Long transposon-rich centromeres in an oomycete reveal divergence of centromere features in Stramenopila-Alveolata-Rhizaria lineages

Yufeng "Francis" Fang[a], Marco A. Coelho[a], Haidong Shu[b], Klaas Schotanus[a], Bhagya C. Thimmappa[c,1], Vikas Yadav[a], Han Chen[b], Ewa P. Malc[d], Jeremy Wang[d], Piotr A. Mieczkowski[d], Brent Kronmiller[e], Brett M. Tyler[e], Kaustuv Sanyal[c], Suomeng Dong[b], Minou Nowrousian[f], and Joseph Heitman[a,2]

[a]Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, North Carolina 27710; [b]College of Plant Protection, Nanjing Agricultural University, Nanjing, China; [c]Molecular Biology and Genetics Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India; [d]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina 27599; [e]Center for Genome Research and Biocomputing and Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon 97331; [f]Lehrstuhl fuer Molekulare und Zellulaere Botanik, Ruhr-Universitaet Bochum, Bochum, Germany

[1]Present address: Department of Biochemistry, Robert-Cedergren Centre for Bioinformatics and Genomics, University of Montreal, 2900 Edouard-Montpetit, Montreal, H3T1J4, QC, Canada.

[2]To whom correspondence should be addressed. Email: heitm001@duke.edu.

Abstract

Centromeres are chromosomal regions that serve as platforms for kinetochore assembly and spindle attachments, ensuring accurate chromosome segregation during cell division. Despite functional conservation, centromere DNA sequences are diverse and often repetitive, making them challenging to assemble and identify. Here, we describe centromeres in an oomycete *Phytophthora sojae* by combining long-read sequencing-based genome assembly and chromatin immunoprecipitation for the centromeric histone CENP-A followed by high-throughput sequencing (ChIP-seq). *P. sojae* centromeres cluster at a single focus at different life stages and during nuclear division. We report an improved genome assembly of the *P. sojae* reference strain, which enabled identification of 15 enriched CENP-A binding regions as putative centromeres. By focusing on a subset of these regions, we demonstrate that centromeres in *P. sojae* are regional, spanning 211 to 356 kb. Most of these regions are transposon-rich, poorly transcribed, and lack the histone modification H3K4me2 but are embedded within regions with the heterochromatin marks H3K9me3 and H3K27me3. Strikingly, we discovered a Copia-like transposon (CoLT) that is highly enriched in the CENP-A chromatin. Similar clustered elements are also found in oomycete relatives of *P. sojae*, and may be applied as a criterion for prediction of oomycete centromeres. This work reveals a divergence of centromere features in oomycetes as compared to other organisms in the Stramenopila-Alveolata-Rhizaria (SAR) supergroup including diatoms and *Plasmodium falciparum* that have relatively short and simple regional centromeres. Identification of *P. sojae* centromeres in turn also augments the genome assembly.

Key words: Stramenopila; genome assembly; CENP-A; Copia-like transposon

Significance Statement

Oomycetes are fungal-like microorganisms that belong to the stramenopiles within the Stramenopila-Alveolata-Rhizaria (SAR) supergroup. The *Phytophthora* oomycetes are infamous as plant killers, threatening crop production worldwide. Because of the highly repetitive nature of their genomes, assembly of oomycete genomes presents challenges that impede identification of centromeres, which are chromosomal sites mediating faithful chromosome segregation. We report long-read sequencing-based genome assembly of the *Phytophthora sojae* reference strain, which facilitated the discovery of centromeres. *P. sojae* harbors large regional centromeres enriched for a Copia-like transposon that is also found in discrete clusters in other oomycetes. This study provides insight into the oomycete genome organization, and broadens our knowledge of the centromere structure, function and evolution in eukaryotes.

Introduction

Accurate segregation of chromosomes during mitosis and meiosis is critical for the development and reproduction of all eukaryotic organisms. Centromeres are specialized regions of chromosomes that mediate kinetochore formation, spindle attachment, and sister chromatid segregation during cell division (1, 2). The DNA coincident with functional centromeres typically consists of unusual sequence composition (e.g. AT-rich) and structure (e.g. repeats, transposable elements), low gene density, and transcription of non-coding RNA (ncRNA) as well as heterochromatic nature (3). However, an active centromere is defined not by DNA sequences but by the deposition of a centromere-associated protein called centromere protein A (CENP-A, also known as CenH3) (1, 4). CENP-A is a histone H3 variant, which replaces the canonical H3 in the nucleosomes at centromeres and provides the foundation for kinetochore assembly (1, 5, 6).

Despite the fact that centromere function is broadly conserved, centromeric sequences vary greatly in size and composition, ranging from “point” centromeres of 125 bp in length to “regional” centromeres consisting of up to megabases of repeated sequences to holocentromeres that extend along the entire length of the chromosome (1, 3). To date, point centromeres have been only reported in the budding yeast *Saccharomyces cerevisiae* and its close relatives, holocentromeres have been identified in some insects, plants and nematodes, represented by *Caenorhabditis elegans*, while regional centromeres are the most common type and found in nearly all eukaryotic phyla (1, 3). Most animals and plants have large regional centromeres composed of satellite sequences that are organized into a variety of different higher order repeats (4, 7, 8). Some plant centromeres also possess a different type of repeat called centromere-specific retroelements (CR) (9). In comparison, all fungal centromeres identified to date do not contain satellite repeats and have diverse organizations. The size of fungal regional centromeres ranges from several kilobases, such as in *Candida albicans*, to hundreds of kilobases in *Neurospora crassa* (10, 11). The centromeric sequences of fungal regional centromeres can be composed of active or inactive clusters of transposable elements and thus very repetitive, such as in *Cryptococcus* spp. and *N. crassa* (12, 13), or can be nonrepetitive and very short, such as in the wheat pathogen *Zymoseptoria tritici* (14) and *C. albicans* (15). Information on centromeres is limited in other eukaryotic lineages. The malaria pathogen *Plasmodium falciparum* and the diatom *Phaeodactylum tricornutum* CENP-A binding regions are characterized by short simple AT-rich sequences (16, 17), while the parasite *Toxoplasma gondii* has a simple centromere without nucleotide bias (18).

Due to their highly repetitive nature, assembly of large regional centromeres presents a significant challenge. Emerging long-read sequencing technologies, such as Pacific Bioscience (PacBio) and Oxford Nanopore Technology (ONT), have led to substantial advances in resolution of chromosomal structures including highly repetitive sequences such as centromeres. Using these technologies, centromeres that were difficult to resolve using short-read sequencing, were defined in various organisms, from fungi (12, 19, 20) to insects (21), plants (22) and humans (23).

Oomycetes are fungal-like organisms but belong to the stramenopila kingdom within the Stramenopila-Alveolata-Rhizaria (SAR) supergroup (24, 25). The SAR supergroup contains a high diversity of lineages that include many important photosynthetic lineages (e.g. diatoms and kelp), and important parasites of animals (e.g., *Plasmodium*, the causative agent of malaria) and plants (e.g., oomycetes, or water molds) (26). *Phytophthora* is a large oomycete genus (>160 species found to date) and contains some of the most devastating plant pathogens that destroy a wide range of plants important in agriculture, forestry, ornamental and recreational plantings, and natural ecosystems (27). One notorious example is *Phytophthora infestans*, which caused the great Irish potato famine of the mid-1840s (28). Today, *Phytophthora* species remain significant threats to major food crops, causing multi-billion US dollars losses annually throughout the world (27, 29). *Phytophthora sojae* is a widespread soil-borne pathogen of soybean. Because of its economic impact, and tractable genetic manipulation (30-32), *P. sojae* has become a model species to study oomycete genetics, biology, and interactions with plants.

To date, the genomes of more than 20 *Phytophthora* species have been sequenced (33). Their genomes are generally large and display complex features: they are diploid, highly heterozygous for heterothallic species, and very repetitive, which makes genome assembly challenging. The most contiguous oomycete genome assembly published to date is of the *P. sojae* reference genome, which was generated based on Sanger random shotgun sequencing and subsequent improvements involving gap closure and BAC sequencing (25, 34). *P. sojae* genome assembly v3.0 (www.jgi.doe.gov) spans ~82 Mb and contains 82 scaffolds; however, there are ~3 Mb of unresolved gaps (N's) persisting in the assembly. Recently, significant progress has been made in genome assemblies of oomycetes based on long-read sequencing (35, 36); however, the identity or the nature of the DNA sequences that form essential chromosomal elements such as centromeres, remain unknown. In this study, using the evolutionarily conserved kinetochore protein CENP-A as a tool, we investigated cellular dynamics of the kinetochore complex in *P. sojae*, and uncovered the nature of the oomycete centromeres with the aid of long-read genome sequencing and ChIP-seq technologies. Our findings suggest that the centromeres of *P. sojae* are divergent from those reported in other SAR lineages, and their features may be used to predict centromeres in other oomycetes.

Results

GFP-tagging of CENP-A in *P. sojae* reveals clustered centromeres in different life stages and throughout hyphal growth

Kinetochore protein homologs have been predicted in diverse eukaryotic lineages including oomycete species (37). To identify kinetochore proteins in *P. sojae*, we conducted BLAST searches against the existing *P. sojae* genome database using the predicted oomycete orthologs as query. Gene models of *P. sojae* kinetochore proteins were examined and corrected based on RNA-seq data when necessary. Protein sequences were verified based on the presence of corresponding motifs (Fig. S1 and Dataset S1).

To examine centromere/kinetochore organization and localization in *P. sojae*, we selected CENP-A, the hallmark of centromere identity in most organisms. The RNA-seq data did not support the gene models of *CENP-A* that was instead verified by 3'-RACE and RT-PCR, followed by Sanger sequencing (Fig. S2 *A* and *B*). *P. sojae* CENP-A (PsCENP-A) has a conserved C-terminus including the "CENP-A targeting domain" (CATD) (Fig. S2*C*). *GFP* was fused to *CENP-A* at the N-terminus and transiently expressed in *P. sojae* transformants with a constitutive promoter derived from the *Bremia lactucae HAM34* gene (Fig. S2*D*). Overexpressed GFP-CENP-A exhibited nuclear localization with a single fluorescent focus in the nucleus (Fig. S2*D*), suggesting that *P. sojae* has a clustered centromere organization.

We also generated GFP labeled CENP-A expressed from the endogenous locus utilizing CRISPR/Cas9-mediated gene replacement (Figs. 1A and S3). Homokaryotic GFP-CENP-A strains exhibited single GFP foci within nuclei from different *P. sojae* life stages (Figs. 1B), confirming that the clustered centromere organization is a feature in *P. sojae.* In addition, we tracked the centromere dynamics during hyphal growth. Intriguingly, the clustered centromere pattern was maintained throughout *P. sojae* nuclear division (Fig. 1C and Movie S1).

