Diatom centromeres suggest a mechanism for nuclear DNA acquisition

Rachel E. Diner[a,b], Chari M. Noddings[c], Nathan C. Lian[c], Anthony K. Kang[c], Jeffrey B. McQuaid[a,b], Jelena Jablanovic[b], Josh L. Espinoza[b], Ngocquynh A. Nguyen[c], Miguel A. Anzelmatti Jr.[b], Jakob Jansson[c], Vincent A. Bielinski[c], Bogumil J. Karas[c,1], Christopher L. Dupont[b], Andrew E. Allen[a,b], and Philip D. Weyman[c,2]

[a]Integrative Oceanography Division, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA 92037; [b]Microbial and Environmental Genomics Group, J. Craig Venter Institute, La Jolla, CA 92037; and [c]Synthetic Biology and Bioenergy Group, J. Craig Venter Institute, La Jolla, CA 92037

Edited by James A. Birchler, Division of Biological Sciences, University of Missouri, Columbia, MO, and approved June 13, 2017 (received for review January 17, 2017)

Centromeres are essential for cell division and growth in all eukaryotes, and knowledge of their sequence and structure guides the development of artificial chromosomes for functional cellular biology studies. Centromeric proteins are conserved among eukaryotes; however, centromeric DNA sequences are highly variable. We combined forward and reverse genetic approaches with chromatin immunoprecipitation to identify centromeres of the model diatom *Phaeodactylum tricornutum*. We observed 25 unique centromere sequences typically occurring once per chromosome, a finding that helps to resolve nuclear genome organization and indicates monocentric regional centromeres. Diatom centromere sequences contain low-GC content regions but lack repeats or other conserved sequence features. Native and foreign sequences with similar GC content to *P. tricornutum* centromeres can maintain episomes and recruit the diatom centromeric histone protein CENH3, suggesting nonnative sequences can also function as diatom centromeres. Thus, simple sequence requirements may enable DNA from foreign sources to persist in the nucleus as extrachromosomal episomes, revealing a potential mechanism for organellar and foreign DNA acquisition.

diatom | *Phaeodactylum tricornutum* | episome | centromere | CENH3

Centromeres play a crucial role in the cellular biology of eukaryotes by acting as a genomic site for kinetochore formation and facilitating effective transmission of replicated nuclear DNA to new cells. Centromere-associated proteins are functionally conserved among eukaryote species (1–3). Nearly all eukaryotes studied to date possess a version of a specialized centromeric histone protein (CENH3, also described as centromere protein A, CENP-A), which binds to centromeric DNA and replaces the histone H3 at the site of kinetochore assembly (4–6). Conversely, the centromeric DNA sequences themselves are extremely variable and appear to evolve rapidly, even among similar organisms (7).

There are three general types of eukaryotic centromeres: point centromeres, holocentromeres, and regional centromeres. Point centromeres are uniquely characterized by specific conserved DNA sequences and are found in limited fungal species including the budding yeast *Saccharomyces cerevisiae* and close relatives (8–10). In holocentromeric organisms, the kinetochore forms along the entire length of each chromosome; a notable example is the model organism *Caenorhabditis elegans* (11). Most eukaryotes have regional centromeres, which are commonly found as a single large DNA region on each chromosome [reviewed in Sullivan et al., 2001 (12) and Torras-Llort et al., 2009 (13)]. Regional centromeres are variable in length and sequence even among closely related species; however, there are often predictable genetic features. For example, human centromeres contain large stretches of repetitive satellite DNA, ranging in size from hundreds of kilobases to megabases (12, 14, 15). Centromeres of several plants and the insect model *Drosophila melanogaster* contain large arrays of satellite repeats interspersed with or adjacent to retrotransposons, which can vary substantially in copy number and organization (16). A common feature of centromeric DNA in many eukaryotes is low-GC content. Centromeres of *Schizosaccharomyces pombe* and other yeast species feature an unconserved core of AT-rich DNA sequence often surrounded by inverted repeats (17–20). The centromeres of the protist *Plasmodium* have no apparent sequence similarity besides being 2–4-kb regions of extremely low-GC content (<3%) (21, 22). Likewise, centromere regions of the red algal species *Cyanidioschyzon merolae* contain 2–3 kb of relatively low-GC content but manifest no other apparent pattern (23, 24).

Centromere identification can also be useful for synthetic biology, enabling further discoveries and biotechnology applications. Artificial chromosomes provide a stable platform for introduction and maintenance of multigene constructs necessary for expression of biosynthetic pathways and large complex proteins (25–28). The experimental identification of eukaryotic centromeres has been extremely useful for developing molecular

Significance

Centromeres are genomic sites facilitating chromosome segregation during cell division. We report our discovery of diatom centromeres and the description of centromere identity in the stramenopile protists. We also show that simple requirements for diatom centromeres permit ecologically relevant foreign DNA molecules to function as diatom centromeres by "hijacking" chromosome maintenance features. Because little is known at the molecular level about chromosome maintenance in diatoms, this paper provides experimental data with broad implications for cellular biology studies and biotechnology applications. The ability to maintain circular artificial chromosomes using foreign DNA sequences is unique among organisms with studied centromeres and opens up fascinating evolutionary questions about the mechanisms of nuclear gene acquisition from the multiple endosymbiotic events characterizing the stramenopile lineage.

Author contributions: R.E.D., C.M.N., N.C.L., A.K.K., J.B.M., J. Jablanovic, J.L.E., V.A.B., B.J.K., C.L.D., A.E.A., and P.D.W. designed research; R.E.D., C.M.N., N.C.L., A.K.K., J.B.M., J. Jablanovic, J.L.E., N.A.N., M.A.A., J. Jansson, V.A.B., and P.D.W. performed research; R.E.D., C.M.N., N.C.L., A.K.K., J.B.M., J. Jablanovic, J.L.E., N.A.N., M.A.A., V.A.B., B.J.K., C.L.D., A.E.A., and P.D.W. analyzed data; and R.E.D., J.L.E., C.L.D., A.E.A., and P.D.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The sequence reported in this paper has been deposited in the National Center for Biotechnology Information Sequence Read Archive (accession no. PRJNA357294).

[1]Present addresses: Department of Biochemistry, Schulich School of Medicine and Dentistry, Western University, London, ON N6A 5C1, Canada; and Designer Microbes Inc., London, ON N6G 4X8, Canada.

[2]To whom correspondence should be addressed. Email: pweyman@jcvi.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1700764114/-/DCSupplemental.

biology tools, particularly in the creation of artificial chromosomes. Circular and/or linear artificial chromosomes based on native centromeres, origins of replication, and in some cases telomeres have been developed for yeast (29), mammalian cells including human cell lines (30), plants (reviewed in ref. 31), and recently the protist *Plasmodium* (32). Despite the great potential for eukaryotic algae in biotechnology, very little is known about algal centromeres, and few resources are available to control gene expression from introduced autonomously replicating genetic constructs. In 1984, autonomously replicating plasmids using chloroplast DNA were described for the green alga *Chlamydomonas reinhardii* (33). However, these vectors were not maintained stably and have not been commonly used. More recently, centromeres have been identified and characterized in the red alga *C. merolae* (23, 24), where each of the 20 chromosomes was found to contain one distinct region recruiting CENH3. However, to our knowledge, these sequences have not yet been used for the construction of artificial chromosomes.

Identifying centromere composition and optimizing artificial chromosome construction would be particularly valuable for diatoms, which are an abundant group of eukaryotic phytoplankton with important ecological significance. Diatom research has facilitated major discoveries in algal physiology and genetics, and several species have been cultivated and genetically manipulated for the development of valuable bioproducts (34–36). In our previous work, we discovered that a region of *S. cerevisiae* DNA containing low-GC content enabled the stable maintenance of autonomously replicating episomes in diatoms (37, 38). The DNA was introduced into the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* by bacterial conjugation, also suggesting a previously unexplored mechanism for horizontal gene transfer from bacteria. Diatom nuclear genomes contain large amounts of DNA derived from nonnuclear sources, including foreign sequences such as bacteria and viruses, and prokaryotic and eukaryotic DNA obtained from endosymbiotic events (e.g., mitochondria, chloroplasts, and additional secondary endosymbioses) (39–42). This genetic complexity and rapid evolution contributes to the ecological success of diatoms. Thus, elucidating mechanisms that may facilitate nuclear gene acquisition and episomal maintenance will advance our knowledge of diatom evolution and enable biotechnological innovation.

Here, we identify centromeric regions of diatom chromosomes using forward and reverse genetics approaches and observe that diatom centromeres are characterized by a simple low-GC signal, which is also found in the previously described synthetic diatom episomes (37, 38). Furthermore, we show that nonnuclear diatom DNA and foreign DNA from a variety of sources with similarly low-GC content can mimic a diatom centromere, suggesting a permissive mechanism for nuclear gene acquisition. This study significantly advances the understanding of diatom genomic features, facilitates the development of diatom molecular tools, and suggests a mechanism for diatom acquisition of foreign genetic material.

