nature genetics

Sequencing of a rice centromere uncovers active genes

Kiyotaka Nagaki[1], Zhukuan Cheng[1,2], Shu Ouyang[3], Paul B Talbert[4], Mary Kim[3], Kristine M Jones[3], Steven Henikoff[4], C Robin Buell[3] & Jiming Jiang[1]

Centromeres are the last frontiers of complex eukaryotic genomes, consisting of highly repetitive sequences that resist mapping, cloning and sequencing. The centromere of rice Chromosome 8 (*Cen8*) has an unusually low abundance of highly repetitive satellite DNA, which allowed us to determine its sequence. A region of ~750 kb in *Cen8* binds rice CENH3, the centromere-specific H3 histone. CENH3 binding is contained within a larger region that has abundant dimethylation of histone H3 at Lys9 (H3-Lys9), consistent with *Cen8* being embedded in heterochromatin. Fourteen predicted and at least four active genes are interspersed in *Cen8*, along with CENH3 binding sites. The retrotransposons located in and outside of the CENH3 binding domain have similar ages and structural dynamics. These results suggest that *Cen8* may represent an intermediate stage in the evolution of centromeres from genic regions, as in human neocentromeres, to fully mature centromeres that accumulate megabases of homogeneous satellite arrays.

Centromeres present an enigma. One and only one is needed on every chromosome for segregation at mitosis and meiosis in all eukaryotes, yet there is no conservation of centromeric sequences: different organisms have markedly different centromeric DNAs[1]. In budding yeast, a 125-bp consensus sequence contains all the information needed for centromere function, whereas in nematodes, the holokinetic centromere spans the entire chromosome. In most multicellular eukaryotes, however, centromeres typically encompass megabases of DNA, often with limited or nonexistent sequence similarity for homologous centromeres between closely related species[2].

Understanding the basis for centromere function in multicellular eukaryotes has been impeded by the sequence content of centromeres, which consist of highly repetitive 'satellite' sequences, interrupted by dispersed transposable elements. For example, the best-studied human centromere comprises a 2-Mb to 4-Mb core of homogeneous 171-bp alpha satellite DNA repeats flanked by ~0.5 Mb of diverged alpha satellite that is densely populated with L1 transposable elements[3]. The enormous tracts of highly repetitive sequences at centromeres have precluded complete sequencing, not only for human centromeres, but also for those from model organisms[4,5]. Indeed, not a single centromere that has been mapped in a multicellular eukaryote has been completely sequenced.

Whereas centromeric sequences have proven nearly intractable, considerable progress has been made in defining the nucleosomal component of centromere-specific chromatin. A centromere-specific histone H3-like protein, referred to as CenH3 (ref. 6), has been found to underlie the kinetochore in species ranging from yeast (Cse4p) to vertebrates (CENP-A) and plants (CENH3; refs. 6–8). CenH3s are essential for chromosome segregation in all organisms tested, and ablation of CenH3 is accompanied by the inability of other kinetochore components to assemble[2].

Plants, like humans, have centromeres comprising megabase-sized satellite sequences, and are similarly refractory to sequence analysis[4,9–11]. In rice, however, some chromosomes have little satellite repeat sequence[12]. The canonical rice centromeric satellite, CentO, lies in all 12 rice functional centromeres, as centromeric misdivisions yield telocentric chromosomes with reduced amounts of CentO[12]. Specifically, a misdivision derivative with a breakpoint in the very short CentO array on rice Chromosome 8 identifies this array as being within the functional Chromosome 8 centromere (*Cen8*). The limited amount of CentO on *Cen8* allowed us to obtain a contiguous tiling of BAC clones that span *Cen8*, to cytologically map CentO arrays to a region in the BAC contig and to sequence the CentO-spanning segment. We also identified a putative gene encoding CENH3 from rice genomic sequence, characterized it molecularly and raised an antibody to the product. We used this antibody to CENH3 in chromatin immunoprecipitation (ChIP) analysis to delimit a 750-kb region that underlies the kinetochore. Notably, this region contains several active genes.

RESULTS

A BAC contig spans *Cen8*

We previously showed that the centers of rice centromeres contain CentO arrays and centromere-specific retrotransposon (CRR) sequences[12–14]. We used CentO and CRR probes to identify BACs from centromeric regions in the rice genome. We screened 36,864 rice BAC clones, which are part of the rice BAC libraries constructed

[1]Department of Horticulture, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA. [2]Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, The People's Republic of China. [3]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA. [4]Howard Hughes Medical Institute, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109-1024, USA. Correspondence should be addressed to C.R.B. (rbuell@tigr.org) or J.J. (jjiang1@wisc.edu).

Published online 11 January 2004; doi:10.1038/ng1289

from *Oryza sativa* spp. *japonica* rice variety Nipponbare[12]. Approximately 3% of the BAC clones contain CentO, CRR or both. We then used all CentO- and CRR-positive clones to search the CUGI rice FPC Database to identify centromere-associated BAC contigs. We identified 58 contigs, each containing at least 50 clones, and selected clones from these contigs for fluorescent *in situ* hybridization (FISH) mapping to anchor them to the 12 rice chromosomes. One contig, containing more than 400 BAC clones, was associated with Chromosome 8, of which a subset of clones in the center of the contig spanned the CentO centromeric satellite repeat.

We mapped clones on either side of the CentO-containing BACs in this contig by FISH to the short-arm or long-arm side of CentO in rice pachytene chromosomes (**Fig. 1a–c**). We constructed a minimal tiling path of BAC clones that extends 5–6 BACs to either side of the CentO-containing BAC a0038J12 (**Fig. 1a**). We used pachytene FISH and fiber FISH to confirm the cytological locations and overlap between the 12 BAC clones (**Fig. 1d–f** and data not shown). Fiber FISH data with all 12 BAC clones indicated concordance between the BAC clones and Chromosome 8, suggesting that this tiling path is a faithful representation of the corresponding chromosomal region. In a more detailed analysis of BAC a0038J12, we compared the relative lengths of the simultaneous hybridization signals from CentO and a0038J12 probes on five Nipponbare genomic fibers and five a0038J12 molecules. The length of the CentO region as a percentage of the a0038J12 hybridization signal was 43.4 ± 1.4% for Nipponbare fibers and 41.5 ± 1.8% for a0038J12 molecules (mean ± s.e.), suggesting that CentO sequences in this clone were stable (**Supplementary Fig. 1** online). Other BAC molecules had similar correspondences to genomic fibers.

We shotgun-sequenced the 12 BACs in the minimal tiling path and then closed gaps and resolved misassemblies. The eight remaining small gaps are attributable to sequencing or assembly difficulties (**Supplementary Table 1** online). We used the sequences of these BACs together with the sequences (including three additional gaps) of four contiguous BACs available from GenBank (**Fig. 1a**) to construct a ~1.65-Mb virtual contig, which we annotated for gene models[15].

We identified repetitive sequences by BLAST-searching[16] the TIGR/Oryza/Repeat Database and the public rice genome sequence (a total of 3,760 BAC/PAC clones representing 515.9 Mb). Approximately 58% (954 kb) of the virtual contig is represented by known repetitive sequences: ~41 kb of CentO, 10 En/Spm-like DNA transposons[17,18] and 162 retrotransposons. An additional 14% of the virtual contig is derived from miniature inverted-repeat transposable elements, unknown repeats and highly truncated DNA transposons and retrotransposons. Thus, repetitive sequences comprise 72% of the 1.65-Mb region. Single-copy sequences include 47 gene models (**Fig. 2a**), of which 19 are similar in sequence to known genes and 28 are predicted solely by *ab initio* gene finders. The loci and their chromatin status are summarized in **Fig. 2a–f**.