Identification of centromeres in a long-read Nanopore-based assembly

To identify *P. sojae* centromeres, we performed native chromatin immunoprecipitation (N-ChIP) using an anti-GFP antibody against the GFP-CENP-A fusion, followed by high-throughput Illumina DNA sequencing. ChIP-seq reads were mapped to the latest Sanger genome assembly (*P. sojae* V3 from JGI), which identified 12 scaffolds that showed relatively concentrated enrichment of CENP-A reads (Fig. S4*A*). CENP-A peaks appeared scattered in Scaffold 1 and

Scaffold 11, while more clustered in the other 10 scaffolds. However, further examination of each CENP-A binding region revealed that most of the regions were interrupted by many sequence gaps, which hampered analysis of the sequence features of the candidate centromeres. Thus, we processed to re-sequence and re-assemble the reference *P. sojae* genome.

To improve the genome assembly of *P. sojae* reference strain P6497, we applied Nanopore long-read sequencing and generated a *de novo* genome assembly with SMARTdenovo together with polishing from PacBio and Sanger reads (Fig. S5*A* and *Appendix SI Text*). The resulting assembly of the nuclear genome (Psojae2019.1) has a size of 86 Mb contained in 70 contigs, with a contig N50 of 2 Mb (Fig. S5*C*). Comparison of Psojae2019.1 to the Sanger assembly indicated that Psojae2019.1 has more repetitive sequences and most regions were colinear (Fig. S5 *B* and *C*, also see *Appendix SI Text* for details). We also checked telomere repeats using a motif proposed for oomycetes (38), and found 13 contigs (versus 7 in Sanger) that harbor telomeric sequences at single ends (*Appendix SI Text* and Dataset S2).

ChIP-seq reads derived from PsCENP-A were mapped to the new genome assembly Psojae2019.1 (Table S2), which initially revealed 16 regions exhibiting CENP-A enrichment. On closer analysis, we found that the unassembled centromere in contig 20 was an artifact caused by inaccurate genome assembly, as this region was duplicated with a centromere-containing region in contig34 (Fig. S6*F*). Of the 15 remaining CENP-A binding regions, 11 regions were assembled within contigs, whereas four regions were disrupted at the edge of contigs (Fig. 2). Long-read coverage analysis verified the integrity of 10 centromeres (Fig. S7), while the CENP-A peaks in Contig 37 and three broken ones (in Contigs 9, 10, 57) lacked sufficient long-read coverage. We focused on the 10 verified CENP-A regions for the further studies (Table 1). RNAseq analysis indicated that all of the 10 CENP-A regions exhibited low transcription, except the region in Contig 11. Contig 11 contained two adjacent ChIP-seq peaks, one was 19 kb and the other was 114 kb, which were interrupted by a 21 kb transcriptionally active region (Fig. 4*C*).

Here, we define it as one centromere (*CEN4*). Among the 10 CENP-A regions, five have a length of ~190 kb, and three are ~160 kb, while *CEN3* and *CEN6* are significant larger (>270 kb) (Table 1). All of these centromeres have a GC content comparable to the whole genome (52.16 - 58.13% vs. 54.6%) (Table 1). Taken together, our CENP-A ChIP-seq analysis utilizing the newly assembled genome indicates that *P. sojae* CENP-A prefers to bind large poorly transcribed genomic regions with no specific DNA sequence bias.

To examine the correlation between the centromere regions identified in the new genome assembly and in the Sanger assembly, we conducted synteny analysis using the genomic regions flanking the centromeres. The locations of CENP-A found in the Psojae2019.1 assembly were highly correlated with those in the Sanger assembly, except *CEN10* (Table 1, Figs. 3 and S6). Contig 51 was colinear with the Sanger scaffold 23; however, no enriched CENP-A signal was detected for this scaffold, probably because the region corresponding to *CEN10* is interrupted by gaps. Notably, the two CENP-A binding regions in Sanger Scaffold 1 were found to correspond to *CEN8* and *CEN9*, and the smaller one was expanded from 20 kb to 188 kb corresponding to *CEN8* (Table 1, Figs. S4*B* and S6*H*). In addition, four contigs of the Psojae2019.1 assembly (contigs 4, 38, 23, and 58) are collinear with Sanger Scaffold 1, and telomere repeats are found at the ends of Contigs 4 and Contig 58, further suggesting that Scaffold 1 of the Sanger genome is assembled incorrectly and should be split into two scaffolds (Fig. S6*H*). Overall, comparison of centromeres identified in the Sanger and Psojae2019.1 assemblies further confirms their authenticity and reflects some misassemblies that are present in the Sanger genome assembly.

P. sojae CENP-A regions are embedded within heterochromatin

To define the epigenetic state of *P. sojae* centromeric regions, we performed ChIP-seq with antibodies against two heterochromatin marks (H3K9me3, trimethylation of lysine 9 of histone H3, and H3K27me3, trimethylation of lysine 27 of histone H3) and one euchromatin mark

(H3K4me2, dimethylation of lysine 4 of histone H3). The distribution of H3K9me3 and H3K27me3 is generally coincident throughout the genome, and both were colocalized with the CENP-A binding regions (Figs. 4 and S7). Intriguingly, the heterochromatic region extended 8kb to 64 kb beyond each CENP-A binding region (Fig. 4A and Table 1), similar to pericentromeric heterochromatin regions described in other species (13, 21, 39). In contrast, the euchromatic mark H3K4me2 was excluded from the CENP-A region and its flanking pericentric regions, and generally overlapped with the mRNA transcriptional profile (Figs. 4 *B*-*C* and S7). Thus, distribution of histone modifications suggests that the CENP-A and heterochromatin regions are not spatially distinct, and we define the latter as pericentric regions.

A Copia-like transposon (CoLT) is highly enriched in the *P. sojae* centromeres

The Psojae2019.1 genome assembly contains 31% repetitive sequences, the majority of which are transposable elements (TEs) (Fig. S5*D*). Our analysis showed that centromeres are also composed of many repetitive elements, mostly LTR-retrotransposons (Figs. 3 and S6). To identify whether the centromeres in *P. sojae* possess any common sequences or repeat elements, all identified CENP-A regions were subject to multiple sequence alignment. This analysis found an ~5 kb sequence that is highly similar (>98%) and shared among 10 centromeres (Fig. S8 and Dataset S3). BLAST analyses with the consensus 5 kb sequence against the genome revealed that although this element is not exclusive to centromeres, it is significantly enriched in centromeres: approximately 90% of all genomic copies of this element localized to centromeres (Fig. 5A). Moreover, this element is present as clusters in centromeric regions, and only sparsely found in other regions of the genome, further strengthening its association with centromeres. Further examination of the sequence indicates that it resembles a Copia transposon-like transposon, and we named it CoLT for Copia Like Transposon (Fig. 5*B*, Dataset S3).

CoLT clusters are conserved in two *P. sojae* oomycete relatives and may be a hallmark of oomycete centromeres

To examine if clustered CoLT elements found in *P. sojae* centromeres are also present in other oomycete genomes, we conducted BLAST searches using the 5 kb consensus sequence derived from *P. sojae* centromeres against the genome assemblies of two *P. sojae* relatives, *Bremia lactucae* (downy mildew, lettuce pathogen) and *Phytophthora citricola* (citrus pathogen), which have relatively contiguous genome assemblies. Interestingly, similar CoLTs clusters were observed in these genomes, and usually appeared once per contig (Figs. 5*C* and S9*A*). To assess if these clustered CoLTs were syntenic with the *P. sojae* centromere-containing contigs, we examined the CoLT clusters that were present within Mb-long scaffolds/contigs. Synteny analysis demonstrated that five regions in the *B. lactucae* genome that had CoLT clusters were collinear with *P. sojae* centromeres (Figs. *5C* and *D).* Unexpectedly, Scaffold 2 (original name, SHOA01000004.1, see Dataset S4 for details) contained two CoLT clusters that were syntenic with *P. sojae CEN3* and *CEN5* (Fig. 5D), indicating that scaffold 2 may be incorrectly assembled (Fig. 5D). It should be noted that the *B. lactucae* genome assembly still has a large percentage of unresolved gaps likely due to its highly heterozygous nature (36). In comparison, all three selected regions that had clustered CoLT clusters within *P. citricola* contigs (PcContigs) were syntenic with *P. sojae* centromeres (*CEN3*/PcContig2, *CEN9*/PcContig1, *CEN5*/PcContig26) (Figs. S9 *B-D*). However, a large number of the CoLT clusters localized at contig ends, or were distributed across the length of short contigs (Fig. S9*A*). This suggests that many of the centromeric regions in *P. citricola* were not fully assembled. Taken together, we propose that the clustered CoLT elements may be used as criteria to predict centromere regions in other *Phytophthora* species and possibly in other oomycetes.

Discussion

In this study, we identified centromeres in the oomycete plant pathogen *P. sojae* by combining long-read sequencing and ChIP-seq with the GFP tagged kinetochore protein CENP-A.Cellular dynamics analysis revealed that *P. sojae* centromeres were clustered within nuclei in different life stages and during vegetative growth. 10 fully assembled and five incompletely assembled CENP-A binding regions were identified. The common features shared by these regions include: a) a low level of transcription; b) a GC content similar to that of the whole genome; c) repetitive sequences; d) enrichment for a specific Copia-like transposon; e) overlapping and surrounding heterochromatin; and f) lack of H3K4me2.

While CENP-A is conserved among different organisms, centromere sequences evolve rapidly (1, 40). Although the filamentous fungal-like oomycetes are classified in the stramenopiles of the SAR supergroup, it is intriguing to observe that the centromeres that we identified in *P. sojae* are much larger and more complex, comparing to those reported in its stramenopile relative, the diatom *P. tricornutum*, and those found in the parasites (*P. falciparum* and *T. gondii*) of the alveolates (Fig. 6). In the latter three cases, all centromeres are composed of non-repetitive sequences. Surprisingly, *P. sojae* centromeres show structural similarity to several, only distantly related, fungal species, such as *N. crassa* (13) and *Cryptococcus neoformans* (12). These features include an enrichment of transposons (or their remnants), and overlap with the constitutive heterochromatin mark H3K9me2/3. Remarkably, the euchromatin mark H3K4me2 has been shown to be associated with centromeres in humans, mouse, *Drosophila*, *S. pombe*, and rice (39, 41-43), but is excluded from other fungal regional centromeres reported to date and in *P. sojae*. In humans and *D. melanogaster*, the CENP-A and pericentromeric heterochromatin domains are spatially distinct, and the CENP-A domain is flanked by but does not overlap with heterochromatin (39, 43, 44). In contrast, the entire centromere of *P. sojae* is embedded in heterochromatin. It is unknown if the distribution of heterochromatin regions affects centromere distribution in *P. sojae*, but heterochromatin has been shown to be important for centromere function and kinetochore assembly in *N. crassa* and *S. pombe* (13, 45, 46). In addition, it is of interest that *P. sojae* H3K9me3 and H3K27me3 fully overlap with the centromeric regions, which have not been observed in centromeres of other species thus far, but was shown in human and mouse pericentromeres (8, 47). On the other hand, these two epigenetic marks generally coexist throughout the entire genome, suggesting it might be just a general profile of H3K27me3 and H3K9me3 in *P. sojae*.

