RESEARCH ARTICLE

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

Yufeng Fang[1], Marco A. Coelho[1], Haidong Shu[2], Klaas Schotanus[1], Bhagya C. Thimmappa[3¤], Vikas Yadav[1], Han Chen[2], Ewa P. Malc[4], Jeremy Wang[4], Piotr A. Mieczkowski[4], Brent Kronmiller[5], Brett M. Tyler[5], Kaustuv Sanyal[3], Suomeng Dong[2], Minou Nowrousian[6], Joseph Heitman[1]*

1 Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, North Carolina, United States of America, **2** College of Plant Protection, Nanjing Agricultural University, Nanjing, China, **3** Molecular Biology and Genetics Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India, **4** Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, United States of America, **5** Center for Genome Research and Biocomputing and Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon, United States of America, **6** Lehrstuhl fuer Molekulare und Zellulaere Botanik, Ruhr-Universitaet Bochum, Bochum, Germany

¤ Current address: Department of Biochemistry, Robert-Cedergren Centre for Bioinformatics and Genomics, University of Montreal, Montreal, Quebec, Canada
* 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 advances the genome assembly.

Citation: Fang Y, Coelho MA, Shu H, Schotanus K, Thimmappa BC, Yadav V, et al. (2020) Long transposon-rich centromeres in an oomycete reveal divergence of centromere features in Stramenopila-Alveolata-Rhizaria lineages. PLoS Genet 16(3): e1008646. https://doi.org/10.1371/journal.pgen.1008646

Editor: Rachel J. O'Neill, University of Connecticut, UNITED STATES

Received: October 3, 2019

Accepted: February 3, 2020

Published: March 9, 2020

Data Availability Statement: The Nanopore sequencing raw reads are available in NCBI under the BioProject PRJNA563922; All raw ChIP-seq reads can be accessed under NCBI's SRA listed in S2 Table. The Psojae2019.1 genome assembly has been deposited at DDBJ/ENA/GenBank under the accession WWEI00000000.1.

Funding: These studies were supported by NIH/NIAID grants R01 AI050113-15 and R37 MERIT award AI039115-21 to J.H., the German Research

Foundation (DFG, grant NO407/7-1) to M.N., and the National Natural Science Foundation of China (grant no. 31772144) to S.D. J.H. is also a co-director and fellow of the CIFAR program, Fungal Kingdom: Threats & Opportunities. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

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 fully embedded in heterochromatin, and 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, broadens our knowledge of centromere structure, function and evolution in eukaryotes, and may help elucidate the high frequency of aneuploidy during oomycete reproduction.

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, 7]. 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, 8, 9]. Some plant centromeres also possess a different type of repeat called centromere-specific retroelements (CR) [10]. 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* [11, 12]. 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* [13, 14], or can be nonrepetitive and very short, such as in the wheat pathogen *Zymoseptoria tritici* [15] and *C. albicans* [16]. 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 [17, 18], while the parasite *Toxoplasma gondii* has a simple centromere without nucleotide bias [19].

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 [13, 20, 21] to insects [22], plants [23] and humans [24].

Oomycetes are fungal-like organisms that belong to the stramenopila kingdom within the Stramenopila-Alveolata-Rhizaria (SAR) supergroup [25, 26]. 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) [27]. *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 [28]. One notorious example is *Phytophthora infestans*, which caused the Great Irish Potato Famine of the mid-1840s [29]. Today, *Phytophthora* species remain significant threats to major food crops, causing multi-billion US dollars losses annually throughout the world [28, 30]. *Phytophthora sojae* is a widespread soil-borne pathogen of soybean. Because of its economic impact, and tractable genetic manipulation [31–33], *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 [34]. 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 [26, 35]. *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 [36, 37]; 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, 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 (chromatin immunoprecipitation followed by high-throughput sequencing) 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 [38]. 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 (S1 Fig and S1 File).

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 model of *CENP-A* that was instead verified by 3'-RACE and RT-PCR, followed by Sanger sequencing (S2A and S2B Fig). *P*. *sojae* CENP-A has a conserved C-terminus including the "CENP-A targeting domain" (CATD) (S2C Fig). *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 (S2D Fig). Overexpressed GFP-CENP-A exhibited nuclear localization with a single fluorescent focus in the nucleus (S2D Fig), 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 (Fig 1A and S3 Fig). Homokaryotic GFP-CENP-A strains exhibited single GFP foci within nuclei from different *P*. *sojae* life stages (Fig 1B), confirming that the clustered centromere organization is a feature in *P*. *sojae*. In addition, we

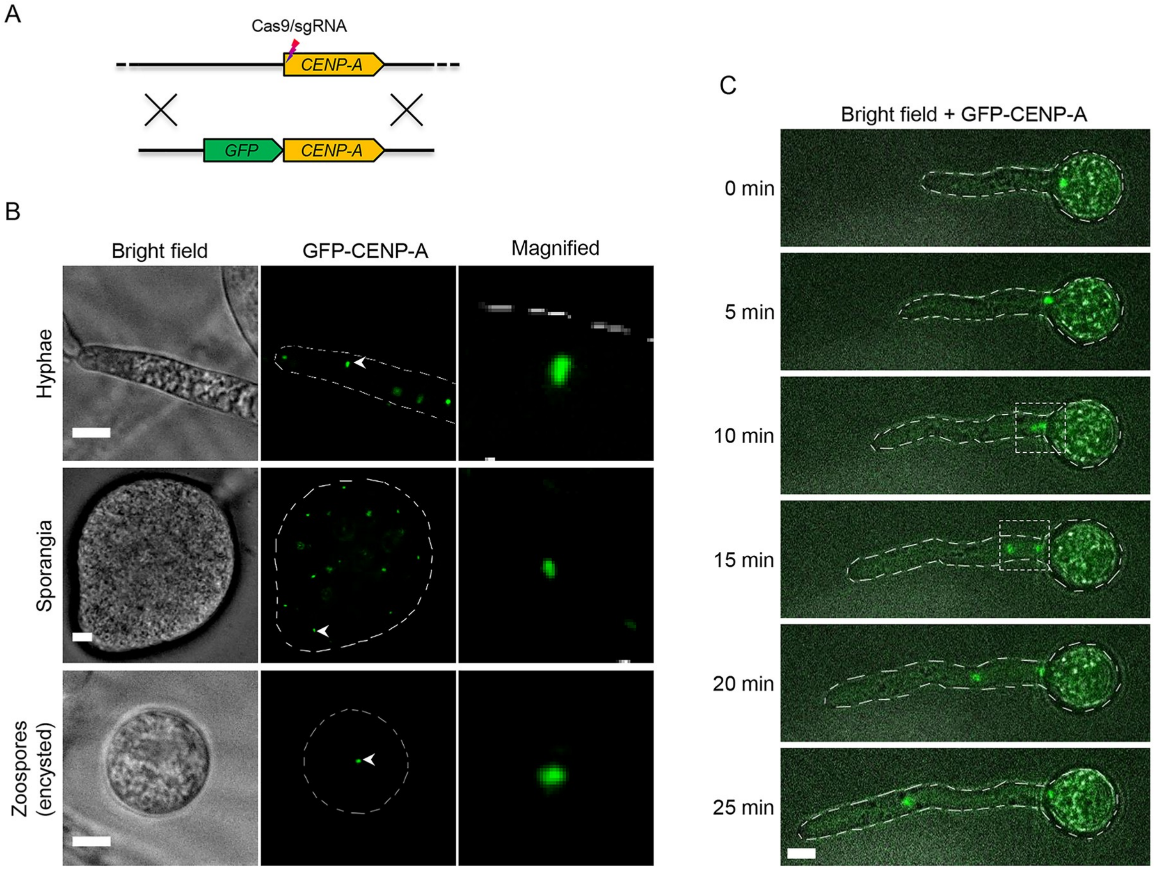

Fig 1. Subcellular localization of CENP-A in *P. sojae* at different life stages and during vegetative growth. (A) A schematic showing the generation of *GFP*-fused *CENP-A* utilizing CRISPR/Cas9 mediated gene replacement. (B) Subcellular localization of GFP-tagged CENP-A (expressed from the endogenous 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.

https://doi.org/10.1371/journal.pgen.1008646.g001

tracked the centromere dynamics during hyphal growth, and found that the clustered centromere pattern was maintained throughout *P. sojae* nuclear division (Fig 1C and S1 Movie).

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 (S4A Fig). 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 (S6 Fig). Thus, we proceeded 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 (S1 Table), and generated a *de novo* genome assembly with SMARTdenovo together with polishing from PacBio and Sanger reads (S5A Fig and S1 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 (S5C Fig). Comparison of Psojae2019.1 to the Sanger assembly indicated that most regions of the two assemblies were collinear, and that Psojae2019.1 has more repetitive sequences (S5B and S5C Fig, also see S1 Text for details). We also checked telomere repeats using a motif proposed for oomycetes [39], and found 13 contigs that harbor telomeric sequences at single ends (versus 6 ends in Sanger, S1 Text and S2 File).

