Lineage-Specific Adaptive Evolution of the Centromeric Protein CENH3 in Diploid and Allotetraploid Oryza Species

Cory D. Hirsch, Yufeng Wu, Huihuang Yan, and Jiming Jiang

Department of Horticulture, University of Wisconsin–Madison

Centromeres in eukaryotic species are defined by the presence of a centromere-specific histone H3 variant, CENH3. CENH3 plays a key role in recruiting other centromeric proteins; thus, it is the central component in kinetochore formation and centromere function. The CENH3 proteins in several plant and animal species were found to be under positive selection, which was hypothesized to respond to the rapid changing of the repetitive DNA sequences associated with the centromeres. Here, we report the expression and evolution of the CenH3 genes in two allotetraploid rice species as well as their representative diploid progenitor species. Both copies of the CenH3 genes were transcribed in the two allotetraploid species and showed a nonpreferential expression pattern. Contrasting positive and stabilizing selection of the CenH3 genes was associated with different diploid Oryza species. This lineage-specific adaptive evolution of CENH3 was maintained in the two allotetraploid species. Thus, we demonstrate that the allopolyploidization events did not alter the expression or evolutionary patterns of the CenH3 genes in the Oryza species.

Introduction

Centromeres govern chromosome segregation and transmission by serving as the docking site for kinetochore assembly. In most higher eukaryotes, centromeres are composed of megabase-sized arrays of satellite repeats (Henikoff et al. 2001; Jiang et al. 2003). Like any typical satellite repeat, centromeric satellite DNA evolves rapidly. Expansion and contraction of satellite repeat arrays caused by unequal crossovers and/or other unknown mechanisms may result in dramatic differences in the amount of satellite repeats located in different centromeres within the same species (Cheng et al. 2002; Jin et al. 2004). Closely related species may contain distinct centromeric satellite repeats due to the rapid turnover of different repeat families (Lee et al. 2005). Such rapid changes of centromeric DNA sequences are in an enigmatic contrast with the highly conserved and stable function of centromeres.

Centromeric chromatin is defined by the presence of a centromere-specific histone H3 variant, CENH3 (centromere protein A in humans and CID in Drosophila) (Henikoff et al. 2001; Allshire and Karpen 2008). Interestingly, CENH3 proteins have been under positive selection in several plant and animal species, including Drosophila (Malik and Henikoff 2001), Arabidopsis (Talbert et al. 2002), and other Brassicaceae species (Cooper and Henikoff 2004). Positive selection on a protein with a highly restricted and conserved function is remarkable because such proteins are often under stabilizing selection, aiding the retention of their utility. The positive selection of CENH3 proteins was proposed to adapt to the rapid changes of centromeric DNA sequences (Malik and Henikoff 2001). Female meiosis is asymmetric, and any dramatic changes of the centromeric satellite that binds CENH3 may break the balance of the random inclusion of the female meiotic products into the next generation. An expanded centromere may provide more microtubule attachment sites and allow preferential inclusion in fertilization (Fishman and Saunders 2008). Such a centromere drive during female meiosis could also lead to increased nondisjunction. Thus, positive selection on CENH3 may counteract such a centromere drive and restore meiotic balance (Henikoff et al. 2001; Malik and Henikoff 2001).

Allopolyploidy formation, which brings two or more distinct genomes, thus multiple CenH3 genes, into the same genetic environment, is a shock to the newly merged genomes, effecting both genes and repetitive DNA sequences (Osborn et al. 2003; Lim et al. 2007). Extensive expression studies of homoeologous gene pairs in both natural and synthetic tetraploid cotton revealed frequent silencing or biased expression of one of the two homoeologous genes (Adams et al. 2003; Flagel et al. 2008). Dramatic turnover of repetitive DNA sequences after a polyploidization event has been well documented in the literature (Kovarik et al. 2005; Lim et al. 2007). We are interested in whether multiple CenH3 genes can survive in the same genetic background and the evolutionary direction of CENH3 in a polyploid setting. We cloned and sequenced the CenH3 genes from two allopolyploid rice species and their three representative diploid progenitor species. We found that the CENH3 proteins are under either positive or stabilizing selection in a lineage-specific manner. The lineage specificity and maintenance of rice CENH3 adaptive evolution is discussed in the context of centromere repeats and polyploidy genome dynamics.