The CentO satellite sequences in the virtual contig are organized into two clusters of three tracts each (**Fig. 2b**). Tract D (18,338 bp) and tract E (7,620 bp) are separated by a truncated CRR element (CRR-5), and tracts E and F (~12,244 bp) are separated by a complete CRR element (CRR-6). Tract F is inverted relative to tracts D and E. We found three shorter CentO tracts of 1,386 bp (tract A), 1,236 bp (tract B) and 147 bp (tract C) on the long-arm side of the virtual contig. Cluster ABC (2.8 kb) has only 7% of the amount of CentO found in cluster DEF (38.3 kb) and is not detectable by FISH, as the single CentO signal corresponds to cluster DEF on the short-arm side of BAC a0095C12 (**Fig. 1c**).

Mapping *Cen8* with CENH3

Our BAC mapping and DNA sequencing of the virtual contig identifies the molecular location of the centromere, which had previously been shown to include CentO repeats based on cytology and the occurrence of centromere misdivisions in this region[12]. But the boundaries of the centromere remained ambiguous, especially given that there are two CentO arrays in the contig that are several hundred kilobases apart. To delimit the entire centromere, we needed a criterion for deciding which sequences are in the *bona fide* centromere and which lie outside. Mapping strategies that have been applied to centromeres include recombination and rearrangement breakpoint mapping to separate a functional centromere from either arm[19,20] and construction of artificial chromosomes using centromeric sequences[3]. Each of these methods for defining centromeres has limitations, because subregions of centromeres with repetitive sequences and redundant functions can lead to ambiguities in mapping.

An alternative centromere mapping method takes advantage of the perfect correspondence between centromeres and CenH3s. High-resolution mapping is achieved using ChIP with antibodies against CenH3s (refs. 8,21). Owing to sequence repetitiveness, the application of ChIP to native satellite-containing centromeres has been limited to

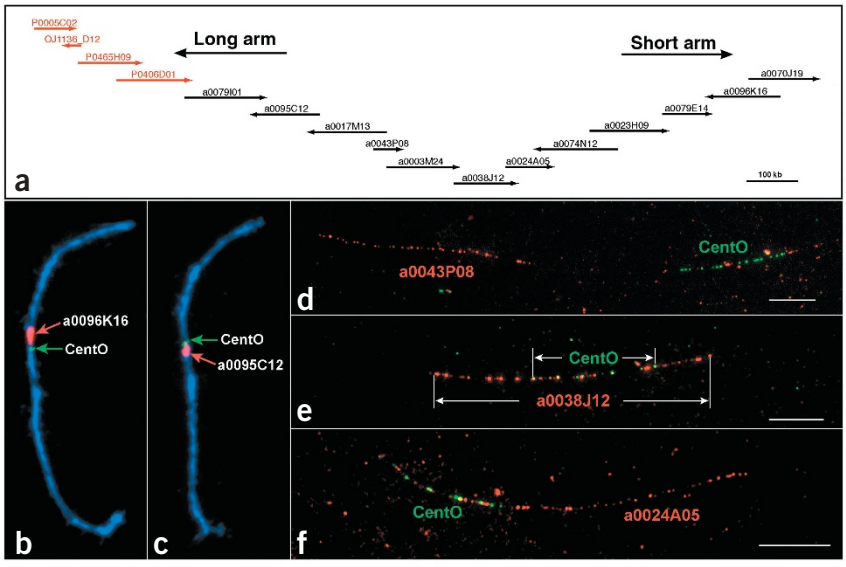

Figure 1 Cytological characterization of a BAC contig spanning rice *Cen8*. (**a**) BAC contig including 12 clones (shown in black) that represents a minimal tiling path centered on CentO and spanning the Chromosome 8 centromere. The sequences derived from the four BACs marked in red were obtained from GenBank. (**b**) BAC a0096K16 (red) is mapped to the short-arm side of the centromeric satellite CentO (green) on rice pachytene Chromosome 8. (**c**) BAC a0095C12 (red) on the contig is mapped to the long-arm side of CentO (green) on Chromosome 8. (**d**) A fiber FISH signal derived from BAC a0043P08 and CentO. (**e**) Reprobing of the fiber shown in **d** using a probe derived from BAC a0038J12 and CentO, showing that BAC a0038J12 spans the CentO signal. (**f**) A fiber FISH signal derived from BAC a0024A05 and CentO. All BACs used in FISH and fiber FISH are included in the minimal tiling path. DAPI staining is shown in blue. Scale bars, 10 m.

identification of the repeat family[22–24]. But ChIP using antibodies against the human CenH3, CENP-A, has been used to map neocentromeres, which are rare functional centromeres that sometimes arise spontaneously in alphoid-free regions of chromosome arms[25–27]. Therefore, in selected cases, ChIP using an organism's CenH3 can delimit the DNA region along the chromosome on which the kinetochore forms, which we refer to as the kinetochore region.

We identified a single gene encoding CENH3 in the rice genome. CenH3s are distinguishable from canonical H3 histones by bioinformatic criteria[28], and a single predicted open reading frame from genomic sequence conformed to the CenH3 profile. We identified an EST corresponding to part of this open reading frame and sequenced the apparently full-length cDNA corresponding to the EST. We raised a rabbit polyclonal antibody to the most N-terminal portion of the putative CENH3 protein. This antibody to CENH3 specifically stained the sister kinetochores at the primary constriction of each of the $2n = 24$ rice chromosomes in mitotic figures (**Fig. 3a–c**), providing a suitable reagent for molecular mapping of *Cen8*.

We first used ChIP to address whether CentO and CRR repeats underlie the kinetochore. Both sequences were enriched in the CENH3-bound fraction in ChIP assays (**Fig. 3d**). As is the case for maize centromeres[23], centromeric satellite is more enriched than centromere-specific retrotransposons. On average, 16.5 ± 3.8% (mean ± s.e., $n = 3$) of the CentO repeat and 7.8 ± 2.2% (mean ± s.e., $n = 3$) of the CRR element were precipitated by the antibody to CENH3. All other tested repeats (5S and 45S rDNA, RIRE3 retrotransposons and Os48 tandem repeats[29]) were not substantially enriched in the precipitated fractions (**Fig. 3d**). We conclude that CentO and CRR are main components of rice centromeres, but because they are repetitive, we cannot determine if they are bound by CENH3 specifically on *Cen8*.