Transposable elements (and their relics) have been known as residents of the centromeres and pericentromeres of many animals, plants, and fungi (48). While animal centromeres are associated with both satellite DNA and retroelements, satellite DNA is usually regarded as the main sequence components (49). Centromeres of many plants, such as maize and rice, are built on centromere-specific retrotransposons (CR), and a certain CR is usually unique to a particular chromosome (7). Centromeres of *N. crassa* (13) and *C. neoformans* (12) are composed of retrotransposons, and the retroelements in *C. neoformans* are centromere-specific (12). In comparison, although *P. sojae* regional centromeres include various transposons, many of these elements are not limited to this region and can also be found in other genomic areas. Our study shows that a specific Copia-like transposon (CoLT) is highly enriched in the *P. sojae* centromeric regions and confines the CENP-A binding regions (Figs. 4 *B-C* and S7). A similar distribution pattern of centromere-associated retrotransposons was recently found in *Drosophila melanogaster* (21). In *D. melanogaster*, a non-LTR retroelement named *G2/Jocky-3* was found to be enriched in CENP-A chromatin, and this element is also associated with centromeres in its sister species *D. simulans* (21). Strikingly, the CoLT elements were found to be clustered in the genomes of *P. sojae* oomycete relatives, and some of those regions were syntenic with *P. sojae* centromeric regions. As most of the oomycete genome assemblies were not based on long-read sequencing technology, and thus are very fragmented, it remains to be seen if the CoLT elements have evolved to be widely utilized by oomycetes as a platform for CENP-A loading.

Due to large genome scales and potentially similar chromosome sizes, the karyotypes of *Phytophthora* species cannot be well resolved by pulse field gel electrophoresis (31, 50). The chromosome number of *P. sojae* is not yet accurately known, but has been estimated to be between 10 and 15 based on an earlier cytological study (51). By comparing the location of centromeres in the Sanger and Psojae2019.1 assembly, we can validate and predict the configuration of 11 centromeres, namely *CEN1-CEN10*, and *CEN_C9* + *CEN_C48* (Table 1 and Table S4). Three centromeres, namely *CEN_C37*, *CEN_C10* and *CEN_C57*, are not fully assembled. Thus, our results offer a new estimate 12-14 chromosomes in *P. sojae*.

N-ChIP was implemented for this study, because several attempts to perform ChIP analysis based on traditional formaldehyde-cross-linking strategies were unsuccessful. Cross-linking with 1% formaldehyde caused degradation of DNA and failure of ChIP. *P. sojae* transformants expressing GFP tagged CENP-A and CENP-C were both used for N-ChIP-seq. However, only the GFP-CENP-A transformant produced significant enrichment, indicating that the binding of CENP-C to chromosomes may be too weak to recover target DNA under native conditions without cross-linking.

Our analysis showed that having an improved reference genome assembly based on long-read sequencing technologies was crucial to the identification and characterization of centromeres in *P. sojae*. Our attempt to characterize centromere sequences using the classical Sanger assembly was not successful because most of the non-coding repetitive regions were not assembled. While the N50 of the new genome assembly Psojae2019.1 is lower than that of the Sanger assembly, the contigs do not contain gaps and many of the gaps present in the Sanger assembly have been closed (Fig. S6). We tried to scaffold the assembly with different scaffolding programs such as npScarf (52), SSPACE (53), LINKS (54) and the optical BioNano mapping

(*Appendix SI Text* and Fig. S10). Although these scaffolders improved the contiguity (up to 35 scaffolds using SSPACE), they also generated multiple conflicts with the Sanger assembly, and most of the joins could not be supported by evidence such as long read coverage (Fig. S11 and Table S3). Thus, we opted to retain the contig-level assembly in our study. However, identification of centromeres helped to resolve several structural problems present in the “classical” *P. sojae* Sanger assembly, and revealed potential structural problems in other oomycete genome assemblies. On basis of the presence of centromeres and predicted telomeres together with synteny analyses, we found that three Sanger scaffolds/contigs may represent full length chromosomes, namely Scaffold 2/Contigs [26+1+35+6] (Fig. S6*A*), Scaffold 5/Contigs [17+36+7+49+45] (Fig. S6*G*); and partial Scaffold 1/Contigs [58+38+4] (Fig. S6*H*). Notably, telomeres appear on the both ends of Sanger Scaffold 5 and its syntenic contigs in Psojae2019 (Fig. S6G). There are five *P. sojae* centromeres that are not fully assembled. With the development of sequencing and assembly technologies, a finalized chromosome-level genome assembly could help to assemble those broken centromeres, and refine the centromere sequences that we identified.

Centromeres and their associated kinetochore network serve critical functions in genome stability and replication. Failures in kinetochore assembly and attachment increase the probability of chromosome mis-segregation leading to aneuploidy (55). While these drastic genome changes can be detrimental to the organism, formation of aneuploidy and polyploidy is an important strategy orchestrated by pathogens to adapt to the environment during periods of stress (56). Polyploidy and aneuploidy are prevalent in *Phytophthora* natural isolates and progeny from sexual reproduction (35, 57-60). Interestingly, plant hosts can induce aneuploidy of the sudden oak death pathogen *P. ramorum*, which enhances its phenotypic diversity and increases its adaption to the environment (59). Recently, a phenomenon termed dynamic extreme aneuploidy (DEA) was described in a vegetable oomycete pathogen, *P. capsici*, in which high variability among progeny produced by asexual spores was caused by ploidy variation (61). However, the mechanisms resulting in oomycete aneuploidy and/or polyploidy is understudied. As centromeres are the functional and structural foundation for kinetochore assembly and proper chromosome segregation, identification of centromeres and kinetochore proteins in *P. sojae* may help to illuminate the mechanisms underlying oomycete genetic, genomic, and phenotypic diversification.

Materials and methods

P. sojae culture and transformation

All the strains used in this study are listed in Table S5. The reference *P. sojae* isolate P6497 (race 2) used in this study was routinely grown and maintained in cleared V8 media at 25 °C in the dark. Transient gene expression assays based on an optimized polyethylene glycol (PEG) mediated protoplast transformation protocol (30) was applied to examine the nuclear localization of CENP-A. Stable and homokaryotic transformants were chosen for ChIP-seq, which were generated by passaging on V8 supplemented with 50 μg/mL G418 (Geneticin, AG Scientific, San Diego, California, USA) for at least 5 times followed by zoospore isolation. Co-transformation was employed to generate strains expressing both H2B-mCherry and GFP-CENP-A. Transformation was performed as previously described (30). Sporangia and zoospores were induced by water flooding according to a method described previously (62).

Construction of plasmids

All the primers used in this study are listed in Table S6. All GFP fusion constructs were generated based on the plasmid backbone pYF3-GFP (63), in which StuI was used for the N-terminal fusions, and HpaI was used for the C-terminal fusions.

3'-RACE was conducted to validate the gene model of CENP-A, according to the manufacturer instruction (Invitrogen, Cat. no. 18373-019). All PCR-amplifications were performed using Phusion High-Fidelity DNA Polymerase (NEB, M0530S).

CRISPR-mediated gene replacement

A sgRNA guide sequence whose PAM sequence overlapped with the start codon of CENP-A was selected as the CRISPR/Cas9 targets. An oligo annealing strategy was used for assembly of the sgRNA expression cassettes according to previously described methods (30).

HDR templates for CENP-A was assembled using NEBuilder® HiFi DNA Assembly. 5’-junction, 3’-junction and spanning diagnostic PCR were performed to genotype mutants, utilizing the primers listed in Table S6.

Microscopy imaging of *P. sojae* transformants

A Zeiss 780 inverted confocal microscope was adopted to examine the subcellular localization of GFP tagged CENP-A driven by strong promoters. Images were captured using a 63 X oil objective with excitation/emission settings (in nm) 488/504-550 for GFP, and 561/605-650 for mCherry. DeltaVision elite deconvolution microscope (Olympus IX-71 base) equipped with Coolsnap HQ2 high resolution CCD camera was employed to examine the subcellular localization of GFP tagged CENP-A produced from the native loci. Images were captured using a 100 X oil objective (100x/1.40 oil UPLSAPO100X0 1-U2B836 WD 120 micron DIC ∞/0.17/FN26.5, UIS2) with an excitation filter, 475/28 and an emission filter, 525/50 for GFP. Time-lapse experiments were performed with 40 X oil objective (40x/0.65-1.35 oil UAPO40XOI3/340 1-UB768R WD 100 micron DIC ∞/0.17/FN22, UIS2, BFP1), with the same filters. Confocal images were edited using microscope’s built-in Zen 2012 software (Blue and/or Black edition according to different purposes). DeltaVision images were edited using Fiji-ImageJ and Photoshop.

High molecular weight genomic DNA extraction and ONT sequencing

High molecular weight (HMW) genomic DNA (gDNA) from *P. sojae* was isolated by the CTAB DNA extraction method. 1 g 3-day old fresh *P. sojae* liquid cultures were collected by filtration and washed twice with sterile water. The resulting damp mycelial pads were frozen immediately in liquid nitrogen in a pre-cooled mortar, then ground by a pestle. Mycelial powder was transferred to a 50 ml Falcon tube and mixed gently with 10 ml room temperature *P. sojae* CTAB extraction buffer (200 mM Tris·HCl pH=8.5, 250 mM NaCl, 25 mM EDTA pH=8.0, 2% SDS, 1% CTAB). The suspension was incubated in 65℃ for 15 minutes with mixing every 5 minutes. An equal volume of phenol/chloroform/isoamyl alcohol (25:24:1, saturated with 10 mM Tris, pH=8.0 and 1 mM EDTA) was added to the suspension and mixed gently by inverting the tube, then centrifuge at4℃, 5000 g for 15min. The supernatant was transferred to a new 50 ml tube and treated with RNase A (final concentration, 100 µg/ml) at 37℃ for about 1 hour, followed by proteinase K treatment (final concentration 200 µg/ml) at 50℃ for 2 hours. An equal volume of chloroform was added to the solution and mixed gently by inverting the tube, then centrifuge, 4℃, 5000 g for 15min. The supernatant was transferred to a new 50 ml Falcon tube and DNA precipitated by addition of an equal volume of isopropanol. The tube was mixed gently and incubated on ice for 6 hours. The resulting white clump of DNA was spooled by a pipette tip and washed once with 70% ethanol. The gDNA was air-dried for 15 minutes at room temperature and dissolved in 100 µl sterile water. The quantity of DNA was examined by Qubit and the quality was checked by pulse field gel electrophoresis (PFGE).

1D Genomic DNA by Ligation kits (SQK-LSK108, for MinION); SQK-LSK109, for GridION) were used to prepare the Oxford Nanopore library. Oxford Nanopore sequencing runs was performed on SpotON R9.4 flow cells with MinKNOW V1.11.5 using MinION or SpotON R9.4.1 flow cells with MinKNOW V3.1.20 using GridION. All of the GridION sequence were basecalled (on GridION, in real time) using Guppy v2.0.5.