Material and Methods

Cloning and Sequencing of the CenH3 Genes in Oryza Species

Five wild Oryza species, consisting of three diploid and two tetraploid species, were used in this study (table 1). Genomic DNA was isolated from leaf tissue from the five Oryza species. CenH3 was amplified from genomic DNA, gel purified using a QIAquick Gel Extraction Kit (Qiagen, Valencia, CA), and cloned using pGEM-T Easy vectors (Promega, Madison, WI). The primers used to amplify the CenH3 genes are listed in supplementary table 1, Supplemental Material online. A specific primer to the N-terminal tail domain of each CenH3 gene was used in concert with a conserved universal primer to the histone fold domain (HFD) to amplify the genomic sequences. To sequence the cloned CenH3 genes, both strands of
the genes were sequenced using ABI Big Dye sequencing on at least three clones. Additional primers were designed to complete the sequencing of the cloned genes (supplementary table 1, Supplementary Material online).

Total RNA was isolated from leaf tissue from the five wild rice species with any residual DNA being removed with TURBO DNA-free (Ambion, Foster City, CA). The RNA was reverse transcribed to cDNA using SuperScript III RT (Invitrogen, Carlsbad, CA). *CenH3* was amplified from the cDNA and cloned using pGEM-T Easy vectors (Promega). The primers used to amplify *CenH3* from the cDNA were designed based on the already obtained genomic DNA sequence. Specific primers to each genome were designed to the N-terminal tail domain and were used together with a conserved primer designed to the HFD to amplify *CenH3* from the cDNA (supplementary table 1, Supplementary Material online). Again, both strands of three clones were sequenced using ABI Big Dye sequencing. To ensure that the sequenced portion would cover both the start and the stop codons, 5' and 3' RACE were applied using the GeneRacer Kit with SuperScript III RT, and amplified products were cloned using TOPO TA Cloning (Invitrogen) according to the manufacturer's instructions. The primers used in the 5' and 3' RACE reactions for each species are listed in supplementary table 1, Supplementary Material online. The relative expression of the two *CenH3* genes in each tetraploid was carried out using conserved primers that would amplify both copies of *CenH3* together (supplementary table 1, Supplementary Material online). The cloning and sequencing was carried out as described above.

Sequence Alignments and Phylogenetic Analysis

Amino acid and coding sequence alignments were performed using clustal (Larkin et al. 2007) and refined manually. Phylogenetic trees based on the coding sequence were constructed using the Neighbor-Joining method applied through MEGA4 using default parameters (Tamura et al. 2007). To construct the trees, all positions containing gaps and missing data were removed from the alignment. This left 444 nucleotide sites in the final alignment to construct the phylogenetic trees. The viewing and editing of the trees was completed in TreeViewX Version 0.5.0 (Page 1996).

K-estimator (Comeron 1999) was used to calculate the rate of synonymous ($K_s$) and nonsynonymous ($K_a$) substitutions in comparisons of the *CenH3* genes from different *Oryza* genomes. Gaps in pairwise amino acid alignments were used to remove noninformative sites from the coding sequences. The removal of gaps resulted in full-length pairwise alignments consisting of a range of 152–166 codons for analysis. When these alignments were sectioned into the N-terminal tail and HFDs for separate analysis, the range of codons analyzed was 57–71 codons and 94–96 codons, respectively. In addition, K-estimator was used to test the significance of selection. To do this, simulations were run where the $K_a$ value is on average set equal to the value estimated for $K_s$ from the dataset. The simulations take into account the divergence values ($K_a$ and $K_s$), the number of codons, the transition to transversion substitution ratio, and the amino acid composition including the G + C content at the third position of codons for the analyzed dataset. Thus, a null distribution for the rate of $K_a/K_s$ is obtained for a condition when the null hypothesis for $K_a/K_s = 1$, allowing likely probabilities to be calculated. To detect the adaptive evolution among lineages, we used the codeml program (PAML package, Version 4) to estimate the $K_a/K_s$ ratios with a free-ratio model (model = 1, codon frequencies = F3 x 4) (Yang 1997).