Results

Identification of Putative Diatom Centromeres in *P. tricornutum* Chromosomes 25 and 26. We hypothesized that a centromeric region of a diatom chromosome would support maintenance of a nuclear episome, as this is a useful experimental method of confirming centromere function for other organisms (32, 43). To identify a diatom centromeric region, we first examined the shortest *P. tricornutum* chromosomes with telomere-to-telomere assembly (25 and 26) (39), which were each previously cloned as five overlapping ∼100-kb (76.6–142.6 kb) DNA fragments (44). In our prior studies (37, 38), sequences supporting episome maintenance in *P. tricornutum* were characterized by greatly improved ex-conjugant colony yield compared with plasmids incapable of episome maintenance. Thus, we predicted 100-kb fragments from a single *P. tricornutum* chromosome that supported episome maintenance would yield similarly increased colony numbers in our standard conjugation assay. Out of the five large fragments spanning each chromosome, one fragment from each chromosome produced increased numbers of ex-conjugant diatom colonies: plasmid Pt25-100kb-1 (containing the first ∼100-kb fragment of chromosome 25) (Fig. 1 *A* and *C*) and plasmid Pt26-100kb-5 (containing the fifth fragment of chromosome 26) (Fig. 1 *B* and *D*). The plasmid containing Pt25-100kb-1 resulted in 14–32-fold more colonies than plasmids containing other 100-kb fragments from chromosome 25 (Fig. 1 *A* and *C*), and plasmid containing Pt26-100kb-5 resulted in 26–100-fold higher colony numbers than other chromosome 26 fragments (Fig. 1 *B* and *D*).

Both Pt25-100kb-1 and Pt26-100kb-5 fragments encompass regions of low-GC content. We calculated the GC content of the genome in 100-bp windows overlapping by 50 bp and found that windows with the lowest GC content were found on fragments enabling episome maintenance (Fig. 1 *E* and *F*). When calculating GC percentage with larger window sizes (10 kb to 0.5 kb), an obvious dip in GC content was not apparent on chromosomes 25 and 26 (*SI Appendix*, Fig. S1). We quantified the number of 100-bp windows less than or equal to 32% GC within a 3-kb larger window and observed clear peaks for chromosomes 25 and 26 (Fig. 1 *G* and *H*).

To clarify whether these specific chromosomal regions enriched in low-GC content enabled episome maintenance, three 10-kb DNA subsequences of Pt25-100kb-1 were cloned into plasmids otherwise incapable of maintenance (pPtPBR2) (38): one sequence encompassing the bioinformatically identified low-GC region (Pt25-10kb-12) (Fig. 1*E*), and two other randomly selected sequences (Pt25-10kb-6 and Pt25-10kb-9) (*SI Appendix*, Fig. S2). Pt25-10kb-12 conjugation led to 85-fold more colonies than the negative control, whereas the other plasmids showed no increase (*SI Appendix*, Fig. S2). We further tested the low-GC region found on Pt25-10kb-12 by assembling a 1-kb subregion containing the lowest GC content region of chromosome 25 into pPtPBR2. This plasmid, Pt25-1kb, yielded 27-fold more colonies than the empty vector control (*SI Appendix*, Fig. S2). Another plasmid containing the 1-kb region encompassing the lowest GC content region of chromosome 26, Pt26-1kb, resulted in 68-fold more colonies than the empty vector control. Thus, for chromosomes 25 and 26, regions containing the lowest GC content were the only regions supporting episome maintenance. To confirm that these plasmids were maintained in the diatoms over extended periods of time, two clones of Pt25-1kb were passaged for 30 d with and without selection. As *P. tricornutum* was observed to divide about once a day in the experimental conditions, this corresponds to roughly 30 generations. Antibiotic-resistant colonies were recovered at percentages similar to prior studies (37, 38), which correspond to high per-generation segregation efficiencies (*SI Appendix*, Table S1) (32). Plasmids were recovered and confirmed by gel electrophoresis after the passaging period (*SI Appendix*, Fig. S3), demonstrating the stable maintenance of episomes in these lines (i.e., not integrated into native chromosomal DNA).

Identification of Diatom Centromeres Using ChIP-Sequencing and Reverse and Forward Genetics. *P. tricornutum* genomic DNA sequences enabled episome maintenance in the diatom, suggesting these regions were functioning as centromeres. Nearly all eukaryotes previously studied incorporate the centromeric histone CENH3 into centromeric nucleosomes, and we tested this in *P. tricornutum* to confirm centromere functionality. We constructed an episome containing the *CEN6-ARSH4-HIS3* maintenance sequence and a translational fusion of *P. tricornutum* CENH3 and yellow fluorescent protein (YFP) regulated by a *P. tricornutum* promoter and terminator. After transfer to *P. tricornutum* using bacterial conjugation (see *SI Appendix*, *Materials and Methods*), we performed chromatin immunoprecipitation (ChIP) assays on

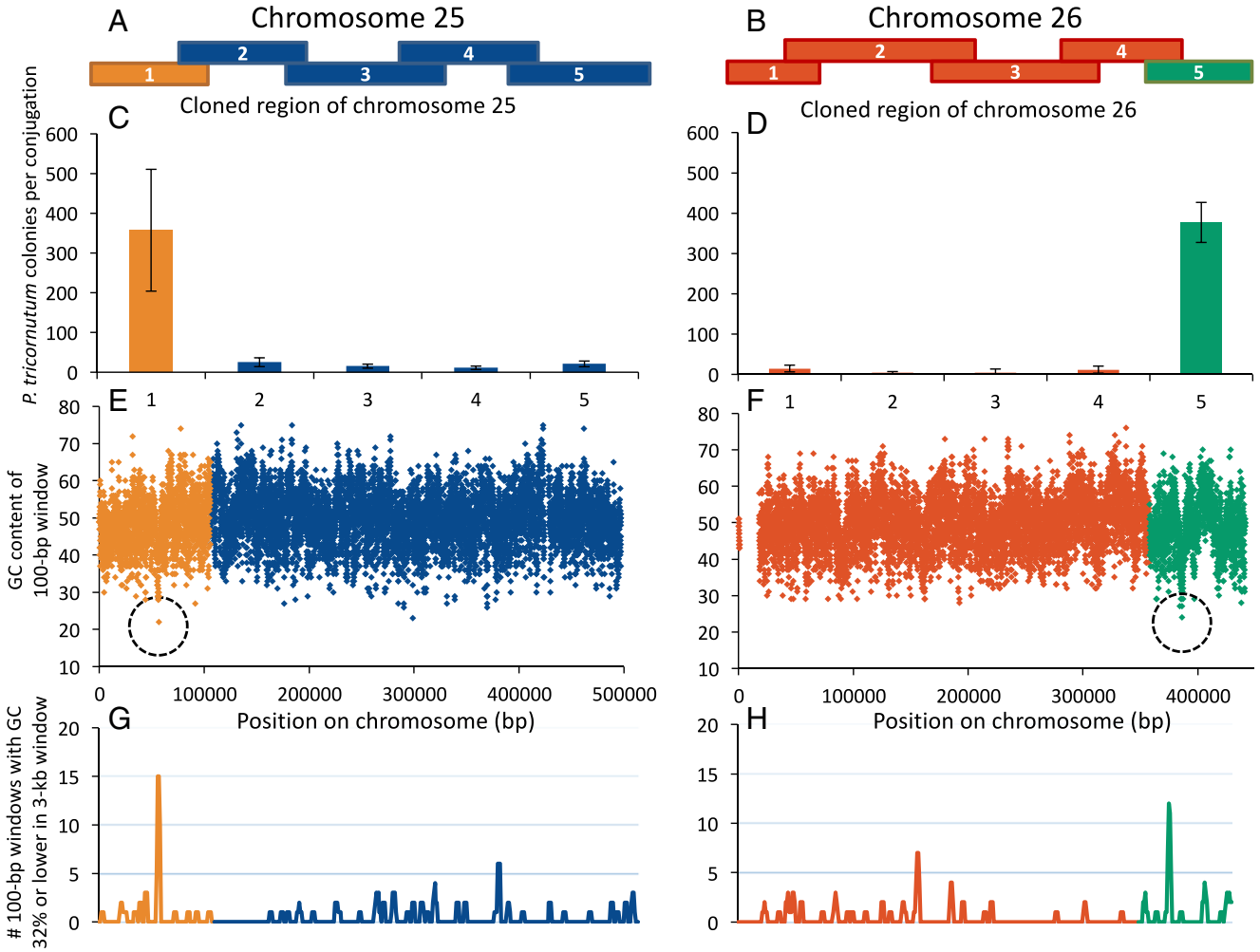

Fig. 1. Regions of *P. tricornutum* chromosomes enriched for low GC support episomal maintenance. (*A* and *B*) Chromosomes 25 and 26 were cloned as five overlapping ∼100-kb fragments. (*C* and *D*) Number of resulting *P. tricornutum* colonies per conjugation for episomes containing the indicated region of chromosome 25 or 26. Error bars indicate SD of four independent conjugation reactions for each fragment. (*E* and *F*) GC content was calculated for chromosomes 25 and 26 in 100-bp sliding windows that overlapped by 50 bp. Dashed circles indicate the lowest GC content for the chromosome in a 100-bp window. (*G* and *H*) Number of 100-bp windows with GC content of 32% or lower within a larger sliding 3-kb window that advanced by 1 kb each step.

ex-conjugant lines using GFP epitope antisera, followed by high-throughput DNA sequencing to identify all *P. tricornutum* genome sequences that recruit the centromeric histone.

