Duplication and Adaptive Evolution of a Key Centromeric Protein in *Mimulus*, a Genus with Female Meiotic Drive

Findley R. Finseth,*[,1] Yuzhu Dong,[1,2] Arpiar Saunders,[1,3] and Lila Fishman[1]

[1]Division of Biological Sciences, University of Montana, Missoula

[2]Key Laboratory of Molecular Epigenetics of the Ministry of Education, Northeast Normal University, Changchun, China

[3]Department of Genetics, Harvard Medical School, Boston, MA

***Corresponding author:** E-mail: findley.finseth@mso.umt.edu.

Associate editor: Michael Nachman

Abstract

The fundamental asymmetry of female meiosis creates an arena for genetic elements to compete for inclusion in the egg, promoting the selfish evolution of centromere variants that maximize their transmission to the future egg. Such "female meiotic drive" has been hypothesized to explain the paradoxically complex and rapidly evolving nature of centromeric DNA and proteins. Although theoretically widespread, few cases of active drive have been observed, thereby limiting the opportunities to directly assess the impact of centromeric drive on molecular variation at centromeres and binding proteins. Here, we characterize the molecular evolutionary patterns of *CENH3*, the centromere-defining histone variant, in *Mimulus* monkeyflowers, a genus with one of the few known cases of active centromere-associated female meiotic drive. First, we identify a novel duplication of *CENH3* in diploid *Mimulus*, including in lineages with actively driving centromeres. Second, we demonstrate long-term adaptive evolution at several sites in the N-terminus of *CENH3*, a region with some meiosis-specific functions that putatively interacts with centromeric DNA. Finally, we infer that the paralogs evolve under different selective regimes; some sites in the N-terminus evolve under positive selection in the pro-orthologs or only one paralog (*CENH3_B*) and the paralogs exhibit significantly different patterns of polymorphism within populations. Our finding of long-term, adaptive evolution at *CENH3* in the context of centromere-associated meiotic drive supports an antagonistic, coevolutionary battle for evolutionary dominance between centromeric DNA and binding proteins.

***Key words:* *CENH3*, female meiotic drive, centromere evolution, gene duplication, genetic conflict, selfish evolution.**

Introduction

The evolution of a gene is generally tied to its effect on the organism, as genes that improve fitness tend to spread within a population. However, genes or chromosomes that distort their own transmission to the next generation can decouple their evolutionary fate from their effects on host fitness. Because these "selfish" genetic elements often increase in frequency despite causing harm to the individual organism, a situation arises where genes within a single organism have opposing interests (Burt and Trivers 2008; Rice 2013). Such genetic conflict can trigger a coevolutionary arms race between selfishly evolving genetic elements and suppressors, the byproduct of which is rapid genetic turnover and a molecular signature of positive selection. One important but poorly understood form of genetic conflict occurs between centromeres, which bias their transmission through meiotic drive, and kinetochore proteins, which evolve to counter driving centromeres (Henikoff et al. 2001). Theoretically, this conflict produces a pattern of rapid evolution at centromeres and associated proteins and may be responsible for the puzzling complexity of some of the core molecular machinery of meiosis.

At every cell division, chromosomes must be accurately redistributed to daughter cells to prevent aneuploidy or cell death. Centromeres mediate chromosome segregation during meiosis and mitosis and thus play an integral role in this process. Despite their vital and conserved role, centromeres display considerable structural and molecular variability, ranging in size from point centromeres of yeast (~125 bp) to greater than 100 kb arrays of tandemly repeated satellite sequence in plants and animals (Malik and Henikoff 2002, 2009; Melters et al. 2013). The primary sequences of centromeric DNA are also diverse and vary across closely related species (Lee et al. 2005), chromosomes of the same species (Kawabe and Nasuda 2004), haplotypes of homologous chromosomes (Wang et al. 2014), and even within a single functional centromere (Neumann et al. 2012). In addition, truncated chromosomes that lack centromeric repeats can form functional centromeres at novel chromosomal sites (Nasuda et al. 2005). The striking lack of homology among centromeric DNA sequence and sequence-independent establishment of centromeres argues that epigenetic, rather than genetic, processes maintain centromere identity. Indeed, nearly all eukaryotic centromeres are now defined by the presence of a universal epigenetic marker, the centromere-specific histone variant, *CENH3* (*CENP-A* in humans; Allshire and Karpen 2008, but see Drinnenberg et al. 2014).

At centromeres, *CENH3* replaces a portion of the canonical histone *H3* and forms the foundation of the kinetochore, the protein complex that links chromatin and spindle

Article

microtubules to coordinate chromosomal movement (Howman et al. 2000; Oegema et al. 2001). In most eukaryotes, *CENH3* is necessary and sufficient for centromere establishment; it is required for kinetochore formation (Howman et al. 2000; Moore and Roth 2001; Oegema et al. 2001), and misincorporation of *CENH3* in noncentromeric regions forms ectopic kinetochores (Heun et al. 2006). Unlike canonical histones, which are typically encoded by many repeated genes and extraordinarily conserved due to strong purifying selection (Rooney et al. 2002), *CENH3* is generally single copy and exhibits extreme sequence divergence. Specifically, an extended N-terminal tail and loop 1 of the histone fold domain (HFD) putatively interact with centromeric DNA (Malik et al. 2002; Vermaak et al. 2002) and show signatures of positive selection in plants (Talbert et al. 2002; Cooper and Henikoff 2004; Hirsch et al. 2009) and animals (Malik and Henikoff 2001; Malik et al. 2002; Schueler et al. 2010; Zedek and Bureš 2012). In contrast, the HFD (outside of loop 1) is generally conserved and under strong purifying selection.

Because the kinetochore is a highly conserved cellular machine with an indispensable and ubiquitous function, the observation that both centromeric DNA and *CENH3* evolve rapidly presents a paradox (Henikoff et al. 2001). Why has *CENH3* not evolved to an optimal and conserved state? In the last decade, an appealing model of "centromeric drive" has emerged and attempts to explain the rapid diversification of both centromeres and kinetochore proteins (Henikoff et al. 2001; Henikoff and Malik 2002; Malik and Henikoff 2002, 2009; Malik and Bayes 2006). Under this model, the fundamental asymmetry of female meiosis favors the evolution of centromere variants that maximize the probability of inclusion in the egg. Costs associated with this "female meiotic drive" of centromeres favor the evolution of suppressors to restore equal segregation, likely at centromere binding proteins such as *CENH3*. Repeated bouts of drive and suppression then produce rapid evolution of centromeric DNA and suppressor proteins within and between species, which could ultimately lead to fitness variation and reproductive incompatibilities (Henikoff and Malik 2002). As theory and circumstantial evidence suggest that female meiotic drive is widespread in nature (Pardo-Manuel de Villena and Sapienza 2001a, 2001b; Birchler et al. 2003; Malik and Bayes 2006), selfish centromeric drive is often invoked as a likely force shaping the evolution of kineotochore proteins generally and *CENH3* in particular (Henikoff et al. 2001; Malik and Henikoff 2001, 2002; Cooper and Henikoff 2004; Malik and Bayes 2006; Hirsch et al. 2009; Talbert et al. 2009; Schueler et al. 2010; Zedek and Bureš 2012). Yet, save a few striking exceptions (Fishman and Saunders 2008; Chmátal et al. 2014), there is scant empirical evidence of centromeric drive, and much remains unknown about the nature and consequences of antagonistic coevolution between the DNA and protein components of the centromere.

In this study, we explore the evolution of *CENH3* in *Mimulus* (monkeyflowers), a genus that includes a well-documented case of centromere-associated female meiotic drive in the common yellow monkeyflower, *Mimulus guttatus*. In heterospecific crosses with *Mimulus nasutus*, a *M. guttatus* allele on LG11 ("D") exhibits near-perfect (98% vs. expected 50%) transmission via female meiosis—a level of distortion only possible by centromeric drive at Meiosis I (Fishman and Willis 2005). Moreover, *D* is genetically linked and physically adjacent to unusually large arrays of the putative centromeric DNA repeat, *Cent728*, providing further evidence that the driving allele is likely the centromere (Fishman and Saunders 2008). *D* is also polymorphic within *M. guttatus* (even within a single focal population), drives weakly in conspecific crosses, and contributes to standing genetic variation for male and female fitness within *M. guttatus* (Fishman and Saunders 2008; Fishman and Kelly 2015). Here, we investigate whether *CENH3* has experienced long-term, recurrent positive selection across *Mimulus*, as expected by the centromeric drive model. We sequenced *CENH3* from 11 species across *Mimulus* and, surprisingly, discovered a duplication of *CENH3* that appears to coincide phylogenetically with a period of genome-wide chromosomal fission resulting in a near doubling of chromosome number (though more sampling is needed to confirm the coincidence of fission and *CENH3* duplication; Fishman et al. 2014). We then used codon substitution models to test whether *CENH3* shows signatures of long-term positive selection, as well as tested for rate heterogeneity among paralogs and the role of selection following *CENH3* duplication. To further link selection acting on *CENH3* to female meiotic drive, we used publicly available resequence data to characterize variation at *CENH3* duplicates within a *M. guttatus* population polymorphic for the driving D allele. Because we find evidence of long-term positive selection acting on *CENH3* in a system with centromere-associated drive, as well as differences among *CENH3* paralogs in their recent selective history, our data support the idea that centromeric DNA and *CENH3* coevolve antagonistically.