Results and Discussion

The Genomic and cDNA Sequences of the *CenH3* Genes in Diploid and Tetraploid *Oryza* Species

We cloned and sequenced the *CenH3* genes from five wild *Oryza* species, including two allotetraploid species, *Oryza minuta* (BBCC, 2$n$ = 4$x$ = 48) and *Oryza alta* (CCDD, 2$n$ = 4$x$ = 48) and three diploid species, *Oryza punctata* (BB, 2$n$ = 2$x$ = 24), *Oryza rhizomatis* (CC, 2$n$ = 2$x$ = 24), and *Oryza australiensis* (EE, 2$n$ = 2$x$ = 24). Each tetraploid species contained two distinct copies of the *CenH3* gene, whereas a single *CenH3* gene was isolated from the diploid species (fig. 1). The *CenH3* genes of

| Table 1 *Oryza* Species Used for Cloning and Sequencing of the *CenH3* Genes |
|-----------------|--------|-----------------|
| Species         | 2$n$   | Genome         | Accession                  |
| *Oryza punctata*| 24     | BB             | A13 (original accession unknown) |
| *Oryza minuta*  | 48     | BBCC           | Acc 101386 (W0045)          |
| *Oryza rhizomatis* | 24     | CC             | PI 105440                   |
| *Oryza alta*    | 48     | CCDD           | Acc 105143                  |
| *Oryza australiensis* | 24     | EE             | Acc 100882 (W0008)          |
the five *Oryza* species were amplified from genomic DNA using polymerase chain reaction (PCR), cloned, and sequenced (see Materials and Methods for details). The lengths of the genomic DNA sequences, from the start to stop codon, which were inferred from alignments between the genomic sequences and cDNA sequences (see below), varied between 1,655 and 2,694 bp among the five species (table 2) (GenBank accessions GQ849328–GQ849334).

We identified two distinct *CenH3* transcripts in both tetraploid species, indicating that both copies of the *CenH3* gene are actively transcribed in the polyploids. We used 5' and 3' RACE to acquire the complete full-length cDNA sequences (GenBank accessions GQ849335–GQ849341). The lengths of the *CenH3* transcripts from the five species (seven genes total) were very similar, ranging from 492 bp to 498 bp (table 2). Based on alignments between the genomic DNA and the cDNA sequences, the seven *CenH3* genes shared the same structure of seven exons and six introns. Variability in length from three of the introns caused the majority of the difference in length of the genomic sequences (supplementary table 2, Supplementary Material online). Sequence analysis suggested that the variation of two of the three introns was likely caused by insertions of Miniature Inverted-Repeat Transposable Elements in the larger introns.

### Table 2

<table>
<thead>
<tr>
<th>Species</th>
<th>Genome</th>
<th>Genomic DNA (bp)</th>
<th>Full-Length CDS (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>O. sativa</em></td>
<td>AA</td>
<td>2163</td>
<td>492</td>
</tr>
<tr>
<td><em>O. punctata</em></td>
<td>BB</td>
<td>2216</td>
<td>498</td>
</tr>
<tr>
<td><em>O. minuta</em></td>
<td>BB</td>
<td>2236</td>
<td>498</td>
</tr>
<tr>
<td><em>O. minuta</em></td>
<td>CC</td>
<td>2694</td>
<td>498</td>
</tr>
<tr>
<td><em>O. rhizomatis</em></td>
<td>CC</td>
<td>2691</td>
<td>498</td>
</tr>
<tr>
<td><em>O. alta</em></td>
<td>CC</td>
<td>2691</td>
<td>498</td>
</tr>
<tr>
<td><em>O. alta</em></td>
<td>DD</td>
<td>1661</td>
<td>492</td>
</tr>
<tr>
<td><em>O. australiensis</em></td>
<td>EE</td>
<td>1655</td>
<td>492</td>
</tr>
</tbody>
</table>

The coding sequences of the *CenH3* genes from the five wild rice species, along with those from cultivated rice (*O. sativa*) and maize (*Zea mays*), were used to construct a phylogenetic tree (fig. 2). The *CenH3* genes from the five wild *Oryza* species were separated into three distinct groups. One of the genes, *CenH3_B* (BBCC) grouped with the *CenH3_B* gene from the diploid *O. punctata* (BB). This group is supported by a bootstrap value of 100. The second gene, *CenH3_C*, from *O. minuta* formed another group with the *CenH3_C* gene from the diploid *O. rhizomatis* (CC), supported by a bootstrap value of 69; this group is closely related, bootstrap value of 96, to *CenH3_C* from *O. alta* (CCDD). These results are in agreement with the conviction that *O. minuta* contains the B genome and both *O. minuta* and *O. alta* contain the C genome (Ge et al. 1999). The *CenH3* genes from the C genome form a clade with the *CenH3_A* gene from *O. sativa* (AA genome), with a supporting bootstrap value of 54. These results, based on *CenH3* gene sequences, support that the wild *Oryza* species with the CC genome are the closest relatives to cultivated rice.