ChIP-seq reads derived from CENP-A were mapped to the new genome assembly Psojae2019.1 (S2 Table), 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 (S6F Fig). 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). We confirmed the integrity of the 10 centromeres assembled within contigs as they were completely covered by long reads (S7 Fig), 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). RNA-seq analysis indicated that all of the 10 CENP-A regions exhibited low transcription, except the region in Contig 11. Contig 11 contained two adjacent CENP-A peaks, one was 18 kb and the other was 114 kb, which were interrupted by a 21 kb transcriptionally active region (S7D Fig). Here, we define the entire region (two CENP-A peaks, and the region between them) as one centromere (*CEN4*). Among the 10 CENP-A regions, five have a length of ~190 kb, and three are relatively smaller, spanning from 115 to 163 kb, while *CEN3* and *CEN7* 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. 54.7%) (Table 1 and S5C Fig). Taken together, our CENP-A ChIP-seq analysis utilizing the newly assembled genome indicates that *P. sojae* CENP-A exhibits preferential binding to 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, Fig 3 and S6 Fig). Contig 51 was collinear with the Sanger scaffold 23; however, no enriched CENP-A

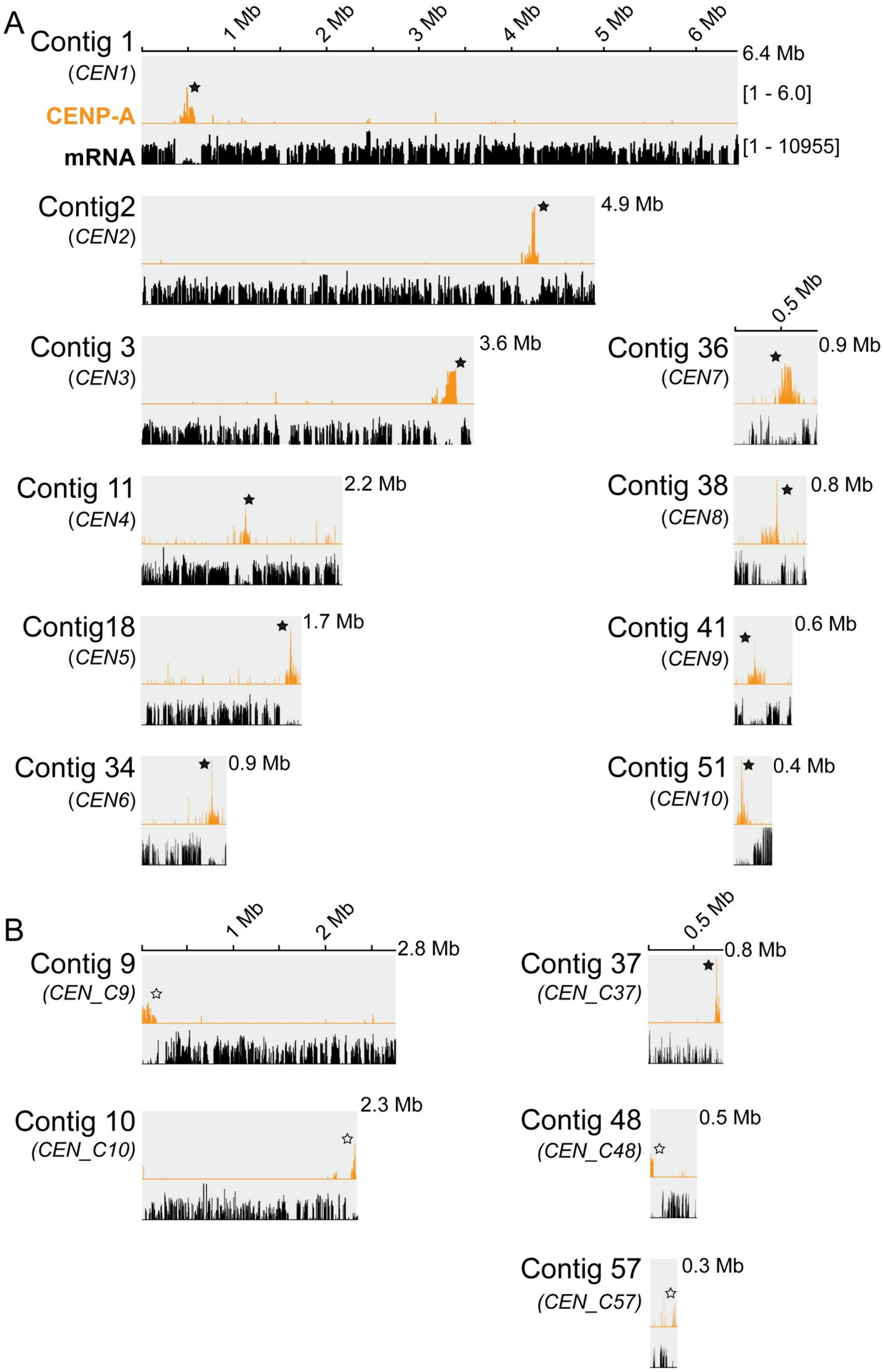

Fig 2. Contigs in the Psojae2019.1 assembly demonstrating CENP-A enrichment based on ChIP-seq. (A) 10 contigs that harbor fully assembled CENP-A binding sites. (B) Five contigs that possess incompletely assembled CENP-A binding regions. All contigs are drawn to scale and the ruler indicates the length 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.

https://doi.org/10.1371/journal.pgen.1008646.g002

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 core centromere (kb)***
CEN1	1	415–578 (163)	56.91	361–415 (54) 579–650 (71)	2	7086–7267
CEN2	2	4102–4295 (193)	55.51	4094–4102 (8) 4296–4305 (9)	8	2374–2420
CEN3	3	3138–3410 (272)	54.81	3223–3138 (15) 3412–3560 (48)	9	3075–3286
CEN4†	11	991–1174 (183)	58.13	944–991 (47) 1175–1205 (30)	4	995–1246
CEN5	18	1556–1706 (150)	57.93	1492–1556 (64) 1709-? (>22)	3	4078–4138
CEN6	34	697–880 (183)	57.65	643–696 (53) 882–904 (22)	6	1521–1726
CEN7	36	432–706 (274)	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	154–342 (188)	57.40	101–153 (52) 342–358 (16)	1	2921–3079
CEN10	51	36–151 (115)	57.93	?-36 (>36)† 152–216 (64)	-	-

* The coordinates of core centromeres in the Psojae2019.1 assembly are defined by peak calling (S7 File). The positions of centromeres in the Sanger assembly are defined by visualization of peaks.

† Contig11 contains a minor (coordinate, 991,465–1,009,035, 18 kb) and a major (coordinate, 1,060,806–1,174,302, 114 kb) peaks that are separated by a 21 kb transcriptionally active region. The core centromere size is defined by the combination of the minor and major CENP-A regions, and the region between the two peaks.

‡ One side of pericentric heterochromatin region is not fully assembled.

https://doi.org/10.1371/journal.pgen.1008646.t001

signal was detected for this scaffold, probably because the region corresponding to *CEN10* is interrupted by gaps. Notably, two CENP-A binding regions in Sanger Scaffold 1 were found to correspond to *CEN8* and *CEN9*, and the smaller one (coordinates: 9,667–9,688 kb) was expanded from 20 kb to 188 kb corresponding to *CEN8* (Table 1, S4B and S6H Figs). 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 (S6H Fig). 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.

Due to large genome scales and potentially similar chromosome sizes, the karyotypes of *Phytophthora* species cannot be well resolved by pulsed-field gel electrophoresis [32, 40]. 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 [41]. By comparing the location of centromeres in the Sanger and Psojae2019.1 assembly, we could validate and predict the configuration of 11 centromeres, namely *CEN1*-*CEN10*, and *CEN_C9* + *CEN_C48* (Table 1 and S4 Table). Three centromeres, namely *CEN_C37*, *CEN_C10* and *CEN_C57*, are not fully assembled. Thus, *P. sojae* is estimated to harbor 12–14 chromosomes, which is in agreement with the previous cytological study.

