG
LEK135	CGTGAAGAGTTCTTGCGCAG	LEK381	CCACAGAGGATCAGGACGC
LEK136	CGCAAGCAACAGAAGCACTT	LEK382	CCATCAGACGGCTAGTGGAG
LEK137	GTTCTTGTGGACACGGCAAC	LEK383	ACCTTCTTTGGGAGACGTCG
LEK138	GCGGACGTGCTCAAAAAGTT	LEK384	TGGCGTCCTGTAGTGTAAGC
LEK202	GCAGAGTGCACAGAACTCG	LEK385	ATGCGAACACTGAGGGTCAA
LEK204	GTCGTCCTGGTTTTCCTCGT	LEK386	GCATCTTGCAGCATGAGTCG
LEK272	TTGGTTTTCATCACGTTACCCG	LEK387	TACGGCCTAAATTTCTCCAC
LEK281	CCCTACASTGCACTGCAAGAAA	LEK388	TTTGATCAGGCGGTGGCT
LEK282	AGTTCACCGAGAATTGTGCG	LEK389	CGTTCGGTCCATGTACCCAA
LEK283	CTAGGGACAGAGTTCACCGAG	LEK390	GGCGTCCTGAAGTGTAAGCT
LEK284	GCGTGGCAGTGAGACCCTAC	PL001	CGATAAGCTCGTGCTGCATGC
LEK285	TGCCACAGCTATTGCCAATG	PL002	GCATCGATCTATTCAGAATACCG
LEK287	GCAAGGAGCGCCAATAATGT	PL003	AGGATTCAATTTCGACAAGTAATGC
		PL004	GTGTAATAAATGATCTATGTAATAGAC

Table 2-6: List of primer pairs for sequencing *Cid* genes.

Species	Subgrou	Gene	Primers	Species	Subgroup	Gene	Primers
D. auraria	montium	Cid1	LEK125/LEK126	D. seguyi	montium	Cid4	LEK130/LEK128
D. barbarae	montium	Cid1	LEK125/LEK126	D. serrata	montium	Cid4	LEK130/LEK128
D. bicornuta	montium	Cid1	LEK125/LEK126	D. triauraria	montium	Cid4	LEK130/LEK128
D. birchii	montium	Cid1	LEK125/LEK126	D. vulcana	montium	Cid4	LEK287/LEK285
D. bocki	montium	Cid1	LEK101/LEK102	D. watanabei	montium	Cid4	LEK130/LEK128
D. diplacantha	montium	Cid1	LEK125/LEK126	D. americana	virilis	Cid1	PL001/PL002
D. kanapiae	montium	Cid1	LEK125/LEK126	D. borealis	virilis	Cid1	PL001/PL002
D. mayri	montium	Cid1	LEK125/LEK126	D. ezoana	virilis	Cid1	LEK085/LEK087
D. nikananu	montium	Cid1	LEK125/LEK126	D. flavomontana	virilis	Cid1	LEK085/LEK087
D. punjabiensis	montium	Cid1	LEK125/LEK127	D. kanekoi	virilis	Cid1	LEK085/LEK087
D. rufa	montium	Cid1	LEK125/LEK126	D. lacicola	virilis	Cid1	PL001/PL002
D. seguyi	montium	Cid1	LEK101/LEK107	D. littoralis	virilis	Cid1	PL001/PL002
D. serrata	montium	Cid1	LEK125/LEK126	D. lummei	virilis	Cid1	PL001/PL002
D. triauraria	montium	Cid1	LEK125/LEK126	D. montana	virilis	Cid1	PL001/PL002
D. vulcana	montium	Cid1	LEK125/LEK126	D. novamexicana	virilis	Cid1	PL001/PL002
D. watanabei	montium	Cid1	LEK101/LEK102	D. texana	virilis	Cid1	PL001/PL002
D. auraria	montium	Cid3	LEK123/LEK124	D. americana	virilis	Cid5	PL003/PL004
D. barbarae	montium	Cid3	LEK114/LEK115	D. borealis	virilis	Cid5	LEK078/LEK076
D. bicornuta	montium	Cid3	LEK123/LEK124	D. ezoana	virilis	Cid5	PL003/PL004
D. birchii	montium	Cid3	LEK114/LEK115	D. flavomontana	virilis	Cid5	PL003/PL004
D. bocki	montium	Cid3	LEK284/LEK283	D. kanekoi	virilis	Cid5	PL003/PL004
D. diplacantha	montium	Cid3	LEK123/LEK124	D. lacicola	virilis	Cid5	PL003/PL004
D. kanapiae	montium	Cid3	LEK281/LEK283	D. littoralis	virilis	Cid5	PL003/PL004
D. mayri	montium	Cid3	LEK123/LEK124	D. lummei	virilis	Cid5	PL003/PL004
D. nikananu	montium	Cid3	LEK114/LEK115	D. montana	virilis	Cid5	PL003/PL004
D. punjabiensis	montium	Cid3	LEK114/LEK115	D. novamexicana	virilis	Cid5	PL003/PL004
D. rufa	montium	Cid3	LEK123/LEK124	D. texana	virilis	Cid5	PL003/PL004
D. seguyi	montium	Cid3	LEK114/LEK115	D. auraria	montium	Cid1	LEK364/LEK365
D. serrata	montium	Cid3	LEK113/LEK116	D. auraria	montium	Cid3	LEK366/LEK367
D. triauraria	montium	Cid3	LEK123/LEK124	D. auraria	montium	Cid4	LEK368/LEK369
D. vulcana	montium	Cid3	LEK123/LEK124	D. virilis	virilis	Cid1	LEK044/LEK045
D. watanabei	montium	Cid3	LEK114/LEK115	D. virilis	virilis	Cid5	LEK047/LEK048
D. auraria	montium	Cid4	LEK130/LEK128				
D. barbarae	montium	Cid4	LEK317/LEK310				
D. bicornuta	montium	Cid4	LEK130/LEK128				
D. birchii	montium	Cid4	LEK317/LEK310				
D. bocki	montium	Cid4	LEK318/LEK310				
D. diplacantha	montium	Cid4	LEK130/LEK128				
D. kanapiae	montium	Cid4	LEK318/LEK310				
D. mayri	montium	Cid4	LEK130/LEK128				
D. nikananu	montium	Cid4	LEK130/LEK128				
D. punjabiensis	montium	Cid4	LEK130/LEK128				
D. rufa	montium	Cid4	LEK318/LEK311				

Phylogenetic analyses

Cid sequences were aligned using the ClustalW[115] 'translation align' function in the Geneious software package (version 6)[116]. Alignments were further refined manually, including removal of gaps and poorly aligned regions. Maximum likelihood phylogenetic trees of *Cid* nucleotide sequences were generated using the HKY85 substitution model in PhyML, implemented in Geneious, using 1000 bootstrap replicates for statistical support. Neighbor-joining trees correcting for multiple substitutions were generated using CLUSTALX[115]. We used the GARD algorithm implemented at datamonkey.org to examine alignments for evidence of recombination[89]. Pairwise percent identity calculations were made in Geneious. Phylogenies were visualized using FigTree (http://tree.bio.ed.ac.uk/software/figtree/) or Dendroscope[117].

Cloning Cid fusion proteins

Cid genes from *D. auraria (Cid1, Cid3* and *Cid4)* and *D. virilis (Cid1* and *Cid5)* were amplified from genomic DNA and cloned into pENTR/D-TOPO (ThermoFisher). We used LR clonase II (ThermoFisher) to directionally recombine each *Cid* gene into a destination vector from the Drosophila Gateway Vector Collection, generating either N-terminal Venus (pHVW) or 3XFLAG (pHFW) fusion under the control of the *D. melanogaster* heat-shock promoter.

Cell culture

Cell lines (*D. auraria* cell line ML83-68 and *D. virilis* cell line WR DV-1) were obtained from the Drosophila Genomics Resource Center in Bloomington, Indiana (https://dgrc.bio.indiana.edu). *D. auraria* cells were grown at room temperature in M3+BPYE + 12.5%FCS and *D. virilis* cells were grown in M3+BPYE + 10%FCS.

Transfection experiments

2 micrograms plasmid DNA was transfected using Xtremegene HP transfection reagent (Roche) according to the manufacturer's instructions. Cells were heat-shocked at 37°C for one hour 24 hours after transfection to induce expression of the *Cid* fusion protein.

Antibody staining and imaging

Cells were transferred to a glass coverslip 48 hours after heatshock. Cells were treated with 0.5% sodium citrate for 10 min and then centrifuged on a Cytospin III (Shandon) at 1900rpm for 1 min to remove cytoplasm. Cells were fixed in 4% PFA for 5 min and blocked with PBSTx (0.3% Triton) plus 3% BSA for 30 minutes at room temperature. Coverslips with cells were incubated with primary antibodies at 4°C overnight at the following concentrations: mouse anti-FLAG (Sigma F3165) 1:1000, chicken anti-GFP (Abcam AB13970) 1:1000, rabbit anti-CENP-C (gift from Aaron Straight) 1:1000. Coverslips with cells were incubated with secondary antibodies for 1 hour at room temperature at the following concentrations: goat anti-rabbit (Invitrogen Alexa Fluor 568, A-11011) 1:2000, goat anti-chicken (Invitrogen Alexa Fluor 488, A-11039) 1:5000, goat anti-mouse (Invitrogen Alexa Fluor 568, A-11031) 1:2000. Images were acquired from the Leica TCS SP5 II confocal microscope with LASAF software.

Expression analyses

RNA was extracted from *D. auraria* cell line ML83-68 and *D. virilis* cell line WR DV-1 using the TRIzol reagent (Invitrogen) according to the manufacturers instructions. To investigate expression profiles in adult tissues, RNA was extracted from whole bodies, and dissected tissues (heads, germline and the remaining carcasses) from *D. auraria, D. rufa, D. kikkawai, D. virilis, D. montana* and *D. mojavensis* flies. All samples were DNase treated (Ambion) and then used for cDNA synthesis (SuperScript III, Invitrogen). During cDNA synthesis, a 'No RT' control

was generated for each RNA extraction in which the reverse transcriptase was excluded from the reaction. For RT-PCR experiments, the presence of genomic DNA contamination was ruled out by performing PCR that amplified the housekeeping gene, *Rp49,* on each cDNA sample as well as each 'No RT' control. 25- (not shown) and 30-cycle PCRs were performed with primers specific to each *Cid* paralog and samples were run on an agarose gel for visualization. RT-qPCR was performed according to the standard curve method using the Platinum SYBR Green reagent (Invitrogen) and primers designed to each *Cid* paralog and to *Rp49*. Reactions were run on an ABI QuantStudio 5 qPCR machine using the following conditions: 50°C for 2 min, 95°C for 2 min, 40 cycles of (95°C for 15s, 60°C for 30s). We ensured that all primer pairs had similar amplification efficiencies using a dilution series of genomic DNA. Three technical replicates were performed for each cDNA sample. Transcript levels of each gene were normalized to *Rp49.* For all primers used in RT-PCR and RT-qPCR experiments, see Table 2-5 and Table 2-7.

Table 2-7: Primer pairs used in expression analyses.

Species	Subgroup/Group	Gene	Primers
D. virilis	virilis	Rp49	LEK065/LEK066
D. auraria	montium	Rp49	LEK370/LEK371
D. virilis	virilis	Cid1	LEK055/LEK056
D. virilis	virilis	Cid5	LEK057/LEK058
D. auraria	montium	Cid1	LEK379/LEK380
D. auraria	montium	Cid3	LEK381/LEK382
D. auraria	montium	Cid4	LEK383/LEK384
D. auraria	montium	Rp49	LEK370/LEK371
D. auraria	montium	Cid1	LEK379/LEK380
D. auraria	montium	Cid3	LEK381/LEK382
D. auraria	montium	Cid4	LEK383/LEK384
D. rufa	montium	Rp49	LEK370/LEK371
D. rufa	montium	Cid1	LEK385/LEK386
D. rufa	montium	Cid3	LEK387/LEK388
D. rufa	montium	Cid4	LEK389/LEK390
D. kikkawai	montium	Rp49	LEK370/LEK371
D. kikkawai	montium	Cid1	LEK358/LEK359
D. kikkawai	montium	Cid3	LEK202/LEK204
D. kikkawai	montium	Cid4	LEK360/LEK361
D. virilis	virilis	Rp49	LEK352/LEK353
D. virilis	virilis	Cid1	LEK354/LEK355
D. virilis	virilis	Cid5	LEK356/LEK357
D. montana	virilis	Rp49	LEK372/LEK373
D. montana	virilis	Cid1	LEK375/LEK376
D. montana	virilis	Cid5	LEK377/LEK378
D. mojavensis	repleta	Rp49	LEK372/LEK373
D. mojavensis	repleta	Cid1	LEK134/LEK135
D. mojavensis	repleta	Cid5	LEK136/LEK137

Motif analyses

Motifs were identified in six different groups of Cid proteins (melanogaster group Cid1s, montium group Cid1, montium group Cid3s, montium group Cid4s, virilis group Cid1s and virilis group Cid5s) using the motif generator algorithm MEME[103] implemented on meme-suite.org. Several motifs identified in different groups were similar to one another. For example, the motif "TDYLEFTTS" appeared in *melanogaster* group Cid1s, *montium* subgroup Cid3s and Cid4s and *virilis* group Cid1s. To determine which motifs were the same, we used the motif search algorithm MAST[104] to search for the top four motifs from each group against all 86 sequences used for motif generation. In total, we found 10 unique motifs (Figure 2-10). The only instance in which the motifs were not totally independent was for motif 2 and motif 9. Motif 2 was contained within motif 9, but motif 9 was significantly longer than motif 2 so we considered it to be an

independent motif. We mapped all 10 motifs to the *Cid* genes in the six groups plus *D. eugracilis* Cid2, *D. mojavensis* and *D. grimshawi Cid1* and *Cid5, D. busckii,* and the outgroup species *P. variegata* Cid1. We considered a motif to be present in a given protein if the MAST p-value was $< 10^{-5}$.

Positive selection analyses

We used the PAML suite of programs[118] to test for positive selection on each *Cid* paralog across deep evolutionary time. Alignments for each *Cid* paralog were generated and manually refined as described above. Alignments and *Cid* gene trees were used as input into the CODEML NSsites model of PAML. To determine whether each *Cid* paralog evolves under positive selection, we compared two models that do not allow dN/dS to exceed 1 (M7 and M8a) to a model that allows dN/dS > 1 (M8). Positively selected sites were classified as those sites with a M8 Bayes Empirical Bayes posterior probability > 95%. We used the McDonald Kreitman (MK) test[119] to look for more recent positive selection at the population level. To implement the MK test for *montium* subgroup *Cid* paralogs we compared *Cid* sequences in 26 strains of *D. auraria* to 10 strains of *D. rufa.* In the *virilis* group, we compared *Cid* sequences in 10 strains of *D. virilis* to 20 strains of *D. montana.*

Chapter 3. Mutually exclusive gametic retention of centromeric histone protein paralogs in *Drosophila virilis*

3.1 Abstract

A single gene can play different roles in males and females. Selection to enhance fitness in males may displace females from their phenotypic fitness optimum and *vice versa*. If both roles are equally important for fitness, a single gene can become 'trapped' for suboptimal function, creating intralocus conflict. Gene duplication followed by specialization could resolve intralocus conflict. Identifying functional specialization between paralogs may provide a means to detect intralocus conflict. We previously discovered multiple, independent gene duplications of the centromeric histone gene, *Cid*, during *Drosophila* evolution. Here, we analyzed the cytological localization of two Cid protein paralogs - Cid1 and Cid5 - in *D. virilis* using specific antibodies and epitope-tagged transgenic strains. We find that Cid1 is detected at centromeres in somatic cells but Cid5 is not. However, both Cid1 and Cid5 are found at centromeres of germ cells in testes and ovaries. Intriguingly, Cid1 is lost in male meiosis but retained throughout oogenesis while Cid5 is lost during female meiosis but retained in mature sperm. Following fertilization, Cid1 rapidly replaces Cid5 during the protamine-to-histone transition. Our studies reveal mutually exclusive gametic specialization of two divergent CenH3 paralogs. We suggest that centromeric histone divergence and acquisition of new protein motifs following gene duplication may allow essential genes involved in chromosome segregation to resolve intralocus conflict and evolve specialized roles.

3.2 Introduction

Sexual genetic conflict arises because of the divergent reproductive strategies of males and females. Classically, sexual conflict plays out as a tit-for-tat arms race between a genetic locus in males and an independent locus in females (interlocus conflict). The male locus evolves to enhance male reproduction, but at the expense of female fitness. This phenomenon places pressure on the female locus to evolve counter-adaptations that restore female fitness[120]. Interlocus conflict has likely driven the evolution of striking reproductive traits including spermataphore spikes and corkscrew genitalia[121,122]. Sexual genetic conflict can also take place within a single genetic locus (intralocus conflict). Much like interlocus conflict, in intralocus sexual conflict, selection on a locus to enhance fitness in males displaces females from their phenotypic fitness optimum and *vice versa*[56]. In this scenario, a single locus plays a different role in males than in females. If both roles are equally important for fitness, a single gene can become 'trapped' for suboptimal function in both roles.

One way that intralocus conflict can be resolved is through gene duplication and specialization[57]. Resolution of intralocus conflict has been invoked to explain the high rate of retention of testis-specific gene duplicates that carry out mitochondrial function[55]. Even though somatic and testis mitochondrial functions are similar, they have different fitness optima. For example, mitochondrial function in sperm might be driven by selection for faster-swimming sperm even at the expense of a higher mutation rate of mitochondrial DNA. A high mitochondrial mutation rate is less consequential in the testis because sperm mitochondria are not passed on to the next generation[123]. However, a high mitochondrial mutation rate would be highly deleterious for somatic tissues and for the female germline. Gene duplications could allow organisms to achieve optimal mitochondrial function simultaneously in somatic tissues and testes[57].

Similar pressures to mitigate intralocus conflict have also driven the retention and specialization of two evolutionarily young paralogs in *Drosophila melanogaster, Apollo and Artemis,* that diverged only 200,000 years ago[124]. *Apollo* is essential for male fertility but detrimental to female fertility, whereas *Artemis* is essential for female fertility, but deleterious to male fertility. To preserve their essential roles but alleviate their deleterious effects, *Apollo* and *Artemis* have dramatically diverged in their expression patterns, being almost exclusively expressed in *D. melanogaster* testes and ovaries, respectively. Thus, intralocus conflict resolution can result in retention and functional specialization of duplicate genes. Identifying cases of functional specialization of paralogs may reveal hidden intralocus conflict, even in genes where such a functional trade-off was not previously anticipated.

Motivated by our recent discovery of multiple centromeric histone (*CenH3)* gene duplication events in Drosophila[125], we considered whether CenH3 genes might also be subject to intralocus conflict. CenH3 is the foundational centromeric protein in most eukaryotes[126,127]. First identified as Cenp-A in mammals[21], CenH3 localizes to centromeric DNA and helps recruit other components of the kinetochore, which mediates chromosome segregation. The loss of CenH3 results in catastrophic chromosome segregation defects and lethality in protists, yeast, flies, nematodes, mice, and plants[32-34,128].

Even though the process of chromosome segregation is highly conserved across eukaryotes, CenH3 could be subject to intralocus conflict in sexual, multicellular organisms. Despite its essential function, CenH3 evolves rapidly in many taxa of plants and animals[36,54,83]. We proposed that this rapid evolution is the result of 'centromere drive'[129-131], in which centromeres compete during female meiosis for inclusion in the egg rather than the polar bodies. Centromeric proteins like CenH3 must evolve to suppress any deleterious consequences of centromere drive. Despite increasing organismal fitness by suppressing centromere drive, the rapid evolution of CenH3 is likely to lead to incompatibilities with other centromeric components. Indeed, centromeric incompatibilities can result in postzygotic

reproductive isolation and hybrid inviability even between closely related species of plants[48,132] and frogs[133]. Therefore, rapid evolution of CenH3 for its drive suppressor function may be at odds with its essential role as a cell division protein, creating intralocus conflict. CenH3 duplications may resolve this conflict by allowing one paralog to rapidly evolve for centromere-drive suppression while keeping the other paralog constant to maintain essential centromeric function.

Similarly, CenH3's function in centromere inheritance through the male germline may be at odds with its function in cycling mitotic or meiotic cells. During spermiogenesis, the spermatid nucleus undergoes a dramatic chromatin repackaging process where nearly all of the histones are removed and replaced by sperm nuclear basic proteins (SNBPs), or protamines[134]. However, in *Drosophila melanogaster*, CenH3 retention on the paternal genome is essential to maintain the centromeric competence of the paternal genome in the early embryo[23]. If sperm do not inherit proper amounts of CenH3, the paternal genome lacks functional centromeres and fails to attach to the mitotic spindle, resulting in early embryonic lethality. In this scenario, selection would favor CenH3 changes that resist eviction during the protamine transition, whereas in somatic cells, it may favor properties of CenH3-nucleosomes that more easily allow eviction upon misincorporation into non-centromeric regions. CenH3's specialized roles in mitosis, meiosis and sperm development are multiple functions that may be at odds with one another.

Some CenH3 paralogs do show signs of tissue-specific specialization in plants. For example, knockdown of one CenH3 paralog in wheat causes growth defects whereas knockdown of the other paralog causes reproductive defects[67]. However, the molecular genetic basis of this specialization is unclear. Moreover, centromeric histone specialization has not been previously observed in animal species. Although an estimated 10% of plant genomes harbor multiple CenH3 paralogs[48,72,75], CenH3 duplications were previously thought to be rare in animals[76,77]. Contrary to this view, based on surveys of 45 Drosophila species and phylogenetic

inference, we recently proposed that the majority *Drosophila* species encode more than one CenH3 (*Cid* in *Drosophila*) gene[125]. Moreover, our initial characterization based on RNA expression suggested that *Cid* paralogs might have independently acquired specialized germline centromeric function; for example, in *D. virilis*, *Cid1* is ubiquitous whereas *Cid5* is specifically expressed in the testes[125].

In this report, we performed cytological analyses of Cid1 and Cid5 in *D. virilis* somatic cells, testes, ovaries and early embryos. Surprisingly, we found that there is mutually exclusive retention of the two Cid proteins in mature male and female gametes, which is achieved by alternate protein loss during meiosis in males and females. We hypothesize that paralog-specific changes in the N-terminal domain have allowed for the functional specialization of Cid1 and Cid5, and may resolve intralocus conflict underlying centromeric function.

3.3 Results

Tools for cytological analysis of Cid paralogs in D. virilis

D. virilis encodes two *Cid* paralogs, *Cid1* and *Cid5*, which diverged nearly 40 million years ago in the Drosophila subgenus[125]. In a previous survey, we found that both paralogs are present in 14/14 examined Drosophila subgenus species, including *D. mojavensis* and *D. grimshawi,* which are two divergent outgroup species to the virilis species group. The ancient retention of *Cid1* and *Cid5* suggests that both paralogs perform important functions[125]. In order to gain insight into the function of Cid1 and Cid5, we investigated their localization in dividing somatic cells, ovaries, and testes of *D. virilis* flies. For this approach, we developed tools to visualize Cid1 and Cid5 *in vivo*. We exploited the high divergence of their N-terminal tails to develop polyclonal antibodies specific to either Cid1 or Cid5 that are non-homologous between the two paralogs and also distinct from other predicted *D. virilis* proteins (Figure 3-1A). We confirmed that these antibodies were specific to each paralog (Figure 3-1B - Figure 3-1C).

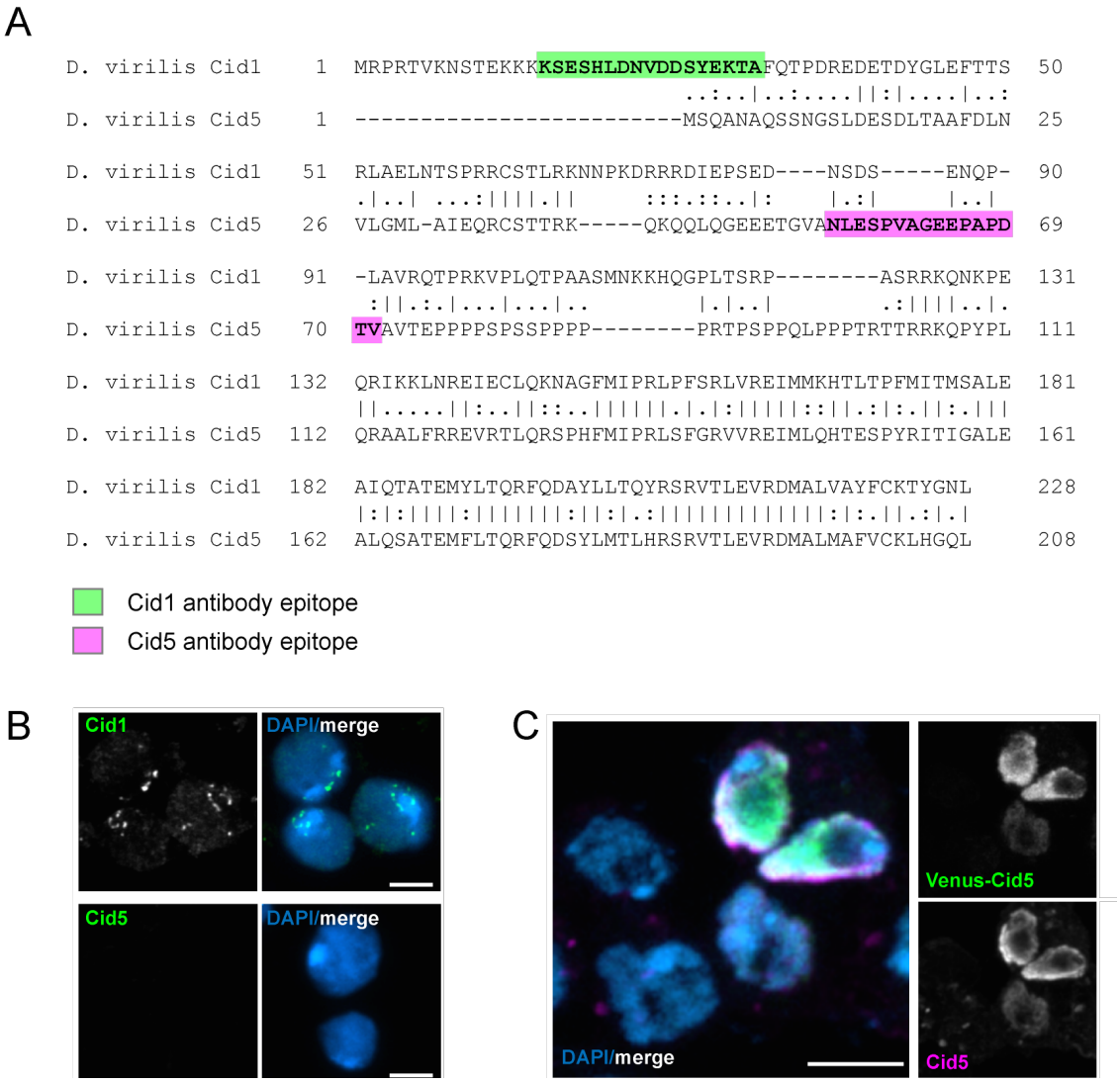


Figure 3-1. Cid1 and Cid5 antibody validation. (A) Protein alignment of *D. virilis* Cid1 and Cid5. Peptides used to generate polyclonal antibodies are highlighted in green (Cid1) and magenta (Cid5). (B) Immunostaining of *D. virilis* larval brain cells with Cid1 (top) or Cid5 (bottom) antibodies. Scale bar = 3μm (C) Image of *D. melanogaster* KC cells with Venus-Cid5 overexpressed under the control of the *D. melanogaster* heat shock promoter. Venus and Cid5 antibody signals are shown in two of the five nuclei. Scale bar = 5μm.

Since antibody occlusion could hamper cytological analyses especially in the male germline[135], we also generated transgenic *D. virilis* flies with *Cid1GFP* or *Cid5mCherry* under the control of their respective native promoters. In *D. melanogaster,* transgenic flies with *Cid-GFP*, where GFP was inserted between the N-terminal tail and histone fold domain, can

complement Cid function[136]; therefore we inserted the fluorescent protein tag between the N-terminal tail and the histone fold domain in both the *D. virilis* Cid1 and Cid5 transgenes.