Results

We obtained *CENH3* sequences from 12 *Mimulus* samples representing 11 species (*M. aurantiacus*, *M. bolanderi*, *M. cardinalis*, *M. dentilobus*, *M. guttatus* Iron Mountain, OR population (IM), *M. guttatus* Florence Dunes, OR population (DUN), *M. jungermannoides*, *M. lewisii*, *M. nasutus*, *M. parishii*, *M. primuloides*, *M. tilingii*; fig. 1). *CENH3* sequences were collected using four different methods—polymerase chain reaction (PCR) amplification and Sanger sequencing of cDNA, PCR amplification and Sanger sequencing of genomic DNA, mapping of Illumina sequences from closely related species to the *M. guttatus* version 2.0 reference genome (http://phytozome.jgi.doe.gov/, last accessed July 1, 2015), or BLAST-based searches of de novo assembled genomes (supplementary table S1, Supplementary Material online). For 10 of the 12 samples, *CENH3* sequences were obtained using both PCR-based and whole-genome approaches, allowing us to independently verify sequences and validate our methodology (supplementary table S1, Supplementary Material online). The overall topology of the species tree made from alignments of third codon position sites largely reconstructed previously described species relationships (fig. 1*a*; Beardsley et al. 2003, 2004).

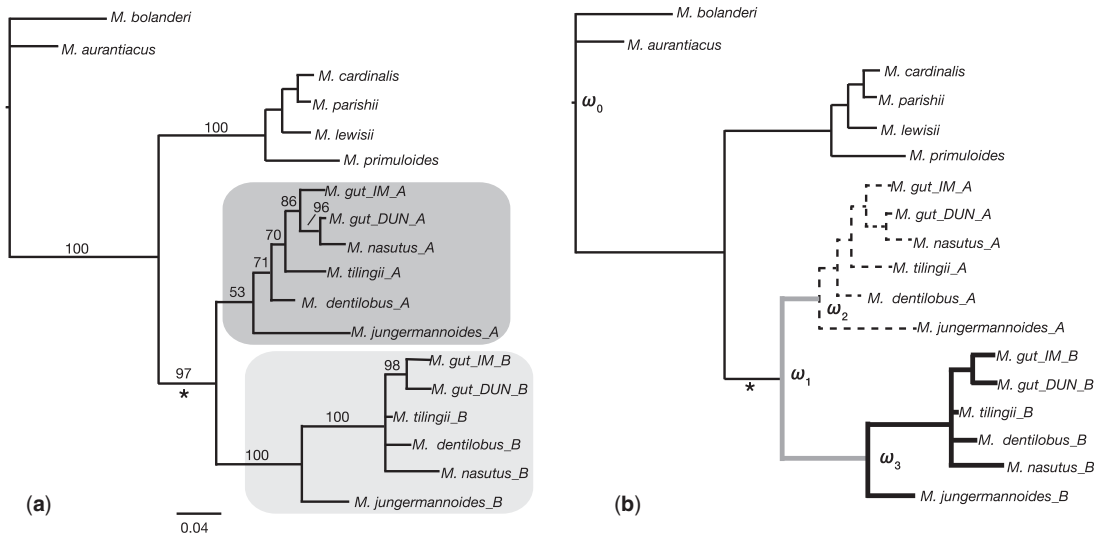

Fig. 1. (*a*) Gene tree based on third codon positions of *CENH3* sequence alignments sampled across *Mimulus*. Bayesian inference was performed with 150,000 MCMC generations under a GTR + Γ nucleotide substitution model. Bayesian posterior probabilities (i.e., clade credibility values) are shown. *CENH3_A* and *CENH3_B* clades are highlighted in dark and light gray boxes, respectively. Centromere-associated meiotic drive has been described for the IM population of *M. guttatus* (*M. gut_IM*). Scale bar represents nucleotides substitutions per site. (*b*) Codon substitution models were used to test various hypotheses of adaptive evolution occurring along certain lineages or clades in a maximum likelihood framework. For branch and branch-sites models, the gene tree was partitioned into pro-orthologs (ω_0; solid), the branches immediately following duplication (ω_1; gray), and paralogs A (ω_2; dashed) and B (ω_3; bold). Asterisk indicates the putative *CENH3* duplication event, as well a genome wide fission event that resulted in a near doubling of chromosome number (Fishman et al. 2014).

Duplication of CENH3

The *Mimulus guttatus* reference genome (IM62 line, from the IM population known to be polymorphic for centromere-associated drive) encodes two unlinked paralogs of *CENH3* (*CENH3_A* on LG14 and *CENH3_B* on LG2). We confirmed this duplication in the *M. guttatus* DUN line, in closely related *M. nasutus* and *M. tilingii*, and in two species from outside section Similous (*M. jungermannoides*, *M. dentilobus*; fig. 1*a*). *CENH3_A* and *CENH3_B* transcripts are present in cDNA libraries prepared from floral buds and RNA-Seq data (http://phytozome.jgi.doe.gov/, last accessed July 1, 2015; Colicchio et al. 2015), suggesting both copies are expressed. We confirmed expression of both paralogs in three different samples, as we were able to amplify two distinct *CENH3* transcripts from cDNA (*M. guttatus* (IM), *M. guttatus* (DUN), *M. nasutus*; supplementary table S1, Supplementary Material online). For all 12 samples, genome scans confirmed *CENH3* copy number (one or two).

CENH3_A and *CENH3_B* are highly divergent at both the amino acid (mean pairwise distance between paralogs = 0.126) and nucleotide (mean pairwise distance between paralogs = 0.129) level. Mean pairwise divergence is much lower within either the A (nucleotide distance = 0.038; amino acid distance = 0.041) or B (nucleotide distance = 0.044; amino acid distance: 0.057) paralog. *CENH3* can be partitioned into two functional regions—the N-terminal tail and the HFD. Most of the divergence between paralogs is concentrated in the N-terminal tail (nucleotide distance = 0.206; amino acid distance = 0.368), rather than the HFD (nucleotide distance = 0.097; amino acid distance = 0.101). *CENH3_B* has one fixed deletion relative to all other *Mimulus CENH3*s (including *CENH3_A*) and *CENH3_A* has one fixed insertion relative to *CENH3_B*. Based on local synteny patterns with *Solanum* and *Populus* (http://phytozome.jgi.doe.gov/, last accessed July 1, 2015), *CENH3_A* appears to be the ancestral copy. Both duplicates retain similar intron–exon structures as the pro-ortholog *CENH3*s, suggesting the duplication event likely did not involve a messenger RNA intermediate.

Accelerated and Adaptive Evolution in the N-Terminal Tail of CENH3

To test for variation in selective constraint across the N-terminal tail and the HFD, we used sites models C and E (Yang and Swanson 2002). Model C constrains the entire gene to a single ω value, whereas model E allows ω to vary across a priori assigned partitions. Using model E to compare all *Mimulus CENH3*s, ω was larger in the N-terminal tail versus the HFD ($\omega_N = 0.869$ vs. $\omega_{HFD} = 0.143$, $P < 0.00001$; table 1), suggesting that N-terminal region has evolved faster than the HFD. Because of these quantitative differences in selective constraint across regions, we treated the N-terminal tail and HFD separately in all downstream analyses.

The high ω values estimated from the entire N-terminal tail are due in part to increased positive selection acting on individual sites. When random-sites models (Nielsen and Yang 1998; Yang et al. 2000, 2005; Swanson et al. 2003; Wong et al. 2004) allowed ω to vary across sites but not lineages, models that allow sites to evolve under positive selection ($\omega > 1$) were well-supported for the N-terminal domain, but not the HFD (M8 or M2a, table 2). Fifteen of 63 sites in the N-terminal domain (23.8%) show greater than

Table 1. Parameters and Log Likelihoods of Fixed-Sites Models.

Partition	Model[a]	Tree Topology	Estimated Parameters	df[b]	Log-Likelihood	-2Δl MC vs. ME[c]
N/HFD	Model C	Figure 1*a*	$\omega_N = \omega_{HFD} = 0.388$	2	−2117.68	79.70***
	Model E	Figure 1*a*	$\omega_N = 0.869$, $\omega_{HFD} = 0.143$		−2077.83	
HFD/Loop 1	Model C	Figure 1*a*	$\omega_{HFD} = \omega_{Loop1} = 0.125$	2	−1097.21	20.07**
	Model E	Figure 1*a*	$\omega_{HFD} = 0.083$, $\omega_{Loop1} = 1.523$		−1087.17	
CENH3_A/*CENH3_B*	Model C	*CENH3_A* topology	$\omega_A = \omega_B = 0.325$	2	−1909.70	3.86
	Model E	*CENH3_A* topology	$\omega_A = 0.255$, $\omega_B = 0.402$		−1907.77	
CENH3_A/*CENH3_B*	Model C	*CENH3_B* topology	$\omega_A = \omega_B = 0.382$	2	−1926.74	2.90
	Model E	*CENH3_B* topology	$\omega_A = 0.314$, $\omega_B = 0.463$		−1925.29	

[a]From Yang and Swanson (2002).
[b]Degrees of freedom.
[c]Significant at $^{**}P < 0.0001$, $^{***}P < 0.00001$.