The second distinct gene, *CenH3_D*, from *O. alta* formed a group with the *CenH3_E* gene from *O. australiensis* (EE), with a supporting bootstrap value of 100. Our result supports an earlier inference that *O. australiensis* is the closest extant relative to the unidentified DD diploid *Oryza* species (Bao and Ge 2004).

### Relative Expression of the *CenH3* Genes in Tetraploid *Oryza* Species

A sequencing-based approach was used to examine the relative transcript abundance of the two *CenH3* genes in each tetraploid *Oryza* species. Briefly, fully conserved primers were designed to amplify both *CenH3* transcripts unbiased. The PCR products were cloned, and clones were randomly selected for sequencing. The ratio of sequenced
clones from the two distinct \textit{CenH3} genes was used as a measurement of relative expression.

Two independent experiments were performed to examine the relative expression of the two \textit{CenH3} genes, \textit{CenH3} \textit{B} \textit{w} and \textit{CenH3} \textit{C} \textit{w}, respectively, in tetraploid \textit{O. minuta}. In the two trials, \textit{CenH3} \textit{B} \textit{w} expression was estimated at 46% and 58%, respectively, whereas \textit{CenH3} \textit{C} \textit{w} expression was 54% and 42% of the total \textit{CenH3} transcripts. Neither of the replications showed significant evidence rejecting the null hypothesis of equal expression of the two \textit{CenH3} genes.

A chi-square test with a null hypothesis of 1:1 expression was used to obtain a \textit{P}-value for each replicate (table 3). Two of the trials showed equal expression of the \textit{CenH3} genes, whereas one trial indicated a higher expression level of \textit{CenH3} \textit{D} \textit{w} than \textit{CenH3} \textit{C} \textit{w}.

It is estimated that up to 70% of the angiosperms are polyploids (Masterson 1994). In addition, whole-genome duplication (WGD) is associated with many plant species that were traditionally classified as diploids, such as \textit{Arabidopsis thaliana} and rice (Simillion et al. 2002; Bowers et al. 2003; Yu et al. 2005). Thus, these diploid plant species are polyploids. Elimination, divergence, or partitioning of function of homeologous genes in polyploids and duplicated genes in polyploids has been an attractive research subject in recent years (Blanc and Wolfe 2004; Maere et al. 2005; Doyle et al. 2008; Ha et al. 2009). It is interesting to note that \textit{A. thaliana}, rice, and maize all contain a single copy of the \textit{CenH3} gene (Talbert et al. 2002; Zhong et al. 2002; Nagaki et al. 2004), although WGD has been well demonstrated in all three species. Thus, the duplicated copies of the \textit{CenH3} genes in these species have been lost since the WGD event.

Our results show that both copies of the \textit{CenH3} genes in the two rice allotetraploids are actively transcribed. The two \textit{CenH3} genes showed a similar level of transcript abundance in both \textit{O. minuta} (BBCC) and \textit{O. alta} (CCDD). A similar result of \textit{CenH3} gene expression has recently been reported in tobacco, an allotetraploid species (Nagaki et al. 2009). Differential expression of homeologous genes is common in both synthetic and natural allopolyploids. The differentially expressed homeologs accounted for \textasciitilde{}10% to 30% of the genes analyzed in recent genome-wide studies (Wang et al. 2006; Hovav et al. 2008; Pumphrey et al. 2009). Our data show that the allopolyploidization events appear to have little effect on the expression of the \textit{CenH3} genes in the two tetraploid rice species.

Divergence of the \textit{Oryza} \textit{CENH3} Proteins

We investigated the divergence of the rice \textit{CENH3} proteins by analyzing the rate of synonymous (\textit{K}_s) and non-synonymous (\textit{K}_a) nucleotide substitutions between the \textit{CenH3} genes. The amino acid sequences of nine \textit{CENH3} proteins, including those from \textit{O. sativa} and \textit{Z. mays}, were aligned (fig. 3). Pairwise alignments of the corresponding cDNA sequences were used to estimate the rate of \textit{K}_a and \textit{K}_s nucleotide substitutions. An \textit{\omega} ratio (\textit{\omega} = \textit{K}_a/\textit{K}_s > 1) indicates adaptive evolution or positive selection, whereas an \textit{\omega} 1 indicates microevolution or neutral evolution.