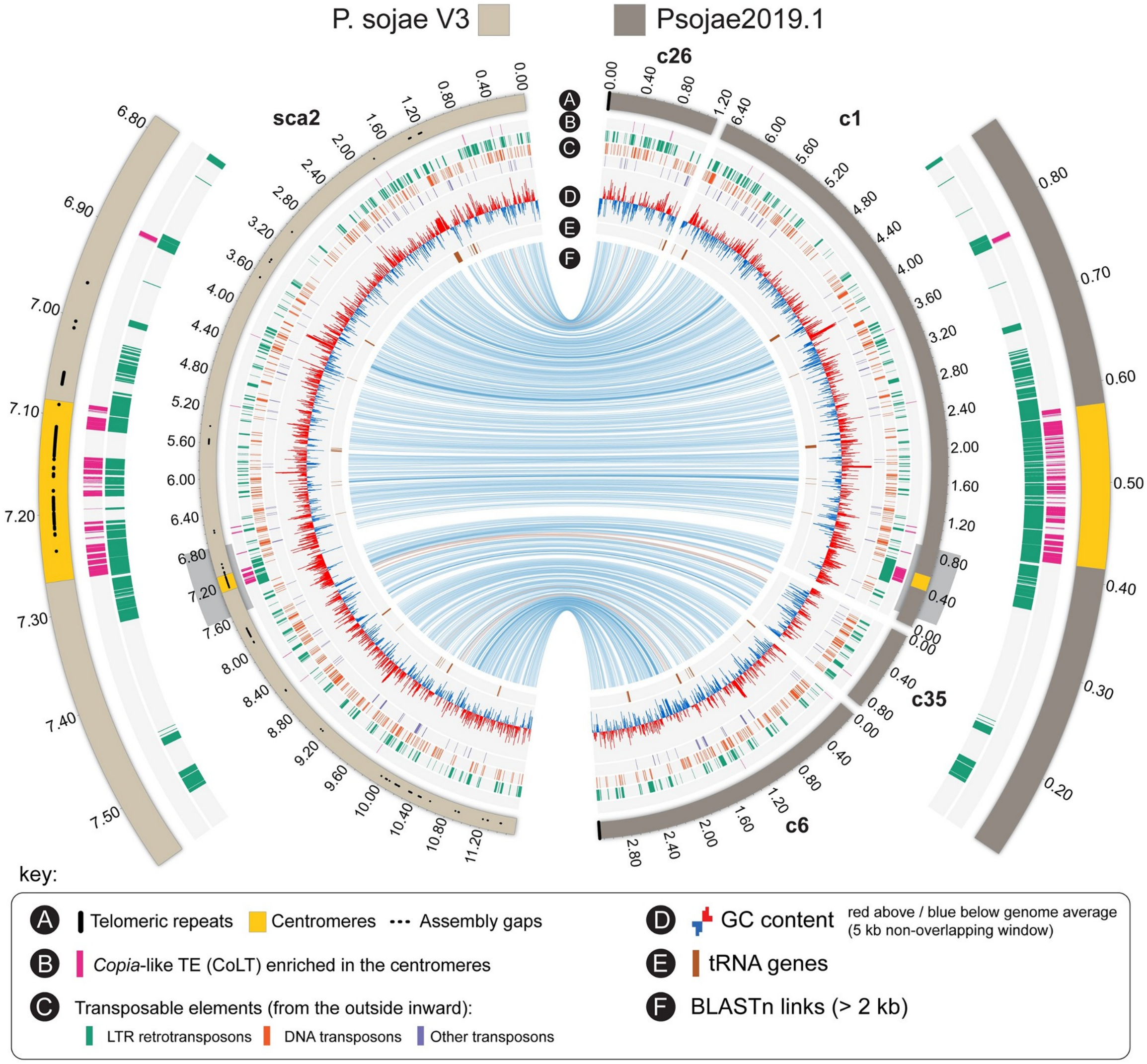

Fig 3. A representative Circos visualization comparing a centromere-containing scaffold in Sanger V3 (Scaffold 2) with its corresponding contigs in the Psojae2019.1 assembly. For the inner circle, track A illustrates assembled contigs (in Psojae2019.1) or scaffold (in P. sojae V3) that are color coded as shown at the top. The locations of centromeres (CENP-A binding regions) are highlighted in yellow. Tracks B-E show the location of other genomic features as given in the key on the bottom. Blue and orange lines in track F link regions with collinearity extending over 2 kb, with orange lines corresponding to inversions. Grey box-shaded centromere-containing regions are magnified for detailed visualization (the two outer arcs).

https://doi.org/10.1371/journal.pgen.1008646.g003

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 (S7 Fig). Intriguingly, the heterochromatic region extended 8 kb to 64 kb beyond each CENP-A binding region (Table 1 and Fig 4A), similar to pericentromeric heterochromatin regions described in other species [14, 22, 42]. In contrast, the euchromatic mark H3K4me2 was excluded from the CENP-A region and its flanking heterochromatic

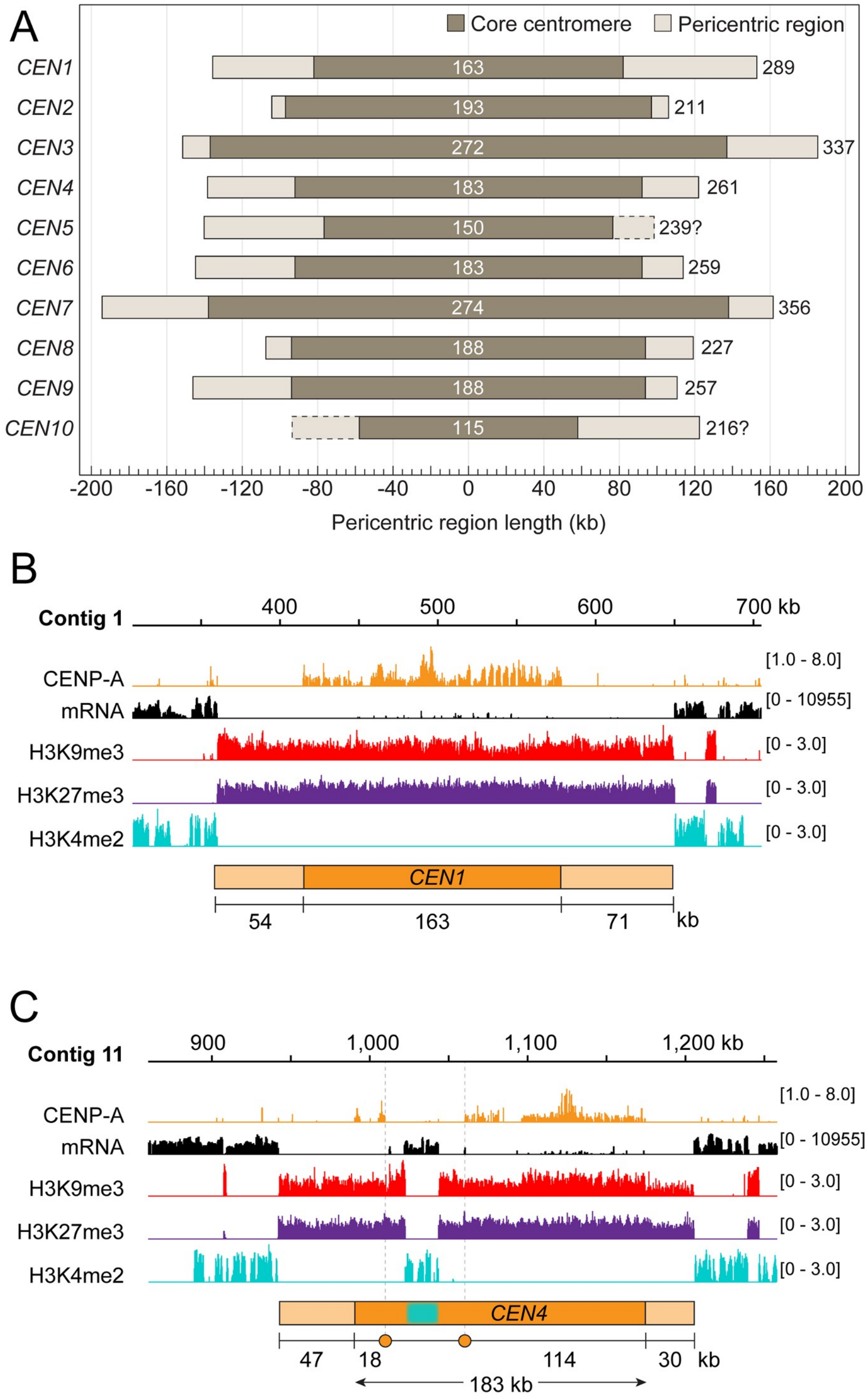

Fig 4. Epigenetic state of *P. sojae* centromeres. (A) Schematics showing the *P. sojae* core centromeres (CENP-A binding regions) and the pericentric regions of various lengths. Dark and light grey bars indicate core centromeric and pericentric regions. Numbers at the center indicate the size of core centromeres; Numbers on the right 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*, indicated by dashed bars, are not fully assembled, and their lengths are 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. A 400 kb region harboring the centromeric region is shown for *CEN1* and *CEN4*. Cyan block, a transcriptionally active region (21 kb) that interrupts *CEN4*. Profiles of CENP-A, H3K9me3, H3K27me3 and H3K4me2 shown were normalized to input. mRNA profiles are shown as log-scales. Orange circles in (C) delimit borders of a CENP-A binding void.

https://doi.org/10.1371/journal.pgen.1008646.g004

regions, and generally overlapped with the mRNA transcriptional profile (Fig4B and 4C and S7 Fig). Thus, distribution of histone modifications suggests that the CENP-A regions are embedded in heterochromatin, and we define the heterochromatic regions adjacent to the CENP-A peaks 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) (S5C and S5D Fig). Our analysis showed that centromeres are structurally organized differently, but all are composed of many repetitive elements, mostly LTR-retrotransposons (Figs 3 and 5A, S6 and S7 Figs). To identify whether the centromeres in *P. sojae* possess any common sequences or repeat elements, all identified CENP-A regions