ChIP-sequencing (ChIP-seq) analysis revealed 25 regions that were enriched for sequence reads (peaks) among the previously reported 33 nuclear chromosome scaffolds (39) (*SI Appendix*, Figs. S4 and S5). The low-GC regions of the chromosomes were specifically enriched for ChIP-seq reads (chromosome 25 shown for reference in Fig. 2; others in *SI Appendix*, Fig. S4). Of the 12 chromosome scaffolds with telomere-to-telomere assembly, all but one (chromosome 11) had ChIP-seq peaks, including chromosomes 25 and 26 (Fig. 3). Two regions recruiting CENH3 were also found within the nonscaffold assemblies ("bottom drawer" sequences) (obtained from the JGI *P. tricornutum* genome website: genome.jgi.doe.gov/Phatr2/Phatr2.home.html) (*SI Appendix*, Fig. S4). A ChIP-seq peak was also identified within the *S. cerevisiae CEN6-ARSH4-HIS3* region on the episome used to express the YFP-CENH3 fusion protein (Fig. 2). No mitochondrial or chloroplast sequences recruited CENH3, which was expected, as these genomes do not contain nucleosomes. Most ChIP-seq peaks on a genome-wide scale colocalized with the presence of at least ten 100-bp windows with GC content less than or equal to 32% GC in a larger 3-kb region (Fig. 2 and *SI Appendix*, Fig. S4).

To verify the ChIP-seq data, we conducted ChIP-qPCR on two regions with ChIP-seq peaks, one in the genome (Pt25-1kb) and one in the episome (*ARSH4*), and a region of genomic and episomal DNA without ChIP-seq peaks as a control (see *Materials and Methods*) (*SI Appendix*, Fig. S6 *A* and *B*). After ChIP, DNA from the low-GC *ARSH4* episomal region was in greater abundance by >50–70-fold compared with the negative control (*SI Appendix*, Fig. S6 *C–E*). Similarly, the Pt25-1kb region was enriched >200–500-fold compared with the genomic DNA negative control (*SI Appendix*, Fig. S6 *C–E*). Thus, ChIP-qPCR confirmed the ChIP-seq results for the CENH3 enriched regions of both episomal and native *P. tricornutum* chromosomal targets.

Of the 25 chromosome scaffolds with ChIP-seq hits, 23 had only one associated ChIP-seq peak that was between 2.4 and 5.6 kb (*SI Appendix*, Fig. S5). Chromosomes 2 and 8 each had two adjacent ChIP-seq peaks (*SI Appendix*, Fig. S5). Both putative centromeres on chromosome 2 (2a and 2b) are contained within a larger direct repeat and separated by a sequencing gap (indicated by Ns in the *P. tricornutum* genome sequence) (*SI Appendix*, Fig. S5). These sequences were highly similar to each other, with ∼2.9 kb aligning along the 3.4-kb sequence at >99% sequence identity. The two putative centromeres on chromosome 8 (8a and 8b, respectively) are each partially contained

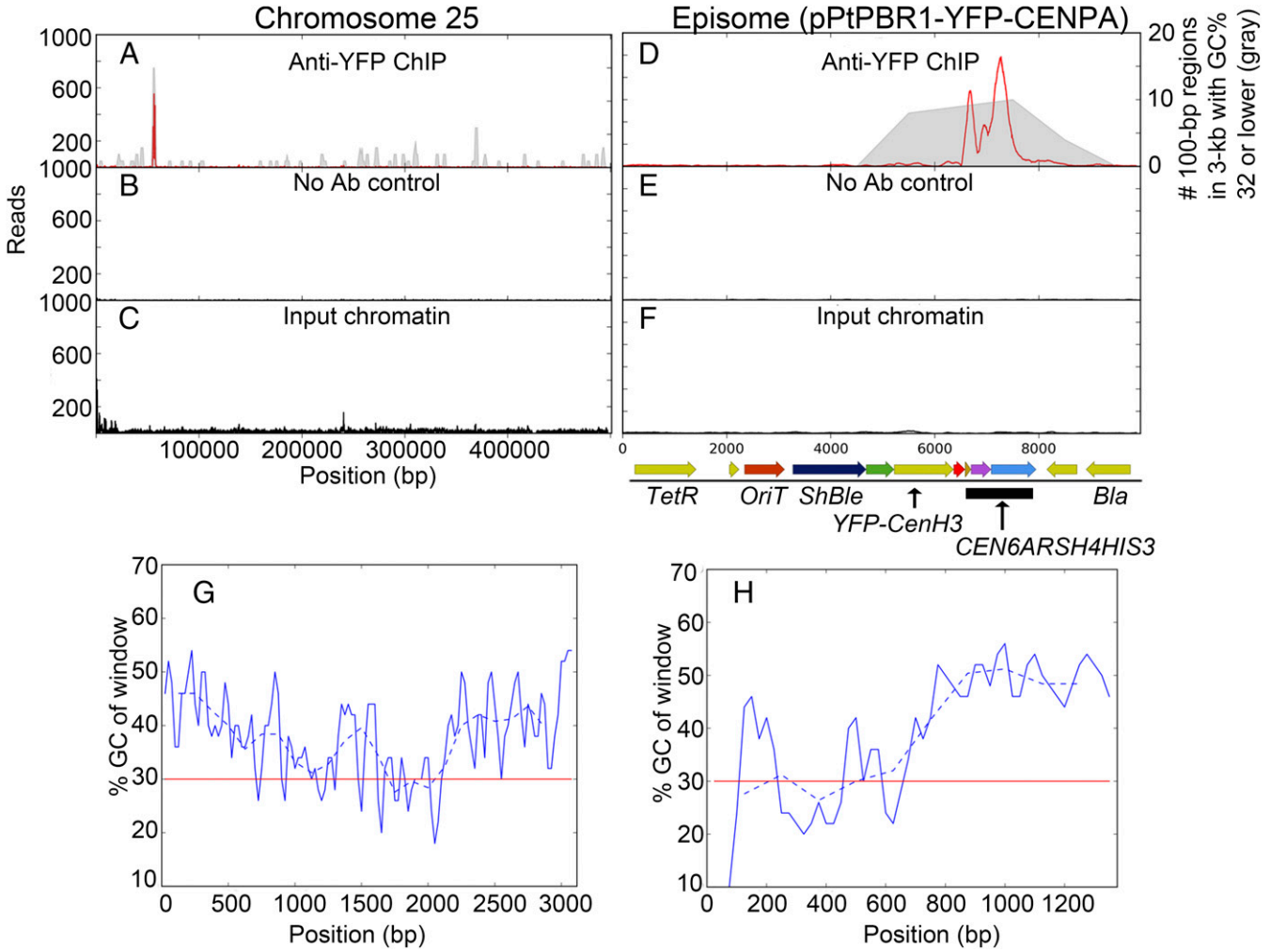

Fig. 2. ChIP-seq and GC data for chromosomee 25 and the episome. For chromosome 25 and the episome, ChIP-seq reads at each position for treatments with the YFP antibody (red) were plotted on the same graph as the number of 100-bp windows with GC 32% or lower in a larger 3-kb window (gray) (*A* and *D*). Graphs of the number of reads for the no-antibody ChIP-seq control (*B* and *E*) and input chromatin (*C* and *F*) were plotted using the same position scale as the anti-YFP ChIP-seq. For the episome, the positions of the genetic features are indicated below the input chromatin (the black bar indicates the *CEN6–ARSH4–HIS3* region). For the peaks identified by the CENH3-YFP ChIP-seq in chromosome 25 and the episome, GC content for 100-bp windows (50 bp overlap, solid blue line) or 250-bp windows (125-bp overlap, dashed blue line), respectively, was plotted with a reference line at 30% in red (*G* and *H*).

within long direct repeats at the 3′ end of the centromere. The 5′ end of 8b is adjacent to a region of unknown sequence (*SI Appendix*, Fig. S5). The 8a and 8b centromere sequences were also highly similar, with alignment across about half of the centromere sequence at 96.5% identity.

Apart from these potentially tandem centromere cases, most *P. tricornutum* centromeres were unique, having no similarity to other centromere sequences, with two exceptions. Predicted centromeres from chromosomes 24 and 29 shared 99.2% sequence identity over the entire 2.4-kb region and differed by only 14 mismatches. Additionally, the centromere from chromosome 30 shared a 1.6-kb region of high identity (97%) to a bottom drawer sequence bd23 × 34, which was one of the two bottom drawer sequences with an associated ChIP-seq hit. Centromeres in *P. tricornutum* were mostly located in intergenic spaces (*SI Appendix*, Fig. S5). Direct repeats were detected in approximately one-third of the centromeres, but the repeat number was low (usually a single sequence found twice) and the repeat period was variable and small (16–400 bp) (*SI Appendix*, Table S2). Genomic coordinates of all predicted centromeres, including ChIP-seq read regions and bioinformatically predicted regions containing low-GC content, are noted in *SI Appendix*, Table S3.

We used forward genetics to test whether sequences in the *P. tricornutum* genome including and in addition to those identified by ChIP-seq could support episomal maintenance. We prepared a *P. tricornutum* genomic library with 2–5-kb inserts using a nonepisome vector (pPtPBR2) and conjugated the library into *P. tricornutum* cells. Episomes were identified by extracting plasmids from antibiotic-resistant *P. tricornutum* colonies and transforming *Escherichia coli*; only DNA maintained as circular episomes in *P. tricornutum* was expected to yield *E. coli* colonies. We amplified and sequenced *P. tricornutum* genomic library inserts from *E. coli* colonies and identified 35 unique insert sequences from 99 recovered plasmids (*SI Appendix*, Table S4). Of these 35 unique insert sequences, 10 mapped to the nuclear genome chromosomal scaffolds and 1 mapped to the unscaffolded bottom drawer assemblies. Eighteen sequences mapped to the chloroplast genome, and 6 mapped to the mitochondrial genome.