![Fig. 3.—\textit{CENH3} amino acid alignment used to estimate \textit{K}_a/\textit{K}_s ratios for pairwise comparisons of the \textit{CenH3} gene coding sequence. Identical residues in all species are shaded dark gray, whereas residues identical in just the \textit{CenH3} genes among the \textit{Oryza} species and not maize are shaded lighter gray.](https://academic.oup.com/mbe/article/26/12/2877/1539376 by guest on 11 March 2022)
ratio <1 indicates stabilizing selection. We first examined the divergence of the full-length *Oryza* CENH3 proteins. The eight CENH3 proteins from the six *Oryza* species contain either 164 or 166 amino acids, whereas maize CENH3 contains 157 amino acids. Of the 36 pairwise comparisons, 11 had $\omega$ values >1, with four showing significance (table 4). The four comparisons showing significant positive selection were between CENH3_C a from *O. alta* and CENH3_Cm from *O. minuta* with CENH3_B m from *O. minuta* and CENH3_B from *O. punctata*. In contrast, significant stabilizing selection was found in 22 of the 36 pairwise comparisons (table 4). Every pairwise comparison that included CENH3_D a from *O. alta*, CENH3_E from *O. australiensis* or CENH3 from maize revealed significant stabilizing selection.

The CENH3 protein can be divided into the HFD and the N-terminal domain (Henikoff et al. 2000). Adaptive evolution of CENH3 (CID in Drosophila) was predominantly located within the N-terminal domain (Malik and Henikoff 2001; Malik et al. 2002). The N tail of rice CENH3 proteins vary between 68 and 71 amino acids (fig. 3). Six of the 36 pairwise comparisons resulted in $\omega$ values indicating significant positive selection in the N-terminal domain (table 5). Interestingly, all six comparisons included CENH3 associated with either the B or the C genome. The comparisons of CENH3_C a and CENH3_C m with CENH3_B m also showed signs of positive selection by having only nonsynonymous substitutions. Although there were no synonymous substitutions ($K_s=0$), of the 73 codons analyzed, 7 (CENH3_C a) or 8 (CENH3_C m) codons have undergone...

### Table 4

$K_s/K_a$ and $\omega$ for the Full-Length CENH3 of Different *Oryza* Species and Maize

<table>
<thead>
<tr>
<th>CENH3_D</th>
<th>CENH3_E</th>
<th>CENH3_maize</th>
<th>CENH3_B</th>
<th>CENH3_C</th>
<th>CENH3_A</th>
<th>CENH3_B</th>
<th>CENH3_C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.043/0.226</td>
<td>0.043/0.251</td>
<td>0.230/0.917</td>
<td>0.027/0.006</td>
<td>0.003/0.000</td>
<td>0.025/0.013</td>
<td>0.031/0.013</td>
<td>0.003/0.010</td>
</tr>
<tr>
<td><strong>0.15</strong></td>
<td><strong>0.17</strong></td>
<td><strong>0.25</strong></td>
<td><strong>4.31</strong></td>
<td><strong>--</strong></td>
<td><strong>1.93</strong></td>
<td><strong>2.36</strong></td>
<td><strong>0.25</strong></td>
</tr>
<tr>
<td><strong>0.38</strong></td>
<td><strong>0.38</strong></td>
<td><strong>0.22</strong></td>
<td><strong>0.18</strong></td>
<td><strong>0.16</strong></td>
<td><strong>0.16</strong></td>
<td><strong>0.19</strong></td>
<td><strong>0.15</strong></td>
</tr>
<tr>
<td>CENH3_E</td>
<td></td>
<td><strong>0.22</strong></td>
<td><strong>0.22</strong></td>
<td><strong>0.18</strong></td>
<td><strong>0.21</strong></td>
<td><strong>0.23</strong></td>
<td><strong>0.17</strong></td>
</tr>
<tr>
<td>CENH3_B m</td>
<td><strong>0.22</strong></td>
<td><strong>0.22</strong></td>
<td><strong>0.22</strong></td>
<td><strong>0.22</strong></td>
<td><strong>0.22</strong></td>
<td><strong>0.22</strong></td>
<td><strong>0.22</strong></td>
</tr>
<tr>
<td>CENH3_C m</td>
<td><strong>0.24</strong></td>
<td><strong>0.25</strong></td>
<td><strong>0.24</strong></td>
<td><strong>0.24</strong></td>
<td><strong>0.24</strong></td>
<td><strong>0.24</strong></td>
<td><strong>0.24</strong></td>
</tr>
<tr>
<td>CENH3_B</td>
<td><strong>0.23</strong></td>
<td><strong>0.23</strong></td>
<td><strong>0.23</strong></td>
<td><strong>0.23</strong></td>
<td><strong>0.23</strong></td>
<td><strong>0.23</strong></td>
<td><strong>0.23</strong></td>
</tr>
<tr>
<td>CENH3_A</td>
<td><strong>0.22</strong></td>
<td><strong>0.22</strong></td>
<td><strong>0.22</strong></td>
<td><strong>0.22</strong></td>
<td><strong>0.22</strong></td>
<td><strong>0.22</strong></td>
<td><strong>0.22</strong></td>
</tr>
</tbody>
</table>