Cid1, but not Cid5, is detectable in tissue culture cells and larval neuroblasts

Our previous expression analyses[125] found that *Cid1* RNA but not *Cid5* RNA is expressed in somatic cells including *D. virilis* tissue culture cells (WR-Dv-1, derived from first instar larvae), heads from male and female *D. virilis* flies and male and female carcasses (de-capitated, gonad-ectomized, animals). To examine protein expression, we looked for Cid1 and Cid5 protein in two types of dividing somatic cells, tissue culture cells and larval neuroblasts. In *D. virilis* tissue culture cells, we could detect endogenous Cid1 protein by both western blot and immunofluorescence analyses (Figure 3-2A, Figure 3-2B). However, we did not detect Cid5 via either method (Figure 3-2A, Figure 3-2C), consistent with our previous finding that *Cid5* is not expressed in these cells[125].

Next, we examined Cid1 and Cid5 localization in larval neuroblasts, a tissue that is enriched with mitotic cells. As expected, we found that Cid1 localized to centromeres in interphase cells and on condensed metaphase chromosomes (Figure 3-2D, Figure 3-2E). As *D. virilis* chromosomes are acrocentric (have their centromeres at one end, next to the telomere) the Cid1 signal was localized to one end of each condensed chromosome. In contrast, we could not detect any Cid5 signal (Figure 3-2D, Figure 3-2E). Our cytological findings using transgenes were confirmed by detection using polyclonal antibodies, reinforcing the validity of our transgene analyses (Figure 3-1C).

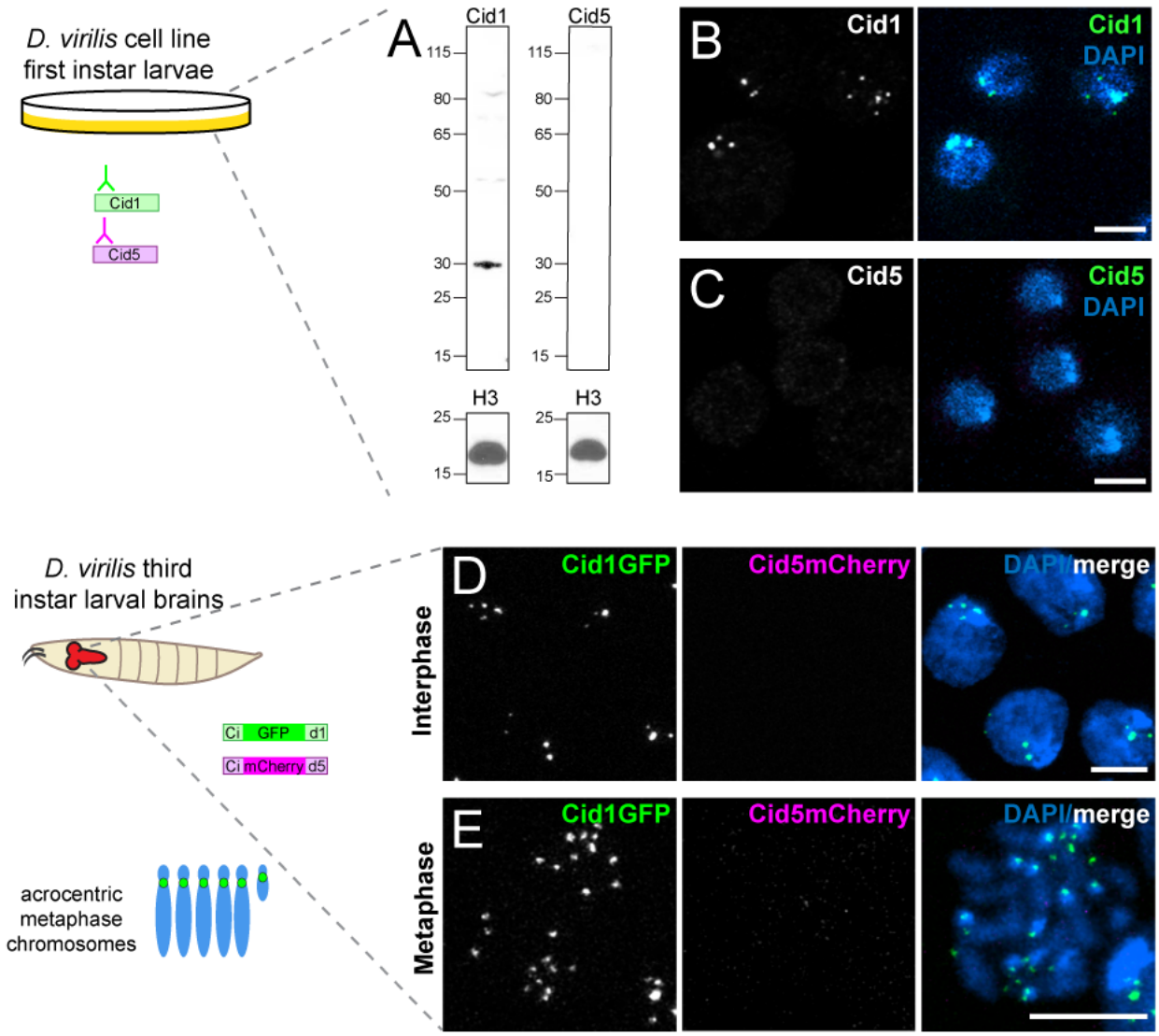


Figure 3-2. Cid1 is the centromeric histone in two dividing somatic cell types. (A) Western blot of Cid1 and Cid5 in *D. virilis* tissue culture cells. A western blot for histone H3 was used as a loading control. This western blot was repeated three times with the same result. (B - C) Immunofluorescence images of *D. virilis* tissue culture cells stained with Cid1 (B) or Cid5 (C) antibodies. (D - E) Images of interphase (D) or metaphase (E) cells from *D. virilis* larval brains dissected from flies containing Cid1GFP and Cid5mCherry transgenes. Scale bar = 5μm.

Differential localization of Cid1 and Cid5 in D. virilis ovaries

Using the same approaches, we also investigated Cid1 and Cid5 protein localization in *D. virilis* ovaries. The *Drosophila* ovary is made up of about 16 ovarioles. At the anterior tip of

each ovariole, germline stem cells divide four times to produce a cyst of 16 interconnected cells, which differentiate into 15 nurse cells (support cells that provide mRNA, protein and other material to the oocyte via a shared cytoplasm) and one oocyte. These interconnected germ cells are surrounded by somatic follicle cells and together form an egg chamber. Egg chamber maturation occurs progressively along the ovariole in a series of defined stages. These stages are referred to as stages 1-14 based on growth and organization of somatic and germline cells. By stage 2, the oocyte has entered into meiotic prophase and reaches pachytene. At stage 5, the oocyte enters primary arrest and remains arrested until stage 13 when the oocyte progresses to secondary arrest in meiosis I metaphase. In the stage 14 egg chamber, no nurse cell nuclei remain and the oocyte is prepared for ovulation[137,138].

Our previous study showed that *Cid1* transcripts were abundant but *Cid5* transcripts were not detectable in RNA extracted from whole ovaries[125]. Moreover, recent work by Teixeira *et al.* examined the expression of *Cid6* (the *Cid1* homolog) and *Cid5* in *D. buzzatii* embryos, pupae, larvae, and adult males and females via RNAseq and found that transcription of *Cid5* was limited to pupae and adult males but *Cid6* was expressed at high levels in all tissues[139]. Based on these characterizations, we expected to find that Cid1 would be the only Cid paralog detectable in *D. virilis* ovaries.

To examine Cid1 and Cid5 protein expression, we used Cid1GFP and Cid5mCherry (Figure 3-3) transgenic flies and the new and the Cid1 and Cid5 antibodies (Figure 3-4) for localization of the both proteins in somatic and germline cells at different stages of egg chamber development. Like in mitotically dividing somatic cells, we detected Cid1 but not Cid5 in somatic follicular cells (Figure 3-3A, Figure 3-3B). However, we were surprised to find that both Cid1 and Cid5 protein were robustly detected in the germline lineage cells of the ovary in egg chamber stages 2 – 9 (Figure 3-3A), including in nurse cells (Figure 3-3C) and the oocyte nucleus (Figure 3-3D). We similarly detected Cid1 and Cid5 protein in germline cells at these stages using and the Cid1 and Cid5 antibodies (Figure 3-4). However, by stage 14, we could only detect Cid1 at

centromeres of metaphase I arrested chromosomes (Figure 3-3E, Figure 3-4G). We note that neither Cid1 nor Cid5 were detectable in Stage 14 oocyte nuclei via antibody staining (Figure 3-4G, Figure 3-4H), likely due to poor antibody accessibility of the oocyte at this stage.

Taken together, these results suggest that both Cid1 and Cid5 are present at centromeres early in oogenesis but only Cid1 remains on centromeres by meiosis I metaphase arrest. Given that removal of centromeric CenH3-containing nucleosomes typically occurs via dilution during DNA replication[140-142], and that bulk DNA replication is completed by these stages, we hypothesize that Cid5 protein is actively removed from the oocyte centromeres and at the onset of meiosis I metaphase arrest. Given that Cid1 is always present throughout oogenesis, it is unclear whether Cid5 performs any function, centromeric or otherwise, in the female germline. However, it is apparent that Cid1 is the only detectable centromeric histone in stage 14 oocytes and is therefore likely to be essential for female fertility and early embryonic mitotic divisions following fertilization (also see below).

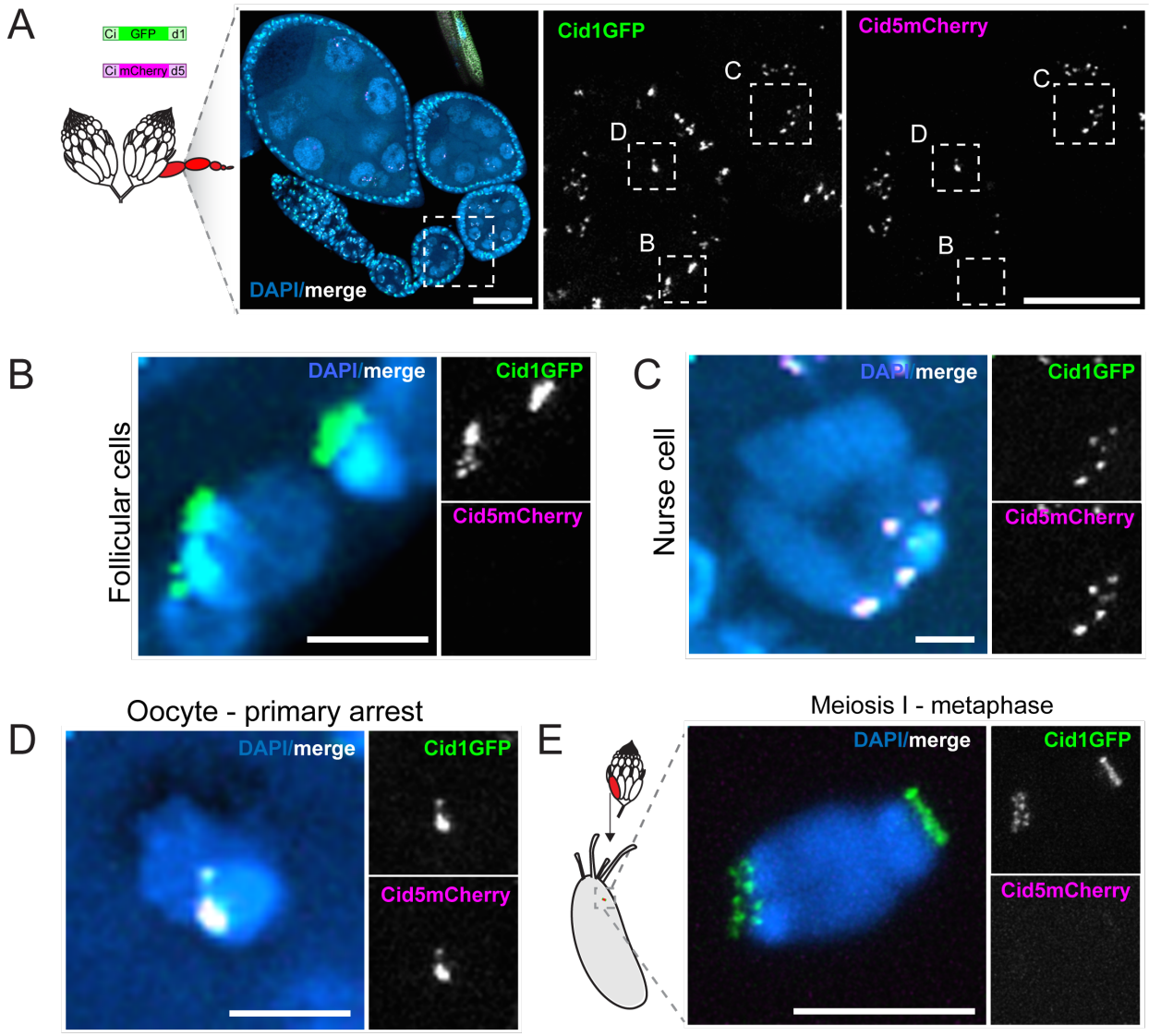


Figure 3-3. Differential localization of Cid1 and Cid5 in ovaries. All images use Cid1GFP and Cid5mCherry to detect Cid1 and Cid5 protein. (A) A whole *D. virilis* ovariole. The two right panels show the boxed region of the left panel at higher magnification. Regions surrounded by boxes in the right two panels are shown at high magnification in subsequent panels. Scale bar = 45μm in left panel, 20μm in the right panel. (B) High magnification image of follicular cells (somatic) from a stage 3 egg chamber. (C) High magnification image of a nurse cell from a stage 4 egg chamber. (D) High magnification image of a stage 3 oocyte in primary arrest. (E) Image of a stage 14 oocyte nucleus in meiosis I metaphase arrest. Scale bars in (B) – (E) = 5μm.

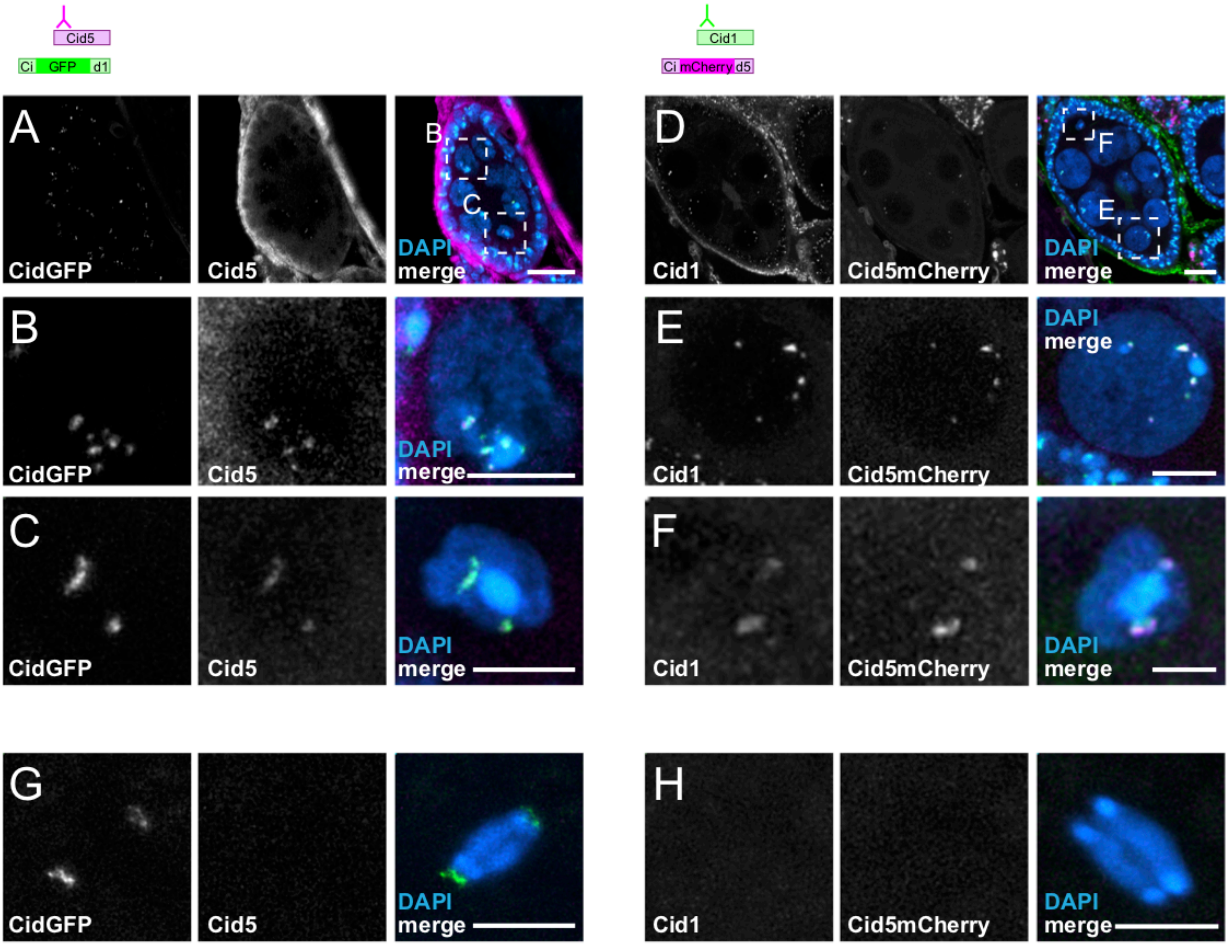


Figure 3-4. Localization of Cid1 and Cid5 in ovaries using antibodies. Panels (A) – (C) and (G) use Cid1GFP and the Cid5 antibody to visualize Cid1 and Cid5 protein. Panels (D) – (F) and (H) use the Cid1 antibody and Cid5mCherry to visualize Cid1 and Cid5 protein. The schematics at the top of the figure reflect which protein detection methods were used. (A) A stage 8 egg chamber. Regions boxed in the right panel of (A) are shown at high magnification in subsequent panels. (B) High magnification image of a nurse cell from the stage 8 egg chamber boxed in (A). (C) High magnification image of the stage 8 oocyte boxed in (A). (D) A stage 8 egg chamber. (E) High magnification image of a nurse cell from the stage 8 egg chamber boxed in (D). (F) High magnification image of the stage 8 oocyte boxed in (D). (G) and (H) Image of a stage 14 oocyte nucleus in meiosis I metaphase arrest. Scale bars in (A) and (D) = 10μm. Scale bars in all other panels = 5μm.

Differential localization of Cid1 and Cid5 in D. virilis testes

Our previous characterization of *Cid1* and *Cid5* mRNA expression in *D. virilis*[125] indicated that both *Cid* paralogs are expressed in testes. We, therefore, examined the cytological localization patterns of Cid1 and Cid5 in the *D. virilis* male germline. In the *Drosophila* male germline, spermatogenesis begins at the apical tip of the testis where the germline stem cells reside. The asymmetric divisions of the germline stem cells replenish the stem cell population and produce gonialblasts. These gonialblasts divide mitotically with incomplete cytokinesis and then enter an extended meiotic prophase. Following this extended period of cell growth, cysts of 16 spermatocytes undergo meiosis and produce bundles of 64 haploid spermatids[143,144]. These spermatids then go through the process of nuclear remodeling resulting in 200-fold compaction of their nuclear volume[144]. During this dramatic nuclear reorganization, nearly all of the histones are removed and are replaced by sperm nuclear basic proteins (SNBPs)[61]. Finally, elongated spermatid bundles go through individualization to produce mature sperm (Figure 3-5A, Figure 3-5B).

Previous studies in *D. melanogaster* have shown that Cid is essential for the mitotic and meiotic divisions in the male germline[59]. Moreover, Cid has also been shown to be critical for transgenerational centromere inheritance through the mature sperm[23]. Therefore, we examined the cytological localization of Cid1 and Cid5 in the mitotic zone, meiotic zone, post-meiotic stages and in mature sperm (Figure 3-5A, Figure 3-5B). We examined testes from Cid5mCherry males and performed antibody staining with the Cid1 antibody and a phospho-histone H3 Serine 10 (PH3S10) antibody to identify condensed chromosomes[145-147]. We found that Cid1 and Cid5 co-localize at centromeres in the mitotic zone of the testis (Figure 3-5C). However, surprisingly, at the onset of metaphase of meiosis I, Cid1 was no longer observed, and we could only detect Cid5 on these chromosomes (Figure 3-5D). We could also detect Cid5 in post-meiotic stages as

a discrete focus on each 'leaf-stage' and 'late-canoe-stage' spermatid nucleus[143], but we never observed Cid1 at these stages (Figure 3-5E, Figure 3-5F).

To make sure our inability to detect Cid1 in meiotic cells with condensing chromosomes and post-meiotic spermatids was not due to antibody accessibility issues, we also examined Cid1GFP in the male germline. The results were nearly identical to the antibody staining. We detected Cid1 at centromeric foci in the mitotic zone (Figure 3-6A) but could only detect faint Cid1 signal in meiotic cells entering metaphase of meiosis I (Figure 3-6B). We could not detect Cid1 at any stage after meiosis, including in mature sperm (Figure 3-6C - Figure 3-6E). Our results are thus consistent between our antibody staining and transgene analysis, except for cells entering meiosis I metaphase, in which Cid1GFP is either slightly more sensitive than the Cid1 antibody, or persists longer than endogenous Cid1. Regardless, these results indicate that metaphase of meiosis I represents a transition state between the presence of Cid1 in mitotic cells and its absence in post-meiotic cells. Like the loss of Cid5 in oocytes, this loss of Cid1 occurs without DNA replication, suggesting an active protein degradation mechanism may be responsible. Interestingly, meiosis I metaphase represent a centromeric transition state in both males and females, except that Cid5 is specifically lost in the female germline and Cid1 is specifically lost in the male germline.

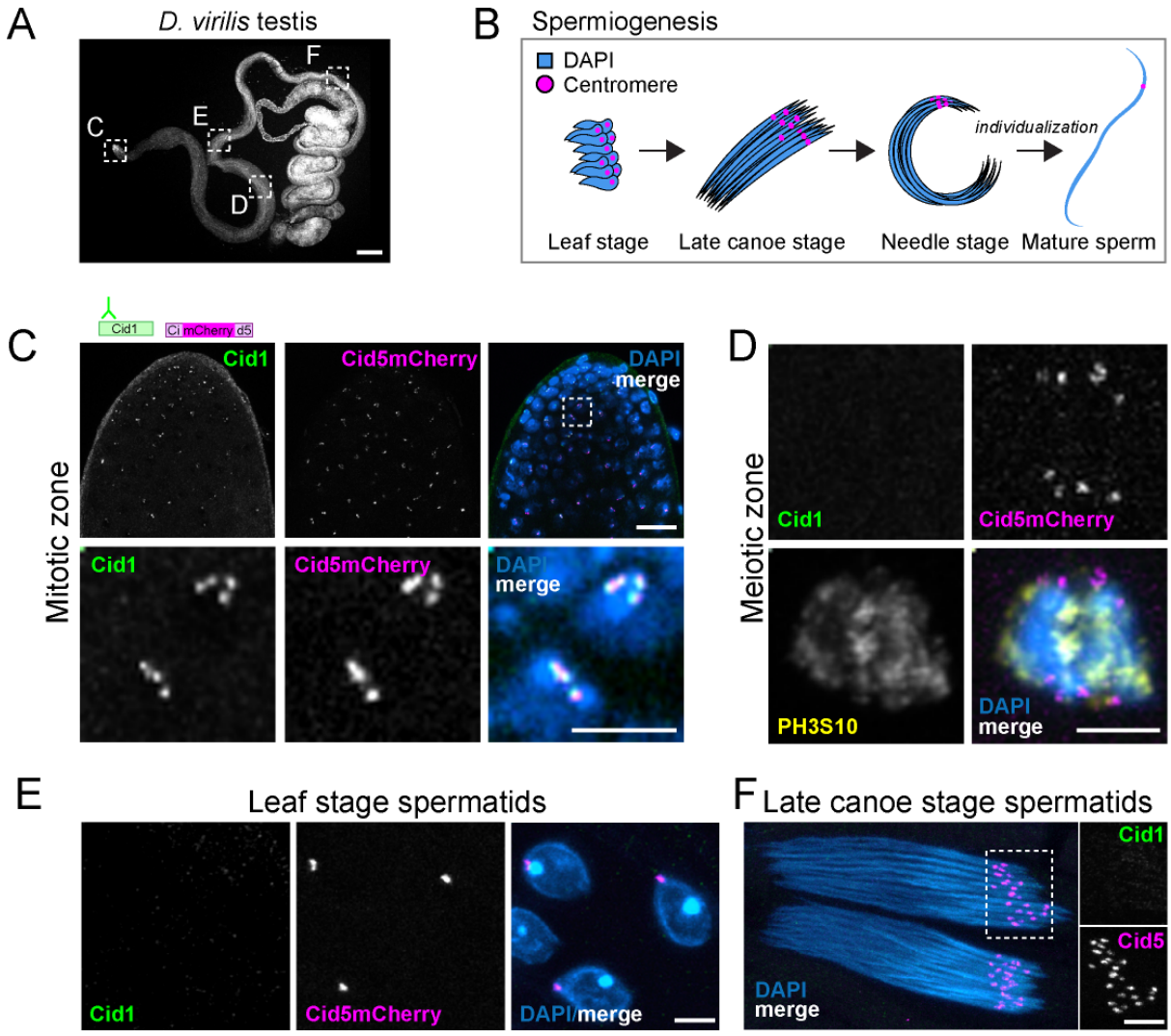


Figure 3-5. Differential localization of Cid1 and Cid5 in testes. (A) Image of a DAPI stained *D. virilis* testis. Boxed regions show the approximate location of panels (C) – (F). Scale bar = 100μm. (B) Schematic showing stages of spermiogenesis. The Cid1 antibody and Cid5mCherry transgene were used to visualize Cid1 and Cid5 in the images in (C) – (F). (C) The apical tip (mitotic zone) of a *D. virilis* testis. The bottom panel shows a high magnification image of the area indicated in the top panel by the dashed box. Scale bar = 25μm in top panel and 10μm in the bottom panel. (D) A single cell with condensing chromosomes in late prometaphase or early metaphase. PH3S10 antibody staining is also shown. (E) Leaf-stage spermatid nuclei. (F) Late-canoe stage spermatid bundles. Scale bars = 5 μm in (D) – (F).