Reverse genetics was used to functionally test whether the sequences identified by ChIP-seq and the *P. tricornutum* forward genetics library could maintain episomes. Forty sequences, including all ChIP-seq peaks, potential ChIP-seq artifacts, and *P. tricornutum* forward genetic library sequences including selected mitochondrial and chloroplast DNA sequences, were

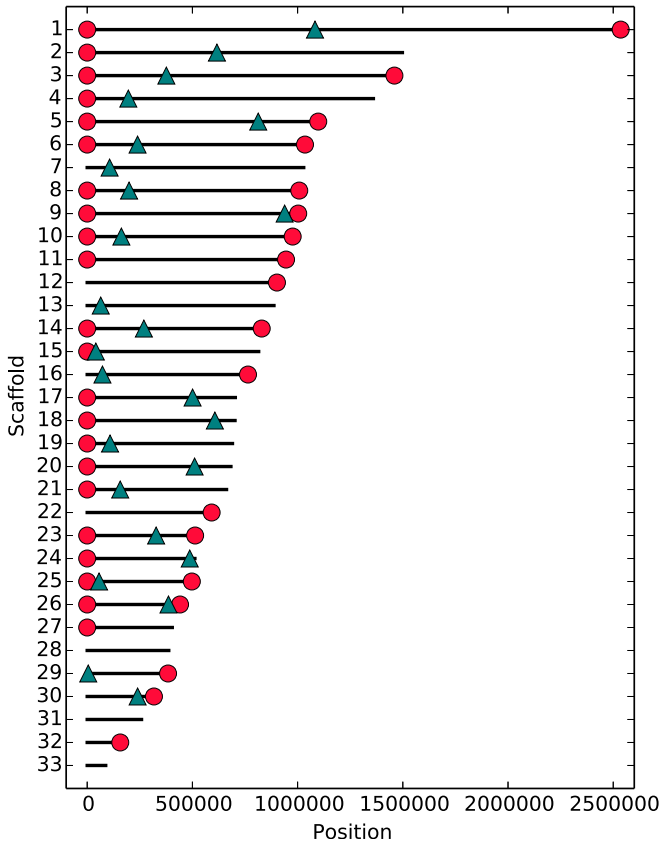

Fig. 3. Centromere and telomere locations on *P. tricornutum* chromosomal-scale scaffolds. Telomeres (crimson circles) and centromeres (teal triangles) are annotated on each scaffold.

cloned into the nonepisomal plasmid pPtPBR2 (see *Materials and Methods*). Most plasmids containing ChIP-seq identified sequences resulted in 7–162-fold more diatom ex-conjugant colonies than the pPtPBR2 negative control (*SI Appendix*, Fig. S7). We also tested random regions of chromosome 1 as negative controls (Test-37, -38, and -39) and regions suspected to be ChIP-seq mapping artifacts based on high read counts in both input and anti-YFP immunoprecipitation treatments (Test-4, -10, and -16). Both classes of sequences were unable to support episome maintenance; ex-conjugant numbers were similar to the negative control and much lower than the positive control pPtPBR1 (*SI Appendix*, Fig. S7). Ex-conjugant colony numbers following conjugation with the pPtPBR1 positive control (containing *CEN6-ARSH4-HIS3*) were not notably different from the episomes containing putative *P. tricornutum* centromeres. One insert sequence from chromosome 11 contained a region of GC content similar to, but slightly higher than, the centromeres (Test-40). However, this region contained no ChIP-seq peak and was unable to maintain an episome (*SI Appendix*, Fig. S7).

We also tested the *P. tricornutum* regions recovered from the forward genetic screen for the ability to maintain episomes. All chloroplast and mitochondrial DNA sequences, the bottom drawer sequence, and 8 of the 10 nuclear genome sequences contained low-GC content of 28–41% (*SI Appendix*, Table S4) across the entire insert region. These 8 nuclear genome sequences and the bottom drawer sequence mapped to identical regions as the ChIP-seq peaks (*SI Appendix*, Fig. S4). The two remaining inserts (Test-18 and Test-20) from the nuclear genome had GC content typical of the *P. tricornutum* nuclear genomic DNA (47%) and did not map to a ChIP-seq peak (*SI Appendix*, Table S3). We retested whether the two high-GC nuclear genome inserts as well as two sequences each from the chloroplast (Test-33 and Test-34) and the mitochondrion (Test-35 and Test-36) could support episome maintenance. Both mitochondrial and both chloroplast sequences supported episomes (*SI Appendix*, Table S4); however, the high-GC nuclear sequences did not, and we predict that their appearance in the library was likely due to plasmid carryover from the initial conjugation (*SI Appendix*, Table S4).

To further examine the minimum sequence size required for centromere function, we combined the information obtained in our initial screening of chromosome 25 with ChIP-seq identification of the full chromosome 25 centromere sequence to test the functionality of various small subsequences (*SI Appendix*, Fig. S8 and Table S5). We designed a series of 19 constructs to test sequences ranging in size from 198 to 1,040 bp. These sequences included a systematic minimization of centromere 25, particularly the 1-kb region (Pt25-1kb) shown above to maintain episomes, as well as a test of low-GC regions of sequentially smaller sizes. When we broke the ∼3-kb region that recruited CENH3 into three equal parts of 1,040 bp, only the middle third supported episomal maintenance. This middle third contained the majority of the Pt25-1kb insert and almost all of the lowest GC 500-bp region. When we further dissected the Pt25-1kb insert, only sequences >500 bp could maintain episomes, and all functional sequences encompassed low-GC DNA sequences, particularly the downstream region of the lowest 500-bp region (*SI Appendix*, Fig. S8).

Foreign DNA Sequences Examined for Episome Maintenance. Because the *CEN6–ARSH4–HIS3* sequence from *S. cerevisiae* supported episome maintenance in *P. tricornutum*, we hypothesized that other foreign DNA sequences with similarly low-GC composition could as well. Deletion analysis of the *CEN6–ARSH4–HIS3* region previously revealed that low-GC regions of >∼500 bp enabled maintenance. To test this pattern in the present study, we examined 24 sequences from *Mycoplasma mycoides* JCVI Syn1.0 (NCBI accession no. CP002027) of various sizes (0.5–1 kb) and GC content (15–50%) for their ability to maintain diatom episomes. All sequences of less than 28% GC content regardless of the size resulted in high numbers of ex-conjugant colonies consistent with episome maintenance (Fig. 4). Most sequences of 28% and 30% GC also resulted in large numbers of *P. tricornutum* ex-conjugant colonies with two exceptions that produced colony numbers similar to the negative control: a 500-bp 28% GC fragment (1.3-fold below control), and a 500-bp 30% GC fragment (1.2-fold above control) (Fig. 4). Additionally, one 700-bp 30% GC fragment produced only 3.3-fold more colonies than the control, a relatively low colony increase. The fragments containing either 40% or 50% GC content sequences produced ex-conjugant colony numbers similar to the negative control. Thus, with a few exceptions (Fig. 5), DNA sequences of ∼30% GC or lower were required and sufficient to support *P. tricornutum* episomes.

The above results suggest that many sequences of at least 500 bp (the smallest fragment tested) of low-GC DNA could maintain an episome in *P. tricornutum*, including sequences with environmental relevance. We examined whether a marine bacterial conjugative plasmid could support episome maintenance by searching the *Alteromonas macleodii* conjugative plasmid pAMDE1 for low-GC content regions (*SI Appendix*, Fig. S9). We then identified and cloned two 500-bp regions, AM-1 and AM-2, with 26.2% and 28.8% GC, respectively; conjugation of plasmids containing either region yielded 6–17-fold more ex-conjugant *P. tricornutum* colonies than the pPtPBR2 negative control with no maintenance sequence elements (*SI Appendix*, Fig. S9). We also tested whether regions of plasmids previously isolated from the diatom *Cylindrotheca fusiformis* (45, 46) could support episomes in *P. tricornutum*. Two plasmids, pCF1 and pCF2, containing low-GC, 560-bp regions (28.9% and 28.4% GC, respectively)

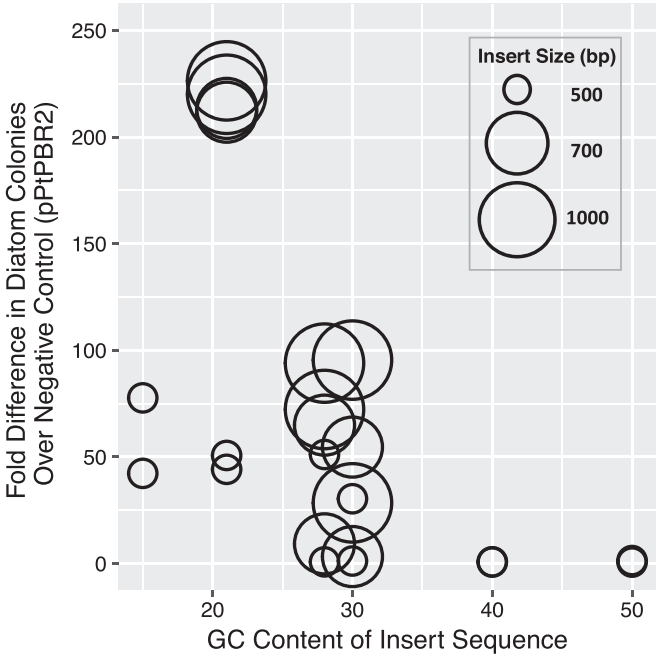

Fig. 4. Maintenance of episomes containing *M. mycoides* DNA sequences. Inserts of various GC content (15%, 21%, 27%, 30%, 40%, and 50% GC) from *M. mycoides* were tested for the ability to support episomal maintenance in *P. tricornutum*. The number of diatom ex-conjugant colonies obtained after conjugation is shown as the fold increase in colony numbers over the pPtPBR2 negative control for plasmids containing inserts of different size and GC content. The size of each circle represents the size of the insert sequence tested: large circles, 1,000 bp; medium circles, 700 bp; and small circles, 500 bp. The center of the circle indicates the data point plotted.

were constructed (see *SI Appendix*, *Materials and Methods*), and each yielded 7–12-fold more *P. tricornutum* ex-conjugant colonies than the pPtPBR2 negative control (*SI Appendix*, Fig. S9).