Next, we used ChIP-PCR to identify the kinetochore region within the 1.65-Mb region. We used 42 single-copy primer sets (**Supplementary Table 2** online) from throughout the *Cen8* virtual contig to determine which sequences are bound by CENH3. The amplified products of 11 sets of primers (10, 12–15, 18, 22–24, 29 and 31) designed between 250 kb and 1,000 kb of the *Cen8* virtual contig

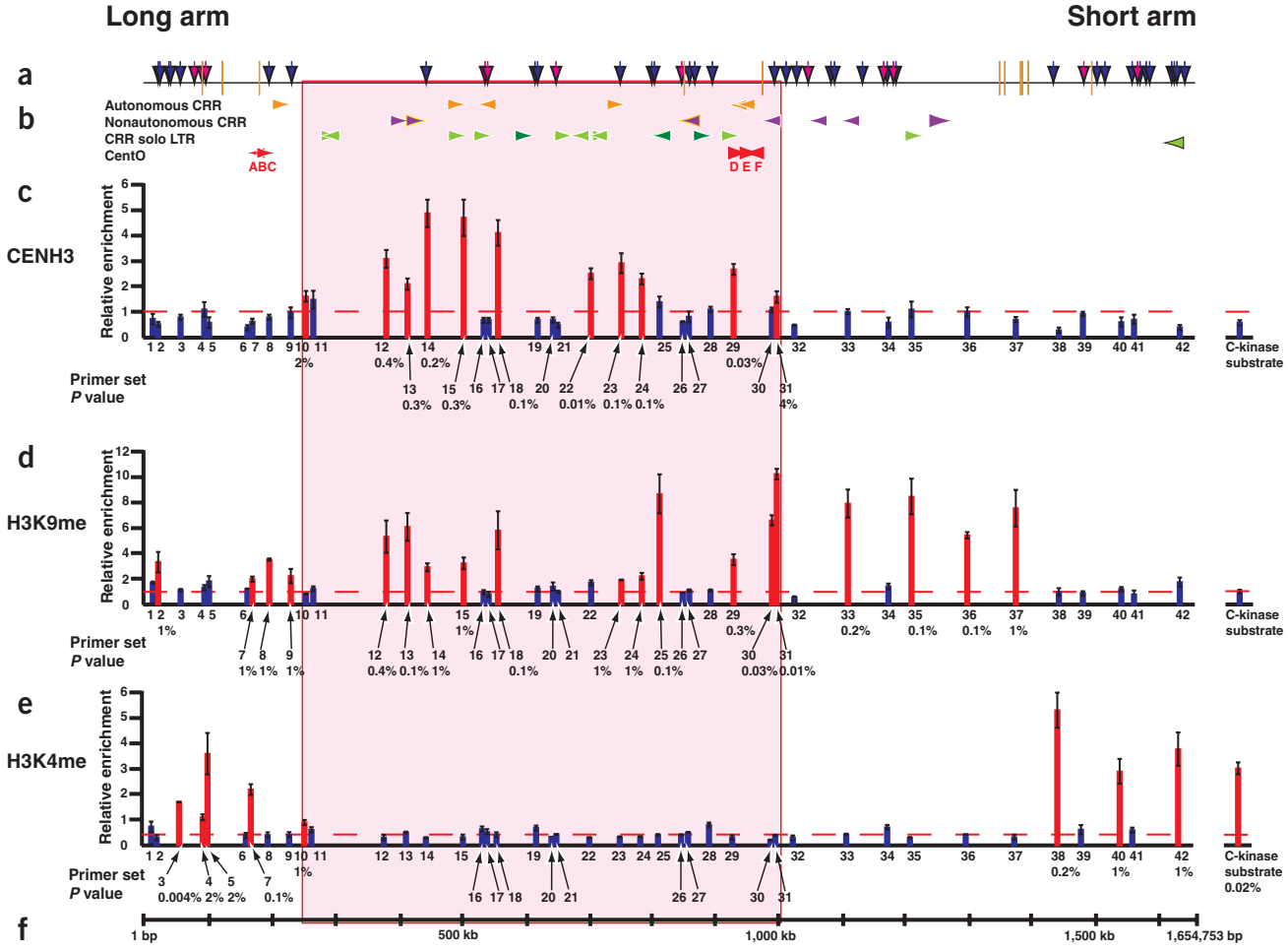

Figure 2 Map of *Cen8*. (**a**) Location of putative genes in the *Cen8* virtual contig. RT-PCR-positive genes are indicated by magenta arrows and RT-PCR-negative genes by blue arrows. Orange bars mark the positions of the remaining sequencing gaps. (**b**) Distributions of the CentO satellite and the CRR elements in the *Cen8* virtual contig. Autonomous CRRs, nonautonomous CRRs, solo LTRs from autonomous CRRs, solo LTRs from nonautonomous CRRs and CentO are indicated by orange, purple, dark green, light green and red arrowheads, respectively. (**c–e**) ChIP-PCR analysis of *Cen8* using (**c**) antibody to CENH3, (**d**) antibody to dimethylated H3-Lys9 (H3K9me) and (**e**) antibody to dimethylated H3-Lys4 (H3K4me). Mean ($n = 3$) relative enrichment levels are shown as histogram bars with standard error. *P* values calculated based on Student's *t*-test are shown as percentages. Red bars indicate significantly greater relative enrichment and blue bars indicate lack of significant enrichment. The red dotted line represents relative enrichment of 1 in **c** and **d**, but is set at 0.4 in **e** because primer pair 15 was used as a baseline rather than the noncentromeric reference gene *Rictpi*, which is enriched in dimethylated H3-Lys4 (cf. **Fig. 4a**). (**f**) A scale bar of the 1.65-Mb *Cen8* virtual contig. Pink shading marks the kinetochore region enriched in CENH3 binding.

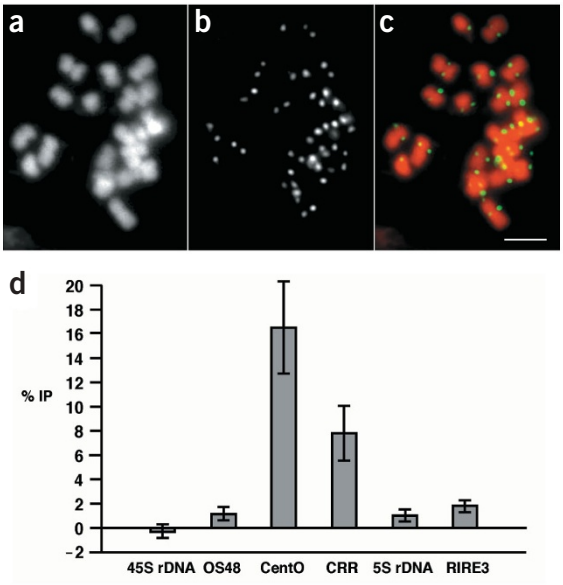

Figure 3 Analysis of CENH3 binding in the rice genome. (**a**) A somatic metaphase cell prepared from root tips. The primary constriction is visible on at least 10 of the 24 chromosomes. (**b**) Staining with antibody to CENH3 in the same cell. (**c**) Merge of **a** and **b**. Staining with antibody to CENH3 (green) is present on both chromatids (red) of all visible primary constrictions. Scale bar, 5 m. (**d**) ChIP using antibody to CENH3 on selected repeat sequences. The vertical axis shows the difference between antibody to CENH3 treatment and mock controls in the percent of the total hybridization signal that was found in the immunoprecipitated chromatin fraction (% IP) when probed with each repetitive sequence class.

had significantly greater relative enrichment in ChIP assays ($P < 0.05$; **Figs. 2c** and **4a** and **Supplementary Fig. 2** online). These positive primer sets surround CentO tracts D, E and F, indicating that these CentO tracts are included in the kinetochore region.