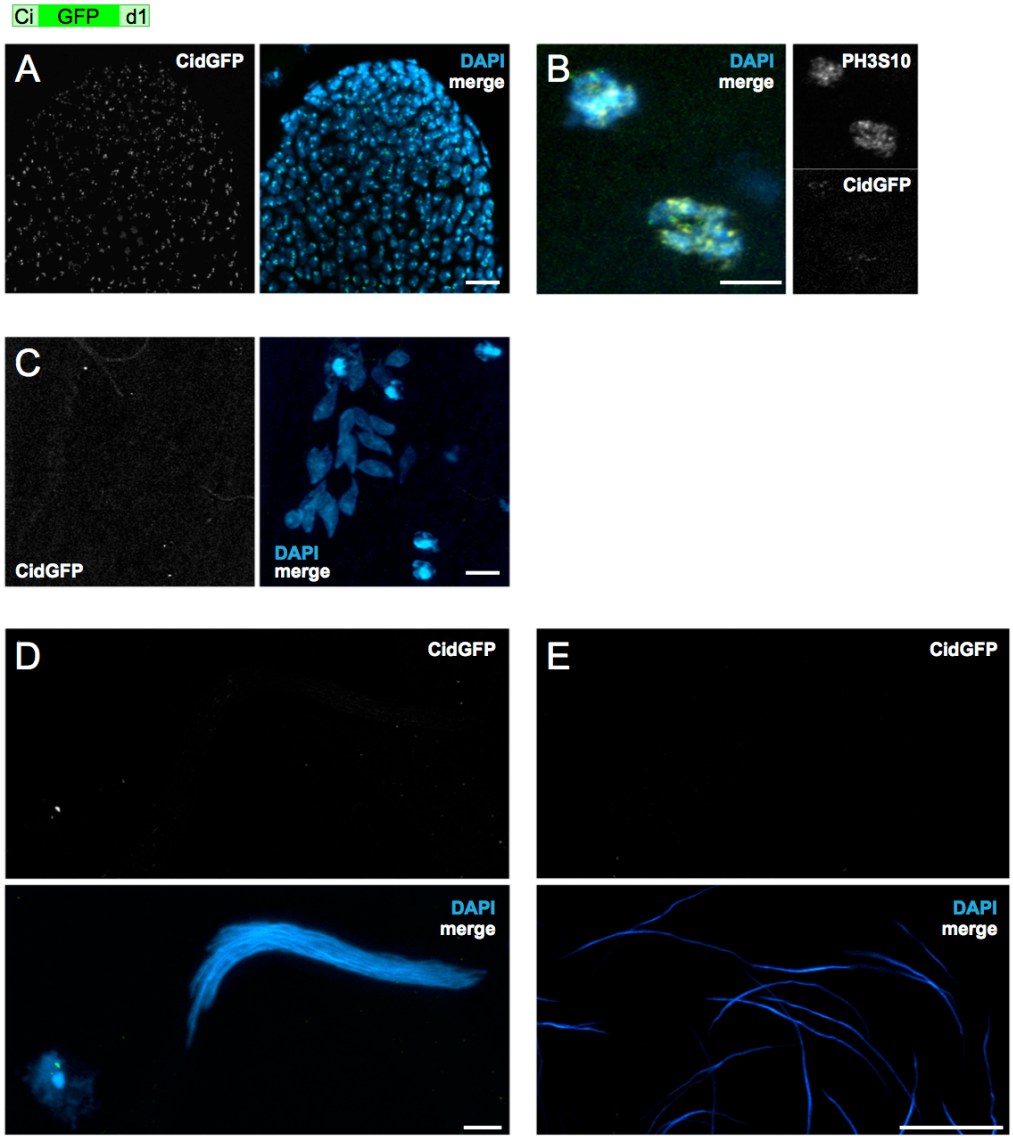


Figure 3-6. Localization of Cid1GFP in testes. Images showing Cid1GFP localization at various stages of spermatogenesis. (A) The apical tip (mitotic zone) of a *D. virilis* testis. Scale bar = 25μm (B) Two cells in late pro-metaphase or early metaphase. PH3S10 antibody staining is also shown. (C) Leaf-stage and (D) late-canoe stage sperm bundles. Scale bars = 5μm in (B) – (D). (E) Individualized mature sperm. Scale bar = 25μm

Our cytological analyses further indicate that Cid5's centromeric localization persists throughout male gametogenesis from early germ cells to sperm. Previous findings have elegantly demonstrated that Cid protein is required for transgenerational inheritance of centromere identity through sperm in *D. melanogaster*[23]. Since Cid1 is not detectable during spermiogenesis, we hypothesized that Cid5 might provide the transgenerational centromeric mark in mature sperm in *D. virilis*. To further investigate Cid5 localization in *D. virilis* sperm and validate its centromeric localization, we employed GFP-Hiphop as a centromere-adjacent marker[148]. Since *D. virilis* flies have acrocentric chromosomes, their centromeric cytological signals should be adjacent to one of the two telomeric, GFP-Hiphop-labeled, cytological signals on each chromosome. Thus, Hiphop localization serves as an additional centromere-adjacent marker in *D. virilis*.

We examined the localization of GFP-HipHop and Cid5mCherry in the testes of flies that contained both transgenes. We observed two primary HipHop foci corresponding to telomeric ends in each spermatid nucleus (Figure 3-7B). We also saw a single Cid5 focus, which always co-localized with one of the two HipHop foci (Figure 3-7C). This localization pattern persisted throughout spermatid development and in mature sperm (Figure 3-7C - Figure 3-7E). These experiments give additional support to the hypothesis that Cid5 provides the transgenerational centromere mark in *D. virilis* – Cid5 is present at centromeres in mature sperm, but Cid1 is not.

Taken together, our cytological examination of Cid1 and Cid5 in the *D. virilis* male germline suggests that Cid5 is the predominant centromeric histone in male meiotic cells after prometaphase I and in developing and mature sperm. Our inability to robustly detect Cid1 in post-prometaphase meiotic cells and post-meiotic spermatids strongly suggests that male meiotic and centromere inheritance function in *D. virilis* flies does not require Cid1, even though the *D. melanogaster Cid1* ortholog, *Cid*, is essential for both processes[23,59]. Thus, male and female gametes alternately retain different Cid protein paralogs in *D. virilis*.

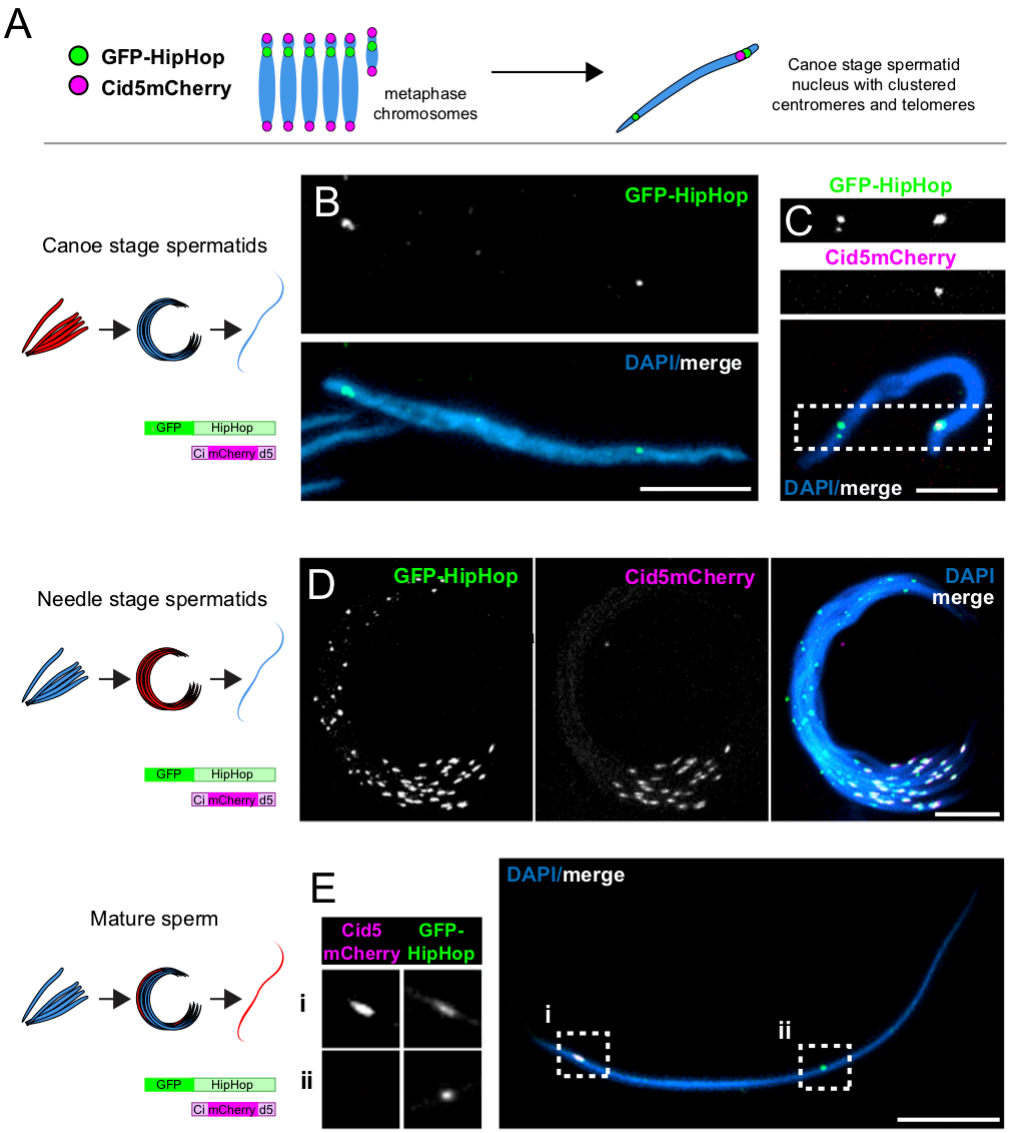


Figure 3-7. Cid5 provides the centromere mark in mature sperm. (A) Schematic showing chromosomes in the RabI configuration in *D. virilis* sperm. (B) A single late canoe stage spermtid from a GFP-HipHop fly. All subsequent panels show images from flies with both GFP-HipHop and Cid5mCherry transgenes. (C) A single late-canoe stage spermatid. (D) Needle-stage spermatid bundle. (E) A single mature sperm nucleus. Boxed regions (i) and (ii) are also shown at slightly higher magnification and as separate channels (left). All scale bars = 10μm. Schematics on the left side of panels (B) – (E) indicate stages of spermiogenesis: canoe stage, needle stage and mature sperm. The stage highlighted in red is shown in the panel on the right.

Maternal Cid1 rapidly replaces paternal Cid5 following fertilization

Our cytological analyses revealed that the mature oocyte nucleus in *D. virilis* only retains Cid1 whereas mature sperm only retain Cid5 (Figure 3-8A). We next investigated how parental genomes with distinct Cid paralogs coordinate chromosomal events following fertilization. One of the most dramatic chromosomal changes is the remodeling of the sperm nucleus, in which SNBPs (protamines in mammals), which package the bulk of sperm chromatin, are replaced with maternally provided canonical and variant histones in a replication-independent manner[149-151]. In *D. melanogaster*, paternal Cid persists on the paternal genome throughout this extensive remodeling and is required for the first embryonic cell divisions, even though the specific molecules of paternal Cid molecules only persist until the third embryonic cell cycle[23]. While paternal chromosome remodeling occurs, female meiosis is completed. Pronuclei then congress towards each other, appose and undergo mitosis synchronously but on separate halves of the first spindle (Figure 3-8A). Defects in this synchronization lead to embryonic lethality[152,153].

Based on the precedent in *D. melanogaster*, we expected that paternally inherited Cid5 would persist on the paternal genome through the first several embryonic cell cycles, whereas Cid1 would define centromeres throughout the completion of female meiosis, co-localize with Cid5 in the early embryo and eventually become the only Cid protein present in the embryo. To test this hypothesis, we examined Cid1GFP and Cid5mCherry in embryos produced by male and female parents bearing both transgenes. Consistent with our previous findings that meiosis I metaphase arrested oocytes only contain Cid1 (Figure 3-3), we found that Cid1, but not Cid5, is detectable on the maternal genome through the completion of meiosis (Figure 3-8B, Figure 3-8C). Were surprised to also detect only Cid1 on the paternal pronucleus, even at very early stages (Figure 3-8B, Figure 3-8C) despite our observation that mature sperm only contain Cid5. Although Cid1 signal was faint on the paternal genome at earlier stages, it reached a level comparable to the Cid1 signal on the maternal genome by the time of the synchronous first

mitosis (Figure 3-8B - Figure 3-8F). Our results suggest that in *D. virilis,* maternal Cid1 rapidly replaces paternal Cid5 during the protamine-to-histone chromatin transition. Thus, the dynamics of paternal Cid in the *D. virilis* embryo are subtly distinct from those previously observed in the *D. melanogaster* embryo[23].

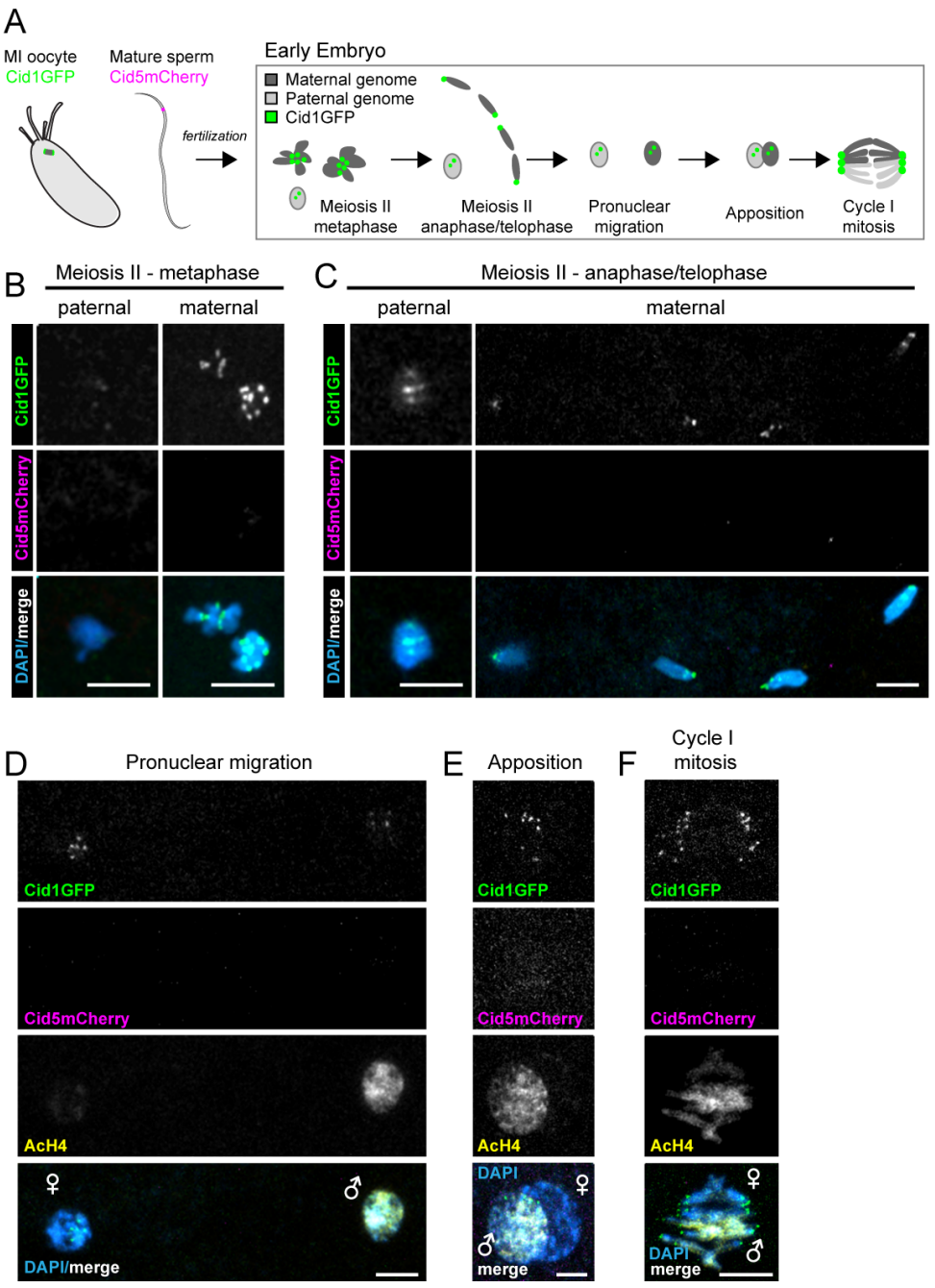


Figure 3-8. Cid1 replaces Cid5 in the early embryo. (A) Schematic of fertilization and the progression of the maternal and paternal genome in the early embryo through the first embryonic mitosis. All other panels are images from *D. virilis* early embryos that were collected from parents that had Cid1GFP and Cid5mCherry transgenes. Paternal and maternal genomes were discerned by nuclear morphology, (B) and (C), or by acetylated histone H4 (AcH4) antibody staining, (D) – (F). AcH4 preferentially stains the paternal genome. (B) Meiosis II metaphase. (C) Meiosis II anaphase/telophase. (D) – (F) Pronuclear migration, apposition and the first embryonic mitotic cell division. All scale bars = 5μm.

3.4 Discussion

Until recently, CenH3 duplications were thought to be rare in animals. However, we previously showed that many *Drosophila* species encode more than one *Cid* gene[125]. We hypothesized that these divergent *Cid* genes were preserved following duplication to carry out specialized germline centromeric roles, possibly to relieve intralocus conflict[57]. To gain insight into the functions of ancient *Cid* paralogs, we investigated the cytological localization of two specific Cid protein paralogs, Cid1 and Cid5 in *D. virilis.* We found that Cid1 localizes to centromeres in the soma and germline, but Cid5 is germline restricted. Moreover, we found that Cid1 and Cid5 have mutually exclusive localization patterns in certain germline cell types. Both Cid1 and Cid5 are present in germ cells in males and females up until metaphase of meiosis I. Then, in the male germline, only Cid5 remains on germ cells during spermiogenesis and in mature sperm. In the female germline, only Cid1 is present for the completion of oogenesis and in the mature oocyte. Our results suggest two levels of specialization of Cid1 and Cid5; Cid1 is the somatic centromeric histone and female meiotic centromeric histone, whereas Cid5 is specialized for male meiosis and transgenerational centromere inheritance through sperm. However, these possible functions suggested by protein localization must be rigorously assessed with further experimentation.

If centromeric proteins achieve functional specialization how might Cid1 and Cid5 do so? We speculate that it may derive from their highly divergent N-terminal tails, however little is known about the specific function of Cid's N-terminal tail in *Drosophila*. What we do know is that the N-terminal tail may be involved in the recruitment of kinetochore proteins[105] but it is not required for Cid localization to centromeres[154], including in male meiosis[155]. Consistent with our results, this suggests that the differences in the Cid1 and Cid5 N-terminal tail should not affect their ability to localize to centromeres. Instead, differences in Cid1 and Cid5's N-terminal tails

could point to different protein-protein interaction partners and cell-type specific kinetochore formation. In theory, these proposals could also be experimentally tested.

A clue as to what the function of the N-terminal tail may be comes from our previous characterization of conserved N-tail motifs (Figure 2-10). All single copy *Cid* genes in *Drosophila* (including from *D. melanogaster*) encode a highly stereotyped set of motifs 1 – 4 in their N-terminal tails[86,125]. While the specific function of each motif remains to be elucidated, motifs 1 – 3 have been implicated in sister centromere cohesion in male meiosis[155] whereas motif 4 has been associated with BubR1 recruitment[105]. Both Cid1 and Cid5 have motif 4, so they both likely are capable of recruiting BubR1 and signaling the spindle assembly checkpoint. Motifs 1-3 were recently implicated in sister chromatid cohesion in *D. melanogaster* male meiosis[155]. Cid1 has maintained motifs 1 – 3 whereas Cid5 has lost motifs 1 and 3 but maintained motif 2. This could suggest that Cid1 still contributes to sister chromatid cohesion in the male germline even though Cid5 is the only detectable centromeric histone later in male germ cell development. Although the detailed molecular functions of the N-terminal tail motifs have yet to be elucidated, our results suggest that motifs 1 and 3 are not required for male meiotic chromosome segregation or centromere inheritance through the male germline. Future proteomic analysis of Cid1 and Cid5 in *D. virilis* may reveal paralog-specific protein-protein interacting partners, which could be attributable to differences in the Cid1 and Cid5 N-terminal tail.

We also speculate that differences in Cid1 and Cid5 N-terminal tails could be responsible for the rapid disappearance of Cid1 from meiotic cells in the male germline. Following duplication, *D. virilis* Cid1 retained motifs 1 – 4 but also acquired motif 8, a new sequence motif present in all Cid1 homologs in the virilis subgroup but entirely absent from all single copy *Cid* genes. Intriguingly, in a previous study by Monen *et al.*, separase cleavage of the N-terminal tail has been implicated in the degradation of one of two CenH3 protein paralogs during *C. elegans* female meiosis[78]. Acquisition of motif 8 in Cid1 may also help explain the sudden loss of Cid1 during male meiosis if it represents a male germline-specific 'degron' motif.

We further predict that knockout or genetic knockdown of *Cid1* and *Cid5* will have different phenotypic consequences, just as previously observed for wheat *CenH3* paralogs[67]. Since Cid1 is the centromeric histone in somatic cells, we predict that flies lacking Cid1 would be inviable, just like *Cid* knockdown *in D. melanogaster*[32]. Furthermore, germline-specific knockdown for *Cid1* and *Cid5* may result in female and male sterility, respectively. However, it is possible they are both important in female and male fertility. For example, Cid1 might be required for loading Cid5 in male germ cells, in which case Cid1 knockdown in the male germline would also result in sterility. Although we do not expect Cid5 to play a non-redundant role during oogenesis, if knockdown of Cid5 in ovaries impairs fertility, this will imply that Cid1 and Cid5 are both required for full female fertility.

Cid paralogs like the two we have described in *D. virilis* in this study likely represent *CenH3* alleles with separable gametic functions that have been vetted over millions of years of natural selection. They allow us to develop concrete hypothesis regarding the specific functional roles of N-terminal motifs that would not be otherwise possible in species carrying only one essential copy of *CenH3*. Furthermore, our study points to the possibility that CenH3 proteins, which are believed to functionally identical across various tissues and cells of the same organism, instead carry out distinct roles in different cell types. At least two other ancient *Cid* duplication events have been found in the montium group of *Drosophila* species. Some of these paralogs also show testis specific expression patterns, just like in *D. virilis*[125]. Moreover, their N-terminal tails show a similar pattern of gain and loss of motifs. Investigations in these species will reveal whether the evolution of *Cid* paralogs follows a convergent trajectory of functional specialization.

3.5 Materials and methods

Cid1 and Cid5 antibody production

We raised an antibody against Cid1 residues 15 – 31 (KSESHLDNVEDSYEKTA) and Cid5 residues 56 - 71 (NLESPVAGEEPAPDTV). These sites were selected because they were in regions where Cid1 and Cid5 shared no apparent homology. Covance Inc. (Princeton, NJ) immunized two rabbits with the conjugated Cid5 peptide by injecting it four times over the course of four months. Covance also immunized two rabbits for the Cid1 peptide by injecting it five times over the course of five months. Our previous analysis of Cid5 population genetics revealed non-synonymous variation in the Cid5 peptide sequence; therefore, we ensured that we used *D. virilis* strains and cell lines that match the Cid5 antibody peptide sequence.

Western blots from D. virilis WR DV-1 cells

D. virilis WR DV-1 cells were collected in RIPA buffer and sonicated. Protein was quantified by Bradford assay and 20ug total protein was analyzed by western blot. We probed the membrane with either rabbit anti-Cid1 (1:2000), rabbit anti-Cid5 (1:2000), or rabbit anti-H3 (1:5000 Abcam ab1791) primary antibodies followed by goat anti-rabbit IgG-HRP (1:5000 Santa Cruz Biotechnologies Inc., Dallas, TX).

Antibody staining D. virilis tissue culture cells

Cells were transferred to coverslips and fixed in 4% PFA for 5 min and blocked with PBSTx (0.3% Triton) plus 3% BSA for 30 minutes at room temperature. Then cells were incubated with primary antibodies at 4°C overnight at the following concentrations: rabbit anti-Cid1 1:5000. Coverslips with cells were incubated with secondary antibodies for 1 hour at room temperature at the following concentrations: goat anti-rabbit (Invitrogen Alexa Fluor 568, A-11011) 1:2000.

Overexpression of Venus-Cid5 in D. melanogaster KC cells

Since *Cid5* is not expressed in *D. virilis* tissue culture cells, we confirmed that the Cid5 antibody works for cytology by overexpressing Venus-Cid5 in *D. melanogaster* KC cells and performing immunostaining with the Cid5 antibody. Venus-Cid5 was cloned into an expression vector from the Drosophila Gateway Collection generating an N-terminal Venus (pHVW) fusion protein under the control of the *D. melanogaster* heat shock promoter. Transfections and antibody staining were performed as follows: two micrograms plasmid DNA was transfected using Xtremegene HP transfection reagent (Roche) according to the manufacturer's instructions. Cells were heat-shocked at 37°C for one hour 24 hours after transfection to induce expression of the *Cid* fusion protein. Cells were transferred to a glass coverslip 24 hours after heatshock. Cells were fixed in 4% PFA for 5 min and blocked with PBSTx (0.3% Triton) plus 3% BSA for 30 minutes at room temperature. Coverslips were then incubated with primary antibodies at 4°C overnight at the following concentrations: rabbit anti-Cid5 1:2500. Coverslips with cells were incubated with secondary antibodies for 1 hour at room temperature at the following concentrations: goat anti-rabbit (Invitrogen Alexa Fluor 568, A-11011) 1:2000.

D. virilis transgenics

Cid1GFP and Cid5mCherry were cloned into a vector backbone containing piggyBac inverted repeats and the miniwhite gene cassette. This vector was generated by first removing 3XP3EGFP from the nosGal4-MW-pBacns plasmid (stock number 1290, Drosophila Genomics Resources Center). The 3XP3EGFP was removed as follows: nosGal4-MW-pBacns was digested with AgeI and AsiSI, run on a gel and the largest band was gel isolated. Overhangs were blunted, and then the vector was ligated to itself to produce nosGal4_MWonly. Then the nanosGal4 casette was removed as follows: nosGal4_MWonly was digested with NotI, run on a gel and the largest bad was gel isolated. Then the vector was ligated to itself to produce

NoPromoter_miniwhite. Cid1GFP and Cid5mCherry plus ~1kb sequence upstream and downstream of Cid1 and Cid5 were inserted between the AvrII and SbfI sites of the NoPromoter_miniwhite plasmid. For both Cid1 and Cid5, fluorescent proteins were inserted immediately 5-prime of the RRRK motif at the beginning of the histone fold domain. Fluorophores were flanked on both sides by three glycine residues to function as flexible linkers. Cid1GFP and Cid5mCherry plasmids were injected along with the piggyBac helper plasmid phspBac[156] into *D. virilis* embryos. Injected flies were screened for red eye color. Injections and screening was performed by Rainbow Transgenic Flies Inc (www.rainbowgene.com).

Cytology: general data collection and presentation practices

For all cytological data, we present representative images acquired from the Leica TCS SP5 II confocal microscope with LASAF software and present maximally projected image files. For protein localization in larval neuroblasts, ovaries, testes and the early embryo, a minimum of five organs and five cell-types were examined for each assay of each stage.

Preparation of larval neuroblasts for immunofluorescence

To assess Cid1 and Cid5 localization in larval brains, we used both Cid1 and Cid5 specific antibodies and Cid1GFP and Cid5mCherry transgenes. Brains from actively crawling third-instar larvae were dissected in PBS and transferred to 0.5% sodium citrate hypotonic solution 10 minutes. We transferred brains to a drop chromosome isolation buffer (120mg MgCl2:6H2O, 1g citric acid, 1mL Triton-X100, distilled H2O to 100mL) on a glass slide and fragmented the brains with needles for four minutes. Next, we lowered a coverslip onto the fragmented brains and squashed the brains under gentle pressure for 30 seconds. We then froze slides in liquid nitrogen. Then, slides were removed from liquid nitrogen and the cover slip was flipped off with a razor blade. Slides were immediately immersed in cold methanol for five minutes, cold

acetone for one minute and PBS for one minute at room temperature. For experiments obtaining fluorescent signal from transgenes only, we removed the PBS and added mounting medium with DAPI. For antibody staining, after incubation in acetone, brains were rinsed once in PBS and then incubated in PBS + 1% TritonX for 10 minutes for permeabilization. Slides were blocked in PBS + 0.1% TritonX + 3% BSA for 30 minutes at room temperature. Slides were incubated with primary antibody overnight at 4°C. Then slides were washed and incubated with secondary antibodies for one hour at room temperature. Primary antibody dilutions were as follows: rabbit anti-Cid1 1:1000, rabbit anti-Cid5 1:1000. Secondary antibody dilutions were Alexa Fluor goat anti-rabbit 568 1:1000.

Testis immunofluorescence

To assess Cid1 and Cid5 localization in testes, we used Cid1 and Cid5 specific antibodies or transgenic flies encoding Cid1GFP or Cid5mCherry (both with internal tags and expressed under the control of their native promoters, as described above). To characterize Cid1 and Cid5 localization without antibody staining, we dissected testes in PBS from sexually mature (~10-day old) Cid1GFP, Cid5mCherry, or Cid1GFP/Cid5mCherry males. Testes were spread out on charged microscope slide, squashed under a coverslip and immediately immersed in liquid nitrogen. Testes were then fixed in 4% paraformaldehyde (PFA) for seven minutes or cold methanol (5 minutes) and acetone (5 minutes). Testes were then mounted in SlowFade Gold with DAPI (Thermo Fisher Scientific). For immunofluorescence, we fixed testes from Cid1GFP or Cid5mCherry transgenic flies in 4% PFA. Testes were permeabilized in PBS + 0.3% TritonX for 30 minutes (two-15 minutes washes) and blocked in PBS + 0.1% TritonX + 3% BSA for 30 minutes at room temperature. Primary antibodies were diluted in block and incubated with testes overnight according to the following dilution: mouse anti-phospho-histone H3 serine 10 (1:1000 Millipore clone 3H10). Secondary antibodies were incubated in block for one hour at room temperature according to the following dilution: Alexa Fluor goat anti-mouse 633 (1:1000).