**Note.**—Underlined numbers indicate $\omega$ greater than 1. Bold numbers indicate $\omega$ value.

$^a$ Significant at the 99% level.

$^b$ Significant at the 95% level.

### Table 5

$K_s/K_a$ and $\omega$ for the N-terminal Tail Domain of CENH3 of Different *Oryza* Species and Maize

<table>
<thead>
<tr>
<th>CENH3_D</th>
<th>CENH3_E</th>
<th>CENH3_maize</th>
<th>CENH3_B</th>
<th>CENH3_C</th>
<th>CENH3_A</th>
<th>CENH3_B</th>
<th>CENH3_C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.050/0.090</td>
<td>0.074/0.097</td>
<td>0.337/0.897</td>
<td>0.045/0.000</td>
<td>0.006/0.000</td>
<td>0.048/0.016</td>
<td>0.052/0.015</td>
<td>0.006/0.022</td>
</tr>
<tr>
<td><strong>0.56</strong></td>
<td><strong>0.77</strong></td>
<td><strong>0.38</strong></td>
<td><strong>--</strong></td>
<td><strong>--</strong></td>
<td><strong>3.05</strong></td>
<td><strong>3.42</strong></td>
<td><strong>0.27</strong></td>
</tr>
<tr>
<td>CENH3_D</td>
<td><strong>0.58</strong></td>
<td><strong>0.35</strong></td>
<td><strong>0.87</strong></td>
<td><strong>0.63</strong></td>
<td><strong>0.76</strong></td>
<td><strong>0.78</strong></td>
<td><strong>0.49</strong></td>
</tr>
<tr>
<td>CENH3_E</td>
<td><strong>0.45</strong></td>
<td><strong>0.386/0.853</strong></td>
<td><strong>0.119/0.010</strong></td>
<td><strong>0.081/0.097</strong></td>
<td><strong>0.112/0.102</strong></td>
<td><strong>0.128/0.119</strong></td>
<td><strong>0.081/0.123</strong></td>
</tr>
<tr>
<td>CENH3_C m</td>
<td><strong>0.33</strong></td>
<td><strong>0.35</strong></td>
<td><strong>0.37</strong></td>
<td><strong>0.37</strong></td>
<td><strong>0.37</strong></td>
<td><strong>0.37</strong></td>
<td><strong>0.37</strong></td>
</tr>
<tr>
<td>CENH3_B m</td>
<td><strong>0.33</strong></td>
<td><strong>0.35</strong></td>
<td><strong>0.37</strong></td>
<td><strong>0.37</strong></td>
<td><strong>0.37</strong></td>
<td><strong>0.37</strong></td>
<td><strong>0.37</strong></td>
</tr>
<tr>
<td>CENH3_B</td>
<td><strong>0.34</strong></td>
<td><strong>0.35</strong></td>
<td><strong>0.35</strong></td>
<td><strong>0.35</strong></td>
<td><strong>0.35</strong></td>
<td><strong>0.35</strong></td>
<td><strong>0.35</strong></td>
</tr>
<tr>
<td>CENH3_A</td>
<td><strong>0.34</strong></td>
<td><strong>0.35</strong></td>
<td><strong>0.35</strong></td>
<td><strong>0.35</strong></td>
<td><strong>0.35</strong></td>
<td><strong>0.35</strong></td>
<td><strong>0.35</strong></td>
</tr>
<tr>
<td>CENH3_B</td>
<td><strong>0.34</strong></td>
<td><strong>0.35</strong></td>
<td><strong>0.35</strong></td>
<td><strong>0.35</strong></td>
<td><strong>0.35</strong></td>
<td><strong>0.35</strong></td>
<td><strong>0.35</strong></td>
</tr>
</tbody>
</table>

**Note.**—Underlined numbers indicate $\omega$ greater than 1. Bold numbers indicate $\omega$ value.

$^a$ Significant at the 99% level.