Primer sets 32–42 to the right of the CentO tract F did not have greater relative enrichment (**Fig. 2c** and **Supplementary Fig. 2** online), placing the short-arm boundary of the kinetochore region between primer sets 31 and 32. Primer sets 1–9 also did not have greater relative enrichment (**Fig. 2c** and **Supplementary Fig. 2** online), placing the long-arm boundary between primer sets 9 and 10. Because the short CentO tracts A, B and C are between primer sets 7 and 8, they do not appear to be included in the CENH3-binding region. Thus, the kinetochore region is ~750 kb, based on the distance between primer intervals 9–10 and 31–32. The relatively low significance of CENH3-binding at the edges of the kinetochore region (primer sets 10 and 31) might reflect a decrease in CENH3 density. Our mapping of the kinetochore region is consistent with the location of the DEF cluster of CentO tracts at the short-arm side of the primary constriction of Chromosome 8 in pachytene FISH (**Supplementary Fig. 1** online).

Distribution of H3 in *Cen8*

Eleven primer sets (11, 16, 17, 19–21, 25–28 and 30) in the kinetochore region did not have significantly greater relative enrichment (**Fig. 2c**), suggesting that CENH3 binding is discontinuous. To ascertain whether these discontinuities represent interspersion of CENH3- and H3-containing nucleosomes, we carried out ChIP-PCR analysis with the same primer sets and nucleosome fractions using antibodies to dimethylated histone H3 at Lys4 (H3-Lys4) and dimethylated H3-Lys9. Levels of dimethylated H3-Lys4 were low throughout the kinetochore region except at primer set 10 at the long-arm edge (**Figs. 2e** and **4a** and **Supplementary Fig. 2** online), whereas levels of dimethylated H3-Lys9 were high throughout the kinetochore region and beyond (**Figs. 2d** and **4a** and **Supplementary Fig. 2** online). Nine of the eleven primer sets with low relative enrichment for CENH3 also had insignificant relative enrichments for dimethylated H3-Lys9. Therefore, these primer sets cannot be used to detect ChIP signals, owing either to inaccessibility of the epitopes or to their absence. We conclude that high levels of CENH3 binding are contained within regions of high dimethylated H3-Lys9 binding.

The nine primer sets that are positive for both CENH3 and dimethylated H3-Lys9 (12–15, 18, 23, 24, 29 and 31) are probably not detecting a uniform population of *Cen8*s with mixed nucleosome arrays. ChIP analysis was done with the mono-, di- and trinucleosome fraction using probes that ranged from 100 bp to 350 bp, so mixed arrays would require an almost exact alternation of CENH3- and H3-containing nucleosomes at all nine sites. Rather, we interpret this result as reflecting heterogeneity among different *Cen8*s in the population. Thus, the alternating arrays of CENH3- and H3-containing nucleosomes expected based on fiber analysis for fruit fly and human cells[30] would average out in our ChIP-PCR study to show similar distributions of the two histone epitopes.

Mobile elements in *Cen8*

Of the 162 retroelements in the 1.65-Mb region, 28 are CRR elements (**Fig. 2b** and **Supplementary Fig. 3** online). Four of the CRRs are full-sized autonomous elements (7.6–7.8 kb), five are full-sized nonautonomous elements (4.4 kb) and the others are truncated elements or solo long terminal repeats (LTRs). Within the virtual contig, CRRs are most abundant in the *Cen8* kinetochore region, which contains 21 of the 28 CRRs, including 5 of the 6 autonomous CRRs and all 3 CRR solo LTRs. In contrast, we found no difference in the density of other elements between the kinetochore region and the rest of the virtual contig (48% of elements in ~45% of length).

Retrotransposons in rice *Cen8* are relatively young, based on age estimates from the number of substitutions per nucleotide site in the originally identical LTRs[11,31]. Using an average substitution rate of 6.5 10^{-9} per synonymous site per year[32], we estimate ages of 0–9.4 million years (**Supplementary Table 3** and **Supplementary Fig. 3** online). Only 6 of the 48 retrotransposons analyzed had transposed more than five million years ago, and 71% of the elements had transposed within the last three million years. We did not observe any differences in the transposition timing of the retrotransposons located in the kinetochore region from those located in the rest of the virtual contig.

Cen8 contains active genes

Of the 47 putative genes in the 1.65-Mb contig, 14 are in the kinetochore region (**Fig. 2a**). This is notable because centromeres of other multicellular eukaryotes are known to be devoid of genes[3,4,33]. To ascertain whether the putative centromeric and flanking genes are expressed, we carried out RT-PCR analysis of the 39 intron-containing apparently unique genes (**Table 1**). Twelve of these genes are expressed in leaf and root tissues (**Fig. 4b**, **Table 1** and **Supplementary Fig. 4** online), including four genes (6729.t00010, 6729.t00009, 6730.t00011 and 6827.t00018) located in the CENH3-binding region. The annotations for these genes did not suggest any notable difference from randomly sampled rice genes.

We carried out Southern-blot hybridizations using the RT-PCR products as probes to confirm that the active genes are unique in the rice genome. Seven of the twelve active genes showed only a single band

of the expected size; the other five showed a strong band of the expected size and a second weak band, indicating the presence of a related gene elsewhere in the genome (**Fig. 4c** and **Supplementary Fig. 4** online). We conclude that at least 12 genes on the virtual contig are expressed and that at least 4 of these active genes lie exclusively in the *Cen8* kinetochore region.

DISCUSSION

The presence of long tracts of satellite repeats have made centromeres the last frontiers of higher eukaryotic genomes, and no native centromere had previously been mapped, cloned and sequenced. A fully functional minichromosome centromere from *Drosophila melanogaster* has been cloned and partially sequenced[5,34].