Ovary immunofluorescence

To assess Cid1 and Cid5 localization in ovaries, we used Cid1 and Cid5 specific antibodies or transgenic flies encoding Cid1GFP or Cid5mCherry (as described above). To characterize Cid1 and Cid5 localization without antibody staining, we dissected ovaries in PBS from sexually mature (~10-day old) Cid1GFP, Cid5mCherry, or Cid1GFP/Cid5mCherry *D. virilis* females. Ovaries were fixed in 1:1 paraPBT:heptane (paraPBT = 4% paraformaldehyde in PBS + 0.1% TritonX) for 10 minutes at room temperature. Then ovaries were washed, including one wash with 1X DAPI and mounted in SlowFade Gold. For immunofluorescence, we performed fixation as above. We then blocked ovaries in PBS + 0.1% TritonX + 3% BSA for 30 minutes at room temperature. Ovaries were incubated with primary antibodies overnight at 4C. Ovaries were then washed and incubated with secondary antibodies for 1 hour at room temperature. Then ovaries were washed and mounted as above. Antibody dilutions were as follows for primary antibodies: Rabbit anti-Cid1 1:1000, rabbit anti-Cid5 1:1000. Secondary antibody dilutions were Alexa Fluor goat anti-rabbit 568 1:1000.

Embryo collection, fixation and immunofluorescence

To characterize Cid1 and Cid5 in the early development we conducted immunofluorescence and DAPI staining on embryos produced from mothers and fathers with both Cid1GFP and Cid5mCherry transgenes. 0–60 min old embryos were collected on grape agar plates. Embryos were incubated in 30% bleach for 2 minutes to remove chorion. Fixation and antibody staining was performed according to Fanti and Pimpinelli method 3[157]. Briefly, embryos were transferred to a 1:1 mixture of heptane and methanol and shaken vigorously for one minute. The heptane layer was removed and embryos were washed twice with ice-cold methanol. Embryos were rehydrated in PBS plus a drop of PBS + 0.1% Triton. Next, embryos were permeablized in PBS + 1% Triton for 30 min at room temperature. Embryos were blocked in PBS + 0.1% TritonX +

3% BSA (BSA block) for one hour at room temperature. We diluted primary antibodies in BSA block and incubated overnight at 4°C. Embryos were washed and then incubated with secondary antibodies diluted in BSA block for two hours at room temperature. We washed embryos again after incubation with secondary antibodies and mounted embryos in wash solution (PBST). Embryos were imaged immediately after mounting. Primary antibody dilutions were the following: rabbit anti-AcH4 (Millipore, Billerica, MA; 1:1000), Alexa-Fluor goat secondary antibodies (Life Technologies) were diluted at 1:1000.

Chapter 4. Ancient paralogs of *Cid* centromeric histones and *Cal1* chaperones in mosquito species

4.1 Introduction

Centromeric proteins represent an evolutionary conundrum. Their critical role in cell division and chromosome segregation make them essential for viability throughout eukaryotic life[32-34]. However, centromeric proteins evolve rapidly in plants and animals[36,54,83] in spite of their essential function possibly due to their involvement in genetic conflict[129,131]. This 'centromere paradox'[37] is exemplified by the centromeric histone, which is the foundational centromeric protein in most eukaryotes. CenH3 is essential for chromosome segregation in protists, fungi, plants and most animals[32-34,128]. Yet, it is subject to rapid evolution in plants and animal species that undergo asymmetric female meiosis[158]. This asymmetry provides an opportunity for centromeres to act as selfish genetic elements and bias their transmission to the next generation (a process termed 'centromere drive'). Rapid evolution of CenH3 proteins has been hypothesized to suppress deleterious cheating behavior of 'selfish' centromeres[37,129-131].

In addition to CenH3's hypothesized role as a drive suppressor in the male germline, CenH3 may also perform specialized germline functions in males and females. For example, CenH3 might require a different amino acid composition for function in cycling somatic cells or the female germline than it would for retention during spermatogenesis, in which bulk chromatin undergoes a histone-to-protamine transition[134]. Our previous research[125] suggested that optimality of all these multiple functions of CenH3 proteins might not be simultaneously achievable by a single CenH3 gene. However, such intralocus conflict could be resolved via gene duplication and specialization[57]. Indeed, 10% of plant species have been shown to encode more than one CenH3 gene[48,72,75], although evidence of functional specialization is still quite sparse even in plants[67].

Evidence for CenH3 duplication is even more rare in animals. Until recently, the only instances of CenH3 duplications in animals were two young duplications in nematode species[77,78], and several CenH3 duplications in *Bovidae* (cows and sheep), most of which have become pseudogenized[159]. In contrast to the view that animal genomes have a single centromeric histone gene, we recently identified four independent gene duplication events of the CenH3 gene, *Cid*, in *Drosophila.* Our phylogenetic analysis indicated that the majority of *Drosophila* species encode two or three *Cid* paralogs (see Chapter 2), including some that have been stably co-retained for over 40 million years[125]. *Cid* paralogs evolve under different evolutionary constraints and some paralogs have germline restricted expression patterns[125]. Moreover, our cytological analysis of Cid1 and Cid5 protein paralogs in *D. virilis* revealed that Cid1 and Cid5 have acquired specialized gametic localization patterns, which may be indicative of specialized functions (Chapter 3).

If selection to resolve intralocus conflict favors retention of specialized CenH3 paralogs, we would expect to find recurrent instances of germline-specialized centromeric histones outside of *Drosophila*, including in other Dipteran species. We took advantage of recent genome sequencing efforts in another Dipteran family, *Culicidae* (mosquitoes)[160,161] to investigate whether we could discover additional instances of *CenH3* duplication and specialization. Surprisingly, we find that most mosquito species encode two *CenH3* paralogs that diverged over 150 million years old, making this the oldest *CenH3* duplication identified so far. From here on, we designate these as *mosquitoCid* (*mosCid)* to distinguish them from *Drosophila Cid* paralogs. We also find that *mosCid* paralogs encode divergent N-terminal tails and evolve under different evolutionary constraints. Furthermore, some *mosCid* paralogs show ovary and early embryo biased expression patterns. Finally, we also report that *Anopheles* genomes encode two paralogs of the centromeric histone chaperone, *CAL1.* Like the duplication of CENP-C seen previously in *Drosophila* species[139], our findings suggest that multiple inner kinetochore proteins (*e.g.,* Cid, CENP-C, CAL1 in Diptera) may be subject to intralocus conflict due to multiple,

incompatible centromeric functions, which can be resolved via gene duplication and specialization[57].

4.2 Results

Mosquito genomes harbor ancient Cid paralogs

Recently, 16 high quality Anopheles genomes were published[160], providing an additional set of densely sampled Dipteran genomes that are well suited to phylogenomic analyses (http://www.vectorbase.net)[161]. To identify mosquito *Cid* (*mosCid)* homologs, we used *D. melanogaster Cid* as a query and performed tBLASTn against 21 mosquito genomes including 18 *Anophelinae* mosquitoes and three *Culicinae* species (two *Aedes* and one *Culex).* We recorded the syntenic location (5-prime and 3-prime neighbor genes) of all BLAST hits. Many of these genes are not yet annotated in the public databases and the *mosCid* open reading frame required manual curation. We found that *Anopheles gambiae* encodes two *mosCid* paralogs (Figure 4-1), both encoded by a single open reading frame in distinct genomic loci. The *An. gambiae mosCid* paralogs are highly divergent and share only 51% amino acid identity in their histone fold domains; their N-terminal tails can barely be aligned. The *mosCid1* paralog is located in the intron of the *mRpL48* gene, whereas *mosCid2* is found between the *Integrator complex subunit 1* (*Integrator 1)* and *Syndapin* genes (Figure 4-1). We found both paralogs in the same syntenic location in nearly all other *Anopheles* mosquitoes. The only exceptions were in *An. albimanus* and *An. darlingi* where we only found the *mosCid2* gene. In these species, we were able to find the shared syntenic locus containing the *mRpL48* gene*,* which was missing *mosCid1*. Moreover, we did not find any other *mosCid* sequences in these genomes, suggesting these two species lack *mosCid1*.

Next we investigated the *mosCid* genes in two *Aedes* species: *Aedes aegypti* and *Ae. albopictus.* We identified three *mosCid* paralogs in these species, starting with *mosCid1* located

in the intron of *mRpL48* (Figure 4-1). We also identified *mosCid3* located in close proximity to *mosCid1.* Finally, we identified a third *Aedes mosCid* paralog located between the *ATP synthase subunit 1* and *CG7083* genes. Although the unique syntenic location suggested an independent *mosCid* gene duplication, phylogenetic analyses (below) confirmed that this gene is likely to be an ortholog of the Anopheles *mosCid2* gene (Figure 4-1). Finally, we examined the *mosCid* genes present in *Culex quinquefasciatus.* We found that *C. quinquefasciatus* also contained *mosCid1* in the *mRpL48* intron (Figure 4-1) and a second *mosCid* gene in a distinct syntenic location adjacent to a gene that shares homology with *D. melanogaster Kibra.* Once again, we relied on phylogenetic analyses to confirm that the second gene corresponds to *mosCid2* in spite of its distinct syntenic location from either the *Anopheles* or *Aedes* orthologs.

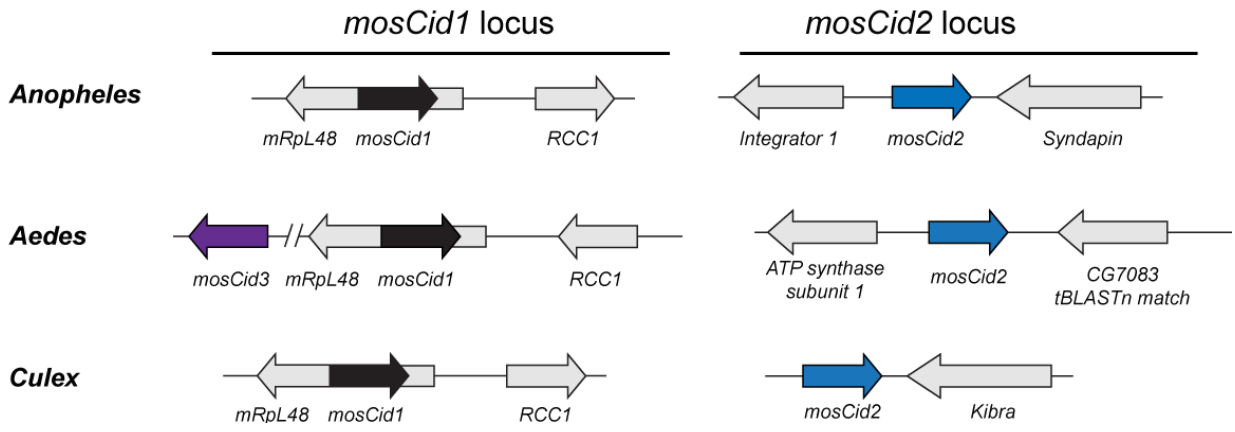


Figure 4-1. Syntenic location of mosquito *Cid* paralogs. The genomic context of representative mosquito *Cid* paralogs identified by tBLASTn is schematized for *Anopheles*, *Aedes* and *Culex*. In total, we found three mosquito Cid genes, *mosquitoCid1* (*mosCid1*) is present in the intron of mRpL48 in *Anopheles, Aedes and Culex* mosquitoes (black arrow, mosCid1 locus). *MosCid2* is found in *Anopheles* between the genes *Integrator 1* and *Syndapin* (blue arrow, *mosCid2* locus column). In *Aedes*, *mosCid2* is located between *ATP synthase subunit 1* and a gene with homology to D. melanogaster *CG7083*. In *Culex*, *mosCid2* is in a genomic locus next to the *Kibra* gene. *MosCid3* (purple arrow) is an *Aedes*-specific paralog that is also present in the *mosCid1* locus. Arrows colored in grey represent genes that define the shared syntenic locus of each paralog. Genes that define each syntenic locus are named based on the *D. melanogaster* gene name.

The presence of *mosCid* genes in a shared syntenic location across species is a strong indicator that they are likely orthologous. Based on this criterion, we predict that all *mosCid1*

genes are orthologs and that *mosCid1* was likely present in the common ancestor of all mosquitoes but was subsequently lost in the ancestor of *An. albimanus* and *An. darlingi.* In contrast to *mosCid1*, we were unable to assign the other *mosCid* genes into orthologous groups based on synteny alone. To clarify their evolutionary relationships to each other and to *mosCid1*, we performed maximum-likelihood based phylogenetic analyses using a nucleotide alignment of the histone fold domain of all *mosCid genes* (Figure 4-2). We found that *mosCid1* and *mosCid3* from the two *Aedes* species group together, suggesting that *mosCid3* arose from a *mosCid1* duplication event in the common ancestor of *Ae. aegypti* and *Ae. albopictus* (Figure 4-2). Furthermore, we found that the *mosCid2* genes from all 21 mosquito species examined are likely to be orthologous despite being found in distinct syntenic contexts (Figure 4-2).

Overall, our synteny-based and phylogenetic analyses identify two independent duplications of *mosCid* genes during the 150 million year old history of mosquito evolution that we have investigated. We conclude that *mosCid1* and *mosCid2* were present in the common ancestor of all examined mosquito species and have been largely co-retained for over 150 million years old, making them the oldest and most diverged *CenH3* paralogs found in the same genome. Due to the high degree of conservation in the histone fold domain and the high divergence to other Dipteran species, we cannot determine if *mosCid1* or *mosCid2* represents the original, ancestral *mosCid* gene. Even though *mosCid1* is retained in the same, shared syntenic context whereas *mosCid2* is not, this does not imply one is older than the other. The numbering of *mosCdi1* and *mosCid2* is thus arbitrary and not an indication of ancestry. Although both ancient *mosCid* paralogs have been co-retained in most species, *mosCid1* was lost once in the common ancestor of *An. albimanus* and *An. darlingi*, suggesting that at least in this pair of species, *mosCid2* is fully capable of carrying out all centromeric functions as a single gene. Finally, *mosCid3* was born via *mosCid1* duplication in the common ancestor of *Aedes* mosquitoes 20 – 60 million years ago (Figure 4-2).

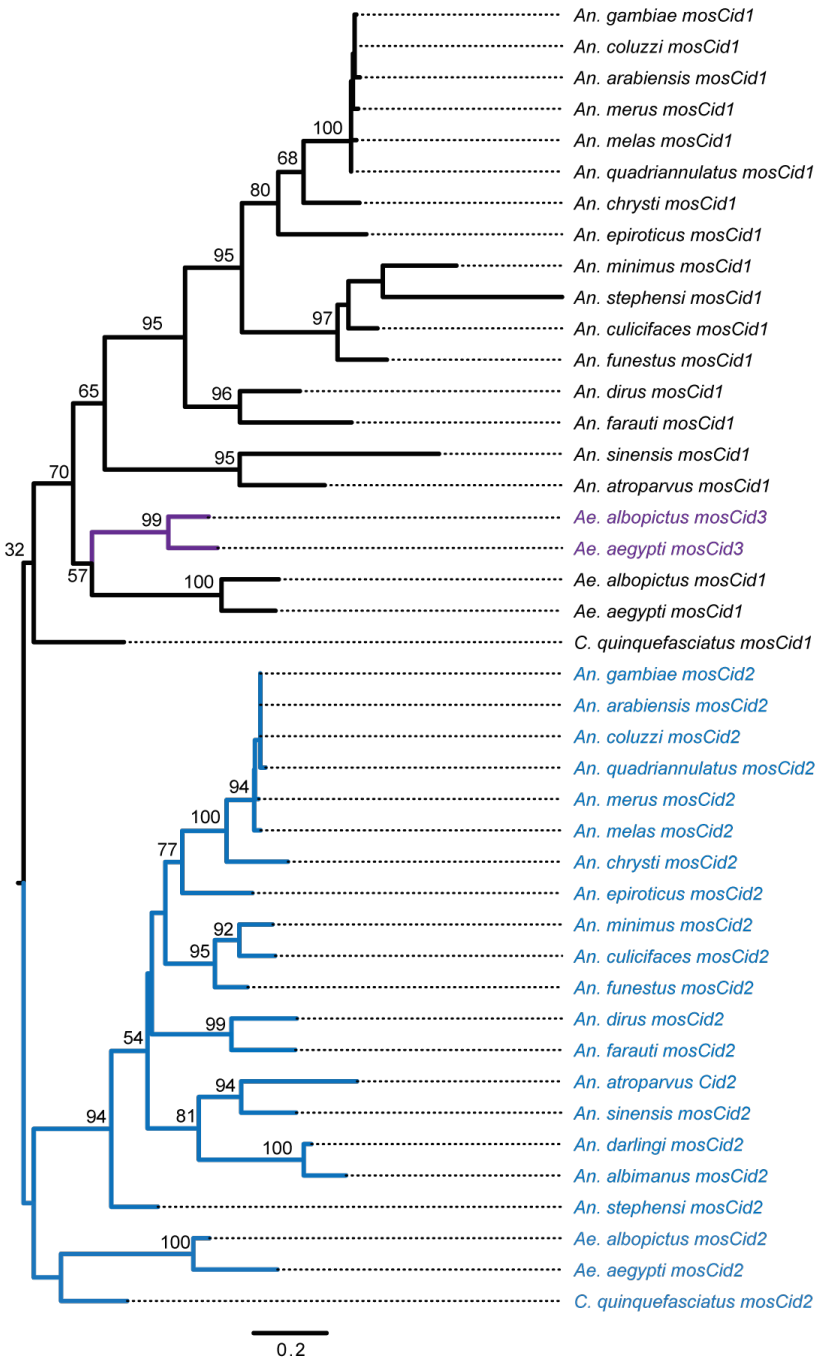


Figure 4-2. Phylogenetic analysis of mosquito *Cid* paralogs. We performed maximum likelihood phylogenetic analyses using PhyML with a nucleotide alignment of the histone fold domain of all *mosCid* paralogs. We found that *mosCid1* (black) forms a monophyletic clade. *mosCis3* (purple) branches within the *mosCid1* clade, indicating that *mosCid3* is derived from a *mosCid1* gene duplication event. All *mosCid2* genes form a monophyletic clade. This suggests that even though *mosCid2* genes are in different syntenic location in *Anopheles, Aedes* and *Culex,* they are likely orthologous. Bootstrap values greater than 50 are shown. The tree is arbitrarily rooted on the common ancestor of *Anopheles, Aedes and Culex* mosquitoes. Scale bar represents number of substitutions per site.

We considered the possibility that at least some of the *mosCid* paralogs no longer encode centromeric proteins and instead have been maintained for another function. To test this possibility, we focused on *Ae. albopictus*, which has three *mosCid* paralogs. We expressed GFP-tagged versions of each of the mosCid paralogs in *Ae. albopictus* cell lines using transient transfections and examined the cytological location of the expressed proteins (Figure 4-3). We found that all three proteins localize to the primary constriction in metaphase chromosomes and to presumed centromeric 'dots' in interphase chromosomes, confirming that all three mosCid paralogs can localize to centromeres.

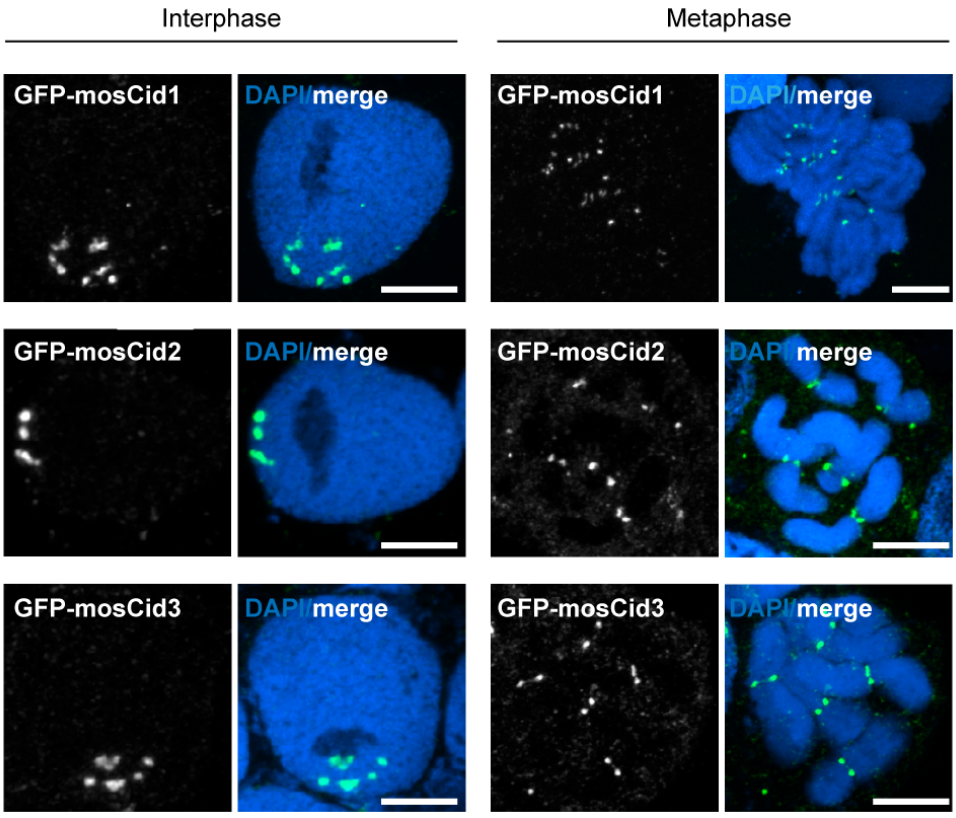


Figure 4-3. Localization of mosCid paralogs in an *Ae. albopictus* cell line. Immunofluorescence images of GFP-tagged mosCid paralogs overexpressed in *Ae. albopictus* cell culture. All Cid paralogs localize to discrete foci in interphase cells and to the primary constrictions on metaphase chromosomes. Scale bar = 2 μm.

Anopheles mosquitoes have two copies of CAL1, the centromeric histone-specific chaperone

Our discovery of ancient *mosCid* paralogs, and previous findings that *Cid* duplication coincided with the duplication of *CENP-C* in some *Drosophila* species[139], motivated us to examine if any other inner kinetochore proteins showed parallel signatures of gene duplication in mosquitoes. Unlike vertebrates, which have a complex network of inner kinetochore proteins[162], *Drosophila* inner kinetochores are relatively less complex, comprised primarily of Cid and CENP-C[58]. Furthermore, Cid physically interacts with and is thought to co-evolve with centromeric histone chaperone, CAL1[51,163]. Although *Drosophila* species harbor multiple *Cid* paralogs without any *CAL1* duplication, we investigated the possibility that the highly divergent *mosCid* paralogs may require different *CAL1* chaperones to aid their deposition.

Cid homologs are relatively easy to identify due to the conservation of their histone fold domains. However, CAL1 homology is less well conserved, and we could obtain only marginal matches to a few mosquito genomes using *D. melanogaster* CAL1 as a BLAST query. We, therefore, adopted an iterative search strategy (see Materials and methods) to successfully identify two previously-known homologs of *CAL1* in *An. gambiae* and *Ae. aegypti*[164]. These genes are both in the same syntenic location and share *Ets97D* as their 5-prime neighbor gene (Figure 4-4). When we extended our search to all *Anopheles, Aedes* and *Culex* genomes, we found *CAL1* in the same syntenic location in all species (Figure 4-4). Surprisingly, we found a second strong BLAST hit but only in *Anopheles* genomes. This *CAL1*-related gene (*CAL1b)* resides in a distinct syntenic location between genes homologous to *D. melanogaster Bruno* and *Chitin binding protein.* We found the presence of *CAL1b* in all *Anopheles* species. We found no additional *CAL1* genes in *Aedes* or *Culex* even with other iterations of BLAST searches in which we used multiple *Anopheles* CAL1 or CAL1b homologs as starting queries. This suggests

that *CAL1b* arose via a gene duplication of *CAL1* in the common ancestor of *Anopheles* species.

Next, we performed phylogenetic analyses on all mosquito *CAL1* and *CAL1b* genes. We made an amino acid based alignment of the conserved N- and C- termini of all CAL1 homologs (the central region of CAL1 cannot be reliably aligned) and used PhyML to make a maximum likelihood phylogeny with 100x resampling (Figure 4-5). We found that *Anopheles* CAL1 and *Anopheles* CAL1b each form monophyletic clades with strong bootstrap support. *Aedes* and *Culex* CAL1 form a well supported outgroup to all *Anopheles* CAL1 and CAL1b proteins. This supports our hypothesis that *CAL1* is the ancestral chaperone and that *CAL1b* arose from a *CAL1* gene duplication event in the common ancestor of *Anopheles* mosquitoes ~100 million years ago (Figure 4-5). Our finding of CAL1 duplication is unprecedented for any centromeric histone chaperone in any eukaryote.

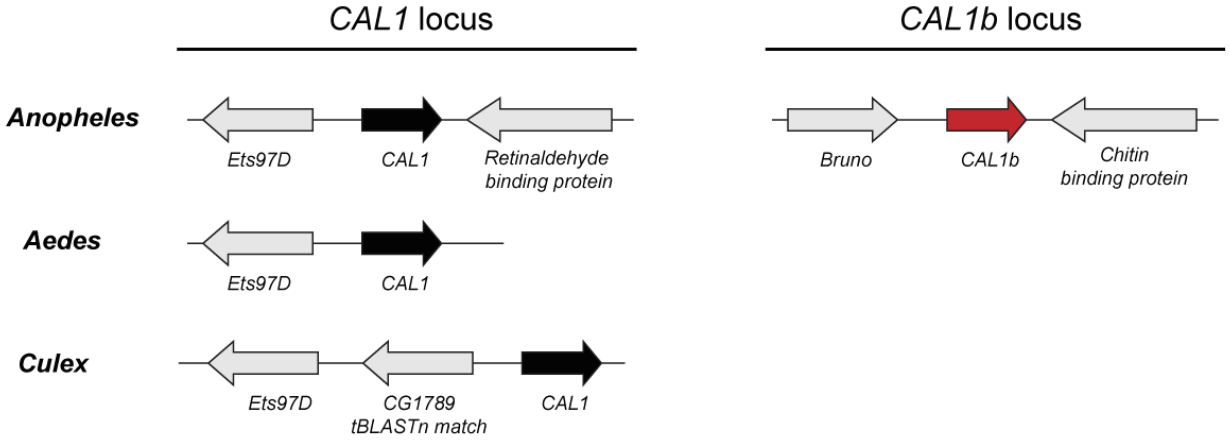


Figure 4-4. Syntenic location of mosquito *CAL1* paralogs. The genomic context of representative mosquito *CAL1* paralogs identified by tBLASTn is schematized for *Anopheles, Aedes* and *Culex.* We found two *CAL1* genes in *Anopheles.* The *CAL1* (black arrow) syntenic locus is defined by the genes *Ets97D* and *Retinaldehyde binding protein.* The *CAL1b* (red arrow) syntenic locus is defined by the genes *Bruno* and *Chitin binding protein.* We only found one *CAL1* gene in *Aedes* and *Culex* present in the conserved *CAL1* locus (black arrows). Arrows colored in grey represent genes that define the shared syntenic locus of each paralog. Genes that define each syntenic locus are named based on the *D. melanogaster* gene name.