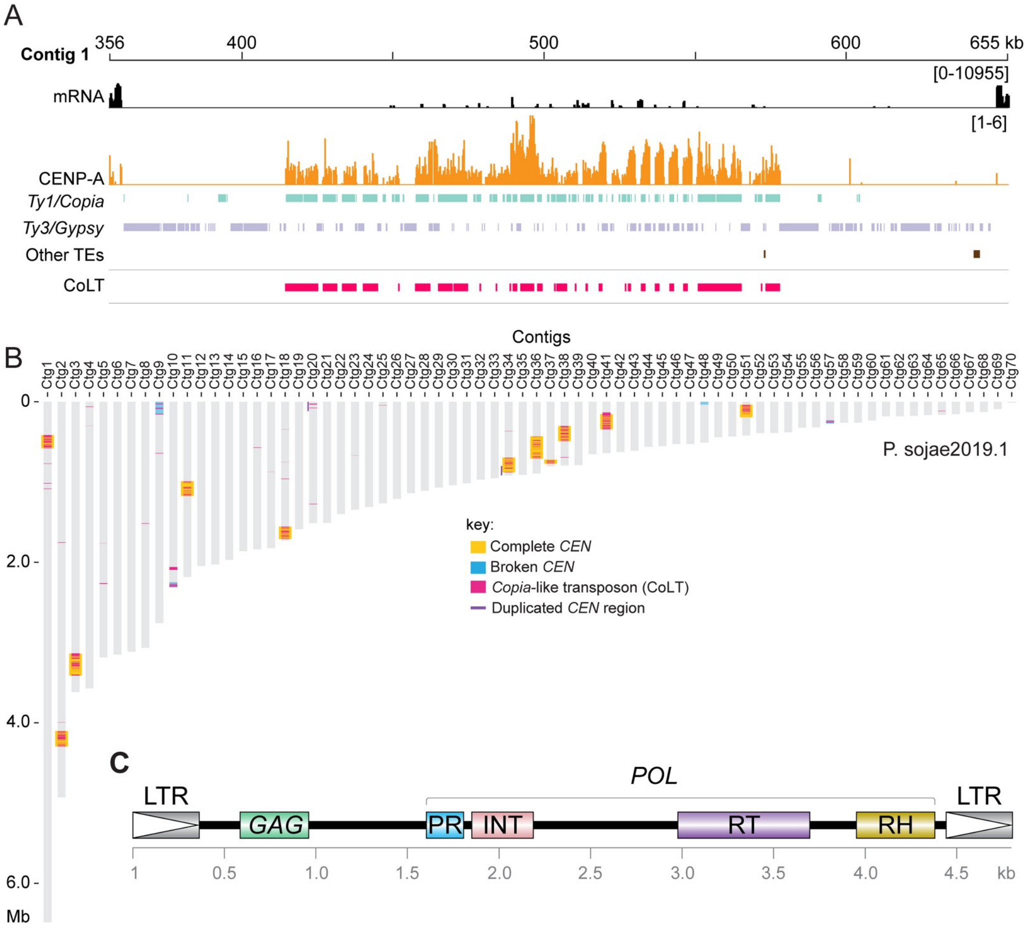

Fig 5. *P. sojae* centromeres are enriched for a *Copia*-like transposon (CoLT). (A) Distribution of transposable elements (TEs) in *CEN1*. TEs were annotated using a *Phytophthora* TE library from Repbase [45]. Tracks of different repeat families are color coded. The track "Other TEs" includes all types of TEs beyond *Gypsy* and *Copia*. CoLT was composed of three elements annotated in Repbase [45], namely Copia-24_PIT-I, Copia-24_PIT-LTR and Gypsy-P17-PR-I. The profile of CENP-A shown was normalized to input. The mRNA track is shown as log-scales and used to define the boundary of pericentric heterochromatin. (B) Location of CoLT elements across all of the Psojae2019.1 contigs. (C) Diagram showing the domain structure of a representative full-length CoLT sequence (see S3 File). The coding domain featured as *Copia* superfamily of retrotransposons, which consists of capsid protein (GAG), Gag-pre-integrase (PR), integrase (INT), reverse transcriptase (RT), and RNase H (RH) domains, and diverges from the *Gypsy* superfamily in the order of the RT and INT domains in their *POL* genes [86].

https://doi.org/10.1371/journal.pgen.1008646.g005

were subject to multiple sequence alignment. This analysis found an ~5 kb sequence that is highly similar (>98%) and shared among 10 centromeres (S8 Fig and S3 File). 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 5B). Moreover, this element is present as clusters in centromeric regions, and only sparsely found in other regions of the genome (Fig 5B), further strengthening its association with centromeres.

In parallel, examination of the 5 kb sequence indicated that it resembled a *Copia*-like transposon that comprises a *GAG* gene, and a *POL* gene encoding PR (protease), INT (integrase), RT (reverse transcriptase), and RH (RNase H) domains in order (Fig 5C and S3 File). To identify the long terminal repeats (LTRs) of CoLT, we analyzed the best BLAST hit of the 5 kb sequence. Examination of its flanking nucleotides and the LTR marks (5' . . .TG-3' and 5' . . .CA-3') enabled us to identify the LTRs that are nearly identical (Fig 5C and S3 File). Hence, we named this retroelement CoLT for *Copia*-Like Transposon. Phylogenetic analysis based on the conserved RT domains of various reported retroelements confirmed the classification of CoLT as a *Ty1*/*Copia* retrotransposon (Fig 6A). Notably, among the selected retroelements shown in Fig 6A, *P*. *sojae* CoLT is distinct from previously studied oomycete *Copia* retrotransposons [43, 44], but clusters together with *P*. *infestans Copia*-24_PIT-I annotated in Repbase [45], and with predicted CoLT sequences found in other two oomycetes as described below. In addition, phylogenetic tree also suggests that CoLT is close to *Copia* sequences found in plants (Fig 6A).

Further BLAST analysis employing the full-length CoLT including LTRs as a query indicated that the Psojae2019.1 assembly, in total, harbors 80 CoLT elements of similar length including 11 full-length elements that possess one long ORF encoding all of the *Copia* domains, and 69 sequences that have frameshift mutations within domains (S4B File). Phylogenetic analysis based on the DNA sequences of these elements revealed that the centromeric CoLT copies cannot be distinguished from those found elsewhere in the genome, indicating that centromeric and non-centromeric CoLT copies do not evolve separately; however, it is compelling that almost all full-length CoLT copies (10 out of 11) are found in centromeres (Fig 6B). Collectively, our analyses suggest that CoLT, a retrotransposon that harbors all predicted functional domains characteristic for a *Copia*-like transposon, is highly enriched in *P*. *sojae* centromeres, and is also the only feature shared by all of the centromeres.

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 7A and 8A). 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 syntenic with *P*. *sojae* centromeres (Figs 7A). Unexpectedly, Scaffold 2 (original name, SHOA01000004.1, see S5 File for details) contained two CoLT clusters that were syntenic with *P*. *sojae CEN3* and *CEN5* (Fig 7B), indicating that Scaffold 2 may be incorrectly assembled. It should be noted that the *B*. *lactucae* genome assembly still has a large percentage of unresolved