Table 1 Expression of genes located in the *Cen8* virtual contig

Name	Protein	Nucleotide position in virtual contig	RT-PCR (Sh)	RT-PCR (R)	EST match
5866.t00010	Hypothetical protein	23,753–25,426	–	–	
5866.t00003	Hypothetical protein	25,450–25,629	NT	NT	
5866.t00005	Hypothetical protein	41,677–42,141	NT	NT	
5866.t00006	Hypothetical protein	43,277–48,449	–	–	
5866.t00007	Putative indole-3-glycerol phosphate synthase	56,931–60,794	–	–	BI813236 (L)
3507.t00003	Putative ribonucleoprotein 1	80,675–84,750	+	+	D48534 (Sh), AU089756 (Sh), AU056736 (L)
3507.t00004	Putative tyrosyl-tRNA synthetase	89,899–95,807	+	+	AU077843 (Sh), CA765926 (P), AU077844 (Sh)
3507.t00005	Putative polyketide cyclase	96,062–104,512	+	+	BE040915
3507.t00018	Hypothetical protein	198,664–199,710	–	–	
5171.t00010	Hypothetical protein	234,369–235,835	–	–	
4832.t00025	Hypothetical protein	448,235–450,480	–	–	
6729.t00010	Putative poly(A)-binding protein	537,498–539,020	+	+	TC84123 (C, E, L, P, Se)
6729.t00009	Putative poly(A)-binding protein	541,460–544,500	+	+	AU075435 (L)
6730.t00016	Putative HGWP repeat containing protein	615,333–616,875	–	–	
6730.t00015	Hypothetical protein	617,868–623,609	–	–	
6730.t00011	Putative CBS domain containing protein	649,284–652,594	+	+	
6827.t00005	Hypothetical protein	753,324–756,515	–	–	
6827.t00028	Hypothetical protein	800,449–803,555	–	–	
6827.t00012	Hypothetical protein	803,755–804,786	–	–	
6827.t00018	Putative ribosomal protein L15	848,317–851,647	+	+	D43490 (C)
6827.t00019	Hypothetical protein	860,960–861,562	–	–	
6827.t00021	Hypothetical protein	866,145–867,448	–	–	
6826.t00006	Putative Rer1 family protein	895,574–896,506	–	–	
6826.t00017	Hypothetical protein	992,729–993,672	–	–	
6735.t00007	Putative DNA repairing protein	1,011,461–1,012,134	–	–	
6735.t00009	Putative kinase	1,024,470–1,041,405	–	–	
6735.t00010	Putative vesicle trafficking protein	1,045,927–1,047,873	+	+	TC115723 (C, L, R)
6731.t00027	Hypothetical protein	1,081,494–1,083,192	–	–	
6731.t00026	Hypothetical protein	1,086,227–1,089,443	–	–	
6731.t00017	Hypothetical protein	1,131,028–1,131,470	–	–	
6731.t00012	Putative cytokinin inducible protein	1,159,199–1,171,740	+	+	AU173140 (R)
6731.t00011	Hypothetical protein	1,172,417–1,173,565	–	–	
6731.t00010	Putative glycosylase	1,17,7249–1,184,924	+	+	
6731.t00009	Hypothetical protein	1,185,672–1,185,905	NT	NT	
6733.t00017	Hypothetical protein	1,431,993–1,432,295	NT	NT	
6733.t00009	Putative sucrose-phosphate synthase	1,475,570–1,484,642	+	+	
6733.t00007	Hypothetical protein	1,503,075–1,503,293	NT	NT	
6733.t00006	Hypothetical protein	1,510,287–1,512,525	–	–	
6828.t00007	Hypothetical protein	1,555,100–1,555,828	NT	NT	
6828.t00008	Putative Ca^{2+}-dependent lipid-binding protein	1,561,752–1,565,171	+	+	C28110 (C)
6828.t00025	Hypothetical protein	1,567,560–1,569,093	–	–	
6828.t00009	Hypothetical protein	1,575,555–1,579,786	–	–	
6828.t00010	Hypothetical protein	1,580,328–1,584,237	–	–	
6828.t00016	Putative transcription factor	1,619,328–1,619,882	NT	NT	
6828.t00030	Hypothetical protein	1,621,951–1,623,671	–	–	
6828.t00027	Hypothetical protein	1,625,800–1,626,416	NT	NT	
6828.t00020	Unknown protein	1,637,979–1,640,087	–	–	BQ908917 (L)

C, callus; E, endosperm; L, leaf; NT, not tested; P, panicle; R, root; Se, seed; Sh, shoot.

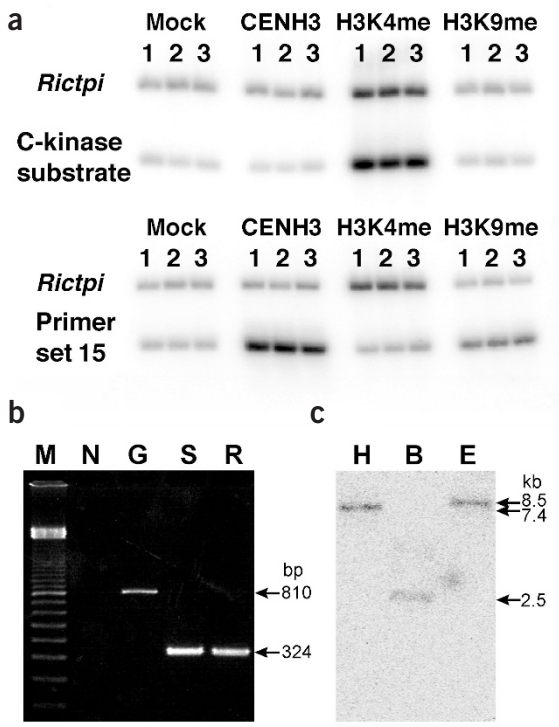

Figure 4 Examples of ChIP targets and gene expression in the CENH3-binding region of *Cen8*. (**a**) The C-kinase substrate gene (a noncentromeric control) and primer set 15 were each amplified and separated by electrophoresis in three replicate experiments together with the noncentromeric gene *Rictpi*, which is used as a reference for calculating relative enrichment. The ChIP-PCR results are negative (relative enrichment not significantly different from 1) for the C-kinase substrate gene and positive for primer set 15 for precipitation with antibody to CENH3 and antibody to dimethylated H3-Lys9 (H3K9me). Both *Rictpi* and C-kinase substrate genes are enriched in fractions precipitated with antibody to dimethylated H3-Lys4 (H3K4me). (**b**) RT-PCR using a cDNA of gene 6729.t00009 from shoot (S) and root (R). M, 100-bp ladder marker; N, negative control (without template); G, genomic DNA template. (**c**) The same RT-PCR product was used to probe a Southern blot of rice genomic DNA digested with *Hin*dIII (H), *Bam*HI (B) and *Eco*RI (E).

In this case, however, the precise relationship between this centromere and the native X-chromosome centromere from which it was derived is unknown, and a satellite block that is essential for minichromosome function is dispensable for X-chromosome segregation[35] and is absent from some wild-type chromosomes[36]. In addition, fully functional human neocentromeres have been cloned and sequenced. These are not native, however, but are extremely rare in the species, with no evidence for their continued propagation over more than a few generations[2]. In contrast, rice *Cen8* is native to the species, found at the same position and containing small CentO arrays in both *indica* and *japonica* cultivars that were independently domesticated ~10,000 years ago.

Except for its low satellite DNA content, *Cen8* is a typical rice centromere. Other rice centromeres are also satellite-poor relative to centromeres of many other eukaryotes. For example, FISH measurements show that rice *Cen4* contains a quantity of CentO that is not significantly higher than that of *Cen8* (ref. 12). Nevertheless, *Cen1* and *Cen11* have megabase-sized arrays of CentO, and other rice centromeres have array sizes that lie in between those of *Cen4* or *Cen8* and *Cen1* or *Cen11*. Therefore, the low level of CentO in *Cen8* places it at one extreme of a continuum of array sizes in this species. Large CentO arrays fall in the size range of satellite DNA arrays found in other multicellular organisms; for example, the human Y-chromosome alpha satellite array ranges from 400 kb to several megabases in normal men[37].

The 750-kb rice *Cen8* kinetochore region is similar in size and chromatin composition to centromeres of other eukaryotes. The *D. melanogaster* minichromosome centromere is 420 kb[33,34], a maize B chromosome centromere lies in a 500-kb region[38] and two human neocentromeres are 330 kb and 460 kb[26,27]. Like other centromeres, rice *Cen8* incorporates both CENH3- and H3-containing nucleosomes. This is expected from models of the centromere, in which the external face of the chromatid consists of CENH3-containing nucleosomes that interact with microtubules and the heterochromatic core promotes sister chromatid cohesion[30,39]. In support of this model, we find that dimethylated H3-Lys9, which is abundant in heterochromatin, is strongly enriched throughout and beyond the kinetochore region, and dimethylated H3-Lys4, which is abundant in euchromatin, is low in the kinetochore region relative to flanking regions. To account for the overlap between CENH3 and dimethylated H3-Lys9 bound to the same DNA segments in a population of *Cen8*s, we propose that there is variation in the distribution of nucleosome types between *Cen8*s. As a result, DNA segments throughout the kinetochore region will be high in both CENH3 and dimethylated H3-Lys9 (compare **Fig. 2c** with **2d**). Variation in the location of CENH3-containing nucleosomes relative to H3-containing nucleosomes is consistent with the interspersion and plasticity of long arrays of CENH3 and H3 observed for individual fibers in flies and humans[30].