$^b$ Seven of 73 codons nonsynonymous.

$^c$ Eight of 73 codons nonsynonymous.

$^d$ Significant at the 95% level.
nonsynonymous substitutions, which supports positive selection. Additionally, six comparisons also had an \( \omega \) value greater than 1, though not significant. In contrast, comparisons between maize CENH3 with any of the eight rice CENH3 proteins showed significant signs of stabilizing selection in the N-terminal domain (table 5).

The HFDs of CENH3 from the Brassicaceae family and Drosophila were also found to be under adaptive evolution (Malik and Henikoff 2001; Malik et al. 2002; Cooper and Henikoff 2004). The HFD of rice CENH3 proteins are 95 or 96 amino acids in length. When \( \omega \) values were calculated for the HFD of the 36 pairwise comparisons, six of the comparisons resulted in an \( \omega \) value greater than 1 but none were significant. Stabilizing selection in the HFD was found in 22 of the 36 pairwise comparisons (table 6). Markedly, all the comparisons showing stabilizing selection included CENH3/E/CENH3/D or the CENH3 from maize for at least one of the two genes in the comparison. The loop 1 region within the HFD of CENH3 is of particular importance due to it being necessary and sufficient for CENH3 localization to the centromeres in Drosophila (Vermaak et al. 2002). Previous research in Drosophila and species in the Brassicaceae family also revealed adaptive evolution in the loop 1 region of the HFD (Malik and Henikoff 2001; Cooper and Henikoff 2004). However, there were no amino acid changes in the loop 1 region of the rice CENH3 proteins and no significant adaptive evolution was found in comparisons with the loop 1 region of the maize CENH3 protein (results not shown).

Talbert et al. (2004) showed that another centromeric protein, CENP-C, is under positive selection in animal and plant species in which CENH3 is associated with stabilizing selection. In mammalian species, the DNA-binding capacity of CENP-C has been well demonstrated (Sugimoto et al. 1994; Yang et al. 1996). CENP-C in humans also binds RNA processed from human centromeric DNA (Wong et al. 2007). The positive selection of CENP-C may replace the role of CENH3 to suppress meiotic drive of centromeres during female meiosis (Talbert et al. 2004). Thus, it will be interesting to know if CENP-C in plants also binds DNA and/or RNA and whether CENP-C from the rice D/E genome has been under a strong positive selection to complement the role of CENH3.

### Lineage Specificity of Adaptive Evolution of the CENH3 Proteins in Rice

It was interesting to note that every pairwise comparison indicating positive selection, be it for the full-length CENH3 or only its N-terminal tail domain, included at least one CENH3 from the B or C genome of rice. To take a closer look at this discovery, phylogenetic trees were constructed with \( K_d \) and \( K_s \) estimations analyzed at every branch. This allowed adaptive evolution to be studied based on phylogenetic positions to investigate if certain rice CENH3 lineages are subjected to positive selection, whereas others are not.

When \( K_d \) and \( K_s \) values were estimated for each branch of the phylogenetic tree using the full-length CENH3 sequences, three branches indicated positive selection (fig. 4a). The lineages showing positive selection were restricted to the B and C genomes of *Oryza*. Positive selection was observed in the B genome common ancestor CENH3 after it diverged from the common ancestor of the A and C genome CENH3. In addition, CENH3_C showed positive selection after the C genome diverged from the A genome. Interestingly, positive selection was observed in the common ancestor CENH3_C of *O. minuta* and *O. rhizomatis* after it diverged from CENH3_Ca from *O. alta* (fig. 4a). Similarly, when the analysis was restricted to the N-terminal tail domain, the same branches showed positive selection as when the full-length sequence was examined (fig. 4b). However, restricting the analysis to the N-terminal tail domain showed positive selection on only one extra branch. Positive selection was seen in the common ancestor CENH3 of the A, B, and C genomes after it split from the D and E genome CENH3 common ancestor (fig. 4b).
The adaptive evolution of the CENH3 (CID) proteins initially discovered in *Drosophila* species is not observed in all eukaryotic species. Talbert et al. (2004) compared the *CenH3* genes from maize and sugarcane and found no evidence of positive selection but rather stabilizing selection. This is in contrast with our discovery of the positive selection associated with several rice *CenH3* genes. The different conclusions drawn from the grass family is most likely due to the different species included in the analysis. In both analyses, the maize CENH3 appears to be under strong stabilizing selection. All the comparisons made between maize and the two sugarcane CENH3 proteins by Talbert et al. (2004) as well as all the comparisons between maize and the eight *Oryza* CENH3 proteins in this study indicate stabilizing selection. Therefore, the strong stabilizing selection acting on maize CENH3 could potentially mask positive selection acting on other grass CENH3 proteins.