Native ChIP-seq

Native ChIP was performed according to the ChIP protocol accompanying Gent, Wang and Dawe (64) with modifications. Briefly, 1-3 mg mycelia were collected from 1-1.5 L of ~3-day culture by filtration system, and ground into fine powder in liquid nitrogen with pre-chilled mortars and pestles. Nuclei were isolated and digested by micrococcal nuclease (M0247S, NEB) at 37 for 6 min. An antibody against GFP (Abcam, ab290) was used to immunoprecipitate single nucleosomes containing the GFP-CENP-A fusion (driven by the strong promoter derived from *HAM34* gene). Antibodies H3K9me3 (Abcam, ab8898), H3K27me3 (Active Motif, 39157), and H3K4me2 (Millipore, 07-030) were used to immunoprecipitate nucleosomes with relevant modifications. ChIP-seq of GFP-CENP-A and H3K27me3 were performed by Genewiz using Illumina NextSeq500 that generated 150 nucleotide paired-end reads; ChIP-seq of H3K9me3 and H3K4me2 were conducted by BGI using Illumina Hiseq 4000 that produced 50 nucleotide single-end reads. Numbers of reads for each sample are listed in Table S2.

Analysis of ChIP-seq and RNA-seq

To map ChIP-seq reads to the genomes, the quality of raw ChIP-seq reads were first assessed by FastQC (v0.11.6). For ChIP-seq of CENP-A, H3K27me3, the resulting reads were trimmed with fastx-clipper and mapped with Bowtie2 with default parameters (65) and aligned to the genome assemblies. For H3K9me3 and H3K4me2, the ChIP-seq reads were polished by BGI prior to be released and thus mapped to the genomes directly using the same Bowtie2 setup. The aligned file (.bam) was sorted and indexed by samtools (version 1.9). Subsequently the ChIP-ed and input samples were analyzed with DeepTools(v3.2.0) ''bamCompare'' to calculate normalized ChIP signal (log2[$\text{ChIP}_{\text{RPKM}}$/$\text{Input}_{\text{RPKM}}$]) and bigwig files were generated . Then .bw files were visualized using the Integrative Genome Viewer (IGV). (https://software. broadinstitute.org/software/igv/). To get profile mRNA, the existing RNA-Seq reads (FungiDB, https://fungidb.org/fungidb/) were aligned to the genomes using HISAT2 (version 2.1.0), and the resulting files (.bam) were sorted and indexed by samtools (version 1.9). The .bam file was converted to .tdf for visualization using IGV.

Genome assembly, analysis of genomic features and synteny comparison

Details of the *de novo* genome assembly is described in SI Text. To predict gene models, first, the assembly Psojae2019.1 was subjected to repeat masking utilizing RepeatMasker (66) based on a library of *de novo*-identified repeat consensus sequences that was generated by RepeatModeler (www.repeatmasker.org/RepeatModeler.html). Next, the repeat-masked assembly was used to predict gene models *ab initio* based on MAKER (v2.31.18) (67) with predicted proteins from available *P. sojae* and *P. infestans* genome annotations as input (25, 68). GC content was calculated in non-overlapping 5-kb windows using a modified Perl script (gcSkew.pl, https://github.com/Geo-omics/scripts/blob/master/AssemblyTools/gcSkew.pl) and plotted as the deviation from the genome average for each contig. Genes encoding ribosomal RNA (18S, 5.8S, 25S, and 5S) and tRNA were inferred and annotated based on RNAmmer (v1.2) (69) and tRNAscan-SE (v2.0) (70), respectively. To find telomeres, a custom-made Perl script was used to search for the sequence "TTTAGGG" that was proposed for oomycetes telomeric sequences (38). Pairwise synteny comparison between the two *P. sojae* genome assemblies (i.e. P. sojae V3 and Psojae2019.1) or between different oomycete species was conducted using BLASTn. BLASTn hits and other genomic features were plotted using Circos (v0.69-6) (71). Whole-genome alignment was computed with MashMap (https://github.com/marbl/MashMap) employing default settings, and was visualized as a dot plot (72).

Bionano mapping

P. sojae protoplasts were generated from 2.5-day old mycelial and were embedded into agarose. Bionano Prep Cell Culture DNA Isolation Protocol was employed for extracting the high molecular weight DNA. DNA labelling with DLE-1 was performed according to the standard protocols provided by Bionano Genomics (Document number 30206, version F). Labelled DNA samples were loaded into two flow cells and run on a Saphyr system (Bionano Genomics). The *de novo* assembly was performed using Bionano Solve 3.3. Standard parameters for Saphyr data were used without "extend and split" and without haplotype refinement in order to create a single map for each allele ("optArguments_nonhaplotype_noES_DLE1_saphyr.xml"). In the process of *de novo* assembly, data generated from two flow cells were merged. An assembly graph was generated during a pairwise comparison of all of the molecules with a p value threshold of 1e-11, and was refined based on molecules aligned to the assembled maps with a p value threshold of 1e-12. After five rounds of extension and refinement, a final refinement was conducted with a p value threshold of 1e-16. Then, the *de novo* assembled map was used to scaffold the sequence assembly. When using the hybrid scaffold module of Bionano Solve 3.3 pipeline, the option of "resolve conflicts" for sequence contigs and Bionano maps was selected. The standard hybrid scaffold settings with a modified parameter (-E 0) was applied to remove discrepancies between sequence assembly and Bionano *de novo* assembly. Sequence contigs were *in silico* digested, based on the recognition sequence (CTTAAG) of DLE-1. Conflicts detection was accomplished by aligning contig maps to Bionano maps with p value threshold of 1e-10. When divergence was identified, the conflicts were resolved by cutting either the contig or the map, depending on the quality of the genome map at the divergent position.

Analysis of transposable elements and identification of CoLT

To identify transposable elements in *P. sojae*, the new genome assembly was subjected to RepeatMasker (Repbase v23.09) analysis and hits were mapped to this genome assembly. The Copia-like transposon (CoLT) element was identified in a stepwise way by multiple sequence alignments followed by extraction of a consensus sequence and BLASTn analyses. Specifically, an approximately 5 kb consensus sequence was identified in the alignment of centromere sequences (including incompletely assembled ones) utilizing the multiple alignment program MAFFT, a plug-in in the Geneious R9 software (http://www.geneious.com), with default parameters. Then the consensus sequence was used as a query to perform a BLASTn search against the Psojae2019.1 genome assembly. The resulting sequence hits were used to map against the genome, and hits longer than 500 bp were used for representing in the figures. The longest sequence hit with highest identity was retrieved, and was used as a query to execute a second round of BLASTn search against the NCBI database to further characterize the sequence. The results of BLASTn analysis indicated that that the sequence was highly similar to a Copia-like transposable element. To define the domains of the CoLT, this sequence was further analyzed by repeat identification (utilizing a bioinformatics software Unipro UGENE(73)), and by searches utilizing the Repbase database (https://www.girinst.org/) and NCBI CD-search (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). Sequences of the 5 kb consensus and the top hit in the Psojae2019.1 genome assembly are shown in Dataset S3.

Prediction of centromeric regions in *P. sojae* closely related species

To predict centromeres of the two oomycete species, namely *Phytophthora citricola* P0716, (Genbank: GCA_007655245.1, with permission of the author) and *Bremia lactucae* SF5, (GenBank: GCA_004359215.1) (36), BLASTn searches were conducted utilizing the *P. sojae* Copia-like transposon (CoLT) as a query. Significant hits (>90% identity and > 500 bp) were retrieved, and were plotted to all scaffolds of the *B. lactucae* assembly and to contigs > 10 kb of the *P. citricola* assembly. For CoLT clusters that were localized within scaffolds or contigs, their collinearities with the Psojae2019.1 assembly were further examined with BLASTn, and visualized by Circos.

Data availability

All raw data of ChIP-seq and Nanopore sequencing and related processed files are available in the NCBI under the BioProject PRJNA563922.

Acknowledgments

We thank BioNano Genomics Support, in particular, Yuanyuan Chang for technical support of BioNano mapping, Beth Sullivan at Duke University for critical reading and comments on the manuscript, and members of the Heitman, Sanyal and Garre labs for helpful discussions. M.N. would like to thank Ulrich Kück and Christopher Grefen for support by the Department of General and Molecular Botany/Molecular and Cellular Botany of Ruhr University. B.M.T. and B.K. were supported by Oregon State University. These studies were supported by NIH/NIAID grants R01 AI050113-15 and R37 MERIT award AI039115-21 to J.H and by the German Research Foundation (DFG, grant NO407/7-1 to MN). J.H. is also a co-director and fellow of the CIFAR program, Fungal Kingdom: Threats & Opportunities.

References

1. L. E. Kursel, H. S. Malik, Centromeres. *Curr Biol* **26**, R487-R490 (2016).
2. K. M. Stimpson, B. A. Sullivan, Epigenomics of centromere assembly and function. *Curr Opin Cell Biol* **22**, 772-780 (2010).
3. A. Buscaino, R. Allshire, A. Pidoux, Building centromeres: home sweet home or a nomadic existence? *Curr Opin Genet Dev* **20**, 118-126 (2010).
4. N. Wang, R. K. Dawe, Centromere size and its relationship to haploid formation in plants. *Mol Plant* **11**, 398-406 (2018).
5. B. E. Black *et al.*, Structural determinants for generating centromeric chromatin. *Nature* **430**, 578-582 (2004).
6. A. Guse, C. W. Carroll, B. Moree, C. J. Fuller, A. F. Straight, In vitro centromere and kinetochore assembly on defined chromatin templates. *Nature* **477**, 354-358 (2011).
7. L. Comai, S. Maheshwari, M. P. A. Marimuthu, Plant centromeres. *Curr Opin Plant Biol* **36**, 158-167 (2017).
8. S. M. McNulty, B. A. Sullivan, Alpha satellite DNA biology: finding function in the recesses of the genome. *Chromosome Res* **26**, 115-138 (2018).
9. W. Jin *et al.*, Maize centromeres: organization and functional adaptation in the genetic background of oat. *Plant Cell* **16**, 571-581 (2004).
10. K. M. Smith, J. M. Galazka, P. A. Phatale, L. R. Connolly, M. Freitag, Centromeres of filamentous fungi. *Chromosome Res* **20**, 635-656 (2012).
11. V. Yadav, L. Sreekumar, K. Guin, K. Sanyal, Five pillars of centromeric chromatin in fungal pathogens. *PLoS Pathog* **14**, e1007150 (2018).
12. V. Yadav *et al.*, RNAi is a critical determinant of centromere evolution in closely related fungi. *Proc Natl Acad Sci U S A* **115**, 3108-3113 (2018).
13. K. M. Smith, P. A. Phatale, C. M. Sullivan, K. R. Pomraning, M. Freitag, Heterochromatin is required for normal distribution of *Neurospora crassa* CenH3. *Mol Cell Biol* **31**, 2528-2542 (2011).
14. K. Schotanus *et al.*, Histone modifications rather than the novel regional centromeres of *Zymoseptoria tritici* distinguish core and accessory chromosomes. *Epigenetics & Chromatin* **8**, 41 (2015).
15. K. Sanyal, M. Baum, J. Carbon, Centromeric DNA sequences in the pathogenic yeast *Candida albicans* are all different and unique. *Proc Natl Acad Sci U S A* **101**, 11374-11379 (2004).
16. W. A. Hoeijmakers *et al.*, *Plasmodium falciparum* centromeres display a unique epigenetic makeup and cluster prior to and during schizogony. *Cell Microbiol* **14**, 1391-1401 (2012).
17. R. E. Diner *et al.*, Diatom centromeres suggest a mechanism for nuclear DNA acquisition. *Proc Natl Acad Sci U S A* **114**, E6015-E6024 (2017).
18. C. F. Brooks *et al.*, *Toxoplasma gondii* sequesters centromeres to a specific nuclear region throughout the cell cycle. *Proc Natl Acad Sci U S A* **108**, 3767-3772 (2011).
19. V. Yadav *et al.*, Cellular dynamics and genomic identity of centromeres in cereal blast fungus. *mBio* **10**, e01581-01519 (2019).
20. M. I. Navarro-Mendoza *et al.*, Early diverging fungus *Mucor circinelloides* lacks centromeric histone CENP-A and displays a mosaic of point and regional centromeres. *bioRxiv*, 706580 (2019).
21. P. B. Becker *et al.*, Islands of retroelements are major components of *Drosophila* centromeres. *PLOS Biology* **17** (2019).