Table 2. Parameters and Log LRTs of Random-Sites Models.

Domain	-2Δl M2a vs. M1a[a,b]	-2Δl M8 vs. M7[a,b]	-2Δl M8 vs. M8a[a,b]	Parameter Estimates from M8[b]	Positively Selected Sites[c]
N-terminal tail	*11.55***	*11.67***	*11.53****	$P = 0.011$, $q = 0.005 P_0 = 0.958$, $P_1 = 0.042$, $\omega = 4.85$	*4K*, *7A*, 12S, 16P, 17A, 22A, *26T*, 33P, *34S*, *35G*, 43A, 52D, 53G, 56E, *57R*
HFD	0.00	0.37	0.00	NA	NA
Loop 1 of HFD	2.76	5.05	2.76	NA	NA

NOTE.—Degrees of freedom = 2 for M1a/2a and M7/8 comparisons, = 1 for M8a/8.
[a]For model comparisons (see text), significance is indicated as $^{**}P < 0.01$, $^{***}P < 0.001$.
[b]Results reported from initial $\omega = 1.0$.
[c]For positively selected residues, those in italic and underline have posterior probabilities > 0.99, italic > 0.95, underline 0.75–0.94, and regular font 0.50–0.74.

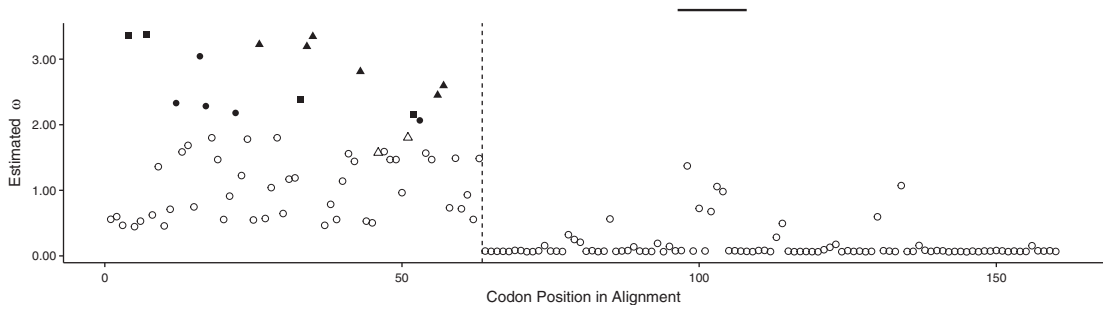

FIG. 2. Distribution of positive selection relative to codon position across *CENH3*. Estimated ω values per codon site in *CENH3* from random-sites models, which allow ω to vary across sites but not lineages. Values are calculated using an empirical Bayes approach as the mean over site classes (M8) weighted by the posterior probabilities for belonging to a particular site class. Sites with significant evidence of positive selection over all lineages are filled (table 2). Squares and triangles represent sites that show signs of positive selection in the pro-orthologs (squares) and the B paralog (triangles) according to branch-sites models (table 3). The N-terminal tail and HFD (separated by the dashed line) were analyzed separately and loop 1 is indicated by the bar above the graph.

50% Bayesian empirical probabilities of belonging to site classes where $\omega > 1$, and five sites were assigned to this site class with greater than 95% posterior probabilities (sites: 4K, 7A, 26T, 34S, 35G; table 2 and fig. 2 and supplementary fig. S1, Supplementary Material online).

There was no evidence of adaptive evolution at any particular site in the HFD, but we did notice a trend where ω estimates for loop 1 of the HFD approached or exceeded one (fig. 2). Therefore, we performed two post hoc tests to explore how selection shaped loop 1 of the HFD. First, we delineated loop 1 of the HFD (amino acids 97–107, following Tachiwana et al. 2011) and asked whether it evolved at different rates than the remainder of the HFD using fixed-sites models C and E (Yang and Swanson 2002). Loop 1 revealed significantly elevated ω values relative to the rest of the HFD, with the regional ω estimate for loop 1 exceeding one ($\omega_{HFD} = 0.083$ vs. $\omega_{Loop1} = 1.523$, $P < 0.0001$; table 1). Random-sites models were then applied specifically to the loop 1 region to investigate whether positive selection was responsible for the observed ω values (Nielsen and Yang 1998; Yang et al. 2000, 2005; Swanson et al. 2003; Wong et al. 2004). For loop 1, models that allowed sites to evolve under positive selection were not significantly better than null models (table 2). Our results are, therefore, consistent with loop 1 of the HFD

evolving under relaxed selective constraint, rather than positive or purifying selection.

Some Sites in the N-Terminal Tail Are under Positive Selection in the Pro-Orthologs and CENH3_B

Across the entire gene, *CENH3_A* and *CENH3_B* do not evolve under different levels of selective constraint (table 1). Fixed-sites models that constrained the paralogs to have a single ω (Model C) were not significantly better than those that allowed ω to vary across paralogs (Model E), regardless of which tree topology (A/B) was used (table 1; Yang and Swanson 2002). Likewise, branch-models showed no evidence that the paralogs evolved under different levels of selective constraint (fig. 1*b* and supplementary table S2, Supplementary Material online). Branch models allow ω to vary among branches in the phylogeny, but hold ω constant among sites on a particular branch (Yang 1998). In all cases, models that constrained all branches to a single ω were not significantly different than models allowing separate ω's for *CENH3_A* and *CENH3_B* (supplementary table S2, Supplementary Material online).

One limitation of fixed-sites and branch models is that they make the unrealistic assumption of among-site homogeneity. However, selection may act differently on particular codons in a subset of lineages. Therefore, we employed branch-sites and clade models to test for variation in ω across sites among prespecified branches of a phylogeny. Branch-sites models test for the signature of positive selection acting on codons along a priori designated lineages (Yang and Nielsen 2002; Swanson et al. 2003; Zhang et al. 2005), whereas clade models account for sites that experienced divergent selective pressures in predefined clades (Bielawski and Yang 2004; Yang et al. 2005; Weadick and Chang 2012).

Using branch-sites models, we uncovered a history of positive selection acting on several sites along the pro-ortholog and *CENH3_B* branches for the N-terminal tail, but not for the HFD (table 3 and fig. 2; supplementary fig. S1, Supplementary Material online). In comparisons of alternative and null branch-sites models from the N-terminal tail, we identified four sites that experienced positive selection in the pro-orthologs but neutral or purifying selection in the background (4K, 7A, 33P, 52D; $\omega_0 = 22.35$; $P = 0.0022$; table 3 and fig. 2). Likewise, eight sites showed evidence of positive selection along the *CENH3_B* paralog branch only (26T, 34S, 35G, 43A, 46P, 51S, 56E, 57R; $\omega_3 = 9.73$; $P = 0.0026$; table 3 and fig. 2). Most sites evolving under positive selection along either the pro-ortholog or *CENH3_B* branch of the N-terminus were also identified as evolving under positive

Table 3. Parameter Estimates (ω's) and LRTs for Branch-Sites Models.

Domain	Foreground Branch[a]	df[b]	Site Class:	0	1	2a	2b	-2Δl MA vs. MA_{null}[c,d]	Positively Selected Sites in Foreground[e]
N	ω_0	1	Proportion:	0.31	0.69	0.00	0.00	9.39**	*4K*, *7A*, 33P, 52D
			Background ω:	0.00	1.00	0.00	1.00		
			Foreground ω:	0.00	1.00	22.35	22.35		
	ω_1	1	Proportion:	0.30	0.46	0.09	0.14	0.52	
			Background ω:	0.03	1.00	0.03	1.00		
			Foreground ω:	0.03	1.00	2.55	2.55		
	ω_2	1	Proportion:	0.37	0.63	0.00	0.00	0.00	
			Background ω:	0.03	1.00	0.03	1.00		
			Foreground ω:	0.03	1.00	1.00	1.00		
	ω_3	1	Proportion:	0.31	0.63	0.02	0.04	9.02**	26T, 34S, *35G*, 43A, 46P, 51S, *56E*, 57R
			Background ω:	0.03	1.00	0.03	1.00		
			Foreground ω:	0.03	1.00	9.73	9.73		
HFD	ω_0	1	Proportion:	0.79	0.16	0.04	0.01	0.00	
			Background ω:	0.04	1.00	0.04	1.00		
			Foreground ω:	0.04	1.00	1.00	1.00		
	ω_1	1	Proportion:	0.82	0.18	0.00	0.00	0.00	
			Background ω:	0.04	1.00	0.04	1.00		
			Foreground ω:	0.04	1.00	1.00	1.00		
	ω_2	1	Proportion:	0.82	0.18	0.00	0.00	0.00	
			Background ω:	0.04	1.00	0.04	1.00		
			Foreground ω:	0.04	1.00	1.00	1.00		
	ω_3	1	Proportion:	0.78	0.15	0.06	0.01	0.07	
			Background ω:	0.04	1.00	0.04	1.00		
			Foreground ω:	0.04	1.00	1.68	1.68		