To investigate if the large genetic distance (i.e., many synonymous substitutions) between rice and maize could mask potential positive selection, we analyzed the *CenH3* gene of barley (*Hordeum vulgare*), which is closer related to rice than maize. We identified a full-length barley EST (CD055401) resembling rice *CenH3*. Overall, barley and rice *CenH3* coding sequences have 68.2% identical sites not including gaps, more specifically the HFD and the N-terminal tail have 72.2% and 54.1% identical sites, respectively. Pairwise alignments of barley *CenH3* with the nine *Oryza* and maize *CenH3* genes were used to investigate the rate of synonymous and nonsynonymous substitutions. When full-length CENH3 proteins were analyzed, as well as just the HFD, all nine comparisons showed stabilizing selection, whereas neutral evolution was observed when the analysis was limited to the N-terminal tail (supplementary table 3, Supplementary Material online). CENH3 pairwise comparisons between the rice species and maize with barley also had numerous synonymous changes (a high $K_s$ value). Therefore, it seems unlikely that the large amount of synonymous substitutions between rice CENH3 and maize CENH3 is due to their divergence time (relatedness) but more likely an attribute of strong stabilizing selection acting on certain cereal *CenH3* genes. This analysis supports that CENH3 adaptive evolution is not universal but specific to certain lineages.

The lineage specificity of CENH3 adaptive evolution was best displayed by pairwise comparisons of CENH3 from different *Oryza* species (fig. 4). Although positive selection is associated with the CENH3 from the B and C genome species, strong stabilizing selection appears to be associated with the CENH3 from the D/E genome species, acting similarly to the maize CENH3. Interestingly, the characteristic of positive or stabilizing selection associated with the diploid species was maintained in the polyploid rice species. Furthermore, the characteristics of stabilizing selection of the CENH3 from the D/E genome was maintained in the CCDD species, even though the CENH3 in the same species is associated with positive selection. Thus, the strong stabilizing selection associated with CENH3 in some specific lineages appears to be controlled by genetic factors and can persist through the myriad of changes brought on by genome merging in an allopolyploid.

The adaptive evolution of the CENH3 proteins was hypothesized to respond to the rapid changing of the satellite repeats at centromeres (Malik and Henikoff 2001). Maintenance of two active CENH3 genes may add another factor for driving the dynamic changes of the centromeric satellite repeats in polyploid species. The 155-bp CentO satellite repeat is the major DNA component of the centromeres of both cultivated rice (A genome) and wild diploid rice species with the B genome (Zhang et al. 2005; Yan et al. 2008). However, the centromeres of diploid rice species with the C genome do not include the CentO satellite repeat but instead contain two different satellite repeats, CentO-C1 and CentO-C2 (Lee et al. 2005). The CentO and CentO-C1 repeats share an 80-bp motif that is also associated with the centromeric satellite repeats in several other grass species, including maize and pearl millet (*Pennisetum glaucum*) (Lee et al. 2005). Interestingly, the centromeres of several chromosomes in the BBCC tetraploid did not hybridize to either CentO or CentO-C1 probes (Lee et al. 2005), indicating that new centromeric repeats have evolved since the polyploidization event. It will be interesting to know if the adaptive evolution of CENH3 has played a role in driving the emergence of new centromeric repeats.
satellite repeats in the polyploid species. Genome-wide characterization of the centromeric DNA sequences in polyploid rice species will shed light on the genomic impact of the presence of multiple CENH3 proteins.

Supplementary Material

Supplementary tables 1–3 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgment

We thank Dr Paul Talbert for critical reading of the manuscript. This research was supported by grant DBI-0603927 from the National Science Foundation.

Literature Cited


Douglas Crawford, Associate Editor
Accepted September 2, 2009