The presence of active genes within the kinetochore region was unexpected, because centromeres are embedded in heterochromatin. But many fruit fly genes have been found to inhabit pericentric heterochromatin at low density[40], where they are expressed despite association with heterochromatin proteins[41]. The same appears to be true for *Arabidopsis*, in which one study mapped genes to a transposon-rich region thought at the time to be within centromeric DNA[20]. More recent work has shown that all five *Arabidopsis* centromeres lie within 178-bp satellite repeat arrays that extend for ~3 Mb[4,24], and there are no unique sequences or active gene candidates in any *Arabidopsis* centromere. Therefore, the contrast between *Arabidopsis* centromeres that lack genes and rice centromeres that have them is attributable to the abundance of satellite arrays.

Incompatibility between genes and centromeres has also been suggested from the disruption of a centromere by transcription in budding yeast[42]. But the yeast centromere is only 125 bp in length with a single CenH3 nucleosome[8], and so it might be easily disrupted, whereas the relatively enormous centromeres of plants and animals might not be noticeably affected by transcription. In addition, kinetochores function only during mitosis and meiosis, whereas transcription occurs only during the remainder of the cell cycle, and so cohabitation of genes and centromeres might be compatible with their disparate functions.

The phenomenon of active genes in rice *Cen8* resembles human neocentromeres that form *de novo* in chromosome arms. Neocentromeres arise in rather ordinary regions that contain genes[26,27]. Active genes were recently found in the functional neocentromere of mardel(10), a human marker chromosome derived from Chromosome 10 (ref. 43). As this marker chromosome is of recent origin and exists in a genome with two normal copies of each gene on the two normal Chromosome 10s, the active genes at this neocentromere are untested by evolution, whereas each gene in rice *Cen8* is fixed in the species. Nevertheless, the parallels between human neocentromeres and *Cen8* suggest that they represent different stages in an evolutionary sequence.

The emergence of centromeres from genic regions is a rare event in karyotype evolution. For example, in most primates, the X-chromosome centromere is monophyletic, consisting of an X-specific subfamily of alpha satellite that is present in multi-megabase arrays[3]. During the evolution of the X chromosome in lemurs, however, new centromeres have emerged twice in ancestrally distal locations[44]. It is unclear whether these rare events represent successful transpositions of other native centromeres or arose *de novo* from genic regions and acquired centromere-specific sequences later[45]. The lack of intermediate stages underscores our ignorance of the process of centromere evolution.

We propose that the earliest stage in centromere evolution is represented by human neocentromeres, in which a kinetochore forms in a genic region with loss or inactivation of the native centromere[37]. In most cases, this neocentromere-containing chromosome would become extinct. In the rare case of fixation, however, sequences would adapt to their new roles in mitosis by acquiring blocks of satellite DNAs and newly inserted centromere-specific transposons[44]. Rice *Cen8* would thus resemble an intermediate stage in the progression from neocentromeres to mature centromeres. During later stages, centromere meiotic drive would favor the expansion of satellite DNA arrays to multi-megabase lengths[1]. Meanwhile, transposons would accumulate in the resulting pericentromeric regions where meiotic recombination is suppressed[40], and the resident genes would gradually adapt to a heterochromatic environment[4,20,46,47]. The CENH3 binding domain of *Cen8* shows no structural, compositional or age differences from its flanking regions, except for the presence of CentO satellite and the enrichment of CRR elements. These observations suggest that *Cen8* may be at an early stage of centromere evolution.

A strength of this evolutionary scenario is that it unites observations made in plants and animals, ancient lineages in which centromeres must have evolved *de novo* many times. This inevitable progression from a diverse genic region to a monotonous block of centromeric repeat arrays is analogous to common ecological processes that result in the dominance of a climax species. For instance, in temperate zones, a new forest evolves into a climax forest dominated by a single species that continually replaces itself. In the case of centromere progression from a genic region, expansion of satellite DNA would lead to the climax state, in which continual homogenization of satellite repeats would maintain the kinetochore over long evolutionary periods.

METHODS

FISH and fiber FISH. We used *O. sativa* spp. *japonica* rice var. Nipponbare for all experiments. We prepared meiotic pachytene chromosomes and carried out chromosomal FISH essentially as described[29]. We carried out fiber FISH on extended genomic DNA fibers and BAC molecules as described[29]. We labeled DNA probes with biotin-dUTP or digoxigenin-dUTP (Boehringer Mannheim) and counterstained chromosomes with propidium iodide or DAPI in an antifade solution, Vectashield (Vector Laboratories). We used plasmid pRCS2 (ref. 14) as a probe to the CentO satellite repeat in FISH and fiber FISH analyses.

Sequencing. We carried out standard high-throughput sequencing as described[15]. Briefly, we constructed a small- and large-insert shotgun library for each BAC clone and end-sequenced clones using dye terminator chemistry on ABI 3700 sequencers (Applied Biosystems). We assembled random sequences for each BAC clone using TIGR Assembler[48] and identified clone linkages between the assemblies using the BAMBUS scaffolding software. We filled sequencing gaps using a combination of alternative chemistries, primer walking, resequencing of PCR products and transposon-mediated sequencing. We assembled highly repetitive areas, such as the CentO repeat region in a0038J12, using transposon-based sequencing with large-insert shotgun clones. Comparison of experimental with predicted restriction enzyme digestion patterns derived by agarose gel electrophoresis with a minimum of one restriction enzyme confirmed the final assembly of the finished BACs, except for an anomalous band in BAC a0070J19, possibly attributable to incomplete digestion.

BAC a0043P08 was only partially sequenced to confirm the overlap of BACs a0017M13 and a0003M24 in the tiling path. Five of the eight sequencing gaps, representing an estimated 3–11.4 kb, are located in the retrotransposons in BAC a0079E14, and this BAC remains unfinished. Two of the other gaps (in a0096K16 and a0003M24) resulted from GC hard stops in the sequence and are estimated to be less than 500 nucleotides each. The final gap (in a0038J12) resulted from assembly difficulties caused by the CentO repeat and was estimated at 7 kb. Examination of overlapping sequence between finished BAC clones that were sequenced independently identified 5 nucleotide differences in 261,119 bases of overlapping sequence (1 error per 52,223 bases). These five differences were a single nucleotide substitution (G→T) and an insertion of four bases (ATAT) in a local AT dinucleotide repeat.