[a]Foreground branches as in figure 1b.
[b]Degrees of freedom.
[c]From Zhang et al. (2005).
[d]Significant at $^{**}P < 0.01$.
[e]For positively selected residues, those in italic and underline have posterior probabilities > 0.99, italic > 0.95, underline 0.75–0.94, and regular font 0.50–0.74.

selection in random-sites models, though two sites were specific to the B branch-sites analysis (46P, 51S; fig. 2). In contrast, we see no evidence for sites in the N-terminus to evolve under positive selection along the branches immediately following duplication ($\omega_1 = 2.55$; $P = 0.518$) or the *CENH3_A* paralog ($\omega_2 = 1.00$; $P = 1.00$; table 3). Because we see sites evolving under positive selection in *CENH3_B* but not *CENH3_A*, it suggests that some codons may have experienced different levels of selective constraint in the two paralogs. For the HFD, we see no support for positive selection acting on any particular site along any foreground branch ($P > 0.05$ in all cases; table 3 and fig. 2).

Unlike branch-sites models, clade models showed no evidence of divergent selection acting on individual sites for either the N-terminal tail or HFD (supplementary table S3, Supplementary Material online). The discrepancy between branch-sites and clade models is likely due to subtle differences in the way selection is detected by the models. Branch-sites models test for signatures of positive selection along designated foreground branches only, thereby disallowing positive selection in background branches or null models. In contrast, clade models freely estimate ω's for each a priori designated clade and permit sites under positive selection in null models. The null models for our clade model analyses include a site class with positive selection acting along all branches and find no support for models that allow selection to act divergently in particular clades (supplementary table S3, Supplementary Material online). Moreover, the divergent site class for all the alternative clade models show positive selection acting in all clades. Taken together, although our branch-sites models suggest that the pro-orthologs and *CENH3_B* clades reveal sites evolving under different constraints than in the *CENH3_A* clade (table 3), these differences fade when positive selection is permitted across the tree (supplementary table S3, Supplementary Material online). Therefore, we interpret our data as providing strong support for adaptive evolution at several sites in the N-terminus (table 2), with weak evidence for positive selection to be particularly strong at certain sites in the pro-orthologs and *CENH3_B* clades (table 3).

No Evidence of a Burst of Positive Selection Following Duplication

Gene duplication events are sometimes followed by a burst of adaptive evolution (e.g., Bielawski and Yang 2001). Because they isolated the branches immediately following duplication (fig. 1*b*, ω_1), our previously described branch and branch-sites models allowed us to test this idea. Models allowing ω to vary among branches were not supported in any case (supplementary table S2, Supplementary Material online). Likewise, models that allowed positive selection at particular sites along the ω_1 branch were also not supported (table 3). In summary, we find no evidence of changes in selective constraint immediately following the gene duplication event. However, strong positive selection occurring in the background may make it difficult to discern minor changes in selective constraint post-duplication.

CENH3_A and CENH3_B Evolve Differently within a Population Polymorphic for Drive

Above, we characterized evolutionary patterns of *CENH3* across the *Mimulus* genus and found compelling evidence that the gene, and the pro-orthologs and *CENH3_B* in particular, evolved under positive selection over deep timescales. To date, however, active female meiotic drive in *Mimulus* has only been documented to occur within a single population (Fishman and Saunders 2008). Therefore, we characterized patterns of intraspecific polymorphism for *CENH3_A* and *CENH3_B* in the IM population, where ongoing drive was originally described (Fishman and Willis 2005). For this analysis, we inferred *CENH3* sequences from previously published whole-genome data of ten resequenced inbred lines deriving from the IM population (Flagel et al. 2014).

Intraspecific polymorphism levels (number of haplotypes, π, and θ_W) are consistently low in *CENH3_A* and intermediate in *CENH3_B* (table 4). Coalescent simulations reveal that *CENH3_A* has significantly lower polymorphism levels than *CENH3_B* for all metrics save haplotype diversity (nonoverlapping confidence intervals (CIs); table 4). Likewise, Hudson, Kreitman, and Aguade's (HKA; 1987) tests show that *CENH3_A* and *CENH3_B* have different

Table 4. Polymorphism Data for *CENH3_A* and *CENH3_B* from Ten Inbred Lines from the IM Population.

	CENH3_A			*CENH3_B*		
Polymorphism metric	**Estimate**	**Lower 95% CI[a]**	**Upper 95% CI**	**Estimate**	**Lower 95% CI**	**Upper 95% CI**
Number of haplotypes	**2**	**1**	**3**	**8**	**4**	**9**
Haplotype diversity	**0.20**	**0.00**	**0.64**	**0.93**	**0.60**	**0.98**
π per gene	**0.20**	**0.00**	**0.96**	**6.33**	**1.51**	**16.62**
θ_w	**0.35**	**0.00**	**1.06**	**5.66**	**1.77**	**13.79**
Tajima's *D*	**−1.11**	**−1.11**	**1.46**	**0.55**	**−1.69**	**1.68**
Fu and Li's *D**	**−1.24**	**−1.24**	**1.03**	**0.56**	**−1.96**	**1.35**
Fu and Li's *F**	**−1.35**	**−1.35**	**1.15**	**0.63**	**−2.16**	**1.47**
HKA test[b]						
Average number of substitutions	**14.0**	**—**	**—**	**17.3**	**—**	**—**
Number of segregating sites	**1**	**—**	**—**	**16**	**—**	**—**
π per site	**0.00044**	**—**	**—**	**0.01406**	**—**	**—**

[a]The 95% CIs were generated by 1,000 coalescent simulations in DNASP v. 5.10.1.

[b]HKA test significant at $P < 0.05$; *M. dentilobus* chosen for outgroup.

ratios of within-species diversity relative to between-species divergence, with *CENH3_A* displaying lower levels of intraspecific polymorphism than expected (table 4). A reduction in polymorphism, as seen in *CENH3_A*, is generally consistent with a history of recent positive selection. However, several tests of selective neutrality (Tajima's *D*, Fu and Li's *D**, Fu and Li's *F*) do not reject neutrality for either paralog (CIs overlap zero in all cases; table 4). While we can conclude that *CENH3_A* and *CENH3_B* evolve under distinct selective dynamics in the short-term, more extensive population genetic sampling is required to link directional selection on *CENH3_A* to the action of centromere-associated drive.

No Evidence for Differential Expression between CENH3 Paralogs

Given divergent selection in both the short- and long-term, we were interested in whether or not *CENH3_A* and *CENH3_B* paralogs are functionally equivalent. To explore this idea, we compared levels of gene expression between *CENH3_A* and *CENH3_B* using RNA-Seq data collected by Colicchio et al. (2015). The RNA for this experiment was collected from leaf tissue of *M. guttatus* progeny of parents that were damaged or undamaged. We compared the number of raw read counts mapped to each gene for both *CENH3_A* and *CENH3_A* and found no significant effect of paralog ($F_{1,18} = 0.016$, $P = 0.902$), treatment ($F_{1,18} = 1.271$, $P = 0.206$), or their interaction ($F_{1,18} = 0.179$, $P = 0.677$) on gene expression levels. Our results suggest that that *CENH3* paralogs are expressed at the same level in mitotic leaf tissue, but cannot speak to expression levels across different tissues, developmental times, or cell cycles.

Discussion

Centromeres mediate faithful segregation of chromosomes, yet both centromeric DNA and kinetochore proteins are highly variable. This paradoxical diversity is thought to result from a coevolutionary arms race between selfishly evolving centromeres that spread via female meiotic drive and kinetochore proteins, like *CENH3*, that adapt to restore equal segregation of centromeres. Here, we characterized the molecular evolutionary patterns of *CENH3* in *Mimulus*, a genus with well-described centromere-associated female meiotic drive (Fishman and Willis 2005; Fishman and Saunders 2008). First, we identified a novel duplication of *CENH3* in the absence of a whole-genome duplication (fig. 1*a*). Second, we found evidence of long-term, recurrent positive selection in the N-terminus of *CENH3*, as predicted by the centromeric drive model (fig. 2 and table 2). Finally, the paralogs appear to evolve under different selective dynamics in both the short- and long-term. At deeper timescales, some sites in the N-terminus of *CENH3_B*, but not *CENH3_A*, showed signatures of positive selection (fig. 2 and table 3); at shorter timescales, *CENH3_A* revealed significantly lower levels of intraspecific polymorphism than *CENH3_B* in a population with drive (table 4). Because evolution occurs within populations, our analyses point to *CENH3_A* as a stronger candidate to suppress active centromere-associated drive than *CENH3_B*.