22. T. Lan *et al.*, Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome. *Proc Natl Acad Sci U S A* **114**, E4435-E4441 (2017).
23. M. Jain *et al.*, Linear assembly of a human centromere on the Y chromosome. *Nat Biotechnol* **36**, 321-323 (2018).
24. P. J. Keeling, F. Burki, Progress towards the Tree of Eukaryotes. *Curr Biol* **29**, R808-R817 (2019).
25. B. M. Tyler *et al.*, *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* **313**, 1261-1266 (2006).
26. J. D. Grattepanche *et al.*, Microbial diversity in the eukaryotic SAR clade: Illuminating the darkness between morphology and molecular data. *Bioessays* **40**, e1700198 (2018).
27. D. C. Erwin, O. K. Ribeiro, *Phytophthora diseases worldwide* (American Phytopathological Society (APS Press), 1996).
28. R. H. Jiang, B. M. Tyler, Mechanisms and evolution of virulence in oomycetes. *Annu Rev Phytopathol* **50**, 295-318 (2012).
29. S. Savary *et al.*, The global burden of pathogens and pests on major food crops. *Nat Ecol Evol* **3**, 430-439 (2019).
30. Y. Fang, L. Cui, B. Gu, F. Arredondo, B. M. Tyler, Efficient genome editing in the oomycete *Phytophthora sojae* using CRISPR/Cas9. *Curr Protoc Microbiol* **44**, 21A-21 (2017).
31. H. S. Judelson, M. D. Coffey, F. R. Arredondo, B. M. Tyler, Transformation of the oomycete pathogen *Phytophthora-megasperma* f.sp. *glycinea* occurs by DNA integration into single or multiple chromosomes. *Current Genetics* **23**, 211-218 (1993).
32. Y. Fang, B. M. Tyler, Efficient disruption and replacement of an effector gene in the oomycete *Phytophthora sojae* using CRISPR/Cas9. *Molecular Plant Pathology* **17**, 127-139 (2016).
33. J. McGowan, D. A. Fitzpatrick, Genomic, network, and phylogenetic analysis of the oomycete effector arsenal. *mSphere* **2**, e00408-00417 (2017).
34. B. M. Tyler, M. Gijzen, "The *Phytophthora sojae* genome sequence: foundation for a revolution" in Genomics of plant-associated fungi and oomycetes: dicot pathogens. (Springer, 2014), pp. 133-157.
35. C. M. Malar *et al.*, Haplotype-phased genome assembly of virulent *Phytophthora ramorum* isolate ND886 facilitated by long-read sequencing reveals effector polymorphisms and copy number variation. *Mol Plant Microbe Interact* **32**, 1047-1060 (2019).
36. K. Fletcher *et al.*, Genomic signatures of heterokaryosis in the oomycete pathogen *Bremia lactucae*. *Nat Commun* **10**, 2645 (2019).
37. J. J. van Hooff, E. Tromer, L. M. van Wijk, B. Snel, G. J. Kops, Evolutionary dynamics of the kinetochore network in eukaryotes as revealed by comparative genomics. *EMBO Rep* **18**, 1559-1571 (2017).
38. J. Fulneckova *et al.*, A broad phylogenetic survey unveils the diversity and evolution of telomeres in eukaryotes. *Genome Biol Evol* **5**, 468-483 (2013).
39. B. A. Sullivan, G. H. Karpen, Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. *Nature Structural & Molecular Biology* **11**, 1076-1083 (2004).
40. S. Henikoff, K. Ahmad, H. S. Malik, The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**, 1098-1102 (2001).
41. T. A. Volpe *et al.*, Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297**, 1833-1837 (2002).
42. X. Y. Li *et al.*, High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression. *Plant Cell* **20**, 259-276 (2008).

43. R. C. Allshire, G. H. Karpen, Epigenetic regulation of centromeric chromatin: old dogs, new tricks? *Nat Rev Genet* **9**, 923-937 (2008).

44. M. D. Blower, B. A. Sullivan, G. H. Karpen, Conserved organization of centromeric chromatin in flies and humans. *Dev Cell* **2**, 319-330 (2002).

45. K. C. Scott, S. L. Merrett, H. F. Willard, A heterochromatin barrier partitions the fission yeast centromere into discrete chromatin domains. *Curr Biol* **16**, 119-129 (2006).

46. R. C. Allshire, H. D. Madhani, Ten principles of heterochromatin formation and function. *Nat Rev Mol Cell Biol* **19**, 229-244 (2018).

47. I. K. Greaves, D. Rangasamy, P. Ridgway, D. J. Tremethick, H2A.Z contributes to the unique 3D structure of the centromere. *Proc Natl Acad Sci U S A* **104**, 525-530 (2007).

48. S. Friedman, M. Freitag, "Evolving centromeres and kinetochores" in Advances in genetics. (Elsevier, 2017), vol. 98, pp. 1-41.

49. J. D. Brown, R. J. O'Neill, The evolution of centromeric DNA sequences. *eLS* (2001).

50. P. W. Tooley, M. M. Carras, Separation of chromosomes of *Phytophthora* species using CHEF gel electrophoresis. *Experimental Mycology* **16**, 188-196 (1992).

51. E. Sansome, C. Brasier, Polyploidy associated with varietal differentiation in the megasperma complex of *Phytophthora*. *Transactions of the British Mycological Society* **63**, 461-IN411 (1974).

52. M. D. Cao *et al.*, Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *Nat Commun* **8**, 14515 (2017).

53. M. Boetzer, W. Pirovano, SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* **15**, 211 (2014).

54. R. L. Warren *et al.*, LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *Gigascience* **4**, 35 (2015).

55. D. A. Compton, Mechanisms of aneuploidy. *Curr Opin Cell Biol* **23**, 109-113 (2011).

56. R. T. Todd, A. Forche, A. Selmecki, Ploidy Variation in Fungi: Polyploidy, Aneuploidy, and Genome Evolution. *Microbiol Spectr* **5**, 599-618 (2017).

57. M. Elliott *et al.*, Characterization of phenotypic variation and genome aberrations observed among *Phytophthora ramorum* isolates from diverse hosts. *BMC Genomics* **19**, 320 (2018).

58. M. P. Dobrowolski, I. C. Tommerup, H. D. Blakeman, P. A. O'Brien, Non-Mendelian inheritance revealed in a genetic analysis of sexual progeny of *Phytophthora cinnamomi* with microsatellite markers. *Fungal Genet Biol* **35**, 197-212 (2002).

59. T. Kasuga *et al.*, Host-induced aneuploidy and phenotypic diversification in the sudden oak death pathogen *Phytophthora ramorum*. *BMC Genomics* **17**, 385 (2016).

60. T. van der Lee, A. Testa, A. Robold, J. van 't Klooster, F. Govers, High-density genetic linkage maps of *Phytophthora infestans* reveal trisomic progeny and chromosomal rearrangements. *Genetics* **167**, 1643-1661 (2004).

61. K. Lamour, S. Shresthra, Y. Zhou, X. Liu, J. Hu, Dynamic extreme aneuploidy (DEA) in the vegetable pathogen *Phytophthora capsici* sheds light on instant evolution and intractability. *bioRxiv*, 297788 (2018).

62. L. Lin *et al.*, The MADS-box transcription factor PsMAD1 is involved in zoosporogenesis and pathogenesis of *Phytophthora sojae*. *Frontiers in Microbiology* **9** (2018).

63. Y. Fang, H. S. Jang, G. W. Watson, D. P. Wellappili, B. M. Tyler, Distinctive nuclear localization signals in the oomycete *Phytophthora sojae*. *Front Microbiol* **8**, 10 (2017).

64. J. I. Gent, N. Wang, R. K. Dawe, Stable centromere positioning in diverse sequence contexts of complex and satellite centromeres of maize and wild relatives. *Genome Biol* **18**, 121 (2017).

65. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

66. A. Smit, R. Hubley, P. Green (2015) RepeatMasker Open-4.0. 2013–2015.

67. B. L. Cantarel *et al.*, MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188-196 (2008).
68. B. J. Haas *et al.*, Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* **461**, 393-398 (2009).
69. K. Lagesen *et al.*, RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35**, 3100-3108 (2007).
70. T. M. Lowe, P. P. Chan, tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res* **44**, W54-57 (2016).
71. M. Krzywinski *et al.*, Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-1645 (2009).
72. C. Jain, S. Koren, A. Dilthey, A. M. Phillippy, S. Aluru, A fast adaptive algorithm for computing whole-genome homology maps. *Bioinformatics* **34**, i748-i756 (2018).
73. K. Okonechnikov, O. Golosova, M. Fursov, U. team, Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* **28**, 1166-1167 (2012).
74. S. Kumar, G. Stecher, M. Suleski, S. B. Hedges, TimeTree: A resource for timelines, timetrees, and divergence times. *Mol Biol Evol* **34**, 1812-1819 (2017).

Figures and tables

Figure 1. Subcellular localization of CENP-A in *P. sojae* at different life stages and during vegetative growth. (A) A schematic showing the generation of *GFP*-fused *CENP-A* utilizing CRISPR/Cas9 mediated gene replacement. (B) Subcellular localization of GFP-tagged CENP-A (expressed from the endogenous locus) in *P. sojae* hyphae, sporangia, and encysted zoospores. (C) Time-lapse images illustrating localization of GFP tagged CENP-A during hyphal growth. Dashed squares denote occurrence of nuclear division. Representative images are shown. Scale bars in all images, 5 µm.