***CenH3* cloning and immunocytochemistry.** We detected clone E30313, consisting of cDNA encoding rice CENH3 inserted into pBluescript II SK+, by TBLASTN searching of the MAFF EST database. This clone was provided in plasmid form by T. Sasaki of the MAFF DNA Bank[49]. A peptide was synthesized to represent the 19 most N-terminal amino acids of the predicted protein followed by a cysteine for conjugation (ARTKHPAVRKSKAEPKKKLC-amide). The peptide was conjugated to keyhole limpet hemocyanin and injected into two rabbits (Biosource International). One of the resulting antisera was affinity purified versus the peptide for immunostaining and ChIP. We used commercial antibodies to detect dimethylated H3-Lys4 and dimethylated H3-Lys9 (Upstate Biochemicals).

We fixed rice roots in PHEMES (60 mM PIPES buffer, 25 mM HEPES buffer, 10 mM EGTA, 2 mM $MgCl_2$, 0.35 M Sorbitol, pH 6.7) containing 3% paraformaldehyde and 0.2% Triton X-100 for 20 min and washed them three times in phosphate-buffered saline (PBS)[50]. Root tips were digested for 30 min at 37 °C with 1% cellulase Onozuka RS (Yakult Honsha) and 0.5% pectinase (Kikkoman) dissolved in PHEMES, washed three times in PBS and squashed onto slides coated with poly-L-lysine (Sigma).

We diluted antiserum to CENH3 1:100 in blocking solution and applied it to tissues on slides. Slides were covered with coverslips, sealed with rubber cement and incubated at 4 °C overnight. We then removed the coverslips and washed the slides three times with PBS. We detected the antibody by applying goat antibodies to rabbit conjugated to fluorescein isothiocyanate (Jackson ImmunoResearch) diluted 1:100 in blocking solution, incubating for 2 h and washing three times in PBS. Vectashield (Vector Laboratories) containing 1 g ml^{-1} DAPI was mounted on the slide. All images were captured digitally using a SenSys charge-coupled device camera (Roper Scientific) attached to an Olympus BX60 epifluorescence microscope. The camera control and image analysis were done using IPLab Spectrum v3.1 software (Signal Analytics).

ChIP. We carried out ChIP analysis as described[24] using two-week-old etiolated rice seedlings. We used normal rabbit serum as a mock treatment. For ChIP analysis of repeat sequences, we separated the immunoprecipitated samples into Sup (unbound) and Pel (bound) fractions. We purified DNA from both bound and unbound fractions and slot-blotted it onto a HybondN+ membrane. We hybridized the membrane with PCR-amplified centromeric and noncentromeric repeats (**Supplementary Table 2** online) or with the plasmids pTa794 (5S rDNA, 410 bp) or pOs48 (Os48, 355 bp). In each case, we subtracted the percent immunoprecipitation (defined as Pel/(Pel + Sup)) of the mock experiments from the percent immunoprecipitation of the antibody to CENH3 treatments. Each experiment was replicated in three independent tubes.

For ChIP-PCR analysis, we designed primers to exclusively single-copy regions of *Cen8* (**Supplementary Table 2** online). Negative controls were the *O. sativa* triosephosphate isomerase gene (*Rictpi*, at 12.2 cM region of chromosome 1) and the myristoylated alanine-rich C-kinase substrate-like gene (at 99.1 cM region of Chromosome 8). We carried out PCR using the primers with the bound fractions from treatments with the antibodies to CENH3, to dimethylated H3-Lys9, to dimethylated H3-Lys4 and mock treatments as templates. PCR conditions were 30 cycles at 94 °C for 30 s, annealing at the specific temperature for each primer set for 30 s, and 72 °C for 1 min. We separated the products by electrophoresis and blotted them on HybondN+ membrane (Amersham). We carried out hybridization, washes and detection as in the RT-PCR procedure (below). We calculated the relative enrichment by comparing

antibody-associated PCR product ratios to product ratios from mock experiments using the following formula: relative enrichment = $(cen8/Rictpi)_{antibody}$ / $(cen8/Rictpi)_{mock}$. The probability (P) that the mock fractions and antibody fractions belong to same group was determined by t-test.

RT-PCR. We germinated sterilized rice seeds in a sterile glass bottle and grew them for ten days in a cycle of 8 h light and 16 h dark at 20–25 °C. We extracted RNA from shoots and roots of the seedlings with AquaPure RNA isolation kits (BIO-RAD Laboratories). We synthesized cDNA with the RNA and ProtoScript First Strand cDNA Synthesis kit (New England Biolabs).

We determined copy numbers of the products in the rice genome by Southern-blot hybridization. Rice genomic DNA was digested with *Bam*HI, *Hin*dIII and *Eco*RI and blotted on HybondN+ membrane. In all hybridization experiments, we incubated membranes at 65 °C overnight and washed them sequentially with 2 saline sodium citrate (SSC) + 0.1% SDS, 0.5 SSC + 0.1% SDS, and 0.1 SSC + 0.1% SDS. Signals were detected by phosphorimaging.

URLs. CUGI rice FPC Database, http://www.genome.clemson.edu/projects/rice/fpc/; TIGR/Oryza/Repeat Database, http://www.tigr.org/tdb/e2k1/plant.repeats/; BAMBUS, http://www.tigr.org/software/bambus/.

GenBank accession numbers. BACs, AY360384–AY360394; CenH3, AY438639.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank Q. Yuan, J. Liu and the members of the TIGR Sequencing Facility and Informatics groups for their assistance and T. Sasaki and MAFF for the *CenH3* cDNA clone. This research was supported in part by grants from the US Department of Energy to J.J. and C.R.B.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 14 October; accepted 11 December 2003
Published online at http://www.nature.com/naturegenetics/