We examined maintenance properties of episomes supported by foreign DNA sequences to identify whether these plasmids were stable over time. *P. tricornutum* ex-conjugant lines were maintained with and without antibiotics for 30 d (see *Materials and Methods*). For all clones, as well as all other experiments where conjugations resulted in a high number of ex-conjugant diatom colonies relative to the negative control, episomes were successfully recovered in *E. coli*, confirming their stable extrachromosomal maintenance. In lines containing plasmids with two different *Mycoplasma* inserts (Myco-15–500bp-2 and Myco-21–500bp-2), between 45% and 75% of cells retained the episome without selection (per-generation segregation efficiency of 97–99%), and 91–95% retained it with selection (>99% segregation efficiency). *A. macleodii* and *C. fusiformis* DNA-containing sequences, with the exception of colony 8 of the AM-2 plasmid, were maintained with retention rates between 24% and 84% in the absence of antibiotics and 77% and 93% with antibiotics, corresponding to segregation efficiencies of >95% and >99%, respectively (*SI Appendix*, Table S1 and Fig. S3). In colony 8 containing the AM-2 episome, only 3% of cells retained the episome after passaging without antibiotic, with 74% retained with antibiotics. Although much lower than the other constructs, this still corresponds to segregation efficiencies of 89% without antibiotics and 99% with antibiotics. With this exception, retention rates of foreign DNA plasmids were similar to maintenance of episomes containing the *CEN6–ARSH4–HIS3* sequence (37, 38) and the native *P. tricornutum* centromere sequence from chromosome 25 (*SI Appendix*, Table S1). Episomes maintained with selection, a scenario more similar to native chromosomes containing essential genes, were maintained at much higher levels in all lines (*SI Appendix*, Table S1).

Bioinformatic Analysis of Episome-Supporting Sequences. Our results indicate that 500-bp sequences can maintain episomes. Thus, we searched within these sequences for the 500-bp subregion with the lowest GC content (*SI Appendix*, Table S6). When viewed together based on ability to maintain an episome, all inserts from the native diatom "Test" series and all foreign DNA inserts examined (including *M. mycoides*, *C. fusiformis*, and *A. macleodii* plasmid pAMDE1 source DNA) indicated a clear pattern of low-GC content supporting episome maintenance regardless of whether the source was foreign or native (Fig. 5). However, three *M. mycoides* DNA sequences (28-500-2, 30-500-2, and 30-700-2) that were predicted to be maintained based on low average GC content produced low numbers of ex-conjugant colonies after conjugation. This suggested that additional signals besides average GC content might be important.

Native *P. tricornutum* centromeres do not have repeats or other structures, and attempts to identify a conserved sequence motif using BLAST (47) and MEME (48) were unsuccessful, so we examined k-mer use to determine if very short sequences were overrepresented in DNA fragments supporting episomes. We chose a k-mer length of 6 because it was the longest string that could still be well-represented in a sequence of 500 bp. We identified unique 6 mers overrepresented in native *P. tricornutum* centromeres by requiring their retention to be statistically significant ($P < 0.001$) compared with randomly selected *P. tricornutum* genomic sequence (47% GC) and randomly generated sequences of 47% GC. Because the overall GC content is lower for centromeric ChIP-seq peaks (39% GC average) compared with the genomic regions (47% GC), we also required the 6 mers to be significantly overrepresented in the centromeres relative to a randomly generated set of 39% GC sequences. This allowed us to identify 6 mers overrepresented in the *P. tricornutum* centromeres that were unexplained by GC content difference from the genomic DNA (*SI Appendix*, Table S7). We then examined the recruitment of this set of centromere-enriched 6 mers in two sets of *Mycoplasma* fragments. One set contained the two 28% GC sequences and one 30% GC sequence that did not support

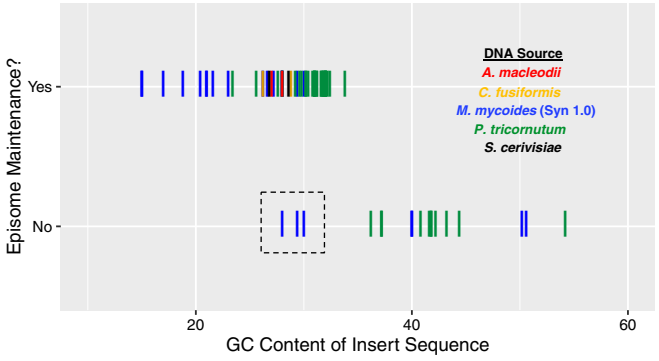

Fig. 5. Relationship between GC content and episome maintenance. The 500-bp subregion with the lowest GC content was identified for each insert sequence in the *P. tricornutum* test series (which includes ChIP-seq peaks, forward genetic library sequences, designed negative controls, and potential ChIP-seq artifacts) and foreign DNA inserts (including *M. mycoides*, *C. fusiformis*, and *A. macleodii* plasmid pAMDE1 source DNA). See *SI Appendix*, Table S6 for data included in the figure. This lowest GC content subregion was plotted as a function of whether the DNA could support episomal maintenance in *P. tricornutum*. DNA sequences are colored by the organism from which each insert sequence originated. The black dotted box indicates three sequences from *M. mycoides* that failed to support episomes despite being in the 28–30% GC range.

episome maintenance despite having a sufficiently low average GC content ("Myco-No" set). The second set comprised the remaining nine *Mycoplasma* sequences with 28% and 30% average GC that successfully supported episome maintenance ("Myco-Yes" set). The 6 mers that were overrepresented in the Myco-Yes set were characterized by very low GC content (i.e., the most abundant 6 mers in the Myco-Yes set were composed entirely of A+T bases) (*SI Appendix*, Fig. S10). When we directly calculated the number of consecutive A+T nucleotides in the *Mycoplasma* sequences that supported episome maintenance compared with those that did not, stretches of 6 or more consecutive A+T bases were more frequent in the *Mycoplasma* fragments that supported episome maintenance (i.e., Myco-Yes; *SI Appendix*, Table S8). The lower distribution of consecutive A+T bases in the Myco-No set was also observed compared with a set of randomly generated sequences of 30% GC (*SI Appendix*, Table S8). Thus, the Myco-No samples that failed to support episome maintenance appear to have fewer long stretches composed of A+T residues despite having the same average GC content as fragments that supported episome maintenance in *P. tricornutum*.

Discussion

Features of Predicted Diatom Centromeres. In this study, we identified native diatom centromere sequences with high resolution. Based on previous studies, we hypothesized that low-GC content would be a common characteristic of diatom centromeres. We deconstructed two *P. tricornutum* chromosomes (25, 26) and found that regions with low-GC content appeared to function as centromeres, whereas adjacent regions did not. We subsequently conducted a genome-wide ChIP-seq screen (confirmed with ChIP-qPCR) and a forward genetic screen to identify centromeres and additional sequences enabling episome maintenance and used reverse genetics to test for function. We discovered 25 unique *P. tricornutum* centromeric DNA sequences: 24 among the nuclear genome scaffolds and 1 in the nonscaffolded genome assemblies. Although there may potentially be more centromeres we did not identify here, if our results are a good estimate of diatom chromosome number, with one unique centromere sequence each, we would predict that the diatom genome contains fewer chromosomes than the 33 predicted previously (39). Centromere sequences may be erroneously missing from the genome assembly. Additionally, some of the *P. tricornutum* chromosome-scale scaffolds lacking telomere-to-telomere assembly may not be individual chromosomes but rather partial chromosomes (Fig. 3). For example, the putative centromeres identified by ChIP-seq from chromosomes 24 and 29 were nearly identical (99%), and each of these two centromeres was positioned near a scaffold terminus lacking a telomere. Thus, chromosome-scale scaffolds 24 and 29 may be two arms of a single chromosome. In any case, the identification of centromeric DNA sequences will help to develop a better model of *P. tricornutum* genome organization.

Our findings suggest that *P. tricornutum* possesses small monocentric regional centromeres. ChIP-seq peaks were typically found only once per chromosome and corresponded to centromere sequences that were unique to each chromosome. This sequence variability precludes categorization as point centromeres, and the presence of a single peak (with rare exceptions, described below) rather than recruitment of CENH3 across the entire chromosome indicates the absence of a holocentric chromosome. The regional centromere structure is found in most eukaryotes, including the closest related organisms with identified centromeres: the protist *Plasmodium falciparum* and the red alga *C. merolae* (21–24). Both organisms have similarly sized centromeric DNA regions (~2–4 kb) and also share low-GC content as a characteristic of their centromeres: ~3% relative to the genome average GC of 21.8% in the case of *P. falciparum* (21, 22), and 48.4% relative to the genome average GC of 55% for *C. merolae* (24). Interestingly, *C. merolae*, which is the only other alga with well-characterized centromeres and the closest relation to *P. tricornutum* of the organisms studied, has centromeres with a GC content that is low only compared with the genome average and not intrinsically, similar to *P. tricornutum* centromeres.