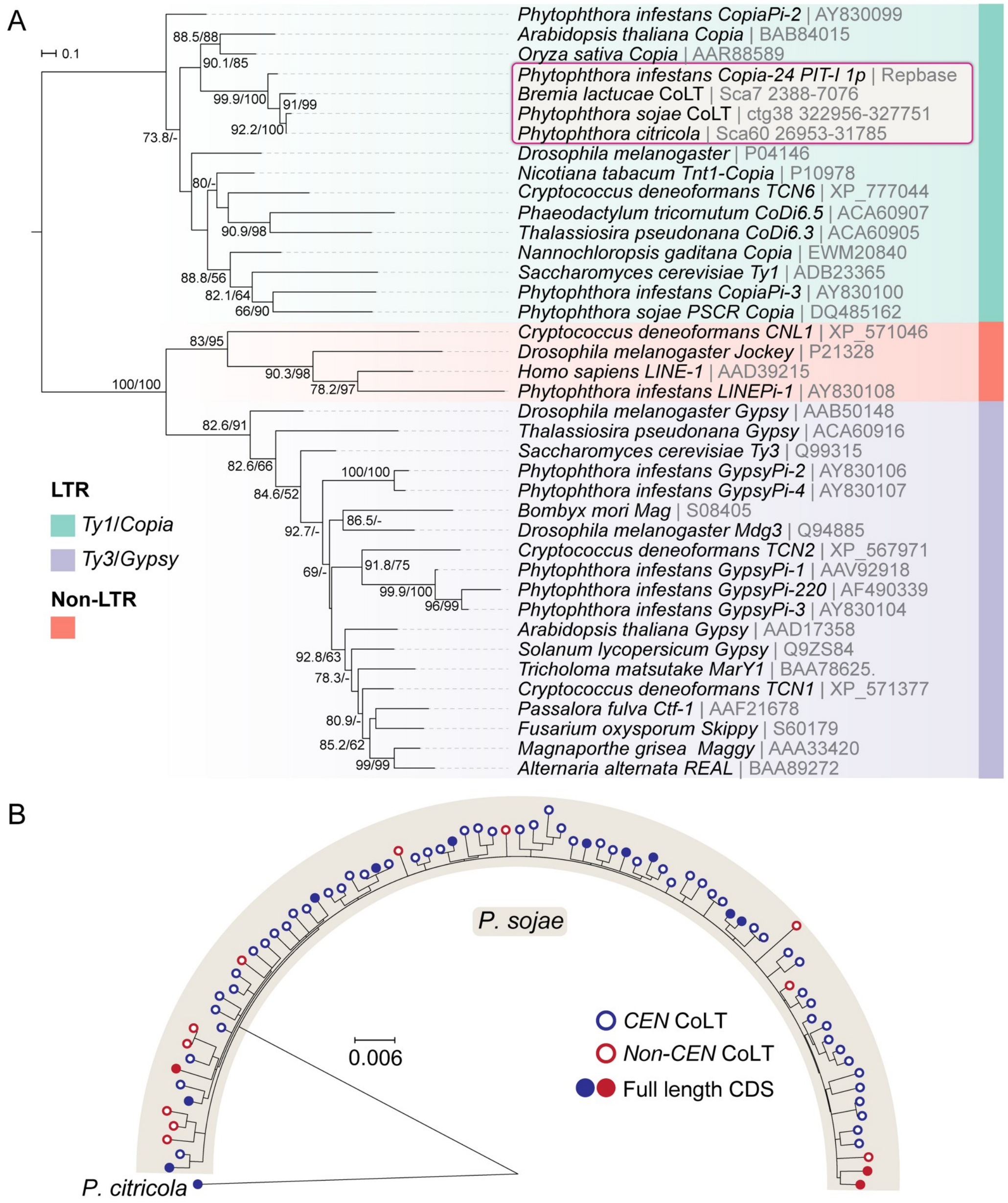
A
0.1
Phytophthora infestans CopiaPi-2 | AY830099
Arabidopsis thaliana Copia | BAB84015
Oryza sativa Copia | AAR88589
Phytophthora infestans Copia-24 PIT-I 1p | Repbase
Bremia lactucae CoLT | Sca7 2388-7076
Phytophthora sojae CoLT | ctg38 322956-327751
Phytophthora citricola | Sca60 26953-31785
Drosophila melanogaster | P04146
Nicotiana tabacum Tnt1-Copia | P10978
Cryptococcus deneoformans TCN6 | XP_777044
Phaeodactylum tricornutum CoDi6.5 | ACA60907
Thalassiosira pseudonana CoDi6.3 | ACA60905
Nannochloropsis gaditana Copia | EWM20840
Saccharomyces cerevisiae Ty1 | ADB23365
Phytophthora infestans CopiaPi-3 | AY830100
Phytophthora sojae PSCR Copia | DQ485162
Cryptococcus deneoformans CNL1 | XP_571046
Drosophila melanogaster Jockey | P21328
Homo sapiens LINE-1 | AAD39215
Phytophthora infestans LINEPi-1 | AY830108
Drosophila melanogaster Gypsy | AAB50148
Thalassiosira pseudonana Gypsy | ACA60916
Saccharomyces cerevisiae Ty3 | Q99315
Phytophthora infestans GypsyPi-2 | AY830106
Phytophthora infestans GypsyPi-4 | AY830107
Bombyx mori Mag | S08405
Drosophila melanogaster Mdg3 | Q94885
Cryptococcus deneoformans TCN2 | XP_567971
Phytophthora infestans GypsyPi-1 | AAV92918
Phytophthora infestans GypsyPi-220 | AF490339
Phytophthora infestans GypsyPi-3 | AY830104
Arabidopsis thaliana Gypsy | AAD17358
Solanum lycopersicum Gypsy | Q9ZS84
Tricholoma matsutake MarY1 | BAA78625.
Cryptococcus deneoformans TCN1 | XP_571377
Passalora fulva Ctf-1 | AAF21678
Fusarium oxysporum Skippy | S60179
Magnaporthe grisea Maggy | AAA33420
Alternaria alternata REAL | BAA89272
88.5/88
90.1/85
99.9/100
91/99
92.2/100
73.8/-
80/-
90.9/98
88.8/56
82.1/64
66/90
83/95
90.3/98
78.2/97
100/100
82.6/91
82.6/66
100/100
84.6/52
86.5/-
92.7/-
91.8/75
69/-
99.9/100
96/99
92.8/63
78.3/-
80.9/-
85.2/62
99/99
LTR
Ty1/Copia
Ty3/Gypsy
Non-LTR
B
P. sojae
0.006
CEN CoLT
Non-CEN CoLT
Full length CDS
P. citricola

Fig 6. Phylogenetic analyses of CoLT. (A) Maximum-likelihood phylogeny of different retroelements. Protein sequences of the reverse transcriptase (RT) domains were used to construct the tree [87] (see S4A File for the complete DNA and protein sequences). The tree was rooted in the midpoint and branch support values (> 50%) shown at the three nodes were determined by 10,000 replicates of both ultrafast bootstrap approximation (UFBoot) and the Shimodaira-Hasegawa approximate likelihood ratio test (SH-aLRT). (B) Maximum-likelihood phylogeny of full-length CoLT elements identified in the Psojae2010.1 genome assembly. CoLT copies located inside and outside of centromeres are denoted by blue and red circles. CoLT elements with full-length coding sequences (CDS) are depicted as filled circles. A CoLT sequence identified in *P. citricola* served as an outgroup. See S4B File for the CoLT copies used for the phylogenetic analysis.

https://doi.org/10.1371/journal.pgen.1008646.g006

gaps likely due to its highly heterozygous nature [37]. 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) (Fig 8). However, a large number of the CoLT clusters localized at contig ends, or were distributed across

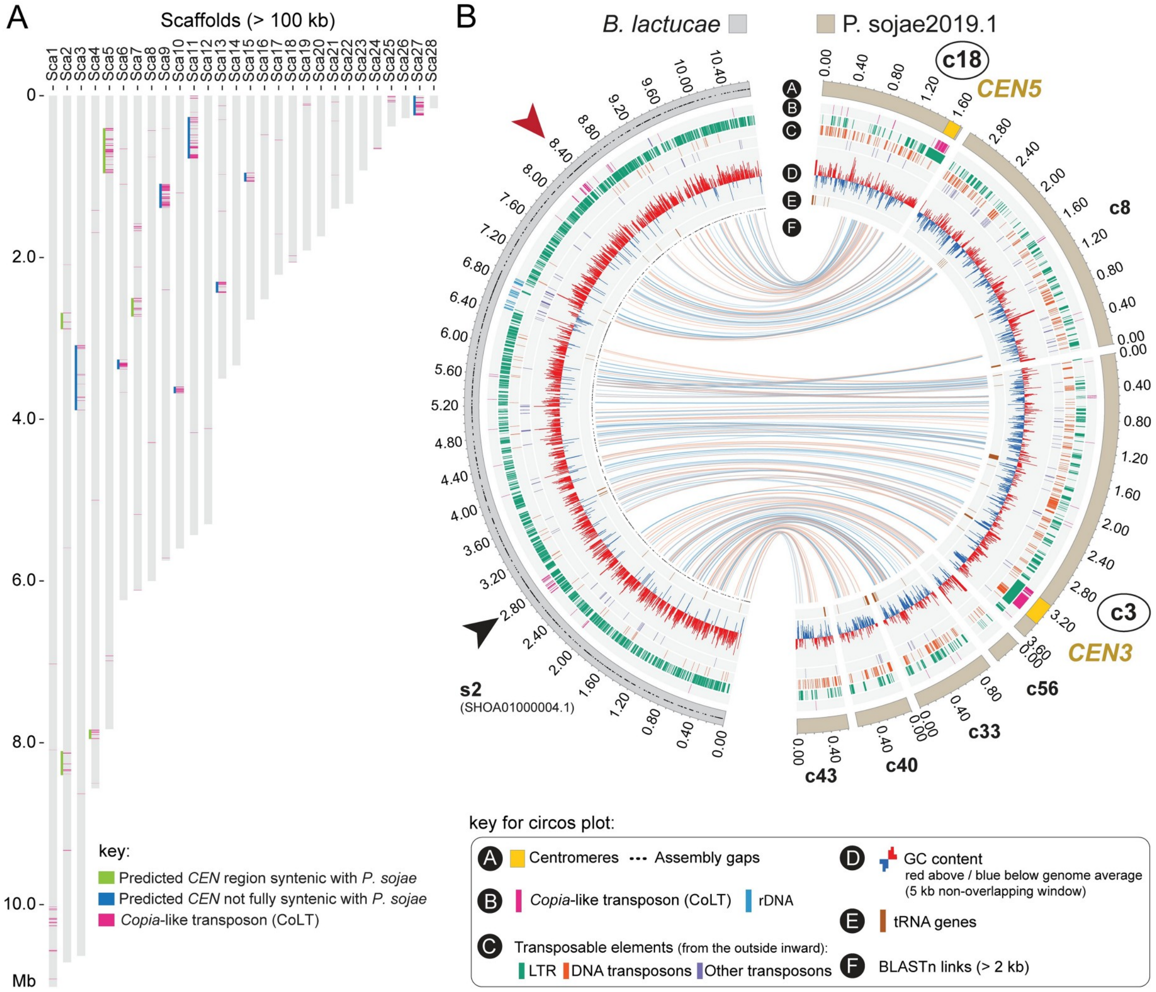

Fig 7. Genomic distribution of CoLT in the *Bremia lactucae* genome. (A) Location of CoLT elements across all *B. lactucae* scaffolds >100 kb. For ease of analysis, scaffolds in the *B. lactucae* assembly were sorted and re-named based on sizes (large to small). See S5 File for the original scaffold names. Regions underlined by green lines indicate that both sides of the CoLT clusters are syntenic with the regions surrounding the *P. sojae* centromeres. Regions underlined by blue indicate only one side of the CoLT clusters was found to be syntenic with *P. sojae* centromere flanking sequences. (B) A representative Circos plot comparing a *B. lactucae* scaffold that has clustered CoLT elements (Scaffold 2)with the corresponding Psojae2019.1 contigs. The outer track illustrates the assembled scaffold (in the sorted *B. lactucae* assembly) or contigs (in Psojae2019.1), which is color coded as shown at the top. Names of contigs possessing *P. sojae* centromeres are enclosed in circles. Yellow regions on the outer tracks indicate the locations of centromeres (CENP-A binding regions). Blue and orange lines in track F link regions with synteny extending over 2 kb, with orange lines corresponding to inversions. Two CoLT clusters are identified in the *B. lactucae* Scaffold 2 (s2; original scaffold, SHOA01000004.1), which are indicated by arrowheads of different colors.

https://doi.org/10.1371/journal.pgen.1008646.g007

the length of short contigs (Fig 8A). 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 a criterion to predict centromere regions in other oomycete species.