A Novel CENH3 Duplication in the Absence of Whole-Genome Duplication

We characterize a novel duplication of *CENH3* in several diploid *Mimulus* lineages, including the IM population of *Mimulus guttatus* where drive was first described and is polymorphic (fig. 1 (IM); Fishman and Willis 2005; Fishman and Saunders 2008). Both unlinked *CENH3* paralogs are expressed, yet they are also highly divergent and show distinct evolutionary histories in the short- and long-term (tables 3 and 4). Intriguingly, our *CENH3* gene tree suggests that the *Mimulus* duplication event coincides with a genome-wide fission event that resulted in a near-doubling of chromosome number, though more sampling is needed to clarify whether fission and duplication perfectly coincide (fig. 1; Fishman et al. 2014). Chromosomal fission has been assumed to be rare in plants, but comparative mapping and genome analysis definitively show that the lineage leading to *M. guttatus* underwent a fissional increase in chromosome number (Clarke 2012; Fishman et al. 2014). Together, our results suggest that *CENH3* duplicated in diploid *Mimulus* in the absence of whole-genome duplication. Alternatively, the paralogs could be retained from one of two ancient paleopolyploidy events that predate the *Mimulus* genus (Clarke 2012). If so, *CENH3_B*, the derived paralog, would have been independently lost a minimum of two times and maintained at least 46 My in lineages with two *CENH3s*. However, Hasegawa–Kishino–Yano distances between *CENH3* paralogs characterized from the focal IM population (0.27; distances normalized to the potato/*Mimulus* speciation event) place the duplication event well inside the youngest of the two paleopolyploidy events, μ (0.66 ± 0.00006; T. Clarke, personal communication; Hasegawa et al. 1985). Thus, *CENH3* duplication most likely represents a local duplication event, not a whole-genome ploidy shift.

Adaptive Evolution of the N-Terminal Tail of CENH3 in the Context of Meiotic Drive

Theory and circumstantial evidence suggest that female meiotic drive is widespread in plants and animals (Pardo-Manuel de Villena and Sapienza 2001a, 2001b; Malik and Bayes 2006). As such, constant coevolutionary conflict between driving centromeres and kinetochore proteins is often argued to produce the nearly ubiquitous observation of rapid, adaptive evolution of *CENH3* in species with female meiosis (Henikoff et al. 2001; Malik and Henikoff 2001, 2002; Cooper and Henikoff 2004; Malik and Bayes 2006; Hirsch et al. 2009; Talbert et al. 2009; Schueler et al. 2010; Zedek and Bureš 2012; but see Elde et al. 2011). Conversely, in *Saccharomyces* budding yeast with only symmetrical meiosis ("male meiosis"), there is no evidence that *CENH3* evolves under positive selection (Talbert et al. 2004). However, because centromeric drive need not be strong to be evolutionarily powerful and because it is predicted to be only transiently polymorphic, few cases of active drive are

known, limiting opportunities to examine the direct effects of drive on centromere protein evolution. Here, we use gene sequence comparisons to infer long-term, recurrent positive selection acting on *CENH3* in a genus known to exhibit centromere-associated female meiotic drive (Fishman and Willis 2005; Fishman and Saunders 2008). Our results suggest that several amino acids in the N-terminal tail of *CENH3* evolve under diversifying selection across the *Mimulus* genus (fig. 2 and table 2; supplementary fig. S1, Supplementary Material online). In addition, by using branch-sites models, we were able to more finely pinpoint the lineages experiencing adaptive evolution. We found that, for many of the sites evolving adaptively in the N-terminus, this signal is likely driven by positive selection acting chiefly in the pro-orthologs and *CENH3_B* in yellow monkeyflowers (fig. 2 and table 3; supplementary fig. S1, Supplementary Material online).

Intriguingly, the N-terminus of *CENH3* may function in some meiosis-specific ways. Localization of N-terminally truncated *CENH3* shows that the HFD of *CENH3* is sufficient for centromere localization in mitosis, but not meiosis (Lermontova et al. 2006, 2011). Moreover, individuals without appropriate N-terminal tails showed error-free mitotic growth, but had reduced fertility due to meiotic defects (Lermontova et al. 2011). Additionally, naturally evolved variation in the N-terminus of *CENH3* can cause segregation errors, genome elimination, and novel genetic rearrangements in crosses of *Arabidopsis thaliana* (though these effects may be due to postzygotic interactions in hybrids rather than meiotic dysfunction per se; Maheshwari et al. 2015). Taken together, meiosis-specific functions of *CENH3* may be compartmentalized in the N-terminus. Given that 1) selfish centromere drive imposes selection in meiosis alone, 2) the N-terminus may function in meiosis-specific ways, 3) the N-terminal tail likely makes extensive contact with linker DNA in centromeric chromatin (Malik et al. 2002), and 4) adaptive evolution is often reported in the N-terminal tail, the N-terminus of *CENH3* is a strong candidate region for coevolving with driving centromeres. The present work strengthens this argument by reporting long-term, recurrent positive selection of the N-terminus of *CENH3* in a system with centromere-associated meiotic drive.

In contrast to the N-terminal domain but similar to findings from previous work, the majority of the HFD in *Mimulus* is conserved and evolves under negative selection (e.g., Cooper and Henikoff 2004; Schueler et al. 2010; Zedek and Bureš 2012; fig. 1). In *Drosophila*, *Arabidopsis*, and *Caenorhabditis*, loop 1 of the HFD has been reported to be evolving under positive selection, consistent with evidence suggesting that loop 1 contacts centromeric DNA and may even provide some *CENH3*-centromeric DNA specificity (Malik and Henikoff 2001; Vermaak et al. 2002; Cooper and Henikoff 2004; Zedek and Bureš 2012). Although we found no significant evidence of adaptive evolution in this region, loop 1 of the HFD in *Mimulus* does evolve under relaxed selective constraint (tables 1 and 2). One possibility is that loop 1 of the HFD is sometimes under positive selection, and other times not, leading to an overall weaker signal of directional selection that we were unable to detect.

CENH3 Paralogs Exhibit Distinct Dynamics within a Population Polymorphic for Drive

Over the long-term, both centromeric DNA and *CENH3* sequences are predicted to turnover due to repeated bouts of female meiotic drive of centromeres and suppression by *CENH3* (Henikoff and Malik 2002; Malik and Henikoff 2002). Our observed strong signal of positive selection in the N-terminus of *CENH3* matches these long-term expectations, with weak evidence that *CENH3_B* historically played a more prominent role in suppression (tables 2 and 3). Yet, drive occurs in populations and, at any given moment, either paralog may be responding to driving centromeres. If either paralog has indeed evolved to suppress drive, we predict to see signatures of recent selection in the interacting duplicate at the population level. Here, we documented distinct patterns of molecular variation for the *CENH3* paralogs in a population that is polymorphic for drive (IM; Fishman and Saunders 2008). Specifically, *CENH3_A* exhibited a significant reduction in intraspecific variation relative to *CENH3_B* (table 4). Although recent selective sweeps produce low levels of polymorphism as seen for *CENH3_A*, we cannot currently reject a hypothesis of neutrality for either paralog (table 4). However, the observed reduction of polymorphism in *CENH3_A* relative to *CENH3_B*, points to *CENH3_A* as the more likely suppressor of active centromere-associated drive and encourages further population genetics surveys.

Possible Mechanisms Maintaining Two Copies of CENH3

Long-term retention of two *CENH3* alleles is puzzling, as putative fitness costs associated with *CENH3* misexpression argue against maintenance of two functional *CENH3s*. Without a single *CENH3*-based kinetochore at each centromere of a sister chromatid (i.e., zero or multiple kinetochores), chromosomes can be detrimentally missegregated (Howman et al. 2000; Oegema et al. 2001; Heun et al. 2006; Lermontova et al. 2011, but see Neumann et al. 2012, 2015). In somatic tissues (i.e., mitosis), such failures drive tumor formation and *CENH3* overexpression has been linked to cancer (Tomonaga et al. 2003; Weaver and Cleveland 2007). In gametic cells (i.e., meiosis), *CENH3* misregulation can result in reduced fertility and aneuploidy (Brar and Amon 2008). Consequently, *CENH3* levels are tightly regulated and the gene is primarily single copy, even in ancestral polyploidy plants (e.g., Zhong et al. 2002); nonetheless, maintenance of two or more *CENH3* proteins is not unheard of in diploid organisms. For example, Maheshwari et al. (2015) recently surveyed publically available *CENH3* sequences from 61 plant species and identified 6 diploid species with duplicated *CENH3*s (Kawabe et al. 2006; Moraes et al. 2010; Sanei et al. 2011; Neumann et al. 2012; Yuan et al. 2014), suggesting the frequency of retained duplicates in diploids may be around 10%.