Two *P. tricornutum* chromosomes, 2 and 8, appeared to deviate from the monocentric model by having two sequences identified by the ChIP-seq analysis. The regions adjacent to the centromeres on the chromosome scaffolds are unresolved DNA sequences, and both centromere regions contained long direct repeats. Thus, sample processing, sequencing, or assembly error could be responsible for the apparent duplication of the centromere on these chromosomes. Alternatively, these may be true centromeres that have simply been duplicated. The presence of a nearby retrotransposon may support this theory and could also confound PCR assays (*SI Appendix*, Fig. S5). Dicentric chromosomes have been noted in several organisms; however, typically only one of the centromeres is active and the other is inactivated (13, 49–52). The presence of two active centromeres typically leads to chromosomal breakage followed by either cell death or two functional monocentric chromosomes. Chromosomes with multiple functional centromeres have been identified. In human cells, two active centromeres were in close proximity, essentially behaving as a single centromere (53). In rice, recombinant centromeres were found to contain two repetitive arrays; both recruited CENH3, whereas an intervening sequence did not (54). Additionally, tricentric chromosomes were identified in wheat where one of the centromeres was large and presumably dominant, and co-occurring centromeres were smaller and weaker (55).

Like *P. tricornutum*, the diatom *T. pseudonana* can also use the yeast-derived *CEN6–ARSH4–HIS3* sequence to maintain episomes (37), which may suggest an overall similarity in DNA maintenance mechanisms. We analyzed the GC content of the *T. pseudonana* genome and found similar regions of low-GC content that were often found once per chromosome-scale scaffold (*SI Appendix*, Fig. S11). Thus, the ability of the yeast *CEN6–ARSH4–HIS3* sequence to support episomal maintenance in both species may be due to similar requirements for low-GC sequences to function as centromeres. It is remarkable that these diatoms may have such similar centromere features, to the degree that the same sequence can function as a centromere in both organisms, given the ancient evolutionary divergence of the centric and pennate diatom lineages (~90 Mya) (39) and the relatively rapid evolution of centromere sequences and structures observed for other groups of organisms (56, 57). Further CENH3 ChIP-seq experiments in *T. pseudonana* will enable centromere identification and comparison with *P. tricornutum*, including an examination of evolutionary implications.

Simple Centromere Requirements Permit Nuclear Maintenance of Nonnuclear DNA Sequences. In this study, by identifying characteristics of native diatom centromere sequences, we have uncovered a mechanism by which foreign DNA can become part of the nuclear DNA repertoire; nonnuclear DNA can act as a centromere, enabling stable maintenance as an extrachromosomal nuclear episome. Maintaining plasmids could expand the diatom's biochemical and physiological potential provided the new DNA acquired the necessary regulatory features over time and may also facilitate permanent integration into the native nuclear chromosomes through chromosomal rearrangements. We previously observed that DNA sequences from the yeast *S. cerevisiae* could enable episome maintenance in *P. tricornutum* (37, 38), and in this study, we confirmed that this sequence does, in fact, recruit the *P. tricornutum* centromeric histone protein CENH3. The recruitment of this centromere-specific histone protein and subsequent maintenance of the episome in diatoms suggests the foreign DNA sequence is using native diatom DNA replication machinery, essentially functioning as a diatom centromere. There are very few examples in eukaryotes of foreign

DNA recruiting host CENH3 to maintain a chromosome. Human centromeres have previously been shown to function in mouse chromosomes (58), and in a recent example, *Arabidopsis* centromeric repeats were shown to recruit human CENH3 and maintain chromosomes in human cells (59). In both cases, the chromosomes maintained by foreign DNA originally derived from chimeric host-donor DNA chromosomes followed by chromosomal breakage and/or rearrangement, resulting in smaller linear chromosomes or "mini-chromosomes." To our knowledge, there are no examples of immediate nuclear genome establishment (i.e., without chimeric intermediates) and maintenance in the host cell as a plasmid. This contrasts with bacteria, where DNA transfer between bacteria and subsequent plasmid establishment is quite common. Our results suggest that nonnuclear DNA can mimic diatom centromeres and, along with colocalized DNA, can immediately establish circular chromosomes in the diatom genome.

Establishment of centromeres in *P. tricornutum* is governed by an apparently simple rule: a small length of sequence (>500 bp) with a GC content less than ∼33%. For foreign DNA sequences examined and for fragments further subdividing the native chromosome 25 centromere, 500 bp was typically a sufficient length for episome maintenance. Despite testing several shorter sequences from *P. tricornutum* chromosome 25 (*SI Appendix*, Fig. S8 and Table S5), only those >500 bp could maintain episomes, a pattern also observed for yeast-derived sequences (38). Notably, 500 bp is a particularly short sequence to enable centromere function compared with previously studied organisms with regional centromeres; most regional centromeres are reported to be thousands of base pairs in length, compared with the relatively small (∼125 bp) point centromeres of some yeast species. Although ChIP-seq peaks for centromeres averaged 39% GC over the entire 2–5-kb sequence, each centromeric ChIP-seq peak contained within it a 500-bp region less than ∼33% GC. Likewise, all foreign sequences that maintained episomes contained a 500-bp region with GC content beneath this threshold. The exceptions were three *Mycoplasma* sequences with GC less than 33% did not support episomal maintenance. We observed that these sequences had lower frequencies of 6 or more consecutive A+T bases, a pattern which persisted compared with a randomly generated set of 30% GC sequences. Although there may be something unique about these *Mycoplasma* DNA sequences generally, it is also possible that the frequency or spacing of longer contiguous A+T sequences or a similar signal may play an important functional role in diatom centromeric DNA, and sequences of <33% GC content usually, but not always, happen to contain these signals. Alternatively, certain sequences, perhaps consisting of high-GC content stretches, may actually interrupt centromere formation in otherwise low-GC content DNA sequences. Thus, rather than A+T stretches defining centromere function, the key feature may be a lack of interrupting sequences. Each of these hypotheses remains to be examined further in future studies.

The permissiveness of sequences that can function as centromeres in our organism may suggest that de novo centromere formation is quite common in diatoms. Two mechanisms of de novo centromere formation are of particular interest in this study: the acquisition of entirely new centromeres from nonnuclear DNA sources, and the potential presence of neocentromeres already present in the genome that can function as centromeres under certain circumstances. Transfer of novel centromere sequences from intracellular (i.e., chloroplast or mitochondrial genome) or extracellular DNA sources into the nucleus could alter the genome in multiple ways. Nonnuclear "centromeric" DNA sequences accompanied by other DNA could possibly form the basis of entire new chromosomes, similar to what we observed with episomes containing both native and foreign DNA sequences. Alternatively, foreign DNA sequences possessing centromeric DNA similarity could integrate into the nuclear chromosomes. Because multiple active centromeres typically lead to chromosomal instability, this could restructure the nuclear genome landscape by causing either a loss or gain in chromosome number as well as other rearrangements during the DNA repair process (reviewed in ref. 13). An alternative possibility is that after integration, one centromere sequence is silenced, which is one mechanism observed to prevent instability in dicentric chromosomes (60, 61). However, our observation of only one centromere-like DNA sequence per chromosome based on ChIP-seq experiments would suggest that this is not the case, at least in recent diatom evolutionary history. Neocentromeres are nuclear genome sequences distinct from centromere sequences but that can become active centromeres and recruit CENH3 when centromeres are inactivated or absent due to chromosomal rearrangements (reviewed in refs. 62–64). We did not observe any noticeable patterns of increased CENH3 recruitment outside of the centromere regions, and when we functionally tested several regions with elevated-background CENH3 recruitment, we found none of them functioned as centromeres. Our data would suggest that if there are in fact neocentromeres in this species, they do not recruit CENH3 before activation and would need to be discovered by altering the centromeric DNA region directly or through identification in genomes of aberrant phenotypes.

The observation that low-GC content sequences can act as *P. tricornutum* centromeres may help to explain the transfer of DNA from diatom endosymbiont or organelle genomes into the nuclear genome, which represent major sources of diatom nuclear DNA throughout evolutionary history (39–41). *P. tricornutum* chloroplast and mitochondrial genomes are low in GC content (32% average GC for the chloroplast, 35% average GC for the mitochondria), and we identified multiple sequences from each that could maintain nuclear episomes (*SI Appendix*, Table S4). Endosymbiotic gene transfer (EGT) from anciently acquired bacterial-derived organelles, namely the mitochondria (derived from a proteobacterium) and chloroplast (derived from a cyanobacterium), make up the majority of horizontally transferred genes found in eukaryotic genomes (42, 65, 66). DNA transfer from plastid and mitochondrial genomes to the nucleus, which include noncoding nuclear plastid DNA's (NUPTs) and nuclear mitochondrial DNAs (NUMTs), occurred at the time of endosymbiont acquisition and later in evolutionary time and likely occurs quite frequently in present time (65, 67–71). This is particularly true for photosynthetic organisms, where the majority of genes required for plastid function are actually housed in the nuclear genome, with a greatly reduced plastid genome from the original acquired state (72, 73). For example, in *Arabidopsis* nearly 18% of nuclear genes were found to be of cyanobacterial origin (74), whereas similar findings of 6–12% cyanobacterial-derived genes have been found to comprise unicellular-algae nuclear genomes (75–77). In these instances, DNA transferred to the recipient nucleus was already inside the host cell. Diatoms, like other stramenopiles, are the result of serial endosymbiotic events, though the precise details are still debated. However, it is generally accepted that the complicated series of whole-organism engulfments by many algal species, followed by transfer of both organelle and nuclear genes to the new host nucleus, has resulted in exceptional complex chimeric nuclear genomes. As new efforts attempt to solve these genomic puzzles, it is also important to understand exactly how these gene transfers occur, a subject that itself is not well understood. Thus, identifying mechanisms that could facilitate nuclear gene acquisition, such as the one proposed in this study, can shed light on algal diversity and evolution.