1. Malik, H.S. & Henikoff, S. Conflict begets complexity: the evolution of centromeres. *Curr. Opin. Genet. Dev.* **12**, 711–718 (2003).
2. Choo, K.H. Domain organization at the centromere and neocentromere. *Dev. Cell* **1**, 165–177 (2001).
3. Schueler, M.G., Higgins, A.W., Rudd, M.K., Gustashaw, K. & Willard, H.F. Genomic and genetic definition of a functional human centromere. *Science* **294**, 104–109 (2001).
4. Hosouchi, T., Kumekawa, N., Tsuruoka, H. & Kotani, H. Physical map-based sizes of the centromeric regions of *Arabidopsis thaliana* chromosomes 1, 2, and 3. *DNA Res.* **9**, 117–121 (2002).
5. Sun, X., Le, H.D., Wahlstrom, J.M. & Karpen, G.H. Sequence analysis of a functional Drosophila centromere. *Genome Res.* **13**, 182–194 (2003).
6. Talbert, P.B., Masuelli, R., Tyagi, A.P., Comai, L. & Henikoff, S. Centromeric localization and adaptive evolution of an Arabidopsis histone H3 variant. *Plant Cell* **14**, 1053–1066 (2002).
7. Palmer, D.K., O'Day, K., Trong, H.L., Charbonneau, H. & Margolis, R.L. Purification of the centromere-specific protein CENP-A and demonstration that it is a distinctive histone. *Proc. Natl. Acad. Sci. USA* **88**, 3734–3748 (1991).
8. Meluh, P.B., Yang, P., Glowczewski, L., Koshland, D. & Smith, M.M. Cse4p is a component of the core centromere of Saccharomyces cerevisiae. *Cell* **94**, 607–613 (1998).
9. Gindullis, F., Desel, C., Galasso, I. & Schmidt, T. The large-scale organization of the centromeric region in Beta species. *Genome Res.* **11**, 253–265 (2001).
10. Hudakova, S. *et al.* Sequence organization of barley centromeres. *Nucleic Acids Res.* **29**, 5029–5035 (2001).
11. Nagaki, K. *et al.* Molecular and cytological analyses of large tracks of centromeric DNA reveal the structure and evolutionary dynamics of maize centromeres. *Genetics* **163**, 759–770 (2003).
12. Cheng, Z. *et al.* Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell* **14**, 1691–1704 (2002).
13. Jiang, J. *et al.* A conserved repetitive DNA element located in the centromeres of cereal chromosomes. *Proc. Natl. Acad. Sci. USA* **93**, 14210–14213 (1996).
14. Dong, F. *et al.* Rice (*Oryza sativa*) centromeric regions consist of complex DNA. *Proc. Natl. Acad. Sci. USA* **95**, 8135–8140 (1998).
15. Yuan, Q. *et al.* Genome sequencing of a 239-kb region of rice chromosome 10L reveals a high frequency of gene duplication and a large chloroplast DNA insertion. *Mol. Genet. Genomics* **267**, 713–720 (2002).
16. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
17. Pereira, A., Cuypers, H., Gierl, A., Schwarz-Sommer, Z.S. & Saedler, H. Molecular analysis of the En/Spm transposable element system of *Zea mays*. *EMBO J.* **5**, 835–841 (1986).
18. Motohashi, R., Ohtsubo, E. & Ohtsubo, H. Identification of Tnr3, a suppressor-mutator/enhancer-like transposable element from rice. *Mol. Gen. Genet.* **250**, 148–152 (1996).
19. Round, E.K., Flowers, S.K. & Richards, E.J. *Arabidopsis thaliana* centromere regions: genetic map positions and repetitive DNA structure. *Genome Res.* **7**, 1045–1053 (1997).
20. Copenhaver, G.P. *et al.* Genetic definition and sequence analysis of Arabidopsis centromeres. *Science* **286**, 2468–2474 (1999).
21. Kniola, B. *et al.* The domain structure of centromeres is conserved from fission yeast to humans. *Mol. Biol. Cell* **12**, 2767–2775 (2001).
22. Vafa, O., Shelby, R.D. & Sullivan, K.F. CENP-A associated complex satellite DNA in the kinetochore of the Indian muntjac. *Chromosoma* **108**, 367–374 (1999).
23. Zhong, C.X. *et al.* Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. *Plant Cell* **14**, 2825–2836 (2002).
24. Nagaki, K. *et al.* Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of *Arabidopsis thaliana* centromeres. *Genetics* **163**, 1221–1225 (2003).
25. Satinover, D.L., Vance, G.H., Van Dyke, D.L. & Schwartz, S. Cytogenetic analysis and construction of a BAC contig across a common neocentromeric region from 9p. *Chromosoma* **110**, 275–283 (2001).
26. Lo, A.W. *et al.* A 330 kb CENP-A binding domain and altered replication timing at a human neocentromere. *EMBO J.* **20**, 2087–2096 (2001).
27. Lo, A.W. *et al.* A novel chromatin immunoprecipitation and array (CIA) analysis identifies a 460-kb CENP-A-binding neocentromeric DNA. *Genome Res.* **11**, 448–457 (2001).
28. Malik, H.S. & Henikoff, S. Phylogenomics of the nucleosome. *Nat. Struct. Biol.* **10**, 882–891 (2003).
29. Cheng, Z., Stupar, R.M., Gu, M. & Jiang, J. A tandemly repeated DNA sequence is associated with both knob-like heterochromatin and a highly decondensed structure in the meiotic pachytene chromosomes of rice. *Chromosoma (Berl.)* **110**, 24–31 (2001).
30. Blower, M.D., Sullivan, B.A. & Karpen, G.H. Conserved organization of centromeric chromatin in flies and humans. *Dev. Cell* **2**, 319–330 (2002).
31. SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. & Bennetzen, J.L. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**, 43–45 (1998).
32. Gaut, B.S., Morton, B.R., McCaig, B.C. & Clegg, M.T. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. *Proc. Natl. Acad. Sci. USA* **93**, 10274–10279 (1996).
33. Sun, X., Wahlstrom, J. & Karpen, G. Molecular structure of a functional Drosophila centromere. *Cell* **91**, 1007–1019 (1997).
34. Murphy, T.D. & Karpen, G.H. Localization of centromere function in a Drosophila minichromosome. *Cell* **82**, 599–609 (1995).
35. Tolchkov, E.V., Rasheva, V.I., Bonaccorsi, S., Westphal, T. & Gvozdev, V.A. The size and internal structure of a heterochromatic block determine its ability to induce position effect variegation in *Drosophila melanogaster*. *Genetics* **154**, 1611–1626 (2000).
36. Talbert, P.B. & Henikoff, S. Mapping the centromere of the X chromosome. *Ann. Dros. Res. Conf.* **41**, 247C (2000).
37. Tyler-Smith, C. *et al.* Transmission of a fully functional human neocentromere through three generations. *Am J. Hum. Genet.* **64**, 1440–1444 (1999).
38. Kaszas, E. & Birchler, J.A. Meiotic transmission rates correlate with physical features of rearranged centromeres in maize. *Genetics* **150**, 1683–1692 (1998).
39. Zinkowski, R.P., Meyne, J. & Brinkley, B.R. The centromere-kinetochore complex: a repeat subunit model. *J. Cell Biol.* **113**, 1091–1110 (1991).
40. Weiler, K.S. & Wakimoto, B.T. Heterochromatin and gene expression in Drosophila. *Annu. Rev. Genet.* **29**, 577–605 (1995).
41. Greil, F. *et al.* Distinct HP1 and Su(var)3-9 complexes bind to sets of developmentally coexpressed genes depending on chromosomal location. *Genes Dev.* **17**, 2825–2838 (2003).
42. Chlebowicz-Sledziewska, E. & Sledziewski, A.Z. Construction of multicopy yeast plasmids with regulated centromere function. *Gene* **39**, 25–31 (1985).
43. Saffery, R. *et al.* Transcription within a functional human centromere. *Mol. Cell* **12**, 509–516 (2003).
44. Ventura, M., Archidiacono, N. & Rocchi, M. Centromere emergence in evolution. *Genome Res.* **11**, 595–599 (2001).
45. Wong, L.H. & Choo, K.H.A. Centromere on the move. *Genome Res.* **11**, 513–516 (2001).
46. Devlin, R.H., Bingham, B. & Wakimoto, B.T. The organization and expression of the *light* gene, a heterochromatic gene of *Drosophila melanogaster*. *Genetics* **125**, 129–140 (1990).
47. Lohe, A.R. & Hilliker, A.J. Return of the H-word (heterochromatin). *Curr. Opin. Genet. Dev.* **5**, 746–755 (1995).
48. Sutton, G.G., White, O., Adams, M.D. & Kerlavage, A.R. TIGR assembler: a new tool for assembling large shotgun sequencing projects. *Genome Sci. Tech.* **1**, 9–19 (1995).
49. Wu, J. *et al.* A comprehensive rice transcript map containing 6591 expressed sequence tag sites. *Plant Cell* **14**, 525–535 (2002).
50. Yu, H.G., Hiatt, E.N., Chan, A., Sweeney, M. & Dawe, R.K. Neocentromere-mediated chromosome movement in maize. *J. Cell Biol.* **139**, 831–840 (1997).