Given the presumed fitness costs of misexpression of *CENH3*, why are two copies maintained in *Mimulus* and

some diploid species? Numerous models exist to explain the evolutionary mechanisms maintaining gene duplicates and can be broadly delimited into neofunctionalization, subfunctionalization, and gene conservation categorizations (Ohno 1970; Hahn 2009; Innan and Kondrashov 2010). Neofunctionalization, or the evolution of a novel function in one paralog, seems unlikely, as both paralogs target to centromeres in other species (Sanei et al. 2011; Neumann et al. 2015) and are expressed at similar levels in leaf tissue in *Mimulus* (this study). However, given the signature of long-term positive selection specific to the derived paralog, *CENH3_B*, it is possible that this copy is involved in some novel function. Subfunctionalization, or the sharing of ancestral function between paralogs, may occur at least partially in some systems. For example, in *Caenorhabditis elegans*, *CENH3* paralogs are expressed at dramatically different levels (Monen et al. 2005) and their presence on centromeres varies in mitosis versus meiosis in wheat (as well as time and space in mitosis; Yuan et al. 2014). Although we did not find that *Mimulus CENH3* paralogs were differentially expressed in mitotic leaf tissue, their distinct molecular evolutionary patterns (particularly within *M. guttatus*), suggest that they may not be functionally equivalent at any given time. One intriguing possibility is that duplication followed by specialization could release *CENH3* from the adaptive conflict imposed by distinct selective pressures in meiosis and mitosis (Hughes 1994). Finally, the conservation of ancestral function between paralogs could maintain duplicate genes through a favorable increase in protein level (i.e., dosage). Such dosage effects may be particularly important when centromeres expand or when *CENH3* interacts with multiple kinds of centromeric repeats. In diploid pea and *Arabidopsis*, *CENH3* duplication coincides with the presence of multiple centromeric repeats across chromosomes, as well as an expansion from monocentric to metapolycentric centromeres in pea (Kawabe and Nasuda 2004; Kawabe et al. 2006; Neumann et al. 2012, 2015). In *Mimulus*, we find that *CENH3* duplication possibly coincides with a genome-wide fission event (fig. 1; Fishman et al. 2014). Preliminary bioinformatics analyses (F. Finseth and L. Fishman, unpublished data) suggest that this radical change in genome architecture was followed by diversification of centromeric repeats within one lineage and massive expansion of a single centromere-associated repeat (Cent728) within the lineage including *M. guttuatus* (Fishman and Saunders 2008; Fishman et al. 2014). Although it is not yet clear whether and how subfunctionalization and dosage effects contribute to the maintenance of divergent (and positively selected) paralogs of *CENH3* in *Mimulus*, this group of taxa is particularly fertile ground for further explorations of *CENH3* evolution and function.

Conclusions

Circumstantial evidence suggests that female meiotic drive promotes variation at *CENH3* across numerous taxa (Henikoff and Malik 2002; Malik and Henikoff 2002), but our work is the first to document molecular variation of *CENH3* in a system with centromere-associated drive (Fishman and Saunders 2008). We provide strong evidence of adaptive evolution in the N-terminus of *CENH3* throughout the genus *Mimulus*, including after a rare diploid duplication and retention of *CENH3*, and also show that paralogs are evolving differently in a population with drive. This work supports the idea that centromeric DNA and *CENH3* evolve rapidly due to constant genetic conflict in meiosis, although we cannot yet isolate the drivers of either rapid protein evolution or duplicate retention. Further work, focused on *CENH3_A* as the most likely interactor with actively driving centromeres, will be necessary to reveal the population genetic processes that must underlie the ubiquitous, long-term pattern of positive selection on *CENH3*.

Materials and Methods

Sequencing

CENH3 sequences were obtained from 12 samples (*M. aurantiacus*, *M. bolanderi*, *M. cardinalis*, *M. dentilobus*, *M. guttatus* IM population, *M. guttatus* DUN population, *M. jungermannoides*, *M. lewisii*, *M. nasutus*, *M. parishii*, *M. primuloides*, *M. tilingii*). We initially designed degenerate primers based on the *M. guttatus* genome assembly (v 1.0; Hellsten et al. 2013) to obtain *CENH3* sequence from genomic DNA of diverse species. Briefly, we extracted genomic DNA from fresh or silica-dried leaves using a modified CTAB-chloroform extraction protocol (Fishman and Willis 2005). *CENH3* was then amplified with various primer pairs using standard touchdown PCR conditions, with annealing temperatures adjusted based on primer melting temperatures. Upon discovering that *CENH3* was duplicated, we also attempted to sequence both paralogs from cDNA for a subset of species. RNA was extracted from floral buds with the RNEasy Plant Mini Kit (Qiagen) and converted into cDNA using the Superscript III First Strand Synthesis System (Invitrogen). For both the genomic and cDNA-based amplifications, PCR products were run on agarose gels, gel-extracted, cloned with the Zero Blunt TOPO PCR Cloning Kit (Invitrogen), and sequenced using standard Sanger sequencing protocols.

For all species, next-generation sequencing data were used to confirm or (in the case of two species) characterize *CENH3* sequence(s). We either used de novo assembled genomes from collaborators or directly analyzed short read data to infer *CENH3* sequence. To obtain short read data, we either downloaded it from the JGI Sequence Read Archive or directly sequenced a species' genome. If sequenced, buds were collected in the greenhouse and immediately frozen on dry ice. DNA libraries were created and barcoded with the Nextera DNA Sample Preparation Kit (Illumina) and 2×150 paired-end sequenced on a HiSeq 2500 (Illumina) by Duke University's Genome Sequencing Resource. We trimmed adaptor and low quality sequences with Trimmomatic version 0.30 (Bolger et al. 2014). For species closely related to *M. guttatus*, sequences were first mapped to the *M. guttatus* v 2.0 genome (http://phytozome.jgi.doe.gov/, last accessed July 1, 2015), with bwa version 0.7.5a (Li and Durbin 2009) and *CENH3* was inferred from the consensus sequence. For species more distantly related to *M. guttatus*, genomes were assembled with Masurca version 2.2 (Zimin et al. 2013) with default

settings, except that jellyfish size was set to 8×10^9. We then identified *CENH3* sequences using BLASTn with *CENH3* sequences from species closely related to the focal species as queries. Because most species' *CENH3* sequence(s) were characterized using at least two approaches, we were able to independently confirm *CENH3* sequence as well as validate our methods. For additional sample and sequencing details including sequencing methodology for each species and primer sequences, see supplementary table S1, Supplementary Material online.

Gene Tree Construction

Eighteen *CENH3* sequences were assembled in Mega version 5.2 (Tamura et al. 2011), aligned using the ClustalW algorithm (Thompson et al. 1994; Larkin et al. 2007), and refined manually, for a total of 483 aligned nucleotides (see supplementary alignment.txt, Supplementary Material online). Mega was also used to calculate mean pairwise amino acid and nucleotide distances within and between paralogs using the Poisson (amino acid) and maximum likelihood composite (nucleotide) methods with default settings. A gene tree was built from third position sites and indels using MrBayes version 3.2.2 (Ronquist et al. 2012). For third position sites, the general time reversible (GTR) + Γ nucleotide substitution model was chosen in MrBayes as determined by Akaike information criterion rank in jModeltest version 2.1.4 (Guindon and Gascuel 2003; Darriba et al. 2012). Indels were coded as binary data and treated as restriction sites according to MrBayes version 3.2.2 (Ronquist et al. 2012). Gene trees were estimated with 150,000 Markov chain Monte Carlo (MCMC) generations sampled every 100 generations. We confirmed convergence and adequate sampling, as the standard-deviation of split frequencies was less than 0.01 at the end of the analysis, and parameter estimate-by-generation plots were stationary. The consensus gene tree was used for downstream analyses of selective constraint and is represented in figure 1*a*. Branch lengths for selection analyses were estimated with the M0 model in the codeml package of the software PAML version 4.7 (Yang 2007).

Codon Substitution Models

We explored patterns of selective constraint using five classes of codon substitution models in the codeml package of PAML version 4.7: fixed-sites, branch, random-sites, branch-sites, and clade models (Yang 2007). Within each model class, we performed likelihood ratio tests (LRTs) to compare the fit of complex models with simpler, nested models. LRT test statistics were computed as twice the difference between log-likelihoods for nested models and compared with a χ^2 distribution with degrees of freedom equal to the number of extra parameters estimated by the complex model. For all analyses, we applied the F3x4 codon model of substitution.

First, fixed-sites models of Yang and Swanson (2002) tested for statistical variation in ω between *CENH3* paralogs and a priori defined functional regions of *CENH3*. For this set of analyses, we treated each paralog (A vs. B) or functional region (N-terminal tail [codons 1–63] versus the HFD [codons 64–160]) as distinct and asked whether a model constrained to a single ω (model C) was significantly better than one that allowed each partition to have separate ω's (model E). For the paralog analysis, only those species with a copy of both *CENH3_A* and *CENH3_B* were included and we ran models with both *CENH3_A* and *CENH3_B* tree topologies. Because model E fit significantly better than model C for functional regions, we treated the N-terminal tail and HFD separately for all downstream analyses.

Second, we employed branch models, which allow ω to vary among branches in the phylogeny, but hold ω constant among sites on a particular branch (Yang 1998). The gene tree was partitioned into pro-orthologs (ω_0), the branches immediately following duplication (ω_1), and paralog A and B clades (ω_2 and ω_3, respectively; fig. 1*b*). Branch models were constrained to one, two, three or four ω ratios and nested models were compared to determine significance.