Although the majority of lateral gene transfer from bacteria to eukaryotes is thought to be via EGT, most unicellular algal genomes studied to date also contain a surprising amount of bacterial and viral DNA. This is particularly true for the diatom *P. tricornutum*, and another well-studied diatom *T. pseudonana*,

which are thought to possess a surprisingly high number of non-EGT horizontally acquired genes (39, 41). This assessment was based on the uniqueness of foreign genes possessed by each of these diatom species as well as gene-specific molecular phylogenies. Non-EGT–derived foreign genes in the *P. tricornutum* genome were found more frequently (by an order of magnitude) than in other free-living eukaryotes and were estimated to comprise about 5% of the nuclear genome, leading to the suggestion that horizontal transfer of bacterial DNA into diatoms may be quite common (39). Little is known about the acquisition of this DNA and how it ultimately integrates into the nuclear genome. Recent and/or transient endosymbiosis may be a possibility. Additionally, the discovery that diatoms are amenable to bacterial conjugation (37), the method of gene transfer used experimentally in the present study, provides a potential mechanism for exogenous DNA transfer, though this has yet to be demonstrated in a natural setting. Foreign DNA can also enter algal cells through viral infection (78, 79), an emerging area of algal research. The presence of this apparently non-EGT DNA in algal genomes further emphasizes the importance of understanding both how new genetic material gets into eukaryotes and how it stays there.

It is unclear why maintenance of foreign DNA in the form of episomes appears to be well tolerated in *P. tricornutum*. One possibility is that transfer of foreign DNA into diatoms, or intracellular transfer of previously acquired nonnuclear genetic material, is not common enough for a defense system to have evolved (such as the production of restriction enzymes in bacteria to destroy foreign DNA). In contrast to bacteria–bacteria DNA transfer, nonnative genes are unlikely to be expressed from a plasmid transferred to a diatom if they are of bacterial origin. Functional gene expression would only occur in the unlikely event that it acquired diatom transcriptional, translational, and subcellular localization signals through further modification. Thus, it is possible there was not strong selection to evolve defense mechanisms against foreign DNA because they were not detrimental to cell fitness and most events were entirely innocuous. If such permissiveness occurs for maintenance of DNA transferred through extracellular mechanisms, it is likely that it would also apply to DNA transferred to the nucleus intracellularly from organelles to the nucleus.

Conclusions

Identifying and characterizing centromeres is essential for understanding cellular biology, as these are critical features for stable DNA maintenance during cell division. These sequences can also advance synthetic biology through the development of new molecular tools. Here, we have used multiple approaches to characterize the centromeres of the diatom *P. tricornutum.* We found very simple sequence requirements for DNA to function as a centromere, namely a moderately low-GC content of <33% across a small region. Although most sequences with a GC content of <33% allowed episomal maintenance, a few sequences did not; these contained a lower frequency of contiguous A+T base stretches compared with functional sequences, indicating that more specific sequence characteristics could potentially play a role in centromere formation. Based on bioinformatic analyses, we predict that these features of centromere identity may be conserved in the distantly related diatom *T. pseudonana*. Although low-GC content has often been identified as a centromeric DNA feature, the diatom centromeres appear to be unique from many other eukaryotes in that they are not composed of repeat regions or other notable primary structures and that the functional centromere region may be quite small. We also show that these simple requirements allow foreign and nonnuclear DNA sequences with these characteristics to act as centromeres in diatoms, enabling establishment as extrachromosomal nuclear episomes. Diatoms possess nuclear genes acquired from many foreign DNA sources including viruses, bacteria, and other eukaryotes, including nuclear DNA acquired as a result of the ancient endosymbiotic acquisition of mitochondria and chloroplasts. Our findings present a host-permissive mechanism by which DNA derived from either external or intracellular genetic pools can persist in the diatom nucleus by using host replication and maintenance machinery. This may ultimately lead to gene integration into diatom genomes and subsequent evolutionary diversification.

Materials and Methods

A description of the strains used in this study, culturing conditions, and detailed explanations of the methods used can be found in *SI Appendix*. Briefly, we used *P. tricornutum* strain CCMP 632 (synonymous with the genome-sequenced strain CCMP2561) (39) and conducted ex-conjugant selection on phleomycin antibiotic. To conduct ChIP-seq and ChIP-qPCR assays, we constructed the plasmid pPtPBR1-YFP-CENH3 to express a YFP-CENH3 fusion protein. We confirmed protein expression and nuclear localization by confocal microscopy and Western blot analysis, respectively. Sample preparation and data analyses for the ChIP-seq experiments were conducted as previously described (80), using the Illumina sequencing platform. Sequences were deposited to the NCBI Sequence Read Archive (SRA) with accession no. PRJNA357294. For the *P. tricornutum* genomic library, purified PCR products of inserts were sequenced using Sanger DNA sequencing. Bacteria to diatom conjugations and episome maintenance analyses were conducted as previously described (37, 38).

ACKNOWLEDGMENTS. We thank John McCrow for many helpful discussions concerning bioinformatics and Sarah Smith for assistance with ChIP-seq. Funding for this work was provided by Gordon and Betty Moore Foundation Grants GBMF5007 (to P.D.W. and C.L.D.) and GBMF3828 and GBMF5006 (to A.E.A.), US Department of Energy Grant DE- SC0008593 (to A.E.A. and C.L.D.), and National Science Foundation Grants NSF-MCB-1129303 (to C.L.D.) and OCE-1136477 to (A.E.A.).

1. Pluta AF, Mackay AM, Ainsztein AM, Goldberg IG, Earnshaw WC (1995) The centromere: Hub of chromosomal activities. *Science* 270:1591–1594.
2. Westermann S, et al. (2003) Architecture of the budding yeast kinetochore reveals a conserved molecular core. *J Cell Biol* 163:215–222.
3. Cheeseman IM, Desai A (2008) Molecular architecture of the kinetochore-microtubule interface. *Nat Rev Mol Cell Biol* 9:33–46.
4. Earnshaw WC, et al. (2013) Esperanto for histones: CENP-A, not CenH3, is the centromeric histone H3 variant. *Chromosome Res* 21:101–106.
5. Westhorpe FG, Straight AF (2014) The centromere: Epigenetic control of chromosome segregation during mitosis. *Cold Spring Harb Perspect Biol* 7:a015818.
6. McKinley KL, Cheeseman IM (2016) The molecular basis for centromere identity and function. *Nat Rev Mol Cell Biol* 17:16–29.
7. Henikoff S, Ahmad K, Malik HS (2001) The centromere paradox: Stable inheritance with rapidly evolving DNA. *Science* 293:1098–1102.
8. Smith KM, Galazka JM, Phatale PA, Connolly LR, Freitag M (2012) Centromeres of filamentous fungi. *Chromosome Res* 20:635–656.
9. Cleveland DW, Mao Y, Sullivan KF (2003) Centromeres and kinetochores: From epigenetics to mitotic checkpoint signaling. *Cell* 112:407–421.
10. Cottarel G, Shero JH, Hieter P, Hegemann JH (1989) A 125-bp *CEN6* DNA fragment is sufficient for complete meiotic and mitotic centromere functions in *Saccharomyces cerevisiae*. *Trends Genet* 5:322–324.
11. Albertson DG, Thomson JN (1982) The kinetochores of *Caenorhabditis elegans*. *Chromosoma* 86:409–428.
12. Sullivan BA, Blower MD, Karpen GH (2001) Determining centromere identity: Cyclical stories and forking paths. *Nat Rev Genet* 2:584–596.
13. Torras-Llort M, Moreno-Moreno O, Azorín F (2009) Focus on the centre: The role of chromatin on the regulation of centromere identity and function. *EMBO J* 28:2337–2348.
14. Willard HF (1998) Centromeres: The missing link in the development of human artificial chromosomes. *Curr Opin Genet Dev* 8:219–225.
15. Tyler-Smith C, et al. (1993) Localization of DNA sequences required for human centromere function through an analysis of rearranged Y chromosomes. *Nat Genet* 5: 368–375.
16. Ma J, Wing RA, Bennetzen JL, Jackson SA (2007) Plant centromere organization: A dynamic structure with conserved functions. *Trends Genet* 23:134–139.
17. Clarke L, Amstutz H, Fishel B, Carbon J (1986) Analysis of centromeric DNA in the fission yeast *Schizosaccharomyces pombe*. *Proc Natl Acad Sci USA* 83:8253–8257.
18. Nakaseko Y, Adachi Y, Funahashi S, Niwa O, Yanagida M (1986) Chromosome walking shows a highly homologous repetitive sequence present in all the centromere regions of fission yeast. *EMBO J* 5:1011–1021.
19. Kapoor S, Zhu L, Froyd C, Liu T, Rusche LN (2015) Regional centromeres in the yeast *Candida lusitaniae* lack pericentromeric heterochromatin. *Proc Natl Acad Sci USA* 112: 12139–12144.