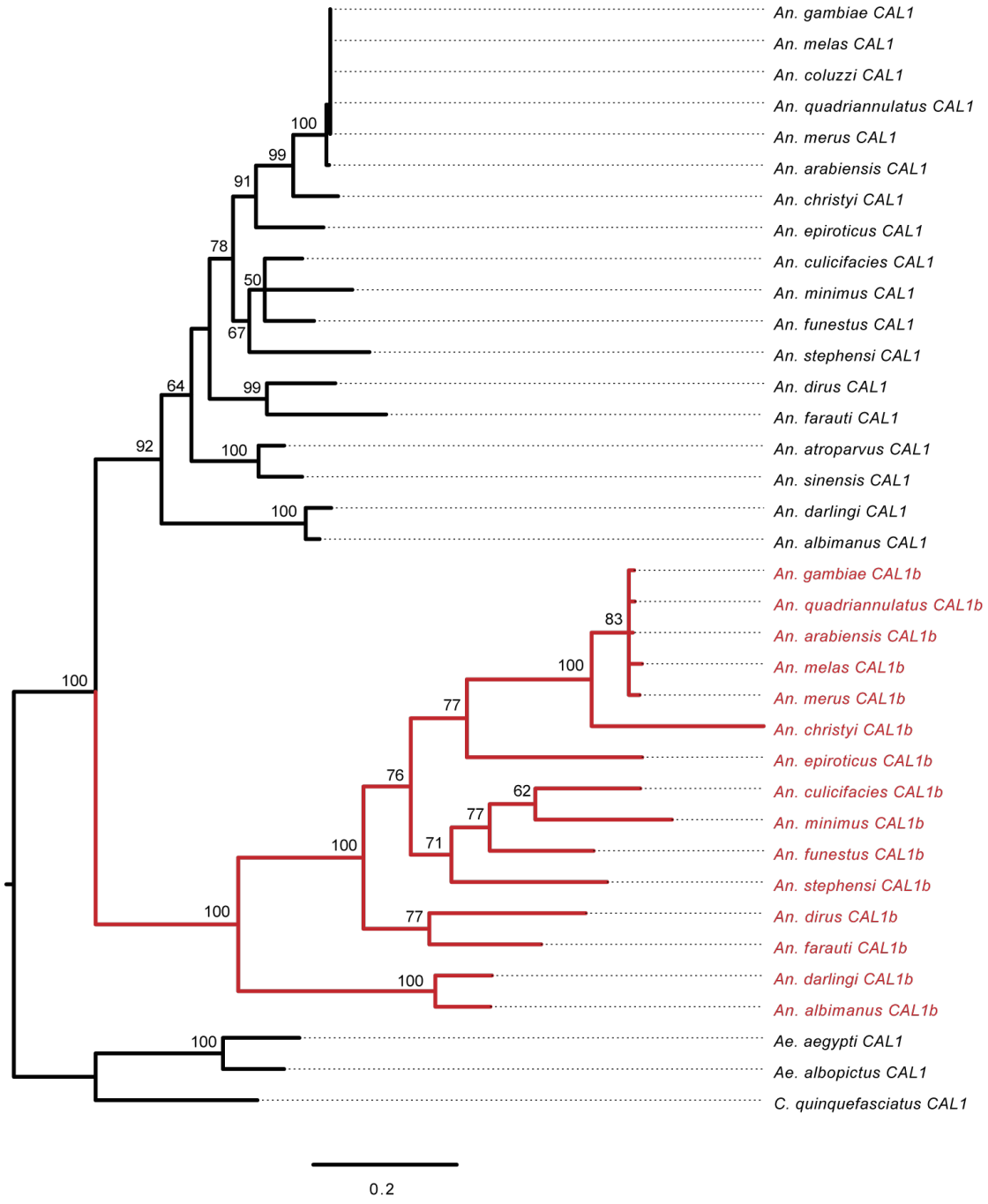


Figure 4-5. Phylogenetic analysis of mosquito *CAL1* paralogs. We performed maximum likelihood phylogenetic analyses using PhyML with an amino acid alignment of the conserved N- and C-terminus of CAL1 and CAL1b. We found that *Anopheles* CAL1 and CAL1b each form well-supported monophyletic clades. *Aedes* and *Culex* CAL1 form a well supported outgroup to Anopheles CAL1 and CAL1b. This suggests that CAL1 was the ancestral chaperone and that CAL1b was born in the common ancestor of *Anopheles* mosquitoes. Bootstrap values greater than 50 are shown. The tree is arbitrarily rooted on the common ancestor of *Anopheles*, *Aedes* and *Culex* mosquitoes. Scale bar represents number of substitutions per site.

Having found paralogs for both *mosCid* and *CAL1* in mosquito genomes, we next examined if they also encoded paralogs of *CENP-C*. At the sequence level, CENP-C is even less conserved CAL1 with only the C-terminal cupin domain being a reliable bioinformatic marker for assigning CENP-C homology[54,165-167]. Therefore, we used the *D. melanogaster* CENP-C cupin domain to identify all mosquito homologs of CENP-C (see Methods). In each case, we were able to find CENP-C orthologs with very high confidence based on the highly conserve cupin domain. However, we discovered no putative paralogs.

Thus, our analyses reveal that the centromeric histone *mosCid* and the centromeric histone chaperone C*AL1* underwent ancient gene duplications in mosquitoes. However, *CENP-C* did not. We compared the duplication histories of *mosCid* and *CAL1* to examine the possibility that two duplications may be causally related (Figure 4-6). If that were the case, we would expect that the *mosCid* and *CAL1* duplications and retention patterns would coincide. In contrast to this expectation, we find that the *mosCid* duplication (in all mosquito species) preceded the *CAL1b* duplication (which is found only in *Anopheles* species). Moreover, even after the loss of *mosCid1* in *An. albimanus* and *An. darlingi*, *CAL1* and *CAL1b* are both retained, arguing against a one-for-one specialization of the two chaperones to correspond with the two *mosCid* paralogs (Figure 4-6).

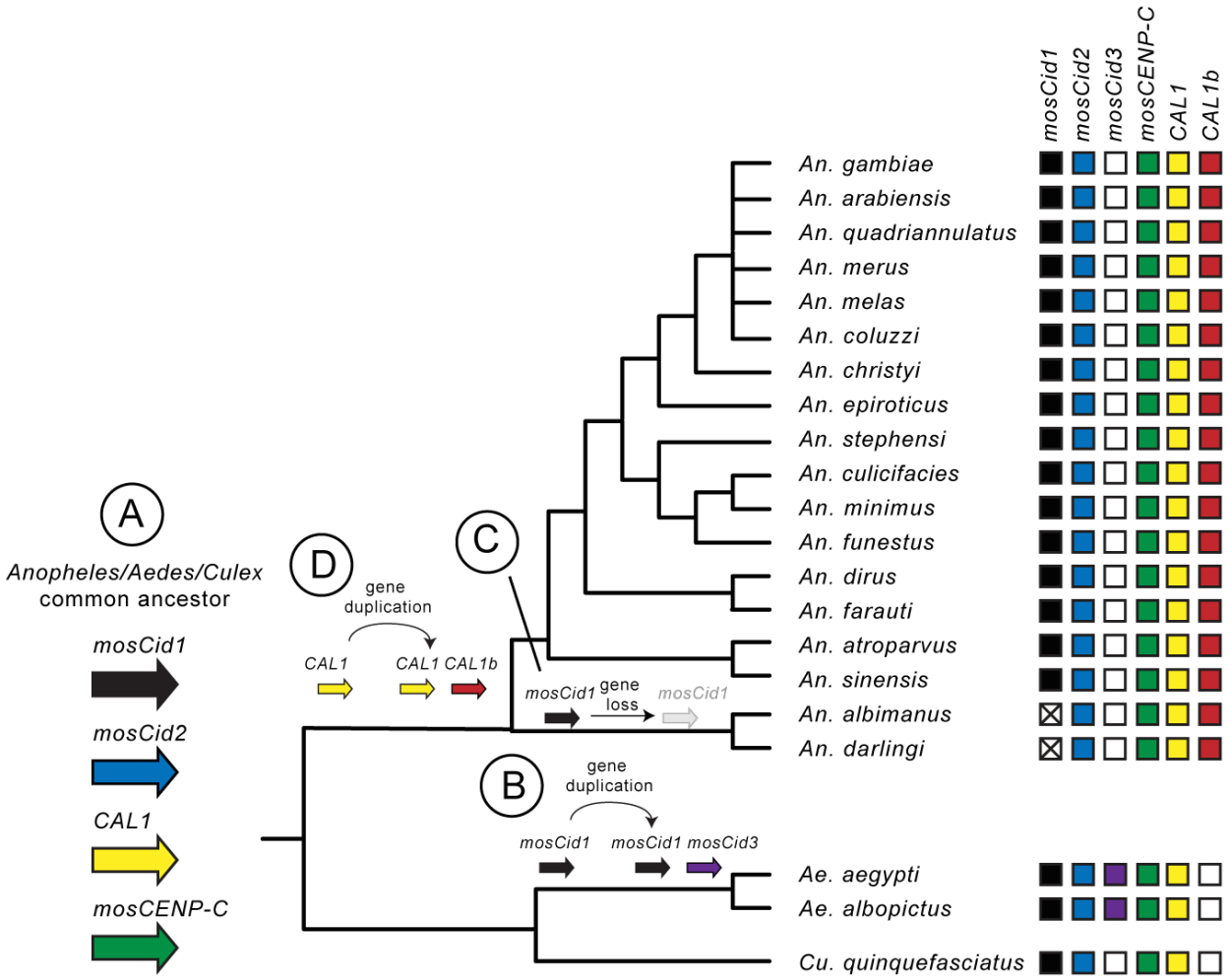


Figure 4-6. Summary of mosquito *Cid, CAL1* and *CENP-C* evolution. A mosquito species tree is presented with boxes to the right of each species indicating the presence (color-filled box) or absence (white box) of each mosquito *Cid*, *CAL1* and *CENP-C* gene. Genes present in the same vertical column are hypothesized to be orthologous. (A) The common ancestor of *Anopheles*, *Aedes* and *Culex* mosquitoes likely had two *mosCid* genes, *mosCid1* (black arrow) and *mosCid2* (blue arrow), making these paralogs over 150 million years old. The *Anopheles/Aedes/Culex* common ancestor also had single copy of *CAL1* (yellow arrow) and *CENP-C* (green arrow). (B) *mosCid1* duplicated in the common ancestor of *Aedes* mosquitoes to give rise to *mosCid3* (purple arrow) 20 – 60 million years ago. (C) *mosCid1* was lost in the common ancestor of *An. albimanus* and *An. darling* (grey arrow with dashed line and white box with an "x" indicates that *mosCid1* is not detectable). (D) *CAL1* (yellow arrow) duplicated in the common ancestor of *Anopheles* mosquitoes to give rise to *CAL1b* (red arrow) approximately 60 million years ago.

Tissue specific expression pattern of mosquito Cid paralogs

The ancient retention of *mosCid* and *CAL1* paralogs in mosquito genomes raised the possibility that these paralogs have acquired specialized functions. In *Drosophila, Cid* paralogs in the montium subgroup and the *Drosophila* subgenus have independently acquired male-germline biased expression patterns[125]. Moreover, our detailed cytological analysis of the two Cid paralogs in *D. viriis* suggested that Cid1 and Cid5 play specialized roles in male and female germlines (Chapter 3). We wondered whether *mosCid* and *CAL1* paralogs have also acquired tissue specific expression patterns, which could be indicative of specialized function. Since we did not have tools to query the cytological localization pattern of the mosquito paralogs *in vivo*, we instead analyzed previously published genome-wide RNA-seq analyses to discern any evidence for tissue-specific specialization.

One of the most useful datasets was published by Biedler *et al.,* who investigated female-specific gene expression in *An. stephensi* at several life stages including the ovary prior to blood feeding, the ovary 24 hours after blood feeding, the early embryo*,* larvae, pupae and adult males and females[168]. We examined the expression of *mosCid1, mosCid2, CAL1,* and *CAL1b* in each of these tissues. We found that *mosCid1, mosCid2, CAL1* and *CAL1b* are all expressed at relatively high levels in ovaries and the early embryo and at fairly low levels in larvae, pupae, and adults (Figure 4-7A). Interestingly, expression of *mosCid2* increases more than 10-fold from the non-blood fed (NBF) ovary to the ovary 24 hours after blood feeding. In mosquitoes, blood feeding induces oogenesis. This might suggest that *mosCid2* plays an important function in female meiosis or female gamete development. In contrast to *mosCid2*, blood feeding did not alter expression of *mosCid1, CAL1,* and *CAL1b*.

Many *Drosophila Cid* paralogs show testis-biased expression. However, the Biedler *et al.* study did not investigate gene expression in testes[168]. Therefore, we decided to examine the expression of *mosCid1* and *mosCid2* by RT-qPCR in dissected adult tissues (including testes

and ovaries) from *An. stephensi* mosquitoes (Figure 4-7B). We found that expression of both *mosCid* genes was highest in the germline tissues but expression of *mosCid2* was nearly 40 times higher than *mosCid1* in testes and approximately 20 times higher than *mosCid1* in ovaries. Based on this analysis, we cannot infer that either paralog has a testes-specific function. However, we cannot rule out the possibility that mosCid1 and mosCid2 are specialized for specific germline cell types, as is the case for Cid1 and Cid5 in *D. virilis.* For example, if mosCid1 were specialized for or retained on sperm, RNA-seq analyses of whole testes would not have the resolution to identify this.

We wished to extend our survey to include a species distantly related to *An. stephensi*. We focused on *Ae. aegypti* because it has three mosCid paralogs: *mosCid1, mosCid2* and the *Aedes*-specific *mosCid3.* Using RNAseq data previously published by Akbari *et al.*[169], we found that *Ae. aegypti mosCid2* expression in the ovary dramatically increased after blood feeding, then gradually decreased over the first 24 hours of embryonic development (Figure 4-7C) just like it did for *mosCid2* in *An. stephensi* (Figure 4-7A). This further supports our conclusion of orthology between these two genes. Concurrent with the decrease in *mosCid2* expression, we found that the expression of the *Aedes*-specific paralog, *mosCid3,* rose rapidly over the first 36 hours of embryonic development and then decreased from 36 to 76 hours after the onset of embryonic development (Figure 4-7C). This suggests that *mosCid3* might have specialized function in embryogenesis. Expression of *Ae. aegypti mosCid1* is quite low at all stages, leaving its potential function unclear. Notably, all *Aedes mosCid* paralogs are expressed at low levels in testes. In summary, unlike in *Drosophila* where *Cid* paralogs have acquired testis-biased expression patterns*, mosCid* paralogs appear to have acquired ovary-biased expression patterns.

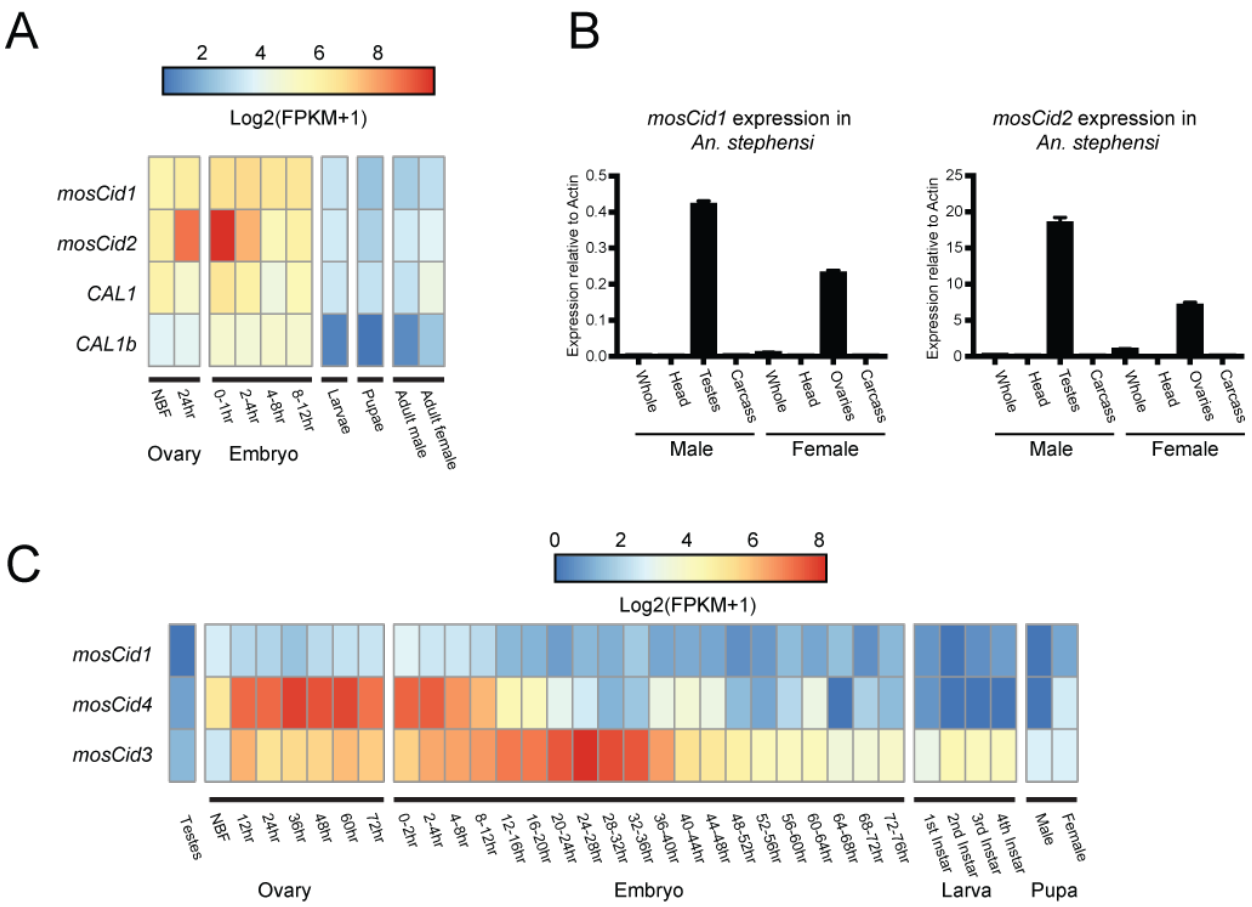


Figure 4-7. Expression of *mosCid* and *CAL1* paralogs. (A) Heatmap representing Log2(FPKM+1) values for *An. stephensi mosCid1, mosCid2, CAL1* and *CAL1b* at various developmental stages. *mosCid2* expression increases in the ovary after blood feeding. (B) RT-qPCR for *mosCid1* and *mosCid2* from dissected tissues from *An. stephensi* males and females revealed that both *mosCid1* and *mosCid2* are expressed in testes and ovaries but *mosCid1* is expressed at higher levels than *mosCid2.* All RT-qPCR was normalized using *Actin* as a control. Error bars represent standard deviation calculated from three technical replicates.

Distinct selective pressures shape the mosCid paralogs

Our finding that both *mosCid* and *CAL1* paralogs have been co-retained for more than 100 million years of mosquito evolution (with one exception, Figure 4-6) suggests that all paralogs perform important, non-redundant functions. If *mosCid* and *CAL1* paralogs are selectively retained to perform specialized roles, we expect that they might evolve under different selective pressures, as we previously found to be the case for some *Cid* paralogs in *Drosophila*[125] (Chapter 2).

We focused our attention to only a subset of *Anopheles* species because these are the most densely sampled in terms of sequence and because of their moderate divergence, which allowed us to evaluate selective constraints without the confounding effect of saturated synonymous site substitutions. We used maximum likelihood methods in the PAML suite of programs to analyze the selective constraints, comparing rates of non-synonymous (dN) to synonymous (dS) substitutions for each codon of *mosCid1, mosCid2, CAL1, and CAL1b*. We did not detect evidence for positive selection in *CAL1* or *CAL1b*. However, we found that *mosCid* paralogs evolve under different evolutionary constraints. Specifically, we found that *mosCid1* evolves rapidly (Table 4-1, mosCid1 M8a vs M8 p-value = 0.004) whereas *mosCid2* does not. Moreover, we found that nearly one-quarter (22%) of the codons in *mosCid1* evolve with a dN/dS of 2.5. Our finding that only *mosCid1* evolves under positive selection is consistent with the hypothesis that *mosCid1* rapidly evolves to suppress centromere drive in mosquito genomes, like has been hypothesized for *Cid* in *Drosophila* species. This finding is also consistent with our intralocus conflict hypothesis, which posits that gene duplication followed by specialization allows one *mosCid* paralog to evolve rapidly without compromising essential centromeric function mediated by the other paralogs. Our observations are highly reminiscent of our previous findings in *Drosophila*, where one of multiple *Cid* paralogs usually showed signatures of positive selection[125].

Table 4-1: PAML tests for positive selection on *mosCid* and *CAL1* paralogs. Summary table of M8a vs. M8 PAML results for *mosCid1, mosCid2, CAL1* and *CAL1b.* P-values less than 0.05 are indicated in bold text.

	Number of sequences	Alignment length (#nts)	M8a vs M8 p-value	Omega (% sites)	Tree length
mosCid1	7	705	**0.004**	2.5 (22%)	2.04
mosCid2	7	777	0.90	n.a.	1.64
CAL1	8	1671	0.92	n.a.	1.86
CAL1b	7	1740	0.99	n.a.	4.33

Another potential means for functional specialization of the *mosCid* paralogs could be via different protein-protein interaction networks. We previously showed that *Drosophila* Cid paralogs acquired and lost N-terminal motifs[125] that may mediate protein-protein interactions with other kinetochore proteins. First, we queried all *Anopheles* mosCid1 and mosCid2 protein sequences using the motifs we previously identified in the N-terminal tail of *D. melanogaster* Cid paralogs. The only modest hit was using *Drosophila* motif 3, primarily due to a stretch of acidic amino acids. Since it was clear that the *Drosophila* Cid motifs were generally not conserved in mosquito Cid paralogs, we investigated whether mosquito Cids have their own unique set of N-terminal tail motifs. One caveat to this approach is that discovery of motifs requires a minimum number of orthologs. Thus, whereas we are able to discover motifs *de novo* using the 18 *Anopheles* genomes, we can only ascertain whether these motifs were conserved in the two *Aedes* and single *Culex* species' mosCid paralogs.

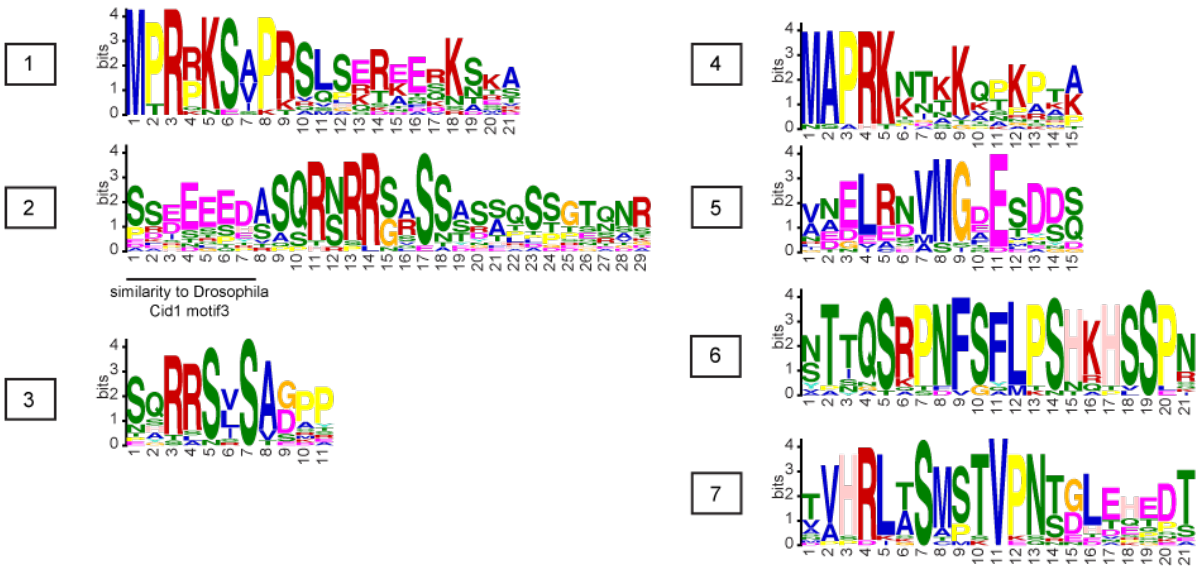
We used the motif generator algorithm MEME to identify regions of conservation in the N-terminal tails of all *Anopheles* mosCid1s and all *Anopheles* mosCid2s (Materials and methods). Next, we used the motif search program MAST to search for mosCid1 motifs in mosCid2 sequences and mosCid2 motifs in mosCid1 sequences. This analysis revealed that mosCid1 and mosCid2 have almost entirely unique N-terminal tails with essentially non-

overlapping motifs. All mosCid1 proteins contain motifs 1 – 3 (Figure 4-8A, Figure 4-8B). The only slight variation on mosCid1 motif content is in *An. chrysti* where motifs 2 and 3 each occur twice in the N-terminal tail, supporting the modularity of these domains. When we searched for mosCid1 motifs in mosCid2s, the only modest hit was from mosCid1 motif2, which also shares similarity to the *D. melanogaster* acidic motif (Figure 4-8B). In general, mosCid2s have their own unique set of N-terminal tails motifs (motifs 4 – 7, Figure 4-8A, Figure 4-8B). The only exceptions were *An. farauti* and *An. dirus,* in which we could only find reliable matches to motif 4 in the N-tail of mosCid2.

Next we looked for all mosCid motifs in *Aedes* and *Culex* mosCid paralogs. This revealed no significant matches to *Aedes* or *Culex* N-terminal tails using the *Anopheles* motifs as a query. *Aedes* and *Culex* almost certainly have their own set of N-tail motifs but we do not have enough sequences to determine what they might be. It is not surprising that the *Anopheles* motifs do not match the *Aedes* or *Culex* mosCid sequences because *Anopheles* and *Aedes* shared a common ancestor over 150 million years ago, far more ancient than the ~60 million year old divergence we previously analyzed in *Drosophila* species[125].

In summary, our motif analysis revealed that mosCid1 and mosCid2 are subject to distinct selective pressures and have highly divergent N-terminal tails. While the function of the N-terminal tail remains mostly unknown, we hypothesize that different N-tails is indicative of different protein-protein interactions, possibly due to functional specialization.

A



B

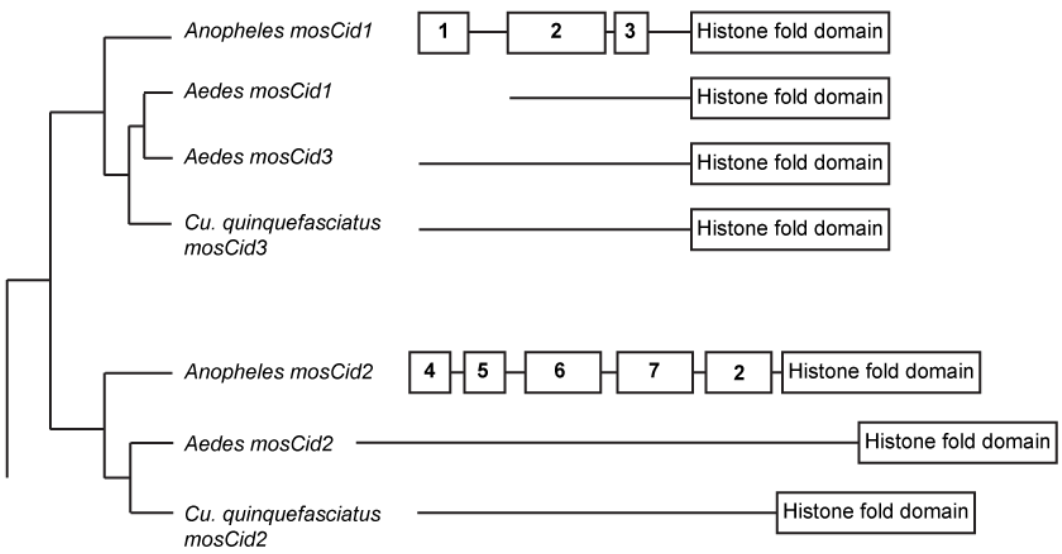


Figure 4-8. Analysis of N-terminal motifs in mosquito Cid proteins. (A) Logos generated by MEME for mosCid consensus motifs 1 – 7. (B) A mosquito species tree with a schematic of N-terminal tail motifs identified by MEME and MAST displayed to right of each species or species group. Each number represents a unique motif that does not statistically match any other motif in the figure.

4.3 Discussion

Although centromeric histone duplications were previously thought to be rare and short-lived in animal species, we previously identified multiple, ancient duplications of *Cid* paralogs during *Drosophila* evolution[125]. Indeed, the majority of *Drosophila* species likely encode more than one *Cid* paralog. In one lineage, both *Cid* and *CENP-C* paralogs have been co-retained for at least 40 million years without loss[139]. These findings suggested that centromeric protein duplications might be both more common as well as more long-lived than previously believed. In the present study, we further confirm this hypothesis by finding that most mosquito species encode two divergent *mosCid* paralogs that diverged at least 150 million years ago. Thus, co-retention of CenH3 paralogs is not an infrequent occurrence in insect genomes. In addition, all *Anopheles* species encode two divergent paralogs of *CAL1*, which is the centromeric histone chaperone in Diptera.

Their ancient co-retention suggests that the *mosCid* paralogs must have diverged in some aspect of their centromeric function. Indeed, just like the *Drosophila* Cid paralogs[125], we find multiple pieces of evidence suggestive of their specialization. First, *mosCid1* evolves under positive selection whereas *mosCid2* does not. Second, we find that both mosCid1 and mosCid2 have different motifs conserved in their N-terminal tails, presumably for different protein-protein interactions with other kinetochore factors. Third, the expression patterns of the two paralogs dramatically vary; *mosCid2* is highly expressed in the germline of both *Aedes* and *Anopheles* species, especially during oogenesis that follows blood-feeding.