Figure 2. Contigs in the Psojae2019.1 assembly demonstrating CENP-A enrichment based on ChIP-seq. (A) 10 fully assembled CENP-A binding sites presented contigs. (B) 5 incompletely assembled CENP-A binding regions. All contigs are drawn to scale and the ruler indicates the length of the contigs. All CENP-A profiles shown were normalized to input DNA. mRNA profiles are shown as log-scales. Solid stars indicate the CENP-A enriched regions within contigs; hollow stars denote broken centromeres at the edge.

Figure 3. A representative Circos visualization comparing centromere-containing genomic regions between the Sanger V3 Scaffold 2 and the Psojae2019.1 assembly. The outer tracks illustrate assembled contigs (in Psojae2019.1) or scaffold (in P. sojae V3) and are color coded as listed in the key on the bottom. Yellow regions on the outer tracks indicate the locations of centromeres (CENP-A binding regions). Blue and orange lines link regions with collinearity across >2 kb, with orange lines corresponding to inversion. Grey box-shaded centromere-containing regions are magnified for detailed visualization.

Figure 4. *P. sojae* centromeres display heterochromatin marks and are enriched for a Copia-like transposon (CoLT). (A) Schematics showing the *P. sojae* core centromeres (CENP-A binding regions) and the pericentric regions of various lengths. Dark and light grey bars indicate core centromeric and pericentric regions. Digits at the center indicate the size of core centromeres; digits on the left denote the full length of the centromeres (a combination of core centromere and pericentromeric region). The right pericentric region of *CEN5* and the left pericentric region of *CEN10* are not fully assembled, and are indicated by dashed bars. Their full lengths labeled with question marks. (B-C) Two centromeres (*CEN1* and *CEN4*) are shown as representatives to compare CENP-A localization to the distributions of modified histones and CoLT elements. A 400 kb region harboring the centromeric region is shown for *CEN1* and *CEN4*. Cyan block, a transcriptionally active region that interrupts *CEN4*. Profiles of CENP-A, H3K9me3, H3K29me3 and H3K4me2 shown were normalized to input. mRNA profiles are shown as log-scales.

Figure 5. Genomic distribution of CoLT in the *P. sojae* and *B. lactucae* genomes. (A) Location of CoLT elements across all Psojae2019.1 contigs. (B) Diagram of a representative CoLT structure. CoLT shown here represents the best hit of BLAST using the consensus sequence obtained from alignment of the centromeres (Dataset S3). CoLT mainly contains two parts annotated (by Repbase) as Copia-24_PIT-LTR, and Copia-24_PIT-I comprising gag, PR (protease), IN (integrase), RT (reverse transcriptase) and RH (RNase H) domains. Other CoLT elements show similar structure, except different lengths of LTR and other domains. (C) Location of CoLT elements across all *B. lactucae* scaffolds >100 kb. For ease of analysis, scaffolds in *B. lactucae* assembly were sorted and re-named based on sizes (large to small). See Dataset S4 for the original scaffold names. (D) A representative Circos plot comparing *B. lactucae* Scaffold 2 that has clustered CoLTs with the corresponding Psojae2019 contigs. Regions underscored by green lines indicate both sides of the CoLT clusters were syntenic with the regions surrounding *P. sojae*

centromeres. Regions underlined by blue indicate only one side of the CoLT clusters were found to be syntenic with *P. sojae* centromere flanking sequences.

Figure 6. Diversity of centromere features within the SAR supergroup. Simplified schematics (not to scale) showing the structure, epigenetic modifications, size and composition of centromeres across SAR lineages. *Epigenetic state was not examined in the diatom centromeric regions; however, several AT-rich DNA sequence can be employed for episome maintenance, suggesting diatom centromere might not be epigenetically dependent. Histone modification H3K27me2 was only tested in *P. sojae*. The phylogeny was constructed using TimeTree (74). *Homo sapiens*, *Arabidopsis thaliana* and *Neurospora crassa* were used as representatives of animals, plants and for the phylogeny analysis, and are used as outgroups illustrating the evolutionary status of the SAR supergroup.

Table 1. Centromeres identified in the Psojae2019.1 assembly and their counterparts in the Sanger assembly

Psojae2019.1					Sanger V3	
Name	Contig	Position of core centromere, kb (size, kb)	GC% of *CEN*	Position of pericentric region, kb (size, kb)	Scaffold	Position of *CEN* (kb)
CEN1	1	415-579 (164)	56.91	361-415 (54) 579-650 (71)	2	7086-7267
CEN2	2	4102-4296 (194)	55.51	4094-4102 (8) 4296-4305 (9)	8	2374-2420
CEN3	3	3138-3412 (274)	54.81	3223-3138 (15) 3412-3560 (48)	9	3075-3286
*CEN4**	11	991-1175 (184)	58.13	944-991 (47) 1175-1205 (30)	4	995-1246
CEN5	18	1556-1709 (153)	57.93	1492-1556 (64) 1709-? (>22)†	3	4078-4138
CEN6	34	696-880 (184)	57.65	643-696 (53) 882-904 (22)	6	1521-1726
CEN7	36	433-708 (275)	52.16	376-433 (57) 708-732 (24)	5	2049-2310
CEN8	38	302-490 (188)	57.01	288-302 (14) 490-515 (25)	1	9667-9688
CEN9	41	153-342 (189)	57.40	101-153 (52) 342-358 (16)	1	2921-3079
CEN10	51	36-152 (116)	57.93	?-36 (>36)† 152-216 (64)	-	-

*Contig11 contains a minor (coordinate, 990,422-1,009,603, 19 kb) and a major (coordinate, 1,060,108-1,174,601, 114 kb) peak that are separated by a 21 kb transcriptionally active region.
†One side of pericentric heterochromatin region is not fully assembled.

Supplemental Information

SI Text: Nanopore sequencing and *de novo* assembly of the reference *P. sojae* genome
Figs. S1. Summary of the presence and absence of putative core kinetochore proteins identified in *P. sojae*
Fig. S2. Identification and expression of CENP-A in *P. sojae*.
Fig. S3. Generation of *P. sojae* strains expressing *GFP* tagged *CENP-A* utilizing CRIPSR/Cas9 mediated genome editing.
Fig. S4. Scaffolds in the Sanger assembly that are suggested to harbor putative centromeres.
Fig. S5. Pipeline used for *de novo* assembly and metrics of the *P. sojae* genome assembly Psojae2019.1.
Fig. S6. Comparison of centromere-containing genomic regions between the Sanger (P. sojae V3) and the Psojae2019.1 assemblies.
Fig. S7. Summary of features of each intact centromere and read coverage analysis of centromere.
Fig. S8. MAFFT-based alignment of CENP-A binding regions reveals a 5 kb sequence that are highly similar among *P. sojae* centromeres.
Fig. S9. Genomic distribution of CoLT in the *P. citricola* genome.
Fig. S10. Representative contigs that are anchored by BioNano mapping and contigs that are suggested to be joined.
Fig. S11. Dot plot comparison of scaffolded assemblies against the original Psojae2019.1 assembly and the Sanger assembly.
Table S1. Metrics of ONT sequencing
Table S2. Statistics of ChIP-seq samples
Table S3. Metrics of scaffolded assemblies and their comparison to the Sanger V3 and Psojae2019.1 assembly.
Table S4. Five incompletely assembled centromeres in the Psojae2019 assembly and their corresponding CENP-A regions mapped in the Sanger assembly
Table S5. *P. sojae* strains used in the study
Table S6. Primers used in this study.
Movie S1 (separate file). Time-lapse experiment showing cellular dynamics of CENP-A during *P. sojae* vegetative growth.
Dataset S1 (separate file). Sequences of kinetochore orthologs identified in *P. sojae.*
Dataset S2 (separate file). 13 telomeres predicted in the Psojae2019.1 assembly.
Dataset S3 (separate file). DNA sequence of the CoLT consensus sequence and the best hit
Dataset S4 (separate file). Original names of the sorted *B. lactucae* scaffolds.
Dataset S5 (separate file). Bionano mapping report.

Figure 1

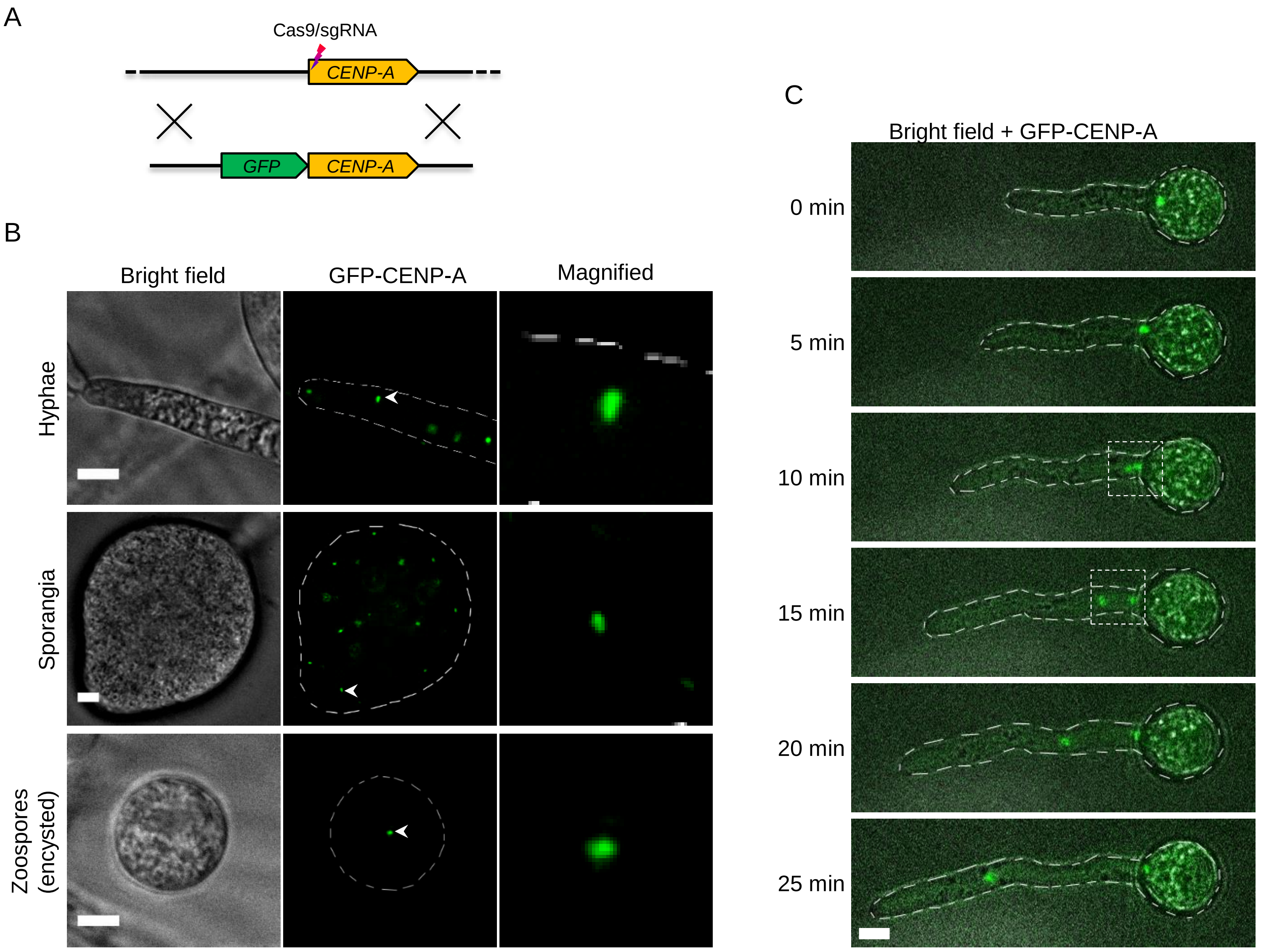

A
Cas9/sgRNA
CENP-A
GFP
CENP-A
B
Bright field
GFP-CENP-A
Magnified
Hyphae
Sporangia
Zoospores (encysted)
C
Bright field + GFP-CENP-A
0 min
5 min
10 min
15 min
20 min
25 min