20. Lynch DB, Logue ME, Butler G, Wolfe KH (2010) Chromosomal G + C content evolution in yeasts: Systematic interspecies differences, and GC-poor troughs at centromeres. *Genome Biol Evol* 2:572–583.
21. Bowman S, et al. (1999) The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* 400:532–538.
22. Iwanaga S, et al. (2010) Functional identification of the *Plasmodium* centromere and generation of a *Plasmodium* artificial chromosome. *Cell Host Microbe* 7:245–255.
23. Maruyama S, et al. (2008) Centromere structures highlighted by the 100%-complete *Cyanidioschyzon merolae* genome. *Plant Signal Behav* 3:140–141.
24. Kanesaki Y, Imamura S, Matsuzaki M, Tanaka K (2015) Identification of centromere regions in chromosomes of a unicellular red alga, *Cyanidioschyzon merolae*. *FEBS Lett* 589:1219–1224.
25. Yu W, Yau YY, Birchler JA (2016) Plant artificial chromosome technology and its potential application in genetic engineering. *Plant Biotechnol J* 14:1175–1182.
26. Kouprina N, Tomilin AN, Masumoto H, Earnshaw WC, Larionov V (2014) Human artificial chromosome-based gene delivery vectors for biomedicine and biotechnology. *Expert Opin Drug Deliv* 11:517–535.
27. Monaco AP, Larin Z (1994) YACs, BACs, PACs and MACs: Artificial chromosomes as research tools. *Trends Biotechnol* 12:280–286.
28. Coudreuse D (2009) Insights from synthetic yeasts. *Yeast* 33:483–492.
29. Murray AW, Szostak JW (1983) Construction of artificial chromosomes in yeast. *Nature* 305:189–193.
30. Harrington JJ, Van Bokkelen G, Mays RW, Gustashaw K, Willard HF (1997) Formation of de novo centromeres and construction of first-generation human artificial microchromosomes. *Nat Genet* 15:345–355.
31. Liu W, Yuan JS, Stewart CN, Jr (2013) Advanced genetic tools for plant biotechnology. *Nat Rev Genet* 14:781–793.
32. Iwanaga S, Kato T, Kaneko I, Yuda M (2012) Centromere plasmid: A new genetic tool for the study of *Plasmodium falciparum*. *PLoS One* 7:e33326.
33. Rochaix JD, van Dillewijn J, Rahire M (1984) Construction and characterization of autonomously replicating plasmids in the green unicellular alga *Chlamydomonas reinhardii*. *Cell* 36:925–931.
34. Lopez PJ, Desclés J, Allen AE, Bowler C (2005) Prospects in diatom research. *Curr Opin Biotechnol* 16:180–186.
35. Bozarth A, Maier UG, Zauner S (2009) Diatoms in biotechnology: Modern tools and applications. *Appl Microbiol Biotechnol* 82:195–201.
36. Fu W, Wichuk K, Brynjólfsson S (2015) Developing diatoms for value-added products: Challenges and opportunities. *N Biotechnol* 32:547–551.
37. Karas BJ, et al. (2015) Designer diatom episomes delivered by bacterial conjugation. *Nat Commun* 6:6925.
38. Diner RE, Bielinski VA, Dupont CL, Allen AE, Weyman PD (2016) Refinement of the diatom episome maintenance sequence and improvement of conjugation-based DNA delivery methods. *Front Bioeng Biotechnol* 4:65.
39. Bowler C, et al. (2008) The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* 456:239–244.
40. Armbrust EV (2009) The life of diatoms in the world's oceans. *Nature* 459:185–192.
41. Armbrust EV, et al. (2004) The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. *Science* 306:79–86.
42. Timmis JN, Ayliffe MA, Huang CY, Martin W (2004) Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet* 5:123–135.
43. Clarke L, Carbon J (1980) Isolation of a yeast centromere and construction of functional small circular chromosomes. *Nature* 287:504–509.
44. Karas BJ, et al. (2013) Assembly of eukaryotic algal chromosomes in yeast. *J Biol Eng* 7:30.
45. Jacobs JD, et al. (1992) Characterization of two circular plasmids from the marine diatom *Cylindrotheca fusiformis*: Plasmids hybridize to chloroplast and nuclear DNA. *Mol Gen Genet* 233:302–310.
46. Hildebrand M, et al. (1992) Nucleotide sequence of diatom plasmids: Identification of open reading frames with similarity to site-specific recombinases. *Plant Mol Biol* 19: 759–770.
47. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.
48. Bailey TL, et al. (2009) MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res* 37:W206–W208.
49. Stimpson KM, Matheny JE, Sullivan BA (2012) Dicentric chromosomes: Unique models to study centromere function and inactivation. *Chromosome Res* 20:595–605.
50. Neumann P, et al. (2012) Stretching the rules: Monocentric chromosomes with multiple centromere domains. *PLoS Genet* 8:e1002777.
51. Cuacos M, H Franklin FC, Heckmann S (2015) Atypical centromeres in plants-what they *can* tell us. *Front Plant Sci* 6:913.
52. Sato H, Masuda F, Takayama Y, Takahashi K, Saitoh S (2012) Epigenetic inactivation and subsequent heterochromatinization of a centromere stabilize dicentric chromosomes. *Curr Biol* 22:658–667.
53. Sullivan BA, Willard HF (1998) Stable dicentric X chromosomes with two functional centromeres. *Nat Genet* 20:227–228.
54. Wang G, Li H, Cheng Z, Jin W (2013) A novel translocation event leads to a recombinant stable chromosome with interrupted centromeric domains in rice. *Chromosoma* 122: 295–303.
55. Zhang W, Friebe B, Gill BS, Jiang J (2010) Centromere inactivation and epigenetic modifications of a plant chromosome with three functional centromeres. *Chromosoma* 119:553–563.
56. Malik HS, Henikoff S (2002) Conflict begets complexity: The evolution of centromeres. *Curr Opin Genet Dev* 12:711–718.
57. Malik HS, Henikoff S (2009) Major evolutionary transitions in centromere complexity. *Cell* 138:1067–1082.
58. Hadlaczky G, et al. (1991) Centromere formation in mouse cells cotransformed with human DNA and a dominant marker gene. *Proc Natl Acad Sci USA* 88:8106–8110.
59. Wada N, et al. (2016) Maintenance and function of a plant chromosome in human cells. *ACS Synth Biol* 6:301–310.
60. Agudo M, et al. (2000) A dicentric chromosome of *Drosophila melanogaster* showing alternate centromere inactivation. *Chromosoma* 109:190–196.
61. Faulkner NE, Vig B, Echeverri CJ, Wordeman L, Vallee RB (1998) Localization of motor-related proteins and associated complexes to active, but not inactive, centromeres. *Hum Mol Genet* 7:671–677.
62. Burrack LS, Berman J (2012) Neocentromeres and epigenetically inherited features of centromeres. *Chromosome Res* 20:607–619.
63. Stimpson KM, Sullivan BA (2010) Epigenomics of centromere assembly and function. *Curr Opin Cell Biol* 22:772–780.
64. Scott KC, Sullivan BA (2014) Neocentromeres: A place for everything and everything in its place. *Trends Genet* 30:66–74.
65. Martin W, Herrmann RG (1998) Gene transfer from organelles to the nucleus: How much, what happens, and Why? *Plant Physiol* 118:9–17.
66. Ku C, et al. (2015) Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* 524:427–432.
67. Ju YS, et al.; ICGC Prostate Cancer Working Group; ICGC Bone Cancer Working Group; ICGC Breast Cancer Working Group (2015) Frequent somatic transfer of mitochondrial DNA into the nuclear genome of human cancer cells. *Genome Res* 25:814–824.
68. Huang CY, Ayliffe MA, Timmis JN (2003) Direct measurement of the transfer rate of chloroplast DNA into the nucleus. *Nature* 422:72–76.
69. Stegemann S, Hartmann S, Ruf S, Bock R (2003) High-frequency gene transfer from the chloroplast genome to the nucleus. *Proc Natl Acad Sci USA* 100:8828–8833.
70. Martin W (2003) Gene transfer from organelles to the nucleus: Frequent and in big chunks. *Proc Natl Acad Sci USA* 100:8612–8614.
71. Kleine T, Maier UG, Leister D (2009) DNA transfer from organelles to the nucleus: The idiosyncratic genetics of endosymbiosis. *Annu Rev Plant Biol* 60:115–138.
72. Archibald JM (2015) Genomic perspectives on the birth and spread of plastids. *Proc Natl Acad Sci USA* 112:10147–10153.
73. Archibald JM (2009) The puzzle of plastid evolution. *Curr Biol* 19:R81–R88.
74. Martin W, et al. (2002) Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci USA* 99:12246–12251.
75. Moustafa A, Bhattacharya D (2008) PhyloSort: A user-friendly phylogenetic sorting tool and its application to estimating the cyanobacterial contribution to the nuclear genome of *Chlamydomonas*. *BMC Evol Biol* 8:6.
76. Price DC, et al. (2012) *Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants. *Science* 335:843–847.
77. Reyes-Prieto A, Hackett JD, Soares MB, Bonaldo MF, Bhattacharya D (2006) Cyanobacterial contribution to algal nuclear genomes is primarily limited to plastid functions. *Curr Biol* 16:2320–2325.
78. Suttle CA (2007) Marine viruses–Major players in the global ecosystem. *Nat Rev Microbiol* 5:801–812.
79. Brussaard CPD (2004) Viral control of phytoplankton populations–A review. *J Eukaryot Microbiol* 51:125–138.
80. Lin X, et al. (2012) Protocol: Chromatin immunoprecipitation (ChIP) methodology to investigate histone modifications in two model diatom species. *Plant Methods* 8:48.