Although it encodes a centromeric protein (Figure 4-3), the function of *mosCid1* remains perplexing for a number of reasons. First, the expression pattern of *mosCid1* is quite low in bulk RNA-seq analysis both through development and in adult tissues. This is in spite of the fact that *mosCid3*, itself a recent duplicate of *mosCid1* in *Aedes*, appears to be abundantly expressed during embryogenesis. However, these observations are not strong evidence of lack of function

since *mosCid1* could be expressed during a critical stage of the mosquito life-cycle that is not captured in bulk tissue RNA-seq analyses. Indeed, it would be hard to explain either the ancient retention of *mosCid1* or its positive selection if it were not performing some important function. Nevertheless, the possibility that its function can be rendered superfluous is bolstered by our finding that *mosCid1* appears to have been lost in at least two species: *An. albimanus and An. darlingi.* Although we cannot entirely rule out that *mosCid1* has transposed to a different syntenic location that is not captured in the genome sequencing efforts, we find this possibility to be unlikely. This suggests that *mosCid2* is capable of performing all CenH3 functions, at least in these two species. Loss of *mosCid1* did not result in any dramatic changes in the *mosCid2* orthologs in these species, relative to other orthologs.

There is precedent for loss of previously essential chromatin proteins that is correlated with changes in chromosome architecture. For example, four insect lineages independently lost otherwise essential CenH3 proteins, coincident with their independent transitions from monocentric to holocentric chromosomes[29]. Similarly, the heterochromatin protein, *HP1E,* which is essential for early embryonic development in *D. melanogaster,* was lost in species in the *D. pseudoobscura* group, concurrent with dramatic karyotypic changes of their sex chromosomes[153]. It will be interesting to evaluate in the future whether similar chromosome configuration changes distinguish *An. albimanus and An. darling* from other mosquito species that could explain their loss of *mosCid1.* This could also provide a cogent hypothesis for the function of mosCid1 in the majority of mosquito species that have retained it.

Ultimately, dissection of the function of the various paralogs will require tools to closely examine their cytological localization and retention patterns as we have done for the Cid1 and Cid5 paralogs in *D. virilis*, and means to genetically knockout individual paralogs and assess function, ideally in multiple mosquito species. The development of robust Cas9- mediated techniques for genetic knockouts in mosquito species[170,171] will facilitate these studies in the future.

Equally uncertain is the functional specialization of the *CAL1* paralogs in *Anopheles* species. Although their initial identification raised the intriguing possibility that each CAL1 paralog may have specialized interactions with each mosCid paralogs, the pattern of retention and duplication is not congruent for both genes (Figure 4-6). For example, *mosCid* duplication preceded *CAL1* duplication by >50 million years, and loss of *mosCid1* did not lead to loss of either *CAL1* paralog in *An. albimanus and An. darlingi.* Again, better cytological and genetic tools will allow us to investigate the functional implications of this unprecedented finding of duplication of this centromeric histone chaperone in mosquitoes.

In sum, our work suggests that mosquito species might present an excellent opportunity to study the functional specialization of centromeric proteins, in turn providing insight into their multiple functions.

4.4 Materials and methods

Identification of mosCid, CENP-C and CAL1 orthologs and paralogs

Mosquito *Cid* genes were identified in previously sequenced genomes using *D. melanogaster* Cid1 histone fold domain to query mosquito genomes using tBLASTn implemented in Vectorbase[161]. Many *mosCid* BLAST hits were not annotated genes or were mis-annotated and required manual identification of the *mosCid* gene open reading frame. For identifying mosquito *CAL1* homologs, we employed an iterative tBLASTn search strategy in which we first used *D. melanogaster* CAL1 to first identify homologs in other intermediate branching Dipteran species including *Glossina morsitans*. Subsequently, we used *G. morsitans* CAL1 to identify CAL1 homologs successfully from all mosquito species using BLASTp and tBLASTn searches. To assign CENP-C homoly, we relied on the C-terminal cupin domain as the only reliable bioinformatic marker. We used the *D. melanogaster* CENP-C cupin domain to do BLASTp searches of the predicted mosquito proteomes to first identify putative CENP-C orthologs in the

well annotated *An. gambiae* and *Ae. aegpytii* genomes. We then used these predicted mosquito proteins as queries in iterative BLASTp and tBLASTn searches to identify all mosquito homologs of CENP-C. We also recorded the syntenic locus (3' and 5' flanking genes) of each gene hit as indicated by the Vectorbase genome browser track and by homology to genes in *D. melanogster*. Each *mosCid, CAL1 and CENP-C* gene was named according to its shared syntenic location and phylogenetic relationship to other paralogs if present.

Phylogenetic analyses

mosCid sequences were aligned using the ClustalW[115] "translation align" function in the Geneious software package (version 6)[116]. Alignments were further refined manually, including removal of poorly aligned regions. Maximum likelihood phylogenetic trees of *Cid* nucleotide sequences were generated using the HKY85 substitution model in PhyML[172], implemented in Geneious, using 100 bootstrap replicates for statistical support. Amino acid alignments of CAL1 and CAL1like were generated using ClustalW function in the Geneious software package. Neighbor-joining phylogenetic trees[173] of CAL1 and CAL1b protein sequences were generated using the Jukes-Cantor model for genetic distance and implemented in the Geneious tree builder in the Geneious software package. Phylogenies were visualized using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Positive selection analyses

We used the PAML suite of programs[174] to test for positive selection on *mosCid1, mosCid2, Cal1 and Cal1b* in *Anopheles.* Alignments for each gene paralog were generated and manually refined as described above. We chose a subset of *Anopheles* species (*An. coluzzi, An. gambiae, An. arabiensis, An. melas, An. merus, An. quadriannulatus* and *An. chrysti*) for these analyses in order to maintain high-confidence alignments across the full length of each gene.

Alignments and gene trees were used as input into the CODEML NSsites model of PAML. To determine whether each *mosCid* or *CAL1* paralog evolves under positive selection, we compared a model that do not allow dN/dS to exceed 1 (M8a) to a model that allows dN/dS > 1 (M8). Positively selected sites were classified as those sites with a M8 Bayes Empirical Bayes posterior probability > 95%.

Heatmaps generated from previously published RNAseq experiments

To visualize the expression of *Ae. aegypti CenH3* paralogs across multiple developmental time points we generated a heatmap in R using FPKM values from Akbari *et al.*[169]. All *Ae. aegypti CenH3* paralogs were already annotated, so we simply looked up the corresponding FPKM values in Akbari *et al.* supplementary data. To examine the expression of *CenH3* and *Cal1* paralogs in *An. stephensi,* we used RNAseq data from Bielder *et al.*[168]. We manually added an annotation for *mosCid1* to the *An. stephensi* GFF3 file and then calculated FPKM values for all four genes (*mosCid1, mosCid2, Cal1 and Cal1b*) using cufflinks[175]. Heatmaps for *Anopheles* expression data were generated in R.

Cloning GFP-mosCid fusion proteins

mosCid genes from *Ae. albopictus (mosCid1, mosCid2* and *mosCid3)* were amplified from genomic DNA and cloned into pENTR/D-TOPO (ThermoFisher). We used LR clonase II (ThermoFisher) to directionally recombine each *mosCid* gene into a destination vector from the Drosophila Gateway Vector Collection, generating N-terminal Venus (pHVW) fusion under the control of the *D. melanogaster* heat-shock promoter.

Transfection and imaging of Ae. albopictus tissue culture cells

The *Ae. albopictus* cell line C6/36 (a gift from Alan Goodman) was used for all transfection experiments. One microgram of plasmid DNA was transfected using Xtremegene HP transfection reagent (Roche) according to the manufacturer's instructions. Cells were heat-shocked at 37°C for one hour 24 hours after transfection to induce expression of the mosCid fusion protein. Cells were transferred to a glass coverslip 24 hours after heatshock. Cells were treated with 0.5% sodium citrate for 10 min and then centrifuged on a Cytospin III (Shandon) at 1900rpm for 1 min to remove cytoplasm. Cells were fixed in 4% PFA for 5 min and blocked with PBSTx (0.3% Triton) plus 3% BSA for 30 minutes at room temperature. Coverslips with cells were incubated with primary antibodies at 4°C overnight at the following concentration: chicken anti-GFP (Abcam AB13970) 1:1000. Coverslips with cells were incubated with secondary antibodies for 1 hour at room temperature at the following concentration: goat anti-chicken (Invitrogen Alexa Fluor 488, A-11039) 1:5000. Images were acquired from the Leica TCS SP5 II confocal microscope with LASAF software.

Chapter 5. Discussion and future prospects

5.1 Summary of dissertation

Despite the fact that all eukaryotes last shared a common ancestor close to a billion years ago, there is a remarkable conservation in their chromosome segregation mechanisms. It was previously believed that all eukaryotes rely on a core set of centromeric proteins like CenH3 that are almost universally found (except in trypanosomes[28,176], holocentric insects[29], and some other lineages); this universality has now been questioned by phylogenomic studies[177,178]. The assumption has been that when present, centromeric proteins perform similar functions. For instance, when present, CenH3 proteins localize to the centromeric DNA, establish a specialized centromeric chromatin, and help recruit kinetochore proteins to mediate chromosome segregation.

Centromeric proteins are assumed to perform the same function in different tissues including in somatic and germline cells. However, the evolution of sex and multicellularity in eukaryotes raised the possibility that all centromeric proteins may not be subject to the same functional constraint in every tissue. For instance, organisms that undergo asymmetric female meiosis, in which only one of four meiotic products is retained, can be subject to centromere-drive, that in turn may drive rapid evolution of centromeric proteins[129]. Moreover, special packaging requirements in male gametes that lead to bulk histones being replaced by protamines or analogous packaging proteins[179] would further affect how centromeric proteins (especially CenH3) are retained. My dissertation research investigating the evolution and cytological localization of divergent centromeric histone paralogs, suggests that somatic and germline or male and female centromeric functions are indeed distinct. I hypothesize that it might be evolutionary advantageous to encode multiple CenH3 paralogs to carry out the multiple CenH3 functions without compromising any individual function.

In my dissertation research, I investigated the molecular evolution and cytological localization of duplicate centromeric histone genes in *Drosophila* and mosquitoes. First, contrary to what was previously believed in animal species, I showed that duplicate centromeric histone genes are common and long-lived in *Drosophila*[125]. I showed that *Cid* experienced at least four independent gene duplications during *Drosophila* evolution and that *Drosophila Cid* paralogs have been almost entirely preserved since birth, suggesting that they perform non-redundant centromeric functions. Intriguingly, I found some evidence that indicates that *Cid* paralogs have acquired specialized functions including the observation most *Cid* duplicates are primarily expressed in the germline and often evolve under distinct selective constraints. Moreover, each paralog encodes a unique N-terminal tail, which may provide the basis for paralog-specific protein-protein interactions. This study[125] challenged the idea that duplicate centromeric histone genes are rare in animals. A subsequent study found a fifth *Cid* duplication in *D. buzzatii* as well as the concomitant duplication of a second essential inner kinetochore protein, CENP-C, in the *Drosophila* subgenus[139], arguing that multiple centromeric proteins may be subject to similar evolutionary dynamics.

Next, I investigated the cytological localization of two specific *Drosophila* Cid paralogs in *D. virilis,* Cid1 and Cid5. I found that Cid1, the paralog most closely related to *D. melanogaster* Cid, is the primary centromeric histone protein in somatic cells. However, the male and female germline contain both Cid1 and Cid5. Strikingly, I found that Cid1 and Cid5 co-localize at centromeres of mitotic germ cells in in ovaries and testes, but then are alternately retained at later stages, suggesting that Cid1 is the primary centromeric histone in female meiosis and the mature oocyte and Cid5 is the primary centromeric histone in male meiosis and mature sperm. This result is especially intriguing because in *D. melanogaster,* paternal Cid (the Cid1 homolog) is absolutely essential. Degradation of Cid in sperm results in embryonic lethality[23]. That means that specialization of Cid1 and Cid5 may have rendered an otherwise essential protein, Cid1, dispensable for transgenerational centromere inheritance through the male germline. This study

highlights the multiple functions of Cid in the soma-versus-germline and suggests that Cid1 and Cid5 have specialized roles in male and female gametes.

Finally, I reasoned that if duplicate *Cid* genes were evolutionarily advantageous, I would expect to find additional instances of *Cid* duplication, even outside of *Drosophila*. Therefore, I expanded my evolutionary analysis to include *mosquito Cid* (*mosCid*) paralogs. This analysis revealed that *mosCid* has duplicated at least twice during mosquito evolution and that, like *Drosophila*, most mosquito genomes encode more than one *mosCid* gene. My phylogenetic analyses indicated that *mosCid* likely duplicated in the common ancestor of *Anopheles* and *Aedes* mosquitoes*,* making the *mosCid* paralogs over 150 million years old, the oldest centromeric histone paralogs known to date. I also found evidence supporting the hypothesis that *mosCid* paralogs have acquired specialized functions based on high divergence of their N-terminal tails and different selective pressures; *mosCid1* evolves rapidly under positive selection but *mosCid2* does not. Furthermore, *mosCid* paralogs show different tissue specific expression patterns; *mosCid2* expression is enhanced during oogenesis but *mosCid1* expression remains constant throughout development. Therefore, mosquito *Cid* paralogs may have also acquired specialized germline functions. In parallel, I also found gene duplications of the histone chaperone CAL1 in Anopheles mosquitoes. Although I did not find strong evidence of tissue-specific specialization or differing selective pressures on the CAL1 paralogs, their retention without loss for 100 million years suggests that they are not functionally redundant.

5.2 Understanding the basis for functional specialization of Cid paralogs

My evolutionary and cell biological studies lead to the tantalizing hypothesis that centromeric function is not identical between the soma and the germline or between females and males. My research provides the foundation for investigating the function of centromeres in different cellular contexts. Since many of the paralogs are not expressed in tissue culture cells,

the most direct way to test for the functional specialization is to genetically knockout or knockdown individual paralogs *in vivo*. I have attempted to do this in *D. virilis* (see Appendix 1) although my efforts were slowed by the dearth of tools in non-model *Drosophila* species. Doing similar knockout or knockdown experiments in *Anopheles gambiae* or *Aedes aegypti*, two model mosquito species, will allow us to evaluate the functional consequences of loss of either of the specialized mosCid or CAL1 paralogs.

It is possible that different expression patterns alone could explain the divergence in *Cid1* and *Cid5* function, and that divergence in their protein sequence is inconsequential. If *Cid1* and *Cid5* knockdown (Appendix A) produce measurable phenotypes, we could address this possibility by trying to rescue *Cid5* knockdown using a *Cid1* recoded transgene driven by a *Cid5* promoter, and vice-versa (Figure 5-1A). If such a rescue does not work, it will confirm that protein divergence has contributed to functional specialization. We could then further refine the specific regions of Cid1 and Cid5 that have contributed to specialization by attempting to rescue knockdown phenotypes with chimeric Cid1-Cid5 transgenes. Given that the histone fold domains of Cid1 and Cid5 proteins are quite closely related but their N-terminal tails are quite diverged, I predict that the histone fold domains may be interchangeable but the N-terminal tails may be required for paralog-specific function (Figure 5-1B). Experimentally, this means that rescue of *Cid1* knockdown with a chimeric protein containing the Cid1 N-terminal tail and the Cid5 histone fold domain might rescue to the same degree as the complete Cid1 protein. If this were the case, it would imply that the functional specialization between *Cid1* and *Cid5* has taken place via divergence of their N-tail domains.

Differences in the Cid1 and Cid5 interaction network

The centromeric histone N-terminal tail has been shown to be important for the recruitment of kinetochore proteins in humans and fission yeast[52,99,100,180]. My evolutionary analysis showed that Cid1 and Cid5 in *D. virilis* share very little homology in their N-terminal

tails, and, instead, they possess distinct sets of conserved sequence motifs. I hypothesize that Cid1 and Cid5 could utilize their distinct N-terminal tails to help recruit specialized sets of kinetochore protein, thereby diverging in function.

To determine Cid1- and Cid5-specific protein-protein interactions we could perform immunoprecipitation followed by mass spectrometry using the Cid1GFP and Cid5mCherry transgenes I have already created in testes and ovaries. As a negative control, we could compare both of these conditions to peptides obtained through IP-MS of UAS-RFP expressed by the Nanos-Gal4 driver in the *D. virilis* germline. We could also compare the results of our proteomic analysis in germline tissues with parallel proteomic analyses in *D. virilis* somatic cell lines. This will identify germline-specific protein-protein interactions. If we find that Cid1 and Cid5 have unique sets of protein-protein interaction partners, this would support the hypothesis that Cid1 and Cid5 have specialized functions. Moreover, it would indicate that centromere composition varies between germline versus somatic tissues in *D. virilis* cells.

Subfunctionalization versus specialization

The Cid paralogs we have observed could be the result of subfunctionalization, in which two daughter genes now carry out functions that were previously performed by a single, ancestral gene. This could occur, for example, because the paralogs are expressed in different tissues in a non-overlapping manner; thus, expression of both paralogs is required for full function. If this were the case, however, one might expect relatively little divergence of the protein coding regions, since all functionality needs to be preserved by both paralogs. This process is simply the result of neutral mutations, in which the initial redundancy of gene function upon gene duplication results in complementary, distinct mutations of functional attrition in each paralog. Instead, we see that together with changes in expression patterns, the long-term co-retention of divergent Cid paralogs in *Drosophila* and mosquito lineages is accompanied by significant changes to the N-terminal tail as well as by differences in selective constraints. We

speculate that this indicates that these Cid paralogs perform specialized functions, which may have divergent protein coding requirements. Moreover, we propose that this process was the result of adaptive evolution rather than neutral mutation.

My evolutionary and cytological characterization of Cid paralogs led me to hypothesize that retention of multiple, specialized Cid paralogs may be advantageous. I propose that single copy *Cid* genes may carry out multiple functions that have divergent fitness optimum (intralocus conflict), and that these multiple functions cannot be simultaneously optimized in one gene. *Cid* duplications can resolve this conflict by allowing multiple functions to be simultaneously optimized in separate genes. This specialization is distinct from subfunctionalization because it is driven by adaptive gain-of-function mutations rather than neutral mutation.

However, distinguishing subfunctionalization from the specialization that would be consistent with intralocus conflict is challenging. One hint that this may be the case is the acquisition of completely new motifs in Cid1 and Cid5 that are distinct from the ancestral copy of the gene. Although it is possible that this acquisition could follow the initial subfunctionalization as each paralogs improves its 'new' function, it is unclear why this adaptive value would not also be beneficial to the ancestral, single copy gene. The fact that new motifs were gained, and that different motifs were gained by Cid1 and Cid5, is more consistent with the hypothesis that the adaptive value of these motifs is offset by the cost of these motifs for other functions. This would not be tolerated if a single Cid gene encoded all functions. It is this 'cost' that can experimentally discriminate subfunctionalization from intralocus conflict. If the intralocus conflict hypothesis is correct, I predict that Cid1 and Cid5 have evolved in ways that might be detrimental in the wrong context. Specifically, the newly evolved sequence motifs in Cid1 and Cid5 N-terminal tails may represent domains that are beneficial to one Cid function but harmful to other Cid functions. For example, the newly gained motif 8 in Cid1 might be detrimental to male meiosis or sperm development and the newly gained motifs 9 and 10 in Cid5 might be detrimental to female germline or somatic function.

The simplest way to test whether the newly acquired motifs have antagonistic functions is via CRISPR-Cas9 mediated allele swaps. To test whether motif 8 (new motif in Cid1) has antagonistic function we could replace endogenous *D. virilis* Cid5 with motif8+Cid5 (Figure 5-1C). If motif8 has antagonistic male germline effects, I predict that motif8+Cid5 will cause male fertility detects. We could also test whether motifs 9 and 10 have a negative impact on somatic and female-germline function by replacing endogenous Cid1 with motif9+motif10+Cid1, preserving all the original Cid1 motifs but appending new motifs 9 and 10 (Figure 5-1C). I predict that motif9+motif10+Cid1 will cause inviability or female fertility defects. These results could reveal potentially incompatible functional optima of essential genes involved in chromosome segregation.

Testing these hypotheses might be quite challenging in *D. virilis* flies because it requires somewhat complex genetic manipulations in a non-model organism. However, at least one of these costs might be more easily measured in *D. melanogaster*, which encodes a single Cid protein that is similar to Cid1 from *D. virilis*, except that it lacks motif8. By appending motif8 to *D. melanogaster* Cid and replacing the endogenous *Cid* gene, we can investigate whether motif8+Cid has a negative impact on *D. melanogaster* male fertility.

Although I have focused my thesis on the *D. virilis* Cid paralogs, functional investigation of the Cid paralogs in *D. auraria* and other members of the *montium* group may reveal a parallel trajectory of acquisition and loss of N-terminal motifs that is accompanied by their germline specialization. This would then represent a remarkable case of convergent evolution that would strengthen our model. Thus, the identification and characterization of Cid paralogs in my thesis have provided an unprecedented opportunity to functionally test the predictions of the intralocus conflict hypothesis, and how it shapes the fundamental process of chromosome segregation in unanticipated ways.

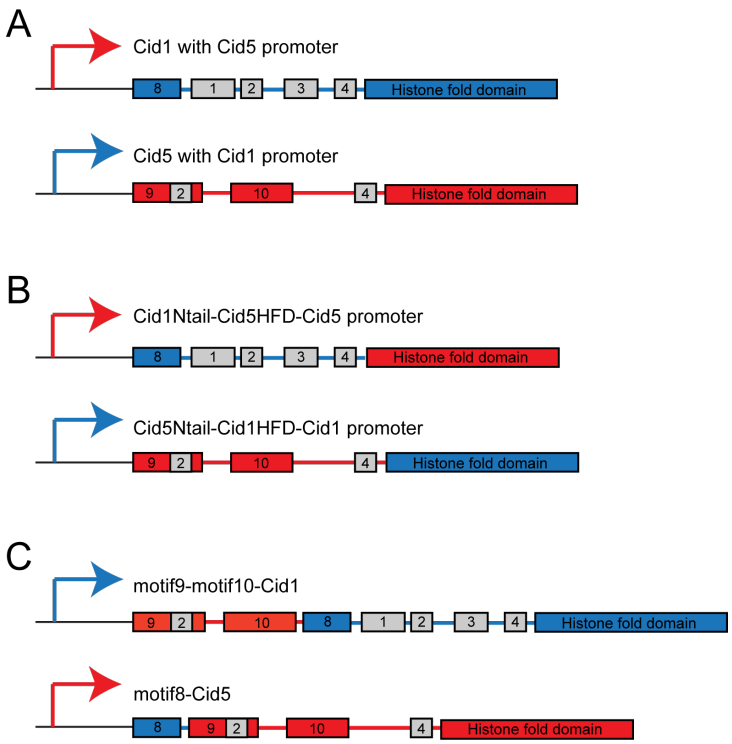


Figure 5-1. Cid1 and Cid5 chimeras could help elucidate means of specialization.

Schematic of Cid1, Cid5 or chimeric transgenes. Blue represents Cid1 promoter or coding sequence and red represents Cid5 promoter or coding sequence. Boxes with numbers represent N-terminal tail motifs (identified in Chapter 2). Grey shaded motifs are "canonical" motifs present in all single copy Cid genes. Red or blue shaded motifs are paralog specific. (A) If Cid1 and Cid5 have different knockdown phenotypes, attempting to rescue knockdown with opposite paralogs (for example, rescue Cid1 knockdown with Cid5 under the control of the Cid1 promoter) will reveal whether protein coding changes between Cid1 and Cid5 have resulted in specialization. (B) If Cid5 does not rescue Cid1 knockdown, we will attempt to rescue Cid1 knockdown with a Cid1-Cid5 chimeric protein. This will reveal which domains (N-terminal tail or histone fold domain) are important for paralog-specific function. (C) Adding Cid1-specific N-terminal tail motifs to Cid5 (and vice versa) may result in male or female infertility, respectively. This would suggest that motifs that are tolerated in one context are detrimental in another context, which would support our intralocus conflict hypothesis.

Appendix A. Elucidating the function of Cid1 and Cid5 in *Drosophila virilis*

My cytological studies of Cid1 and Cid5 in *D. virilis* revealed that Cid1, the paralog most closely related to *D. melanogaster* Cid, is the primary centromeric histone protein in somatic cells. However, the male and female germline contain both Cid1 and Cid5. Whereas both Cid1 and Cid5 co-localize at centromeres of mitotic germ cells in in ovaries and testes, they are alternately retained for meiosis and germ cell maturation. Cid1 is the primary centromeric histone in female meiosis and the mature oocyte and Cid5 is the primary centromeric histone in male meiosis and mature sperm. This suggests that *Cid* paralogs have specialized functions in *D. virilis.* Based on these findings, I would expect Cid1 and Cid5 to have different knockdown phenotypes. To test this hypothesis, I have put forth significant effort on two fronts: CRISPR knockout of *Cid5* and germline knockdown of *Cid1* and *Cid5* in *D. virilis.*

A.1 Attempting to knockout Cid5 with CRISPR/Cas9

First, I attempted to knockout *Cid5* using CRISPR/Cas9 technology by replacing the *Cid5* open reading frame with a dominant eye marker, 3XP3dsRED. I focused my efforts on *Cid5* because I predicted that *Cid1* knockout would be lethal given *Cid1's* centromeric localization pattern in somatic cells. I collaborated with Rob Harrell at the University of Maryland Insect Transformation Facility who performed the injections in *D. virilis* embryos. Unfortunately, we generated no transgenic animals after our first attempts. I think that the most likely reason that CRISPR failed is due to low numbers of G0 flies screened. If repeating attempts at CRISPR in *D. virilis,* I would ideally screen at least ~200 embryos per targeted gene replacement. It might also be worth using piggyBac transposase to generate a *D. virilis* line with germ-line expressed Cas9 since this has been shown to increase efficiency in *D. melanogaster*[181] and *Ae. aegypti*[182]. Another approach that might be worth trying would be the injection of Cas9 protein

into the hemolymph of adult females, with the Cas9 protein tagged such that it trafficks to and becomes concentrated in the ovaries[171].

A.2 Cid1 and Cid5 knockdown using paralog-specific miRNAs

In parallel, I developed reagents for miRNA-based knockdown of *Cid1* and *Cid5*. I designed miRNAs that that were specific to *Cid1* or *Cid5* and I tested their specificity and their ability to reduce gene expression in *D. melanogaster* tissue culture cells (Figure A-1, Figure A-2). I found that co-expression of *Cid1GFP* and a Cid1-targeting miRNA resulted in almost no detectable GFP fluorescence. However, if I co-expressed *Cid1GFP* and a non-targeting miRNA (luciferase_miRNA), GFP signal was easily detectable (Figure A-2A). Similarly, co-expression of *Cid5GFP* and a Cid5-targeting miRNA produced no GFP fluorescent cells but co-expression of Cid5GFP and a non-targeting miRNA allowed for expression of Cid5GFP (Figure A-2B). Furthermore, I confirmed that the miRNAs were specific to Cid1 or Cid5 by co-expressing Cid1GFP with the Cid5miRNAs and *vice versa.* Co-expression of the mis-matched miRNA did not reduce GFP fluorescent signal. These experiments, performed in a heterologous system (*D. melanogaster* tissue culture), suggested that the miRNAs I had designed were specific and potent at least in tissue culture.