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

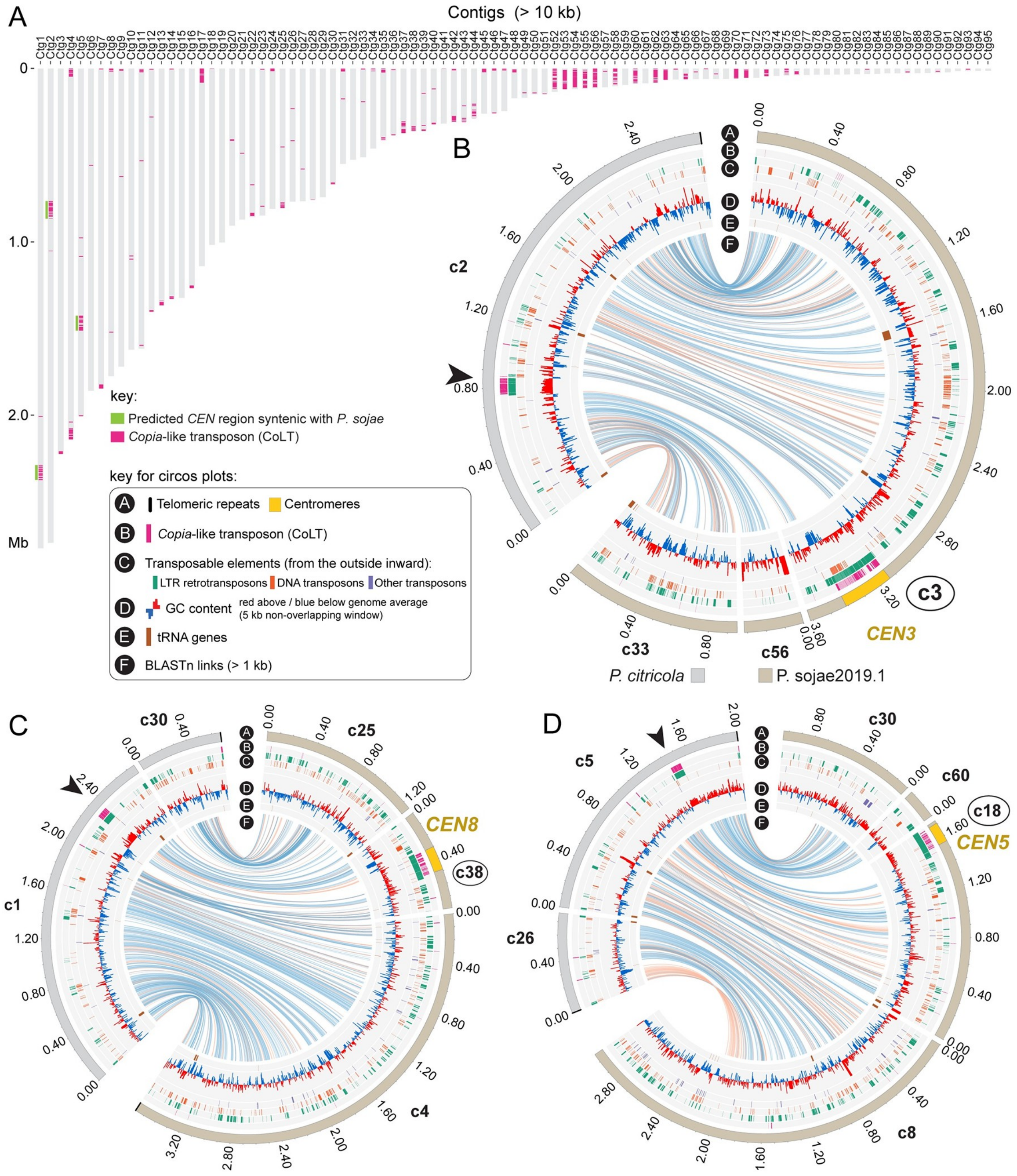

Fig 8. Genomic distribution of CoLT in the *Phytophthora citricola* genome. (A) Location of CoLT across all *P. citricola* contigs >100 kb. (B-D) Circos plots comparing three *P. citricola* contigs that have CoLT clusters with the corresponding Psojae2019.1 contigs. In each panel, the outer track (bars) illustrates the contigs in the *P. citricola* or the Psojae2019.1 genome assemblies, which is color coded as shown below the Circos plot in panel B. Yellow regions indicate the locations of centromeres (CENP-A binding regions). Names of contigs harboring *P. sojae* centromeres are enclosed in circles. Blue and orange lines in track F link regions with synteny extending over 2 kb, with orange lines corresponding to inversions. CoLT clusters present in the *P. citricola* contigs are indicated by arrowheads. The flanking sequences of each CoLT cluster are syntenic with the regions surrounding the *P. sojae* centromere.

https://doi.org/10.1371/journal.pgen.1008646.g008

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, 46]. 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 9). 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* [14] and *Cryptococcus neoformans* [13]. 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 [42, 47–49], 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 [42, 49, 50]. 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* [14, 51, 52]. 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 [9, 53], and in the centromeres of

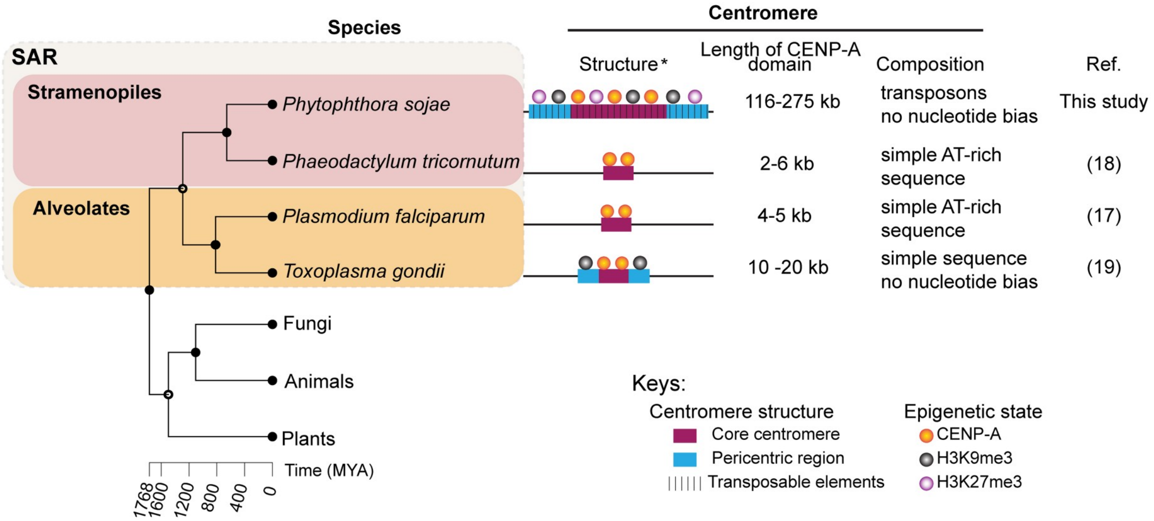

Fig 9. Diversity of centromere features within the Stramenopila-Alveolata-Rhizaria (SAR) supergroup. Simplified schematics (not to scale) showing the structures, epigenetic modifications, sizes and compositions of reported centromeres across the SAR lineages. Asterisk, epigenetic state was not examined in the diatom centromeric regions; however, several AT-rich DNA sequences can be employed for episome maintenance, suggesting diatom centromere might not be epigenetically dependent. The histone modification H3K27me3 was only tested in *P*. *sojae*. Phylogenetic tree was constructed using TimeTree [88]. *Homo sapiens*, *Arabidopsis thaliana* and *Neurospora crassa* were used as representatives of animals, plants and fungi for the phylogenetic analysis, and are used as outgroups illustrating the evolutionary status of the SAR supergroup.

https://doi.org/10.1371/journal.pgen.1008646.g009

Z. tritici accessory chromosomes (as their whole accessory chromosomes are enriched for H3K27me3) [15]. 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 [54]. While animal centromeres are associated with both satellite DNA and retroelements, satellite DNA is usually regarded as the main sequence component [55]. 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 [8]. Centromeres of *N. crassa* [14] and *C. neoformans* [13] are composed of retrotransposons, and the retroelements in *C. neoformans* are centromere-specific [13]. 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 (Fig 5A and S7 Fig). We identified 11 CoLT homologs possessing intact domains that are typical for an active *Ty1*/*Copia* retrotransposon (Fig 5B), indicating that they may be still active. High similarity of LTRs of each CoLT indicate the transposition may have occurred recently (S3 File). Interestingly, two independent studies coincidentally employed the central and the 3'- terminal parts of CoLT as probes (S3 File) for DNA fingerprinting of *P. sojae* isolates, and polymorphisms were observed by Southern blotting among strains isolated from different geographic locations [56, 57]. This suggests that CoLT may be mobile. In contrast, RNA-seq and histone modification data derived from the mycelial stage demonstrated that all of the 11 CoLT copies were poorly transcribed and were heterochromatic (S7 Fig), indicating that the element may be inactive at least during mycelial growth. Therefore, it remains to be determined in which life stages or conditions CoLT becomes active. A similar distribution pattern of centromere-associated retrotransposons was recently found in *Drosophila melanogaster* [22]. In *D. melanogaster*, a non-LTR retroelement named *G2*/*Jockey-3* was found to be enriched in CENP-A chromatin, and this element is also associated with centromeres in its sister species *D. simulans* [22]. 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.