Random-sites and branch-sites models were used to test for positive selection on particular sites during the evolution of *CENH3*. Random-site models allow ω to vary among sites but not across lineages (Nielsen and Yang 1998; Yang et al. 2000, 2005; Swanson et al. 2003; Wong et al. 2004). First, the nearly neutral model M1a specifies two site classes, conserved ($0 < \omega < 1$) and neutral ($\omega = 1$). This model was compared with model M2a, allowing an additional class of codons under positive selection ($\omega > 1$). Second, the neutral model M7 that limits ω to a beta distribution between 0 and 1 was compared with model M8 that has an additional site class of codons with $\omega > 1$. Our final sites model LRT compared M8 to a similar neutral model, M8a, that has an additional class of codons with ω constrained to 1. The modified branch-sites Model A was compared with Model A_{null} to examine whether particular sites evolved under positive selection along a priori specified branches (Zhang et al. 2005). The branches representing pro-orthologs (ω_0) and the A or B paralogs (ω_2 and ω_3) and the branch immediately following duplication (ω_1) were specified as foreground branches for these tests (fig. 1*b*).

Clade model C (CmC) allowed us to test for divergent selection on particular sites among a priori designated lineages (Bielawski and Yang 2004; Yang et al. 2005). The modified null model of CmC (M2a_rel) assumes sites fall into three classes; sites either experienced purifying selection ($0 < \omega < 1$), neutral evolution ($\omega = 1$), or positive selection ($\omega > 1$) across the entire phylogeny (Weadick and Chang 2012). For the alternative model CmC, the third site class allows the estimated ω for a site to diverge across a priori assigned branches (e.g., $\omega > 1$ in a "foreground" branch, and $\omega < 1$ in a "background" branch). We first treated either paralog A or paralog B clades as foreground branches in two separate tests. We then used an extended version of CmC that allows ω among to vary across more than two branches and partitioned the gene tree into pro-orthologs (ω_0), paralog A clades (ω_A) and paralog B clades (ω_B; Yoshida et al. 2011). These designations were nearly identical to those in figure 1*b*, with the exception that ω_1 was split and assigned to either ω_A or ω_B as appropriate. To assess significance, LRTs were performed on nested models

(supplementary table S3, Supplementary Material online). For branch, sites, and clade models, three initial ω values (0.5, 1, 3) were run to identify and avoid multiple local optima. For branch-sites models, two initial ω values (1.5, 3) were run.

Within Population Polymorphism of CENH3_A and CENH3_B

To evaluate population-level variation of *CENH3_A* and *CENH3_B*, we obtained genomic data originally generated by Flagel et al. (2014) for ten resequenced inbred lines from the IM population. *CENH3* sequences were inferred and aligned as described above. We treated sequences from each paralog separately and calculated the number of haplotypes, haplotype diversity, π, θ_W, Tajima's *D* (Tajima 1989), Fu and Li's *D**, and Fu and Li's *F** (Fu and Li 1993) in DnaSP v5.10.1 (Librado and Rozas 2009). We compared patterns of summary statistics in A versus B by generating 95% CI using 1,000 coalescent simulations in DnaSP. CIs were also used to determine statistical significance of *D*, *D**, and *F**. We performed coalescent simulations given θ and segregating sites, but only report data with θ as results were biologically similar both ways. HKA tests were performed with *M. dentilobus* as an outgroup using the direct mode in DnaSP (Hudson et al. 1987).

Gene Expression of CENH3_A and CENH3_B

To compare gene expression among *CENH3* paralogs, we obtained RNA-Seq data from leaf tissue of progeny of damaged and control *M. guttatus* individuals collected in Colicchio et al. (2015). A two-way analysis of variance was applied to assess the effect of paralog (*CENH3_A*/*CENH3_B*) and treatment (damaged/control) on variation in the number of raw read counts mapped to each gene. For these analyses, paralog was nested within subject to pair read counts by individual. Analyses were performed in R version 3.1.2 (R Core Team 2014).

Supplementary Material

Supplementary figure S1, tables S1–S3, and alignment.txt are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors would like to thank Josh Puzey, John Kelly, and John Willis for genome library preparation and sequencing, M. Streisfeld and Y. Yuan for access to de novo assembled genomes, A. Sweigart for *M. tilingii* genome sequence, and J. Colicchio and J.Kelly for RNA-Seq data. This work was supported by the National Science Foundation (NSF DEB-0846089 to L.F).

References

Allshire RC, Karpen GH. 2008. Epigenetic regulation of centromeric chromatin: old dogs, new tricks? Nat Rev Genet. 9:923–937.

Beardsley PM, Schoenig SE, Whittall JB, Olmstead RG. 2004. Patterns of evolution in western North American *Mimulus* (Phrymaceae). *Am J Bot.* 91:474–489.

Beardsley PM, Yen A, Olmstead RG. 2003. AFLP phylogeny of *Mimulus* section Erythranthe and the evolution of hummingbird pollination. *Evolution* 57:1397–1410.

Bielawski JP, Yang Z. 2001. Positive and negative selection in the *DAZ* gene family. *Mol Biol Evol.* 18:523–529.

Bielawski JP, Yang Z. 2004. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. *J Mol Evol.* 59:121–132.

Birchler JA, Dawe RK, Doebley JF. 2003. Marcus Rhoades, preferential segregation and meiotic drive. *Genetics* 164:835–841.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120.

Brar GA, Amon A. 2008. Emerging roles for centromeres in meiosis I chromosome segregation. *Nat Rev Genet.* 9:899–910.

Burt A, Trivers R. 2008. Genes in conflict. Cambridge (MA): Belknap Press.

Chmátal L, Gabriel SI, Mitsainas GP, Martínez-Vargas J, Ventura J, Searle JB, Schultz RM, Lampson MA. 2014. Centromere strength provides the cell biological basis for meiotic drive and karyotype evolution in mice. *Curr Biol.* 24:2295–2300.

Clarke TH. 2012. Doubling of chromosome number in *Mimulus* genus not caused by polyploidy. Chapter 3. [Ph.D. Thesis]. University of North Carolina at Chapel Hill.

Colicchio JM, Monnahan PJ, Kelly JK, Hileman LC. 2015. Gene expression plasticity resulting from parental leaf damage in *Mimulus guttatus*. *New Phytol.* 205:894-906.

Cooper JL, Henikoff S. 2004. Adaptive evolution of the histone fold domain in centromeric histones. *Mol Biol Evol.* 21:1712–1718.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods.* 9:772.

Drinnenberg IA, deYoung D, Henikoff S, Malik HS. 2014. Recurrent loss of *CenH3* is associated with independent transitions to holocentricity in insects. *Elife* 3:e03676.

Elde NC, Roach KC, Yao M-C, Malik HS. 2011. Absence of positive selection on centromeric histones in Tetrahymena suggests unsuppressed centromere-drive in lineages lacking male meiosis. *J Mol Evol.* 72:510–520.

Fishman L, Kelly JK. 2015. Centromere-associated meiotic drive and female fitness variation in *Mimulus*. *Evolution* 69:1208–1218.

Fishman L, Saunders A. 2008. Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers. *Science* 322:1559–1562.

Fishman L, Willis JH. 2005. A novel meiotic drive locus almost completely distorts segregation in *Mimulus* (monkeyflower) hybrids. *Genetics* 169:347–353.

Fishman L, Willis JH, Wu CA, Lee YW. 2014. Comparative linkage maps suggest that fission, not polyploidy, underlies near-doubling of chromosome number within monkeyflowers (*Mimulus*; Phrymaceae). *Heredity* 112:565-568.

Flagel LE, Willis JH, Vision TJ. 2014. The standing pool of genomic structural variation in a natural population of *Mimulus guttatus*. *Genome Biol Evol.* 6:53-64.

Fu Y-X, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* 133:693-709.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704.

Hahn MW. 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. *J Hered.* 100:605–617.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol.* 22:160-174.

Hellsten U, Wright KM, Jenkins J, Shu S, Yuan Y, Wessler SR, Schmutz J, Willis JH, Rokhsar DS. 2013. Fine-scale variation in meiotic recombination in *Mimulus* inferred from population shotgun sequencing. *Proc Natl Acad Sci U S A.* 110:19478–19482.

Henikoff S, Ahmad K, Malik HS. 2001. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* 293:1098–1102.