Figure 2

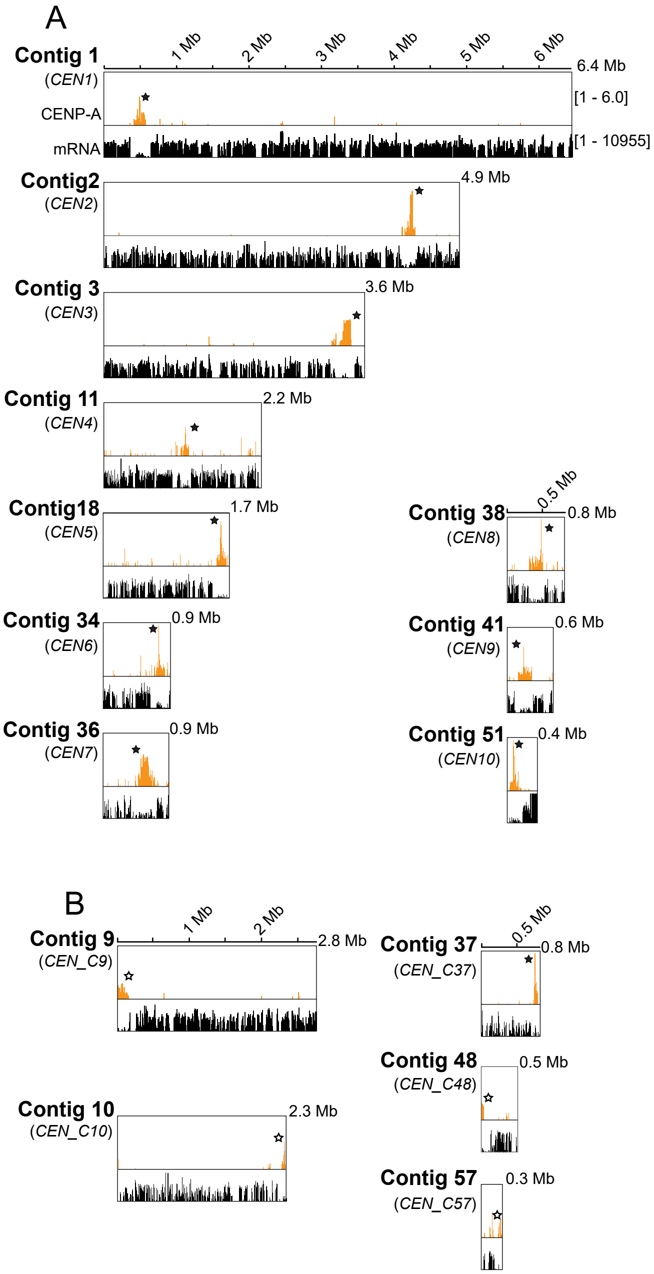
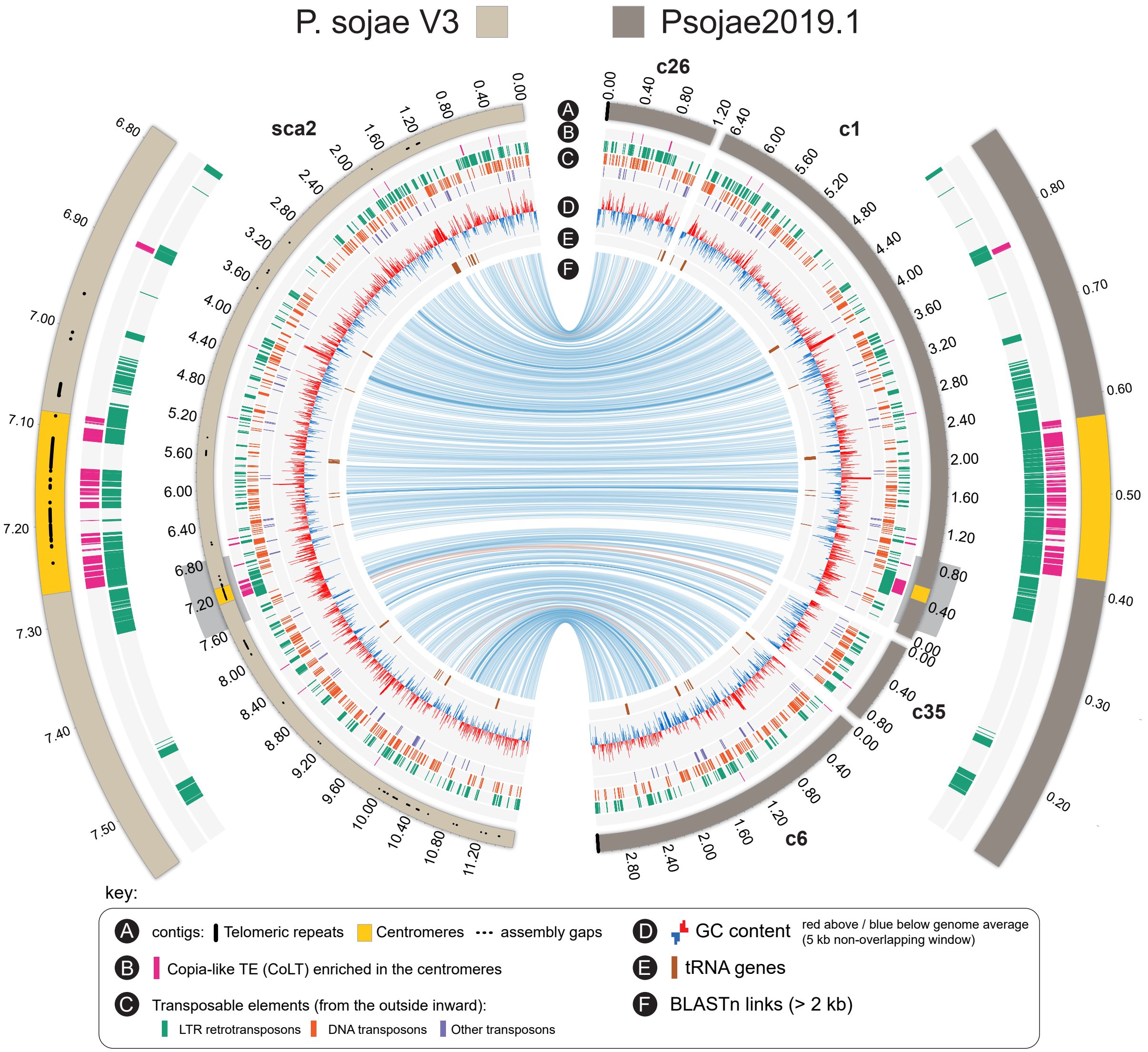

A
Contig 1
(CEN1)
CENP-A
mRNA
1 Mb
2 Mb
3 Mb
4 Mb
5 Mb
6 Mb
6.4 Mb
[1 - 6.0]
[1 - 10955]
Contig2
(CEN2)
4.9 Mb
Contig 3
(CEN3)
3.6 Mb
Contig 11
(CEN4)
2.2 Mb
Contig18
(CEN5)
1.7 Mb
Contig 34
(CEN6)
0.9 Mb
Contig 36
(CEN7)
0.9 Mb
Contig 38
(CEN8)
0.5 Mb
0.8 Mb
Contig 41
(CEN9)
0.6 Mb
Contig 51
(CEN10)
0.4 Mb
B
Contig 9
(CEN_C9)
1 Mb
2 Mb
2.8 Mb
Contig 10
(CEN_C10)
2.3 Mb
Contig 37
(CEN_C37)
0.5 Mb
0.8 Mb
Contig 48
(CEN_C48)
0.5 Mb
Contig 57
(CEN_C57)
0.3 Mb

Figure 3



P. sojae V3
Psojae2019.1
sca2
c26
c1
c35
c6
A
B
C
D
E
F
0.00
0.40
0.80
1.20
1.60
2.00
2.40
2.80
3.20
3.60
4.00
4.40
4.80
5.20
5.60
6.00
6.40
6.80
7.20
7.60
8.00
8.40
8.80
9.20
9.60
10.00
10.40
10.80
11.20
6.80
6.90
7.00
7.10
7.20
7.30
7.40
7.50
0.80
0.70
0.60
0.50
0.40
0.30
0.20
key:
A contigs: Telomeric repeats Centromeres assembly gaps
B Copia-like TE (CoLT) enriched in the centromeres
C Transposable elements (from the outside inward):
LTR retrotransposons DNA transposons Other transposons
D GC content red above / blue below genome average (5 kb non-overlapping window)
E tRNA genes
F BLASTn links (> 2 kb)