>D. virilis Cid1 coding sequence

ATGCGTCCACGCACTGTAAAAAATTCAACTGAAAAAAAGAAGAAATCAGAATCGCATTTAGA
TAATGTTGACGATTCATATGAGAAAACAGCATTTCAAACACCGGATCGTGAAGACGAAACCG
ACTACGGCTTGGAGTTTACCACCAGCCGTTTGGCTGAATTGAACACATCTCCACGTCGGTGC
TCTACGCTACGCAAAAACAATCCAAAAGACCGCCGTCGTGATATAGAACCATCCGAAGACAA
CAGTGATTCAGAGAATCAGCCACTGGCAGTACGACAAACGCCCCGAAAAGTGCCGCTGCAAA
CACCCGCAGCGAGTATGAATAAGAAACATCAGGGGCCACTAACGTCAAGACCTGCGTCGAGA
CGCAAACAAAATAAACCGGAGCAACGTATAAAAAAATTGAACCGAGAAATTGAATGTTTACA
AAAGAATGCAGGCTTCATGATACCGCGTTTACCTTTCTCGCGTTTGGTGCGCGAAATTATGA
TGAAACATACTTTAACGCCCTTTATGATAACTATGAGCGCCCTGGAGGCTATACAGACCGCG
ACAGAAATGTACTTAACCCAGCGCTTCCAGGATGCCTATTTACTTACTCAGTATCGCAGCCG
TGTCACGCTAGAGGTGCGCGACATGGCGTTGGTGGCATATTTCTGCAAAACCTATGGTAATC
TTTGA

>D. virilis Cid5 coding sequence

ATGAGTCAAGCTAATGCACAGAGCTCCAATGGATCCCTGGATGAATCAGACTTAACGGCGGC
ATTTGATTTGAACGTTCTGGGTATGTTGGCCATTGAACAACGCTGCTCGACGACACGCAAGC
AGAAGCAACAATTGCAAGGCGAAGAGGAGACGGGTGTGGCAAATTTGGAGTCGCCAGTTGCA
GGCGAGGAACCAGCACCTGATACCGTCGCTGTCACGGAACCACCGCCACCGTCACCGTCATC
GCCACCGCCACCGCCACGGACACCGTCGCCGCCACAGTTACCGCCACCTACCCGAACAACAC
GCCGTAAACAGCCGTATCCTTTGCAGCGTGCCGCACTGTTCAGGCGCGAGGTGCGAACGCTG
CAGCGTTCACCGCATTTTATGATACCGCGTTTGTCCTTTGGGCGCGTGGTCCGTGAGATTAT
GCTGCAGCACACCGAATCGCCCTATCGGATCACCATTGGCGCTCTGGAGGCCCTACAGTCGG
CCACGGAGATGTTTCTAACGCAACGCTTTCAGGACTCCTACCTGATGACCCTGCATCGCAGT
CGGGTGACCCTAGAGGTGCGCGACATGGCCCTAATGGCATTCGTGTGCAAATTGCACGGACA
ACTCTGA

Figure A-1. Targeting locations of Cid1 and Cid5 miRNAs. *D. virilis Cid1* and *Cid5* coding sequence are shown with miRNA target sequences designate with bold lines. I tested two Cid1 specific miRNAs, Cid1_149 (red underline) and Cid1_213 (blue underline). I also tested two Cid5 specific miRNAs, Cid5_031 (black underline) and Cid5_034 (grey line).

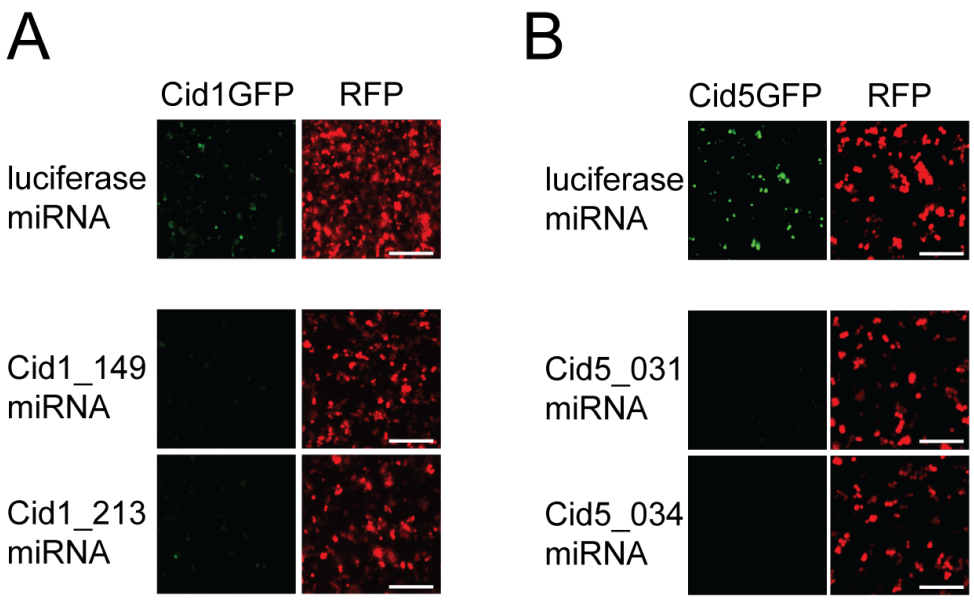


Figure A-2. Testing Cid1 and Cid5 miRNAs in tissue culture.(A) Live images of *D. melanogaster* KC cells transfected with Cid1GFP, RFP (to control for transfection efficiency) and miRNAs that target luciferase (top panel, serves as a control) or Cid1 (bottom two panels). GFP signal is reduced to when Cid1 miRNAs are co transfected with Cid1GFP indicating that the miRNAs decrease expression of Cid1GFP. (B) As in (A), except with Cid5GFP and Cid5-targeting miRNAs. GFP signal is reduced with co-transfection of Cid5GFP and Cid5 miRNAs indicating that the Cid5 miRNAs reduce expression of Cid5GFP. Scale bar = 50 μm.

Encouraged by the tissue culture results, I moved forward with generating transgenic *D. virilis* lines that contain *Cid1* or *Cid5* miRNAs under the control of an inducible promoter (UAS). I cloned the miRNAs into a vector that had piggyBac inverted repeats for random insertion into the *D. virilis* genome. Rainbow Transgenics injected these piggyBac miRNA constructs and screened for transformed flies. Ultimately, we acquired 5 lines for knockdown of *Cid1* targeting two different sequences in the *Cid1* open reading frame (miRNA_149#1-#3 and miRNA_213#1 and #2) and 3 lines for knockdown of *Cid5* targeting a single sequence in the *Cid5* open reading frame (miRNA_034#1-#3). I then determined the genomic location of the insertion for each of the miRNA lines. For two of the lines (Cid1miRNA_149#1 and Cid5miRNA_034#3), I have verified that the miRNA sequence is as expected in the inserted locus.

Recently, I have begun to assess whether these two verified miRNA lines are effective in reducing expression of *Cid1* or *Cid5 in vivo.* Conveniently, a *D. virilis* "driver" line that expresses Gal4 under the control of the germline specific promoter, Nanos, had previously been generated[183]. I crossed the Nanos driver line to flies that were heterozygous for Cid1_miRNA149#1 and measured Cid1 expression in ovaries and testes of these miRNA expressing flies or their non-miRNA expressing siblings (control flies). I found that *Cid1* expression is reduced to 58% of control levels with induction of Cid1miRNA_149#1 in testes (Figure A-3A, p-value = 0.0137), and reduced to 80% of the control in ovaries although the knockdown was not significant in ovaries (Figure A-3B, p-value = 0.1475). Fertility tests are underway to assess the impact of knockdown for these lines.

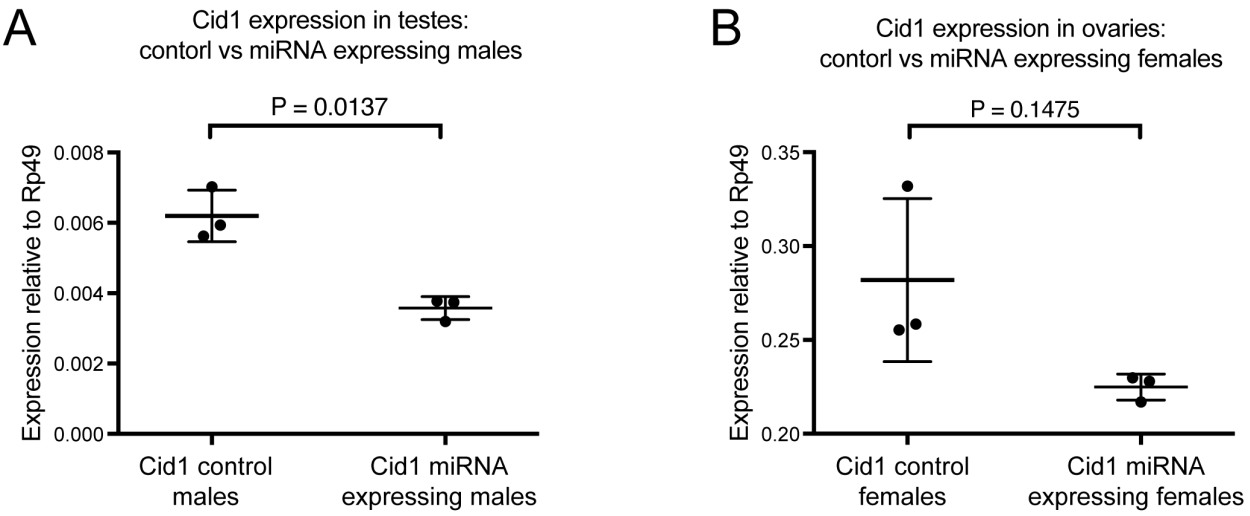


Figure A-3. Measuring *Cid1* knockdown in *D. virilis* testes and ovaries. (A) RT-qPCR showing Cid1 expression relative to Rp49 in testes from control males or testes from males expressing a Cid1miRNA (149#1) under the control of Nanos-GAL4. Cid1 expression was reduced to 58% in Cid1 knockdown testes (p=0.0137). (B) RT-qPCR showing Cid1 expression relative to Rp49 in ovaries from control females or ovaries from females expressing a Cid1 miRNA (149#1) under the control of Nanos-Gal4. Cid1 expression was reduced to 80% in Cid1 knockdown ovaries, although this difference was not significant. Error bars represent standard error from three biological replicates.

While I have made headway in determining the function of *Cid1* and *Cid5* by generating tools for genetic knockdown in *D. virilis,* a significant amount of work remains to be done to validate these tools and to generate additional reagents to serve as controls in knockdown experiments. First, the remaining miRNA lines need to be verified by sequencing and tested for their ability to reduce *Cid1* and *Cid5* expression in the germline. We will also test whether reduction in *Cid1* and *Cid5* mRNA results in a decrease in Cid1 and Cid5 protein by performing western blots from testes and ovaries or by performing quantitative microscopy using Cid1 and Cid5 antibodies. If we see significant knockdown and a phenotypic consequence of this knockdown, we will also generate *D. virilis* lines containing RNAi resistant versions of *Cid1* and *Cid5* so that we can attempt to rescue any phenotypic defects we observe with *Cid1* or *Cid5* knockdown. Finally, the luciferase miRNA used in tissue culture experiments would serve as a good non-targeting control and would allow us to conclude that any effect we see in knockdown experiments is specific to *Cid1* or *Cid5* knockdown and not simply due to the expression of a miRNA. Development of these additional tools will be critical for interpreting the results of knockdown experiments.

A.3 Predictions for Cid1 and Cid5 knockdown

Given what we already know about Cid1 and Cid5 localization in *D. virilis,* I can make some predictions about the possible outcomes of *Cid1* and *Cid5* knockdown. Broadly, I anticipate that germline knockdown of *Cid1* and *Cid5* will have divergent outcomes on male and female fertility. Given that Cid1 is the only detectable centromeric histone in the mature oocyte and that Cid5 is the only detectable centromeric histone in mature sperm, I predict that *Cid1* knockdown will cause reduced female fertility or complete sterility, whereas *Cid5* knockdown will cause reduced male fertility. More specifically, *Cid1* knockdown in the ovary may cause meiotic defects such as lagging chromosomes. Another possibility is that that *Cid1* ovary knockdown may cause maternal effect lethality if *Cid1* knockdown severely reduces the maternal loading of

Cid1 mRNA resulting in an embryonic Cid1 deficiency. In male gametogenesis, I anticipate that *Cid5* knockdown may cause meiotic defects and/or paternal effect lethality due to lack of Cid5 in mature sperm and loss of centromeric identity of the paternal pronucleus[23]. These results would indicate that Cid1 and Cid5 have acquired specialized roles in gametogenesis.

I also predict that germline knockdown of *Cid1* will not reduce male fertility and that germline knockdown of *Cid5* will not reduce female fertility. However, it is also possible that Cid1 is required for male fertility and that Cid5 is required for female fertility, even though neither protein persists for meiosis or mature germ cell development in these contexts. In this scenario, Cid1 and Cid5 might both be required early in male and female gametogenesis to establish some sort of favorable centromeric chromatin environment. Another possibility is that Cid1 and Cid5 are co-dependent for their germline localization. If this is the case, we expect to find impaired Cid5 recruitment upon Cid1 knockdown, or impaired Cid1 recruitment upon Cid5 knockdown, leading to male or female sterility.

A.4 Materials and methods

Designing miRNAs specific for Cid1 or Cid5

BLOCK-iT RNAi designer (Invitrogen) and RNAi Central (Hannon lab) were used to generate a list of candidate miRNA sequences for targeting Cid1 and Cid5. Each candidate miRNA was mapped onto a Cid1 and Cid5 nucleotide alignment to ensure no cross reactivity due to sequence similarity. Finally, each candidate miRNA was the query in a BLASTn search of the *D. virilis* genome in order to minimize the possibility of off target effects. Final miRNA sequences were chosen based on their low numbers of potential off target hits and the lack of sequence similarity to the opposite paralog.

Cloning miRNAs, dsRed, Cid1GFP and Cid5 GFP for experiments in D. melanogaster tissue culture cells

A cassette containing a U6 promoter and an EGFP miRNA was cloned into pUC19 at HindIII and NdeI sites to create pUCmiRNA_EGFP. All other miRNAs were ordered as "top" and "bottom" strand oligos with EcoRI and NheI overhangs for cloning (Table A-1). Top and bottom strands of each miRNA were annealed to each other. pUCmiRNA_EGFP was digested to removed the EGFP miRNA and each annealed miRNA oligo was cloned in to create pUCmiRNA_luciferase, pUCmiRNA_Cid1.149, pUCmiRNA_Cid1.213, pUCmiRNA_Cid5.031 and pUCmiRNA_Cid5.034. Cid1, Cid5 and dsRed open reading frames were cloned as described previously (Chapter 3, Chapter 4) into Gateway destination vector pAFW (dsRed) or pAGW (Cid1 and Cid5), generating an N-terminal 3XFLAG fusion to dsRed and N-terminal GFP fusions to Cid1 and Cid5, all under the control of the *D. melanogaster* actin promoter (pAFW_dsRed, pAGW_Cid1, pAGW_Cid5). The loop region of all miRNAs is from *D. melanogaster* Mir-1.

Table A-1: Oligo sequences for miRNA cloning. miRNA fragments used for cloning into pUCmiRNA were generated by annealing "Top" and ""Bottom" strand oligos, generating a double-stranded fragment with EcoRI and NheI overhangs.

miRNA or oligo name	miRNA or oligo sequence
Luciferase Bottom	AATTCGCGTTCGTCACATCTCATCTACCTATGCTTGAATATAACTAGATAGATGAGATATAACGAACACTG
Luciferase Top	CTAGCAGTGTTCGTTATATCTCATCTATCTAGTTATATTCAAGCATAGGTAGATGAGATGTGACGAACGCG
Cid1_149 Bottom	AATTCGCGCCGTTTGGCTGAATTGAACATATGCTTGAATATAACTATGTTCAATTCAGCCAAACGGCACTG
Cid1_149 Top	CTAGCAGTGCCGTTTGGCTGAATTGAACATAGTTATATTCAAGCATATGTTCAATTCAGCCAAACGGCGCG
Cid1_213 Bottom	AATTCGCGACCGCCGTCGTGATATAGAACTATGCTTGAATATAACTAGTTCTATATCACGACGGCGGTTACTG
Cid1_213 Top	CTAGCAGTAACCGCCGTCGTGATATAGAACTAGTTATATTCAAGCATAGTTCTATATCACGACGGCGGTCGCG
Cid5_031 Bottom	AATTCGCGGATCCCTGGATGAATCAGACTATGCTTGAATATAACTAGTCTGATTCATCCAGGGATCCACTG
Cid5_031 Top	CTAGCAGTGGATCCCTGGATGAATCAGACTAGTTATATTCAAGCATAGTCTGATTCATCCAGGGATCCGCG
Cid5_034 Bottom	AATTCGCCCCTGGATGAATCAGACTTAACTATGCTTGAATATAACTAGTTAAGTCTGATTCATCCAGGTACTG
Cid5_34 Top	CTAGCAGTACCTGGATGAATCAGACTTAACTAGTTATATTCAAGCATAGTTAAGTCTGATTCATCCAGGGGCG

Testing miRNA efficiency and specificity in tissue cultures cells

pAFW_dsRed (1ug), pAGW_Cid1 or pAGW_Cid5 (0.5ug) and a miRNA (0.5ug) were co transfected into *D. melanogaster* KC cells using the Fugene transfection reagent (Promega) according to the manufacture's instructions. The miRNA against luciferase served as a negative control and the miRNA against EGFP served as a positive control. Images of GFP and dsRed fluorescence were taken three days after transfection. Efficiency of each miRNA to reduce expression of pAGW_Cid1 or pAGW_Cid5 was measured by comparing GFP fluorescence in the luciferase_miRNA condition to GFP fluorescence when a Cid-specific miRNA was co-transfected.

Cloning Cid1 and Cid5 miRNAs into vectors for generating transgenic D. virilis

Annealed miRNA oligos (Table A-1) were each cloned into pWALIUM20 (Harvard Medical School Transgenic RNAi Project) creating pWALIUM20_Cid1.149, pWALIUM20_Cid1.213, pWALIUM20_Cid5.031 and pWALIUM20_Cid5.034. Each of these pWALIUM20 miRNA-containing plasmids was then digested with StuI and SapI and run on a gel. The lower band containing the miRNA expression cassette was gel extracted and the Sap1 overhang was blunted with T4 polymerase. The UASpBacNPF_3XP3EGFP vector (Drosophila Genomics Resource Center) was digested with BbvCI and SpeI to remove the 2XGAGA sites, 7X UAS and K10 UTR. This fragment was replaced with a short "replacement fragment" containing a StuI restriction enzyme site creating the pBac3XP3_intermediate plasmid. pBac3XP3_intermediate was digested with StuI to linearize the vector. Then, the linearized pBac3XP3_intermediate was ligated to each of the digested pWALIUM_hairpin plasmids to place the miRNA expression cassette into a piggyBac backbone. The resulting plasmid also contains 3XP3EGFP to serve as a transformation marker.

Measuring Cid1 knockdown in testes and ovaries

Female flies heterozygous for the UAS_Cid1_149#1 miRNA were crossed to male flies homozygous for the Nanos-Gal4 driver in three independent crosses. From each of these three crosses, miRNA-expressing or non-expressing males and females were collected as virgins and maintained separately in vials for 10-days to ensure that all flies were sexually mature. Then, RNA was extracted from dissected testes and ovaries from each condition. cDNA was synthesized as described previously (Chapter 2). RT-qPCR was performed according to the standard curve method using the Platinum SYBR Green reagent (Invitrogen) and primers designed to each *Cid* paralog and to *Rp49*. Reactions were run on an ABI QuantStudio 5 qPCR machine using the following conditions: 50°C for 2 min, 95°C for 2 min, 40 cycles of (95°C for 15s, 60°C for 30s). We ensured that all primer pairs had similar amplification efficiencies using a dilution series of genomic DNA. Three technical replicates were performed for each cDNA sample. Transcript levels of each gene were normalized to *Rp49.* P-values were calculated via a two-tailed T-test.

Bibliography

1 Flemming, W. *Zellsubstanz, kern und zelltheilung*. (Vogel, 1882).

2 Seeliger, O. Giebt es geschlechtlich erzeugte Organismen ohne mütterliche Eigenschaften? *Archiv für Entwicklungsmechanik der Organismen* 1, 203-223 (1894).

3 Clarke, L. & Carbon, J. Isolation of a yeast centromere and construction of functional small circular chromosomes. *Nature* 287, 504 (1980).

4 Fitzgerald-Hayes, M., Clarke, L. & Carbon, J. Nucleotide sequence comparisons and functional analysis of yeast centromere DNAs. *Cell* 29, 235-244 (1982).

5 Hieter, P. *et al.* Functional selection and analysis of yeast centromeric DNA. *Cell* 42, 913-921 (1985).

6 Clarke, L. & Carbon, J. The structure and function of yeast centromeres. *Annual review of genetics* 19, 29-55, doi:10.1146/annurev.ge.19.120185.000333 (1985).

7 Lechner, J. & Carbon, J. A 240 kd multisubunit protein complex, CBF3, is a major component of the budding yeast centromere. *Cell* 64, 717-725 (1991).

8 Doheny, K. F. *et al.* Identification of essential components of the S. cerevisiae kinetochore. *Cell* 73, 761-774 (1993).

9 Baum, M., Ngan, V. K. & Clarke, L. The centromeric K-type repeat and the central core are together sufficient to establish a functional Schizosaccharomyces pombe centromere. *Mol Biol Cell* 5, 747-761 (1994).

10 Murakami, S., Matsumoto, T., Niwa, O. & Yanagida, M. Structure of the fission yeast centromere cen3: direct analysis of the reiterated inverted region. *Chromosoma* 101, 214-221 (1991).

11 Bernard, P. *et al.* Requirement of heterochromatin for cohesion at centromeres. *Science* 294, 2539-2542, doi:10.1126/science.1064027 (2001).

12 Bernard, P. & Allshire, R. C. Centromeres become unstuck without heterochromatin. *Trends Cell Biol* 12, 419-424, doi:Pii S0962-8924(02)02344-9 Doi 10.1016/S0962-8924(02)02344-9 (2002).

13 Schueler, M. G., Higgins, A. W., Rudd, M. K., Gustashaw, K. & Willard, H. F. Genomic and genetic definition of a functional human centromere. *Science* 294, 109-115, doi:10.1126/science.1065042 (2001).

14 Sun, X., Wahlstrom, J. & Karpen, G. Molecular structure of a functional Drosophila centromere. *Cell* 91, 1007-1019 (1997).

15 Nagaki, K. *et al.* Sequencing of a rice centromere uncovers active genes. *Nat Genet* 36, 138-145, doi:10.1038/ng1289 (2004).

16 Marshall, O. J., Chueh, A. C., Wong, L. H. & Choo, K. H. Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution. *Am J Hum Genet* 82, 261-282, doi:10.1016/j.ajhg.2007.11.009 (2008).

17 Roach, K. C., Ross, B. D. & Malik, H. S. Rapid evolution of centromeres and centromeric/kinetochore proteins. *Rapidly Evolving Genes and Genetic Systems; Singh, RS, Xu, J., Kulathinal, RJ, Eds*, 83-93 (2012).

18 Bram, R. & Kornberg, R. Isolation of a Saccharomyces cerevisiae centromere DNA-binding protein, its human homolog, and its possible role as a transcription factor. *Molecular and cellular biology* 7, 403-409 (1987).

19 Cole, H. A., Howard, B. H. & Clark, D. J. The centromeric nucleosome of budding yeast is perfectly positioned and covers the entire centromere. *Proceedings of the National Academy of Sciences* 108, 12687-12692 (2011).

20 Karpen, G. H. & Allshire, R. C. The case for epigenetic effects on centromere identity and function. *Trends Genet* 13, 489-496 (1997).

21 Earnshaw, W. C. & Rothfield, N. Identification of a Family of Human Centromere Proteins Using Autoimmune Sera from Patients with Scleroderma. *Chromosoma* 91, 313-321, doi:Doi 10.1007/Bf00328227 (1985).

22 Palmer, D. K., O'Day, K., Trong, H. L., Charbonneau, H. & Margolis, R. L. Purification of the centromere-specific protein CENP-A and demonstration that it is a distinctive histone. *Proc Natl Acad Sci U S A* 88, 3734-3738 (1991).

23 Raychaudhuri, N. *et al.* Transgenerational propagation and quantitative maintenance of paternal centromeres depends on Cid/Cenp-A presence in Drosophila sperm. *PLoS biology* 10, e1001434, doi:10.1371/journal.pbio.1001434 (2012).

24 Palmer, D. K., O'Day, K. & Margolis, R. L. The centromere specific histone CENP-A is selectively retained in discrete foci in mammalian sperm nuclei. *Chromosoma* 100, 32-36 (1990).

25 Mendiburo, M. J., Padeken, J., Fulop, S., Schepers, A. & Heun, P. Drosophila CENH3 is sufficient for centromere formation. *Science* 334, 686-690, doi:10.1126/science.1206880 (2011).

26 Hewawasam, G. *et al.* Psh1 is an E3 ubiquitin ligase that targets the centromeric histone variant Cse4. *Molecular cell* 40, 444-454 (2010).

27 Moreno-Moreno, O., Medina-Giró, S., Torras-Llort, M. & Azorín, F. The F box protein partner of paired regulates stability of Drosophila centromeric histone H3, CenH3 CID. *Current Biology* 21, 1488-1493 (2011).

28 Akiyoshi, B. & Gull, K. Discovery of unconventional kinetochores in kinetoplastids. *Cell* 156, 1247-1258, doi:10.1016/j.cell.2014.01.049 (2014).

29 Drinnenberg, I. A., deYoung, D., Henikoff, S. & Malik, H. S. Recurrent loss of CenH3 is associated with independent transitions to holocentricity in insects. *Elife* 3, doi:10.7554/eLife.03676 (2014).

30 Kursel, L. E. & Malik, H. S. Centromeres. *Curr Biol* 26, R487-490, doi:10.1016/j.cub.2016.05.031 (2016).

31 Henikoff, S., Ahmad, K., Platero, J. S. & van Steensel, B. Heterochromatic deposition of centromeric histone H3-like proteins. *P Natl Acad Sci USA* 97, 716-721, doi:DOI 10.1073/pnas.97.2.716 (2000).

32 Blower, M. D. & Karpen, G. H. The role of Drosophila CID in kinetochore formation, cell-cycle progression and heterochromatin interactions. *Nature cell biology* 3, 730-739, doi:10.1038/35087045 (2001).

33 Stoler, S., Keith, K. C., Curnick, K. E. & Fitzgeraldhayes, M. A Mutation in Cse4, an Essential Gene Encoding a Novel Chromatin-Associated Protein in Yeast, Causes Chromosome Nondisjunction and Cell-Cycle Arrest at Mitosis. *Gene Dev* 9, 573-586, doi:DOI 10.1101/gad.9.5.573 (1995).

34 Howman, E. V. *et al.* Early disruption of centromeric chromatin organization in centromere protein A (Cenpa) null mice. *P Natl Acad Sci USA* 97, 1148-1153, doi:DOI 10.1073/pnas.97.3.1148 (2000).

35 Talbert, P. B., Masuelli, R., Tyagi, A. P., Comai, L. & Henikoff, S. Centromeric localization and adaptive evolution of an Arabidopsis histone H3 variant. *Plant Cell* 14, 1053-1066, doi:10.1105/tpc.010425 (2002).

36 Malik, H. S. & Henikoff, S. Adaptive evolution of Cid, a centromere-specific histone in Drosophila. *Genetics* 157, 1293-1298 (2001).

37 Henikoff, S., Ahmad, K. & Malik, H. S. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* 293, 1098-1102, doi:10.1126/science.1062939 (2001).

38 Wei, K. H. *et al.* A Pooled Sequencing Approach Identifies a Candidate Meiotic Driver in Drosophila. *Genetics* 206, 451-465, doi:10.1534/genetics.116.197335 (2017).

39 Chmatal, L. *et al.* Centromere strength provides the cell biological basis for meiotic drive and karyotype evolution in mice. *Curr Biol* 24, 2295-2300, doi:10.1016/j.cub.2014.08.017 (2014).

40 Akera, T. *et al.* Spindle asymmetry drives non-Mendelian chromosome segregation. *Science* 358, 668-672 (2017).

41 Iwata-Otsubo, A. *et al.* Expanded satellite repeats amplify a discrete CENP-A nucleosome assembly site on chromosomes that drive in female meiosis. *Current Biology* 27, 2365-2373. e2368 (2017).

42 Daniel, A. Distortion of female meiotic segregation and reduced male fertility in human Robertsonian translocations: consistent with the centromere model of co-evolving

centromere DNA/centromeric histone (CENP-A). *Am J Med Genet* 111, 450-452, doi:10.1002/ajmg.10618 (2002).

43 Wallace, B. M., Searle, J. B. & Everett, C. A. The effect of multiple simple Robertsonian heterozygosity on chromosome pairing and fertility of wild-stock house mice (Mus musculus domesticus). *Cytogenet Genome Res* 96, 276-286, doi:10.1159/000063054 (2002).

44 Elde, N. C., Roach, K. C., Yao, M. C. & Malik, H. S. Absence of positive selection on centromeric histones in Tetrahymena suggests unsuppressed centromere: drive in lineages lacking male meiosis. *J Mol Evol* 72, 510-520, doi:10.1007/s00239-011-9449-0 (2011).

45 Zedek, F. & Bures, P. CenH3 evolution reflects meiotic symmetry as predicted by the centromere drive model. *Sci Rep* 6, 33308, doi:10.1038/srep33308 (2016).

46 Huang, Y. C. *et al.* Evolution of long centromeres in fire ants. *BMC Evol Biol* 16, 189, doi:10.1186/s12862-016-0760-7 (2016).