Henikoff S, Malik HS. 2002. Centromeres: selfish drivers. *Nature* 417:227–227.
Heun P, Erhardt S, Blower MD, Weiss S, Skora AD, Karpen GH. 2006. Mislocalization of the *Drosophila* centromere-specific histone *CID* promotes formation of functional ectopic kinetochores. *Dev Cell.* 10:303–315.
Hirsch CD, Wu Y, Yan H, Jiang J. 2009. Lineage-specific adaptive evolution of the centromeric protein *CENH3* in diploid and allotetraploid *Oryza* species. *Mol Biol Evol.* 26:2877–2885.
Hudson RR, Kreitman M, Aguade. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116.
Howman EV, Fowler KJ, Newson AJ, Redward S, MacDonald AC, Kalitsis P, Choo KA. 2000. Early disruption of centromeric chromatin organization in centromere protein A (*CenPA*) null mice. *Proc Natl Acad Sci U S A.* 97:1148–1153.
Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. *Proc R Soc Lond B Biol Sci.* 256:119-124.
Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet.* 11:97–108.
Kawabe A, Nasuda S. 2004. Structure and genomic organization of centromeric repeats in *Arabidopsis* species. *Mol Genet Genomics.* 272:593–602.
Kawabe A, Nasuda S, Charlesworth D. 2006. Duplication of centromeric histone H3 (*HTR12*) gene in *Arabidopsis halleri* and *A. lyrata*, plant species with multiple centromeric satellite sequences. *Genetics* 174:2021–2032.
Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947–2948.
Lee H-R, Zhang W, Langdon T, Jin W, Yan H, Cheng Z, Jiang J. 2005. Chromatin immunoprecipitation cloning reveals rapid evolutionary patterns of centromeric DNA in *Oryza* species. *Proc Natl Acad Sci U S A.* 102:11793–11798.
Lermontova I, Koroleva O, Rutten T, Fuchs J, Schubert V, Moraes I, Koszegi D, Schubert I. 2011. Knockdown of *CENH3* in *Arabidopsis* reduces mitotic divisions and causes sterility by disturbed meiotic chromosome segregation. *Plant J.* 68:40–50.
Lermontova I, Schubert V, Fuchs J, Klatte S, Macas J, Schubert I. 2006. Loading of *Arabidopsis* centromeric histone *CENH3* occurs mainly during G2 and requires the presence of the histone fold domain. *Plant Cell* 18:2443–2451.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451-1452.
Maheshwari A, Tan EH, West A, Franklin FCF, Comain L, Chan SWL. 2015. Naturally occurring differences in *CENH3* affect chromosome segregation in zygotic mitosis in hybrids. *PLoS Genet.* 11(1):e1004970
Malik HS, Bayes JJ. 2006. Genetic conflicts during meiosis and the evolutionary origins of centromere complexity. *Biochem Soc Trans.* 34:569–573.
Malik HS, Henikoff S. 2001. Adaptive evolution of *Cid*, a centromere-specific histone in *Drosophila*. *Genetics* 157:1293–1298.
Malik HS, Henikoff S. 2002. Conflict begets complexity: the evolution of centromeres. *Curr Opin Genet Dev.* 12:711–718.
Malik HS, Henikoff S. 2009. Major evolutionary transitions in centromere complexity. *Cell* 138:1067–1082.
Malik HS, Vermaak D, Henikoff S. 2002. Recurrent evolution of DNA-binding motifs in the *Drosophila* centromeric histone. *Proc Natl Acad Sci U S A.* 99:1449–1454.
Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D, et al. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol.* 14:R10.
Monen J, Maddox PS, Hyndman F, Oegema K, Desai A. 2005. Differential role of *CENP-A* in the segregation of holocentric *C. elegans* chromosomes during meiosis and mitosis. *Nat Cell Biol.* 7:1248–1255.
Moore LL, Roth MB. 2001. HCP-4, a *CENP-C*-like protein in *Caenorhabditis elegans*, is required for resolution of sister centromeres. *J Cell Biol.* 153:1199–1208.
Moraes ICR, Lermontova I, Schubert I. 2010. Recognition of *A. thaliana* centromeres by heterologous *CENH3* requires high similarity to the endogenous protein. *Plant Mol Biol.* 75:253–261.
Nasuda S, Hudakova S, Schubert I, Houben A, Endo TR. 2005. Stable barley chromosomes without centromeric repeats. *Proc Natl Acad Sci U S A.* 102:9842–9847.
Neumann P, Navrátilová A, Schroeder-Reiter E, Koblížková A, Steinbauerová V, Chocholová E, Novák P, Wanner G, Macas J. 2012. Stretching the rules: monocentric chromosomes with multiple centromere domains. *PLoS Genet.* 8:e1002777.
Neumann P, Pavlíková Z, Koblížková A, Fuková I, Jedličková V, Novák P, Macas J. 2015. Centromeres off the hook: massive changes in centromere size and structure following duplication of *CenH3* gene in *Fabeae* species. *Mol Biol Evol.* 1862–1879.
Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929–936.
Oegema K, Desai A, Rybina S, Kirkham M, Hyman AA. 2001. Functional analysis of kinetochore assembly in *Caenorhabditis elegans*. *J Cell Biol.* 153:1209–1226.
Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag
Pardo-Manuel de Villena F, Sapienza C. 2001a. Nonrandom segregation during meiosis: the unfairness of females. *Mamm Genome.* 12:331–339.
Pardo-Manuel de Villena F, Sapienza C. 2001b. Female meiosis drives karyotypic evolution in mammals. *Genetics* 159:1179–1189.
R Core Team. 2014. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. Available from: http://www.R-project.org/.
Rice WR. 2013. Nothing in genetics makes sense except in light of genomic conflict. *Annu Rev Ecol Evol Syst.* 44:217–237.
Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol.* 61:539–542.
Rooney AP, Piontkivska H, Nei M. 2002. Molecular evolution of the nontandemly repeated genes of the histone 3 multigene family. *Mol Biol Evol.* 19:68–75.
Sanei M, Pickering R, Kumke K, Nasuda S, Houben A. 2011. 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–E505.
Schueler MG, Swanson W, Thomas PJ, Comparative Sequencing Program NISC, Green ED. 2010. Adaptive evolution of foundation kinetochore proteins in primates. *Mol Biol Evol.* 27:1585–1597.
Swanson WJ, Nielsen R, Yang Q. 2003. Pervasive adaptive evolution in mammalian fertilization proteins. 20:18–20. *Mol Bol Evol.* 20:18-20.
Tachiwana H, Kagaw W, Shiga T, Osakabe A, Miya Y, Saito K, Hayashi-Takanaka Y, Oda T, Sato M, Park S-Y, et al. 2011. Crystal structure of the human centromeric nucleosome containing *CENP-A*. *Nature* 476: 232-235.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585-595.
Talbert PB, Bayes JJ, Henikoff S. 2009. Evolution of centromeres and kinetochores: a two-part fugue. In: De Wulf P, Earnshaw WC, editors. The kinetochore. Berlin: Springer. p. 193–229.
Talbert PB, Bryson TD, Henikoff S. 2004. Adaptive evolution of centromere proteins in plants and animals. *J Biol.* 3:1–17.
Talbert PB, Masuelli R, Tyagi AP, Comai L, Henikoff S. 2002. Centromeric localization and adaptive evolution of an *Arabidopsis* histone H3 variant. *Plant Cell* 14:1053–1066.
Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* 28:2731–2739.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673–4680.

Tomonaga T, Matsushita K, Yamaguchi S, Oohashi T, Shimada H, Ochiai T, Yoda K, Nomura F. 2003. Overexpression and mistargeting of centromere protein-A in human primary colorectal cancer. *Cancer Res.* 63:3511–3516.

Vermaak D, Hayden HS, Henikoff S. 2002. Centromere targeting element within the histone fold domain of *Cid*. *Mol Cell Biol.* 22:7553–7561.

Wang L, Zeng Z, Zhang W, Jiang J. 2014. Three potato centromeres are associated with distinct haplotypes with or without megabase-sized satellite repeat arrays. *Genetics* 196:397–401.

Weadick CJ, Chang BSW. 2012. An improved likelihood ratio test for detecting site-specific functional divergence among clades of protein-coding genes. *Mol Biol Evol.* 29:1297–1300.

Weaver BAA, Cleveland DW. 2007. Aneuploidy: instigator and inhibitor of tumorigenesis. *Cancer Res.* 67:10103–10105.

Wong WSW, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* 168:1041–1051.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol.* 15:568–573.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol.* 19:908–917.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.

Yang Z, Swanson WJ. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol.* 19:49–57.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol.* 22:1107–1118.

Yoshida I, Sugiura W, Shibata J, Ren F, Yang Z, Tanaka H. 2011. Change of positive selection pressure on HIV-1 envelope gene inferred by early and recent samples. *PLoS One* 6:e18630.

Yuan J, Guo X, Hu J, Lv A, Han F. 2014. Characterization of the *CENH3* genes and their roles in wheat evolution. *New Phytol.* 206:839-851.

Zedek F, Bureš P. 2012. Evidence for centromere drive in the holocentric chromosomes of *Caenorhabditis*. *PLoS One* 7:e30496.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol.* 22:2472–2479.

Zhong CX, Marshall JB, Topp C, Mroczek R, Kato A, Nagaki K, Birchler JA, Jiang J, Dawe RK. 2002. Centromeric retroelements and satellites interact with maize kinetochore protein *CENH3*. *Plant Cell* 14:2825–2836.

Zimin AV, Marcais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. *Bioinformatics* 29:2669–2677.