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. Despite the fact that most of the ChIP-seq reads of CENP-A are clustered and define the centromeres, we found that additional reads were distributed sparsely out of centromeres. This could be caused by the mapping algorithm, because elements present in the centromeres may also be found in non-centromeric regions, for instance, CoLT. In fact, we observed that all of the CoLT elements including the truncated copies that are present outside of centromeres have peaks of their chromatin marks, and thus are indistinguishable from the centromere-associated ones. Future studies independent of genome assembly, such as fluorescence *in situ* hybridization (FISH) and Hi-C will probably overcome these technical challenges and confirm the identity of the centromere regions mapped by ChIP-seq.

Our analysis showed that having an improved reference genome assembly based on long-read sequencing technologies was crucial for 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 (S6 Fig). We tried to scaffold the assembly with different scaffolding programs such as npScarf [58], SSPACE [59], LINKS [60] and the optical Bionano mapping (S1 Text and S9 Fig). 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 (S3 Table and S10 Fig). Thus, we opted to retain the contig-level assembly in our study. 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 the basis of the presence of centromeres and predicted telomeres together with synteny analyses, we found that three Sanger scaffolds/Psojae2019.1 contigs may represent full-length chromosomes, namely Scaffold 2/Contigs [26+1+35+6] (S6A Fig), Scaffold 5/Contigs [17+36+7+49+45] (S6G Fig); and partial Scaffold 1/Contigs [58+38+4] (S6H Fig). Notably, telomeres appear on the both ends of Sanger Scaffold 5 and its syntenic contigs in Psojae2019.1 (S6G Fig). 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 [61]. 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 [62]. Polyploidy and aneuploidy are prevalent in *Phytophthora* natural isolates and in progeny from sexual reproduction [36, 63–66]. 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 [65]. Recently, a phenomenon termed dynamic extreme aneuploidy (DEA) was described in a vegetable oomycete pathogen, *Phytophthora capsici*, in which high variability among progeny produced by asexual spores was caused by ploidy variation [67]. However, the mechanisms resulting in oomycete aneuploidy and/or polyploidy are 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 S5 Table. 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 [31] was applied to examine the nuclear localization of CENP-A. Stable and homokaryotic transformants expressing GFP tagged CENP-A (driven by strong promoters derived from the *HAM34* gene) 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 [31]. Sporangia and zoospores were induced by water flooding according to a method described previously [68].

Construction of plasmids

All the primers used in this study are listed in S6 Table. All GFP fusion constructs were generated based on the plasmid backbone pYF3-GFP [69], 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, 18373–019). All PCR-amplifications were performed using Phusion High-Fidelity DNA Polymerase (NEB, M0530S).

CRISPR-mediated gene replacement

An sgRNA guide sequence whose PAM sequence overlapped with the start codon of *CENP-A* was selected as the CRISPR/Cas9 target. An oligo annealing strategy was used for assembly of the sgRNA expression cassettes according to previously described methods [31]. HDR templates for *CENP-A* was assembled using NEBuilder HiFi DNA Assembly (NEB, E2621S). 5'-junction, 3'-junction and spanning diagnostic PCR were performed to genotype mutants, utilizing the primers listed in S6 Table.

Microscopy imaging of *P. sojae* transformants

A Zeiss 780 inverted confocal microscope was used 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 using 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°C 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 centrifuged at 4˚C, 5000 g for 15 minutes. The supernatant was transferred to a new 50 ml tube and treated with RNase A (final concentration, 100 μg/ml) at 37˚C for about 1 hour, followed by proteinase K treatment (final concentration 200 μg/ml) at 50˚C for 2 hours. An equal volume of chloroform was added to the solution and mixed gently by inverting the tube, then centrifuged at 4˚C, 5000 g for 15minutes. The supernatant was transferred to a new 50 ml Falcon tube and gDNA was 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 pulsed 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) employing Guppy v2.0.5. See S1 Table for the metrics of the ONT reads.

Native ChIP-seq

Native ChIP was performed according to the ChIP protocol accompanying Gent, Wang [70] 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 then digested by micrococcal nuclease (MNase, M0247S, NEB) at 37˚C for 6 min. An antibody against GFP (Abcam, ab290) was used to immunoprecipitate single nucleosomes containing the GFP-CENP-A fusion. 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 (80%-85% mappability, S2 Table); ChIP-seq of H3K9me3 and H3K4me2 were conducted by BGI using Illumina Hiseq 4000 that produced 50 nucleotide single-end reads (~98% mappability, S2 Table). Numbers of reads for each sample are listed in S2 Table.

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 and H3K27me3, the resulting reads were trimmed by fastx-clipper and mapped with Bowtie2 using default parameters [71], and aligned to the genome assemblies. For H3K9me3 and H3K4me2, the ChIP-seq reads were polished by BGI prior to be released, and thus were mapped to the genomes directly using the same Bowtie2 setup. In case of replicates, the ChIP-seq reads were mapped randomly among all replicates. 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 signals (log2[$\text{ChIP}_{\text{RPKM}}$/$\text{Input}_{\text{RPKM}}$]), and bigwig files (.bw) were generated. Then .bw files were visualized using the Integrative Genome Viewer (IGV). (https://software.broadinstitute.org/software/igv/). Core centromeres (CENP-A binding regions) were defined by one continuous stretch of ChIP-seq peaks (> 5 kb) in scaffolds or contigs in the Sanger or Psojae2019.1 genome assembly. The boundaries of the 10 fully assembled CENP-A regions were defined by peaking calling employing MACS v2.2.5 (–broad -q 0.00001—broad-cutoff 0.00001). Peaks with enrichment fold smaller than 2 were filtered out (S7 File). The boundaries of pericentric heterochromatin were defined by the start of both RNA-seq and H3K4me2 peaks. To generate mRNA profiles, the existing RNA-seq reads (SRA: SRR10283202) derived from the mycelial stage 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 are described in S1 Text. To predict gene models, first, the assembly Psojae2019.1 was subjected to repeat masking utilizing RepeatMasker [72] 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) [73] with predicted proteins from available *P*. *sojae* and *P*. *infestans* genome annotations as input [26, 74]. 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) [75] and tRNAscan-SE (v2.0) [76], respectively. To find telomeres, a custom-made Perl script was used to search for the sequence "TTTAGGG" that was proposed for oomycetes telomeric sequences [39]. 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) [77]. Whole-genome alignment was computed by MashMap (https://github.com/marbl/MashMap) employing default settings, and was visualized as a dot plot [78].

Bionano mapping

P. *sojae* protoplasts were generated from 2.5-day old mycelia 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. See S6 File for the metrics of the Bionano mapping.

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 fashion 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 the sequence was highly similar to a *Copia*-like transposable element. To define the domains of the CoLT element, this sequence was further analyzed by repeat identification (utilizing a bioinformatics software Unipro UGENE [79]), 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).

To validate the classification of CoLT as a *Copia* retrotransposon, phylogenetic analyses were performed based on the alignment of the reverse transcriptase (RT) domains of the identified *P. sojae* CoLT, and of LTR (*Gypsy* and *Copia*) and LINE retrotransposons previously characterized in other organisms. NCBI accession number of each TE sequence is shown in Fig 6A. Representative sequences and the deduced domains of *P. sojae*, *B. lactucae* and *P. citricola* CoLT homologs are shown in S3 File. A total of 39 protein sequences were aligned with MAFFT v7.310 [80] employing the L-INS-i strategy (- -localpair- -maxiterate 1000) and poorly aligned regions were removed with TrimAl (-gappyout) [81]. A maximum-likelihood phylogeny was inferred using the LG+I+G4 model of amino acid substitution in IQ-TREE v1.6.5 [82]. Branch support values were obtained from 10,000 replicates of both the ultrafast bootstrap approximation (UFboot) [83] and the SH-like approximate likelihood ratio test (SH-aLRT) [84]. Additionally, to investigate if the centromeric CoLT sequences have a common origin, BLASTn was conducted using one full-length CoLT (See S3 File for the sequence). All of the full-length CoLT copies were extracted, and hits with less than 95% sequence identity and shorter than 95% of the query length were removed. The obtained genomic coordinates were sorted, merged using 'bedtools merge', and were employed to retrieve the corresponding nucleotide sequences with the aid of 'bedtools getfasta' from the bedtools package (https://bedtools.readthedocs.io/en/latest/) [71]. CoLT nucleotide sequences were subsequently aligned using MAFFT, as described above, and the alignments were manually inspected. A ML phylogenetic tree was constructed based on 80 *P. sojae* CoLT sequences (66 located within the defined centromeric regions and 14 located elsewhere) using IQ-TREE and the TN+F+R2 model of DNA substitution. A full-length CoLT sequence retrieved from the *P. citricola* genome using a similar procedure was adopted as outgroup (S4 File). Truncated CoLT elements (i.e. without a full-length coding sequences) were inferred upon automatic translation of the predicted coding region (see S4 File). All trees were plotted with iTOL v4.3.3 [85].