Figure 4

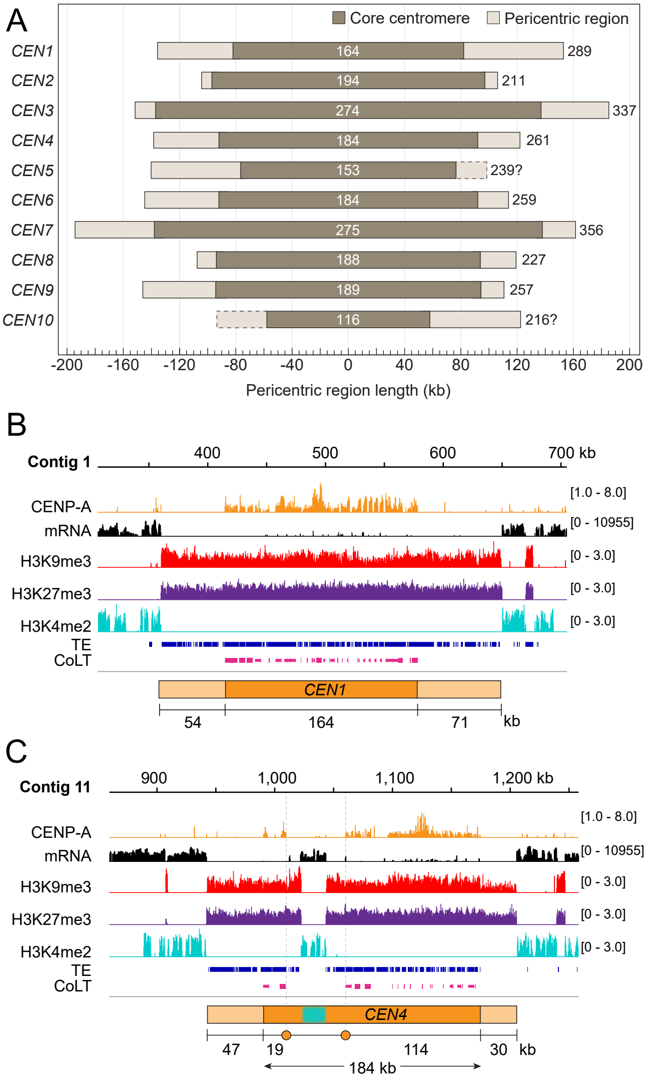

A
Core centromere
Pericentric region
CEN1
CEN2
CEN3
CEN4
CEN5
CEN6
CEN7
CEN8
CEN9
CEN10
164
289
194
211
274
337
184
261
153
239?
184
259
275
356
188
227
189
257
116
216?
-200
-160
-120
-80
-40
0
40
80
120
160
200
Pericentric region length (kb)
B
Contig 1
400
500
600
700 kb
CENP-A
mRNA
H3K9me3
H3K27me3
H3K4me2
TE
CoLT
[1.0 - 8.0]
[0 - 10955]
[0 - 3.0]
[0 - 3.0]
[0 - 3.0]
CEN1
54
164
71
kb
C
Contig 11
900
1,000
1,100
1,200 kb
CENP-A
mRNA
H3K9me3
H3K27me3
H3K4me2
TE
CoLT
[1.0 - 8.0]
[0 - 10955]
[0 - 3.0]
[0 - 3.0]
[0 - 3.0]
CEN4
47
19
114
30
kb
184 kb

Figure 5

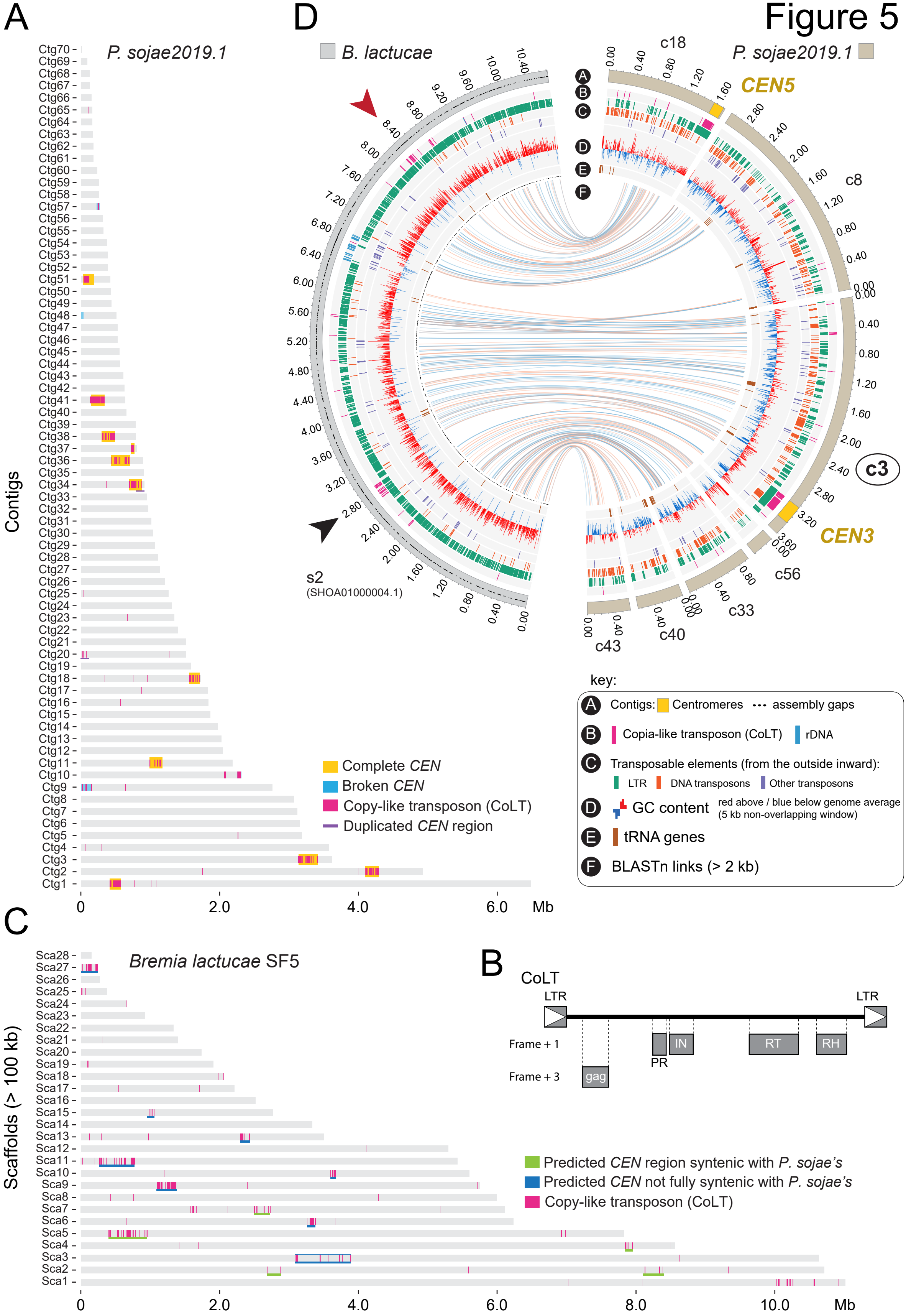
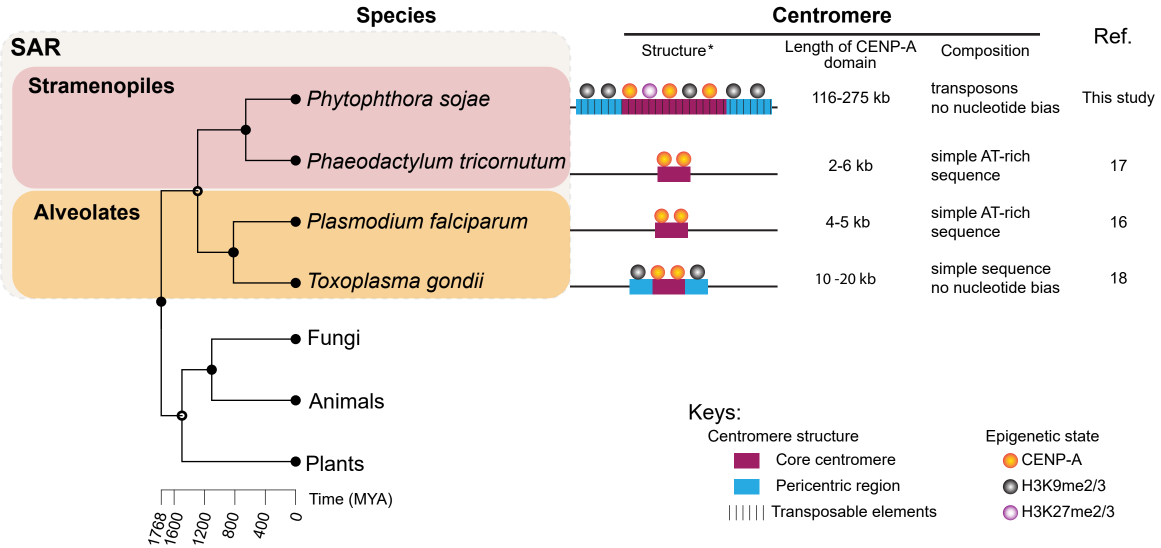

A
P. sojae2019.1
Ctg70
Ctg69
Ctg68
Ctg67
Ctg66
Ctg65
Ctg64
Ctg63
Ctg62
Ctg61
Ctg60
Ctg59
Ctg58
Ctg57
Ctg56
Ctg55
Ctg54
Ctg53
Ctg52
Ctg51
Ctg50
Ctg49
Ctg48
Ctg47
Ctg46
Ctg45
Ctg44
Ctg43
Ctg42
Ctg41
Ctg40
Ctg39
Ctg38
Ctg37
Ctg36
Ctg35
Ctg34
Ctg33
Ctg32
Ctg31
Ctg30
Ctg29
Ctg28
Ctg27
Ctg26
Ctg25
Ctg24
Ctg23
Ctg22
Ctg21
Ctg20
Ctg19
Ctg18
Ctg17
Ctg16
Ctg15
Ctg14
Ctg13
Ctg12
Ctg11
Ctg10
Ctg9
Ctg8
Ctg7
Ctg6
Ctg5
Ctg4
Ctg3
Ctg2
Ctg1
Contigs
0
2.0
4.0
6.0
Mb
🟨 Complete CEN
🟦 Broken CEN
🟥 Copy-like transposon (CoLT)
Duplicated CEN region
D
B. lactucae
P. sojae2019.1
c18
CEN5
c8
c3
CEN3
c56
c33
c40
c43
s2
(SHOA01000004.1)
key:
A Contigs: 🟨 Centromeres ··· assembly gaps
B Copia-like transposon (CoLT) rDNA
C Transposable elements (from the outside inward):
LTR DNA transposons Other transposons
D GC content red above / blue below genome average (5 kb non-overlapping window)
E tRNA genes
F BLASTn links (> 2 kb)
C
Bremia lactucae SF5
Sca28
Sca27
Sca26
Sca25
Sca24
Sca23
Sca22
Sca21
Sca20
Sca19
Sca18
Sca17
Sca16
Sca15
Sca14
Sca13
Sca12
Sca11
Sca10
Sca9
Sca8
Sca7
Sca6
Sca5
Sca4
Sca3
Sca2
Sca1
Scaffolds (> 100 kb)
0
2.0
4.0
6.0
8.0
10.0
Mb
🟩 Predicted CEN region syntenic with P. sojae's
🟦 Predicted CEN not fully syntenic with P. sojae's
🟥 Copy-like transposon (CoLT)
B
CoLT
LTR
LTR
Frame + 1
Frame + 3
gag
PR
IN
RT
RH

Figure 6



Species	Length of CENP-A domain	Composition	Ref.
Phytophthora sojae	116-275 kb	transposons no nucleotide bias	This study
Phaeodactylum tricornutum	2-6 kb	simple AT-rich sequence	17
Plasmodium falciparum	4-5 kb	simple AT-rich sequence	16
Toxoplasma gondii	10 -20 kb	simple sequence no nucleotide bias	18