47 Fishman, L. & Kelly, J. K. Centromere-associated meiotic drive and female fitness variation in Mimulus. *Evolution* 69, 1208-1218, doi:10.1111/evo.12661 (2015).

48 Maheshwari, S. *et al.* Naturally occurring differences in CENH3 affect chromosome segregation in zygotic mitosis of hybrids. *PLoS Genet* 11, e1004970, doi:10.1371/journal.pgen.1004970 (2015).

49 Maheshwari, S., Ishii, T., Brown, C. T., Houben, A. & Comai, L. Centromere location in Arabidopsis is unaltered by extreme divergence in CENH3 protein sequence. *Genome Res* 27, 471-478, doi:10.1101/gr.214619.116 (2017).

50 Comai, L., Maheshwari, S. & Marimuthu, M. P. A. Plant centromeres. *Curr Opin Plant Biol* 36, 158-167, doi:10.1016/j.pbi.2017.03.003 (2017).

51 Rosin, L. & Mellone, B. G. Co-evolving CENP-A and CAL1 Domains Mediate Centromeric CENP-A Deposition across Drosophila Species. *Dev Cell* 37, 136-147, doi:10.1016/j.devcel.2016.03.021 (2016).

52 Logsdon, G. A. *et al.* Both tails and the centromere targeting domain of CENP-A are required for centromere establishment. *J Cell Biol* 208, 521-531, doi:10.1083/jcb.201412011 (2015).

53 Schueler, M. G., Swanson, W., Thomas, P. J., Green, E. D. & Progra, N. C. S. Adaptive Evolution of Foundation Kinetochore Proteins in Primates. *Mol Biol Evol* 27, 1585-1597, doi:10.1093/molbev/msq043 (2010).

54 Talbert, P. B., Bryson, T. D. & Henikoff, S. Adaptive evolution of centromere proteins in plants and animals. *J Biol* 3, 18, doi:10.1186/jbiol11 (2004).

55 Gallach, M., Chandrasekaran, C. & Betran, E. Analyses of nuclearly encoded mitochondrial genes suggest gene duplication as a mechanism for resolving intralocus

sexually antagonistic conflict in Drosophila. *Genome Biol Evol* 2, 835-850, doi:10.1093/gbe/evq069 (2010).

56 Bonduriansky, R. & Chenoweth, S. F. Intralocus sexual conflict. *Trends in Ecology & Evolution* 24, 280-288, doi:https://doi.org/10.1016/j.tree.2008.12.005 (2009).

57 Gallach, M. & Betran, E. Intralocus sexual conflict resolved through gene duplication. *Trends Ecol Evol* 26, 222-228, doi:10.1016/j.tree.2011.02.004 (2011).

58 Mellone, B. G. *et al.* Assembly of Drosophila centromeric chromatin proteins during mitosis. *PLoS Genet* 7, e1002068, doi:10.1371/journal.pgen.1002068 (2011).

59 Dunleavy, E. M. *et al.* The cell cycle timing of centromeric chromatin assembly in Drosophila meiosis is distinct from mitosis yet requires CAL1 and CENP-C. *PLoS biology* 10, e1001460, doi:10.1371/journal.pbio.1001460 (2012).

60 Braun, R. E. Packaging paternal chromosomes with protamine. *Nat Genet* 28, 10-12, doi:10.1038/88194 (2001).

61 Renkawitz-Pohl, R., Hempel, L., Hollmann, M. & Schäfer, M. Spermatogenesis. (2005).

62 Oliva, R. & Dixon, G. H. Vertebrate protamine genes and the histone-to-protamine replacement reaction. *Prog Nucleic Acid Res Mol Biol* 40, 25-94 (1991).

63 Smoak, E. M., Stein, P., Schultz, R. M., Lampson, M. A. & Black, B. E. Long-term retention of CENP-A nucleosomes in mammalian oocytes underpins transgenerational inheritance of centromere identity. *Current Biology* 26, 1110-1116 (2016).

64 Von Stetina, J. R. & Orr-Weaver, T. L. Developmental control of oocyte maturation and egg activation in metazoan models. *Cold Spring Harbor perspectives in biology*, a005553 (2011).

65 Jin, W. *et al.* Maize centromeres: organization and functional adaptation in the genetic background of oat. *Plant Cell* 16, 571-581, doi:10.1105/tpc.018937 (2004).

66 Sanei, M., Pickering, R., Kumke, K., Nasuda, S. & Houben, A. Loss of centromeric histone H3 (CENH3) from centromeres precedes uniparental chromosome elimination in interspecific barley hybrids. *Proc Natl Acad Sci U S A* 108, E498-505, doi:10.1073/pnas.1103190108 (2011).

67 Yuan, J., Guo, X., Hu, J., Lv, Z. & Han, F. Characterization of two CENH 3 genes and their roles in wheat evolution. *New Phytol* 206, 839-851 (2015).

68 Ishii, T. *et al.* The differential loading of two barley CENH3 variants into distinct centromeric substructures is cell type- and development-specific. *Chromosome Res* 23, 277-284, doi:10.1007/s10577-015-9466-8 (2015).

69 Finseth, F. R., Dong, Y. Z., Saunders, A. & Fishman, L. Duplication and Adaptive Evolution of a Key Centromeric Protein in Mimulus, a Genus with Female Meiotic Drive. *Mol Biol Evol* 32, 2694-2706, doi:10.1093/molbev/msv145 (2015).

70 Neumann, P. *et al.* Stretching the rules: monocentric chromosomes with multiple centromere domains. *PLoS Genet* 8, e1002777, doi:10.1371/journal.pgen.1002777 (2012).

71 Neumann, P. *et al.* Centromeres Off the Hook: Massive Changes in Centromere Size and Structure Following Duplication of CenH3 Gene in Fabeae Species. *Mol Biol Evol* 32, 1862-1879, doi:10.1093/molbev/msv070 (2015).

72 Kawabe, A., Nasuda, S. & Charlesworth, D. Duplication of centromeric histone H3 (HTR12) gene in Arabidopsis halleri and A. lyrata, plant species with multiple centromeric satellite sequences. *Genetics* 174, 2021-2032, doi:10.1534/genetics.106.063628 (2006).

73 Moraes, I. C., Lermontova, I. & Schubert, I. Recognition of A. thaliana centromeres by heterologous CENH3 requires high similarity to the endogenous protein. *Plant Mol Biol* 75, 253-261, doi:10.1007/s11103-010-9723-3 (2011).

74 Zedek, F. & Bures, P. Absence of positive selection on CenH3 in Luzula suggests that holokinetic chromosomes may suppress centromere drive. *Ann Bot-London* 118, 1347-1352, doi:10.1093/aob/mcw186 (2016).

75 Finseth, F. R., Dong, Y., Saunders, A. & Fishman, L. Duplication and Adaptive Evolution of a Key Centromeric Protein in Mimulus, a Genus with Female Meiotic Drive. *Mol Biol Evol* 32, 2694-2706, doi:10.1093/molbev/msv145 (2015).

76 Li, Y. & Huang, J. F. Identification and molecular evolution of cow CENP-A gene family. *Mamm Genome* 19, 139-143, doi:10.1007/s00335-007-9083-8 (2008).

77 Monen, J., Maddox, P. S., Hyndman, F., Oegema, K. & Desai, A. Differential role of CENP-A in the segregation of holocentric C-elegans chromosomes during meiosis and mitosis. *Nature cell biology* 7, 1248-1255, doi:10.1038/ncb1331 (2005).

78 Monen, J. *et al.* Separase Cleaves the N-Tail of the CENP-A Related Protein CPAR-1 at the Meiosis I Metaphase-Anaphase Transition in C. elegans. *Plos One* 10, e0125382, doi:10.1371/journal.pone.0125382 (2015).

79 Hassold, T. & Hunt, P. To err (meiotically) is human: the genesis of human aneuploidy. *Nat Rev Genet* 2, 280-291, doi:10.1038/35066065 (2001).

80 McClintock, B. The Behavior in Successive Nuclear Divisions of a Chromosome Broken at Meiosis. *Proc Natl Acad Sci U S A* 25, 405-416 (1939).

81 Lohe, A. R. & Brutlag, D. L. Identical satellite DNA sequences in sibling species of Drosophila. *J Mol Biol* 194, 161-170 (1987).

82 Lee, H. R. *et al.* Chromatin immunoprecipitation cloning reveals rapid evolutionary patterns of centromeric DNA in Oryza species. *Proc Natl Acad Sci U S A* 102, 11793-11798, doi:10.1073/pnas.0503863102 (2005).

83 Schueler, M. G., Swanson, W., Thomas, P. J., Program, N. C. S. & Green, E. D. Adaptive evolution of foundation kinetochore proteins in primates. *Mol Biol Evol* 27, 1585-1597, doi:10.1093/molbev/msq043 (2010).

84 Fishman, L. & Saunders, A. Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers. *Science* 322, 1559-1562, doi:10.1126/science.1161406 (2008).

85 Malik, H. S. & Henikoff, S. Phylogenomics of the nucleosome. *Nat Struct Biol* 10, 882-891, doi:10.1038/nsb996 (2003).

86 Malik, H. S., Vermaak, D. & Henikoff, S. Recurrent evolution of DNA-binding motifs in the Drosophila centromeric histone. *Proc Natl Acad Sci U S A* 99, 1449-1454, doi:10.1073/pnas.032664299 (2002).

87 Russo, C. A. M., Mello, B., Frazao, A. & Voloch, C. M. Phylogenetic analysis and a time tree for a large drosophilid data set (Diptera: Drosophilidae). . *Zool J Linn Soc-Lond* 169, 765-775 (2013).

88 Schildkraut, E., Miller, C. A. & Nickoloff, J. A. Gene conversion and deletion frequencies during double-strand break repair in human cells are controlled by the distance between direct repeats. *Nucleic Acids Res* 33, 1574-1580, doi:10.1093/nar/gki295 (2005).

89 Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H. & Frost, S. D. Automated phylogenetic detection of recombination using a genetic algorithm. *Mol Biol Evol* 23, 1891-1901, doi:10.1093/molbev/msl051 (2006).

90 Vermaak, D., Hayden, H. S. & Henikoff, S. Centromere targeting element within the histone fold domain of Cid. *Mol Cell Biol* 22, 7553-7561 (2002).

91 Tachiwana, H. *et al.* Crystal structure of the human centromeric nucleosome containing CENP-A. *Nature* 476, 232-235, doi:10.1038/nature10258 (2011).

92 Black, B. E. *et al.* Centromere identity maintained by nucleosomes assembled with histone H3 containing the CENP-A targeting domain. *Mol Cell* 25, 309-322, doi:10.1016/j.molcel.2006.12.018 (2007).

93 Rosin, L. & Mellone, B. G. Centromeres drive a hard bargain. *Trends in Genetics* in press (2017).

94 Aul, R. B. & Oko, R. J. The major subacrosomal occupant of bull spermatozoa is a novel histone H2B variant associated with the forming acrosome during spermiogenesis. *Dev Biol* 239, 376-387, doi:10.1006/dbio.2001.0427 (2001).

95 van Steensel, B. & Henikoff, S. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat Biotechnol* 18, 424-428, doi:10.1038/74487 (2000).

96 Lynch, M. & Force, A. The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154, 459-473 (2000).

97 Force, A. *et al.* Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151, 1531-1545 (1999).

98 Dorus, S., Gilbert, S. L., Forster, M. L., Barndt, R. J. & Lahn, B. T. The CDY-related gene family: coordinated evolution in copy number, expression profile and protein sequence. *Hum Mol Genet* 12, 1643-1650 (2003).

99 Folco, H. D. *et al.* The CENP-A N-tail confers epigenetic stability to centromeres via the CENP-T branch of the CCAN in fission yeast. *Curr Biol* 25, 348-356, doi:10.1016/j.cub.2014.11.060 (2015).

100 Fachinetti, D. *et al.* A two-step mechanism for epigenetic specification of centromere identity and function. *Nature cell biology* 15, 1056-+, doi:10.1038/ncb2805 (2013).

101 Goutte-Gattat, D. *et al.* Phosphorylation of the CENP-A amino-terminus in mitotic centromeric chromatin is required for kinetochore function. *P Natl Acad Sci USA* 110, 8579-8584, doi:10.1073/pnas.1302955110 (2013).

102 Bailey, A. O. *et al.* Posttranslational modification of CENP-A influences the conformation of centromeric chromatin. *Proc Natl Acad Sci U S A* 110, 11827-11832, doi:10.1073/pnas.1300325110 (2013).

103 Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 2, 28-36 (1994).

104 Bailey, T. L. & Gribskov, M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* 14, 48-54, doi:DOI 10.1093/bioinformatics/14.1.48 (1998).

105 Torras-Llort, M., Medina-Giro, S., Moreno-Moreno, O. & Azorin, F. A conserved arginine-rich motif within the hypervariable N-domain of Drosophila centromeric histone H3 (CenH3) mediates BubR1 recruitment. *Plos One* 5, e13747, doi:10.1371/journal.pone.0013747 (2010).

106 Brake, L. & Baechli, G. *Drosophilidae (Diptera).* (2008).

107 Daugherty, M. D., Schaller, A. M., Geballe, A. P. & Malik, H. S. Evolution-guided functional analyses reveal diverse antiviral specificities encoded by IFIT1 genes in mammals. *Elife* 5, doi:10.7554/eLife.14228 (2016).

108 Petrov, D. A., Lozovskaya, E. R. & Hartl, D. L. High intrinsic rate of DNA loss in Drosophila. *Nature* 384, 346-349, doi:10.1038/384346a0 (1996).

109 Ancliff, M. & Park, J. M. Evolution dynamics of a model for gene duplication under adaptive conflict. *Physical review. E, Statistical, nonlinear, and soft matter physics* 89, 062702, doi:10.1103/PhysRevE.89.062702 (2014).

110 Heun, P. *et al.* Mislocalization of the Drosophila centromere-specific histone CID promotes formation of functional ectopic kinetochores. *Dev Cell* 10, 303-315, doi:10.1016/j.devcel.2006.01.014 (2006).

111 Schuh, M., Lehner, C. F. & Heidmann, S. Incorporation of Drosophila CID/CENP-A and CENP-C into centromeres during early embryonic anaphase. *Curr Biol* 17, 237-243, doi:10.1016/j.cub.2006.11.051 (2007).

112 Schittenhelm, R. B., Althoff, F., Heidmann, S. & Lehner, C. F. 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, doi:10.1242/jcs.067934 (2010).

113 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-3402 (1997).

114 Attrill, H. *et al.* FlyBase: establishing a Gene Group resource for Drosophila melanogaster. *Nucleic Acids Res* 44, D786-792, doi:10.1093/nar/gkv1046 (2016).

115 Larkin, M. A. *et al.* Clustal W and clustal X version 2.0. *Bioinformatics* 23, 2947-2948, doi:10.1093/bioinformatics/btm404 (2007).

116 Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647-1649, doi:10.1093/bioinformatics/bts199 (2012).

117 Huson, D. H. *et al.* Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* 8, 460, doi:10.1186/1471-2105-8-460 (2007).

118 Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13, 555-556 (1997).

119 McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351, 652-654, doi:10.1038/351652a0 (1991).

120 Parker, G. (Academic Press: New York, 1979).

121 Arnqvist, G. & Rowe, L. *Sexual conflict*. (Princeton University Press, 2013).

122 Brennan, P. L., Clark, C. J. & Prum, R. O. Explosive eversion and functional morphology of the duck penis supports sexual conflict in waterfowl genitalia. *Proceedings of the Royal Society of London B: Biological Sciences* 277, 1309-1314 (2010).

123 DeLuca, S. Z. & O'Farrell, P. H. Barriers to male transmission of mitochondrial DNA in sperm development. *Dev Cell* 22, 660-668, doi:10.1016/j.devcel.2011.12.021 (2012).

124 VanKuren, N. W. & Long, M. Gene duplicates resolving sexual conflict rapidly evolved essential gametogenesis functions. *Nat Ecol Evol* 2, 705-712, doi:10.1038/s41559-018-0471-0 (2018).

125 Kursel, L. E. & Malik, H. S. Recurrent Gene Duplication Leads to Diverse Repertoires of Centromeric Histones in Drosophila Species. *Mol Biol Evol* 34, 1445-1462, doi:10.1093/molbev/msx091 (2017).

126 Sullivan, K. F., Hechenberger, M. & Masri, K. Human CENP-A contains a histone H3 related histone fold domain that is required for targeting to the centromere. *J Cell Biol* 127, 581-592 (1994).

127 Yoda, K. *et al.* Human centromere protein A (CENP-A) can replace histone H3 in nucleosome reconstitution in vitro. *Proc Natl Acad Sci U S A* 97, 7266-7271, doi:10.1073/pnas.130189697 (2000).

128 Buchwitz, B. J., Ahmad, K., Moore, L. L., Roth, M. B. & Henikoff, S. A histone-H3-like protein in C. elegans. *Nature* 401, 547-548 (1999).

129 Henikoff, S. & Malik, H. S. Centromeres: selfish drivers. *Nature* 417, 227, doi:10.1038/417227a (2002).

130 Kursel, L. E. & Malik, H. S. The cellular mechanisms and consequences of centromere drive. *Curr Opin Cell Biol* 52, 58-65, doi:10.1016/j.ceb.2018.01.011 (2018).

131 Malik, H. S. The centromere-drive hypothesis: a simple basis for centromere complexity. *Prog Mol Subcell Biol* 48, 33-52, doi:10.1007/978-3-642-00182-6_2 (2009).

132 Kuppu, S. *et al.* Point Mutations in Centromeric Histone Induce Post-zygotic Incompatibility and Uniparental Inheritance. *PLoS Genet* 11, e1005494, doi:10.1371/journal.pgen.1005494 (2015).

133 Gibeaux, R. *et al.* Paternal chromosome loss and metabolic crisis contribute to hybrid inviability in Xenopus. *Nature* 553, 337-341, doi:10.1038/nature25188 (2018).

134 Gaucher, J. *et al.* From meiosis to postmeiotic events: the secrets of histone disappearance. *FEBS J* 277, 599-604, doi:10.1111/j.1742-4658.2009.07504.x (2010).

135 Bonnefoy, E., Orsi, G. A., Couble, P. & Loppin, B. The essential role of Drosophila HIRA for de novo assembly of paternal chromatin at fertilization. *PLoS genetics* 3, e182 (2007).

136 Schuh, M., Lehner, C. F. & Heidmann, S. Incorporation of Drosophila CID/CENP-A and CENP-C into centromeres during early embryonic anaphase. *Current Biology* 17, 237-243 (2007).

137 Spradling, A. Developmental genetics of oogenesis. *The development of Drosophila melanogaster* (1993).

138 King, R., Rubinson, A. C. & Smith, R. Oogenesis in adult Drosophila melanogaster. *Growth* 20, 121 (1956).

139 Teixeira, J. R., Dias, G. B., Svartman, M., Ruiz, A. & Kuhn, G. C. S. Concurrent Duplication of Drosophila Cid and Cenp-C Genes Resulted in Accelerated Evolution and Male Germline-Biased Expression of the New Copies. *J Mol Evol*, doi:10.1007/s00239-018-9851-y (2018).

140 Bodor, D. L. *et al.* The quantitative architecture of centromeric chromatin. *Elife* 3, e02137, doi:10.7554/eLife.02137 (2014).

141 Dunleavy, E. M., Almouzni, G. & Karpen, G. H. H3.3 is deposited at centromeres in S phase as a placeholder for newly assembled CENP-A in G(1) phase. *Nucleus-Austin* 2, 146-157, doi:10.4161/nucl.2.2.15211 (2011).

142 Jansen, L. E. T., Black, B. E., Foltz, D. R. & Cleveland, D. W. Propagation of centromeric chromatin requires exit from mitosis. *J Cell Biol* 176, 795-805, doi:10.1083/jcb.20070166 (2007).

143 Fabian, L. & Brill, J. A. Drosophila spermiogenesis: Big things come from little packages. *Spermatogenesis* 2, 197-212, doi:10.4161/spmg.21798 (2012).

144 Fuller, M. Spermatogenesis. *Development of Drosophila*, 71-147 (1993).

145 Hendzel, M. J. *et al.* Mitosis-specific phosphorylation of histone H3 initiates primarily within pericentromeric heterochromatin during G2 and spreads in an ordered fashion coincident with mitotic chromosome condensation. *Chromosoma* 106, 348-360 (1997).

146 Ivanovska, I. & Orr-Weaver, T. L. Histone modifications and the chromatin scaffold for meiotic chromosome architecture. *Cell Cycle* 5, 2064-2071 (2006).

147 Tang, T. T.-L., Bickel, S. E., Young, L. M. & Orr-Weaver, T. L. Maintenance of sister-chromatid cohesion at the centromere by the Drosophila MEI-S332 protein. *Gene Dev* 12, 3843-3856 (1998).

148 Gao, G., Cheng, Y., Wesolowska, N. & Rong, Y. S. Paternal imprint essential for the inheritance of telomere identity in Drosophila. *Proceedings of the National Academy of Sciences* 108, 4932-4937 (2011).

149 Loppin, B., Berger, F. & Couble, P. The Drosophila maternal gene sesame is required for sperm chromatin remodeling at fertilization. *Chromosoma* 110, 430-440, doi:10.1007/s004120100161 (2001).

150 Loppin, B. *et al.* The histone H3.3 chaperone HIRA is essential for chromatin assembly in the male pronucleus. *Nature* 437, 1386-1390, doi:10.1038/nature04059 (2005).

151 Loppin, B., Docquier, M., Bonneton, F. & Couble, P. The maternal effect mutation sesame affects the formation of the male pronucleus in Drosophila melanogaster. *Dev Biol* 222, 392-404, doi:10.1006/dbio.2000.9718 (2000).

152 Landmann, F., Orsi, G. A., Loppin, B. & Sullivan, W. Wolbachia-mediated cytoplasmic incompatibility is associated with impaired histone deposition in the male pronucleus. *PLoS Pathog* 5, e1000343, doi:10.1371/journal.ppat.1000343 (2009).

153 Levine, M. T., Vander Wende, H. M. & Malik, H. S. Mitotic fidelity requires transgenerational action of a testis-restricted HP1. *Elife* 4, e07378, doi:10.7554/eLife.07378 (2015).

154 Vermaak, D., Hayden, H. S. & Henikoff, S. Centromere targeting element within the histone fold domain of Cid. *Mol Cell Biol* 22, 7553-7561, doi:10.1128/Mcb.22.21.7553-7561.2002 (2002).

155 Collins, C. M., Malacrida, B., Burke, C., Kiely, P. A. & Dunleavy, E. M. ATP synthase F1 subunits recruited to centromeres by CENP-A are required for male meiosis. *Nat Commun* 9, 2702, doi:10.1038/s41467-018-05093-9 (2018).

156 Handler, A. M. & Harrell, R. A., 2nd. Germline transformation of Drosophila melanogaster with the piggyBac transposon vector. *Insect Mol Biol* 8, 449-457 (1999).

157 Fanti, L. & Pimpinelli, S. Immunostaining of squash preparations of chromosomes of larval brains. *Methods Mol Biol* 247, 353-361 (2004).

158 Zedek, F. & Bures, P. CenH3 evolution reflects meiotic symmetry as predicted by the centromere drive model. *Scientific reports* 6, doi:ARTN 3330810.1038/srep33308 (2016).

159 Li, Y. & Huang, J. F. Identification and molecular evolution of cow CENP-A gene family. *Mammalian genome : official journal of the International Mammalian Genome Society* 19, 139-143, doi:10.1007/s00335-007-9083-8 (2008).

160 Neafsey, D. E. *et al.* Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. *Science* 347, 1258522, doi:10.1126/science.1258522 (2015).

161 Giraldo-Calderon, G. I. *et al.* VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res* 43, D707-713, doi:10.1093/nar/gku1117 (2015).

162 Hori, T. *et al.* CCAN makes multiple contacts with centromeric DNA to provide distinct pathways to the outer kinetochore. *Cell* 135, 1039-1052, doi:10.1016/j.cell.2008.10.019 (2008).

163 Chen, C. C. *et al.* CAL1 is the Drosophila CENP-A assembly factor. *J Cell Biol* 204, 313-329, doi:10.1083/jcb.201305036 (2014).

164 Phansalkar, R., Lapierre, P. & Mellone, B. G. Evolutionary insights into the role of the essential centromere protein CAL1 in Drosophila. *Chromosome Res* 20, 493-504, doi:10.1007/s10577-012-9299-7 (2012).

165 Orr, B. & Sunkel, C. E. Drosophila CENP-C is essential for centromere identity. *Chromosoma* 120, 83-96, doi:10.1007/s00412-010-0293-6 (2011).

166 Cohen, R. L. *et al.* Structural and functional dissection of Mif2p, a conserved DNA-binding kinetochore protein. *Mol Biol Cell* 19, 4480-4491, doi:10.1091/mbc.E08-03-0297 (2008).

167 Kral, L. Possible identification of CENP-C in fish and the presence of the CENP-C motif in M18BP1 of vertebrates. *F1000Res* 4, 474, doi:10.12688/f1000research.6823.2 (2015).

168 Biedler, J. K. *et al.* Maternal germline-specific genes in the Asian malaria mosquito Anopheles stephensi: characterization and application for disease control. *G3 (Bethesda)* 5, 157-166, doi:10.1534/g3.114.015578 (2014).

169 Akbari, O. S. *et al.* The developmental transcriptome of the mosquito Aedes aegypti, an invasive species and major arbovirus vector. *G3 (Bethesda)* 3, 1493-1509, doi:10.1534/g3.113.006742 (2013).

170 Kistler, K. E., Vosshall, L. B. & Matthews, B. J. Genome engineering with CRISPR-Cas9 in the mosquito Aedes aegypti. *Cell Rep* 11, 51-60, doi:10.1016/j.celrep.2015.03.009 (2015).

171 Chaverra-Rodriguez, D. *et al.* Targeted delivery of CRISPR-Cas9 ribonucleoprotein into arthropod ovaries for heritable germline gene editing. *Nat Commun* 9, 3008, doi:10.1038/s41467-018-05425-9 (2018).

172 Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52, 696-704 (2003).

173 Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4, 406-425, doi:10.1093/oxfordjournals.molbev.a040454 (1987).

174 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24, 1586-1591, doi:10.1093/molbev/msm088 (2007).

175 Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7, 562-578, doi:10.1038/nprot.2012.016 (2012).

176 Akiyoshi, B. & Gull, K. Evolutionary cell biology of chromosome segregation: insights from trypanosomes. *Open Biol* 3, 130023, doi:10.1098/rsob.130023 (2013).

177 Drinnenberg, I. A., Henikoff, S. & Malik, H. S. Evolutionary Turnover of Kinetochore Proteins: A Ship of Theseus? *Trends Cell Biol* 26, 498-510, doi:10.1016/j.tcb.2016.01.005 (2016).

178 van Hooff, J. J., Tromer, E., van Wijk, L. M., Snel, B. & Kops, G. J. Evolutionary dynamics of the kinetochore network in eukaryotes as revealed by comparative genomics. *EMBO Rep* 18, 1559-1571, doi:10.15252/embr.201744102 (2017).

179 Rathke, C., Baarends, W. M., Awe, S. & Renkawitz-Pohl, R. Chromatin dynamics during spermiogenesis. *Biochim Biophys Acta* 1839, 155-168, doi:10.1016/j.bbagrm.2013.08.004 (2014).

180 Chen, Y. H. *et al.* The N terminus of the centromere H3-like protein Cse4p performs an essential function distinct from that of the histone fold domain. *Molecular and Cellular Biology* 20, 7037-7048, doi:Doi 10.1128/Mcb.20.18.7037-7048.2000 (2000).

181 Kondo, S. & Ueda, R. Highly improved gene targeting by germline-specific Cas9 expression in Drosophila. *Genetics* 195, 715-721, doi:10.1534/genetics.113.156737 (2013).

182 Li, M. *et al.* Germline Cas9 expression yields highly efficient genome engineering in a major worldwide disease vector, Aedes aegypti. *Proc Natl Acad Sci U S A* 114, E10540-E10549, doi:10.1073/pnas.1711538114 (2017).

183 Holtzman, S. *et al.* Transgenic tools for members of the genus Drosophila with sequenced genomes. *Fly* 4, 349-362 (2010).

Vita

Lisa grew up in Oregon, Wisconsin. She obtained a Bachelor of Science degree in biology and music performance (cello) at the University of Wisconsin – Madison. After completing her bachelors' degree, Lisa worked as a research specialist for two years in the laboratory of John Doebley. Lisa moved to Seattle in 2012 to begin her PhD at the University of Washington – Seattle.