Prediction of centromeric regions in other oomycete species

To predict centromeres of the two oomycete species, namely *Phytophthora citricola* P0716, (Genbank: GCA_007655245.1, with permission of the submitter) and *Bremia lactucae* SF5, (GenBank: GCA_004359215.1) [37], 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 synteny with the Psojae2019.1 assembly were further examined with BLASTn, and visualized by Circos.

Supporting information

S1 Fig. Summary of the presence and absence of putative core kinetochore proteins identified in *P. sojae*. Kinetochore orthologs were identified based on BLAST searches. *P. sojae* CENP-A (in bold) was selected to track subcellular localization of centromere/kinetochore and profile centromere sequences. Sequences of *P. sojae* putative core kinetochore proteins are listed in S1 File.
(TIF)

S2 Fig. Identification and expression of CENP-A in *P. sojae*. (A) Image of *CENP-A* gene model and RNA-seq log scale coverage taken from FungiDB. The left right arrow and triangle denote an erroneous *P. sojae CENP-A* gene model caused by an intron that was missed in the gene model prediction. (B) Electrophoresis image showing 3'-RACE result of *CENP-A*. 5'-primer JOHE45057 (not shown in scale) served as a gene-specific primer for 3'-RACE (See S6 Table). (C) Alignment of *P. sojae* CENP-A with orthologs from different organisms. Ps, *P. sojae*; Pt, *Phaeodactylum tricornutum* (diatom); Sc, *Saccharomyces cerevisiae;* Cn, *Cryptococcus neoformans; Hs, Homo sapiens;* At, *Arabidopsis thaliana*; Dm, *Drosophila melanogaster*. (D) Transient expression of *GFP* tagged *CENP-A* in *P. sojae* transformants. Upper panel, a plasmid constructed for transient expression of *CENP-A*. Expression of *P. sojae CENP-A* (*PsCENP-A*) is driven by a constitutive promoter derived from the *B. lactucae HAM34* gene. Lower panel, subcellular localization of GFP-tagged CENP-A in the *P. sojae* transformants based on the constructs shown in the upper panel.
(TIF)

S3 Fig. Generation of *P. sojae* strains expressing *GFP* tagged *CENP-A* utilizing CRISPR/Cas9 mediated genome editing. (A) Schematic of gene replacement of the endogenous *CENP-A* with *GFP-CENP-A*. Lightning bolts, an sgRNA guide sequence was designed overlapping the start codon of *CENP-A*. Primer pairs, JOHE50062/JOHE45358, JOHE45420/JOHE50063, JOHE50062/JOHE50063 were used for 5'-junction, 3'-junction, and spanning diagnostic PCR screening *GFP-CENP-A* mutants. See S6 Table for the primer information. (B) Representative genotyping results of zoospore isolated (homokaryotic) *GFP-CENP-A* strains (YFP10a1 and YFP10b1, see S5 Table for their genetic backgrounds). Products observed in the 5'- and 3'-junction PCR of wild type (WT) are non-specific amplicons.
(TIF)

S4 Fig. Scaffolds in the Sanger assembly that are suggested to harbor putative centromeres. (A) CENP-A enrichment identified in 12 scaffolds. Solid and hollow stars denote CENP-A enrichment regions that are sequence-gap free or contain gaps, respectively. All CENP-A profiles shown have been normalized to input. mRNA profiles are shown as log-scales. (B) Two putative centromeric regions in Scaffold 1 were identified by poor transcription and the syntenic regions in the Psojae2019.1 assembly.
(TIF)

S5 Fig. Pipeline used for the *de novo* genome assembly and metrics of Psojae2019.1. (A) Pipeline used to generate the assembly of Psojae2019.1. Box in grey, different scaffolding

programs were employed to enhance the contiguity of the assembly (See more details in S10 Fig and S3 Table). As some of them generated conflict and sequence gaps, we opted to use the contig-level assembly for the centromere study. (B) Dotplot comparison of the long-read Nanopore assembly Psojae2019.1 against the Sanger assembly. (C) Genome assembly metrics. *Statistics for both the Psojae2019.1 and Sanger V3 assemblies were calculated by QUAST [89]; †Annotation based on the repeat-masked assembly (See Method). ‡Annotation obtained from FungiDB release 33 (https://fungidb.org/fungidb/). §Measured by RepeatModeler (See Methods). (D) Pie chart summarizing retroelements (LINE, SINE and LTR), DNA transposons and other repeat sequences predicted in the Psoaje2019.1 assembly.
(TIF)

S6 Fig. Comparison of centromere-containing genomic regions between the Sanger (P. sojae V3) and the Psojae2019.1 assemblies. In all panels (A-H), the outer tracks (track A) illustrate assembled contigs (in Psojae2019.1) or scaffolds (in P. sojae V3), and are color coded as given in key at the top. Yellow regions on the outer tracks indicate the location of centromeres (CENP-A binding regions). Tracks B-E show the location of other genomic features as given in the key on the bottom. Blue and orange lines in track F link regions with synteny extending over 2 kb, with orange lines corresponding to inversions. To demonstrate sequence gaps in the Sanger assembly more clearly, black dots representing assembly gaps are also shown between track E and F. Names of contigs that contain *P. sojae* centromeres are enclosed in circles. Arrowheads indicate the shrunk centromeres present in the Sanger scaffolds. (A) Comparison of Sanger Scaffold 2 (sca2) and its syntenic contigs in the Psojae2019.1 assembly. (B) Comparison of Sanger Scaffold 8 (sca8) and its syntenic contigs in the Psojae2019.1 assembly. (C) Comparison of Sanger Scaffold 9 (sca9) and Scaffold 12 (sca12), and their syntenic contigs in the Psojae2019.1 assembly. (D) Comparison of Sanger Scaffold 4 (sca4) and its syntenic contigs in the Psojae2019.1 assembly. Dots under *CEN4* indicate the regions showing CENP-A peaks, as a transcriptionally active region is found in *CEN4*. (E) Comparison of Sanger Scaffold 3 (sca3) and its syntenic contigs in the Psojae2019.1 assembly. (F) Comparison of Sanger Scaffold 6 (sca6) and its syntenic contigs in the Psojae2019.1 assembly. Dashed lines under Psojae2019.1 Contig 20 and Contig 34 indicate duplicated regions. (G) Comparison of Sanger Scaffold 5 (sca5) and its syntenic contigs in the Psojae2019.1 assembly. (H) Comparison of Sanger Scaffold 1 (sca1) and its syntenic contigs in the Psojae2019.1 assembly. The two centromeric regions in Scaffold 1 of the Sanger assembly (P. sojae V3) suggested by CENP-A ChIP-seq are indicated by arrows. Part of Contig 4 is inverted in the Psojae2019.1 assembly compared to the Sanger assembly (inversion breakpoint indicated by an asterisk). Intriguingly, both versions can be supported by some of the mapped reads, indicating that both ends of the inverted region are repeat-rich and composed of the same types of repeats, so that reads map to both versions. Alternatively, this may represent a structural rearrangement between the two haplotypes of the diploid genome, or it could also represent a real difference between the P6497 culture used for the Sanger sequence and the one used for the ONT assembly.
(PDF)

S7 Fig. Features and read coverage analyses of each intact centromere. For each panel (A-J), upper, an overview of the contig exhibiting enrichments of both CENP-A and histone modifications (H3K9me3, H3K27me3 and H3K4me2), and distribution of transposable elements. TE, transposable elements. CoLT, *Copia*-like transposon. Middle, a magnified view of the region shaded in the contig shown above; a diagram displaying the correlation of core centromere (CENP-A binding region) and heterochromatic regions. A 400 kb region is shown for each centromere that has assembled pericentromeric regions on both sides. A 270 kb region is shown for *CEN5*, as it is missing one side of the intact pericentric region. The length of Contig

51 is close to 400 kb, only the entire contig is shown (J). Bottom, an image showing long-read coverage. Canu-corrected Nanopore reads were mapped to all centromeres using Minimap2 [90], except *CEN7*, which was verified by all Nanopore and PacBio reads (without Canu-correction) using GraphMap [91]. For better visualization, indels of reads were masked. Asterisk in (B), regions (underlined) that have high CENP-A enrichment but show no CoLT and very low TE density contains unknown repetitive sequences unique to *CEN2*.
(PDF)

S8 Fig. MAFFT-based alignment of CENP-A binding regions reveals a 5 kb consensus sequence. Upper panel, alignment of CENP-A binding regions. Lower panel, magnified view of the conserved 5 kb region. Images were adapted from the alignment result generated by Geneious.
(TIF)

S9 Fig. Representative contigs that are anchored by Bionano mapping and contigs that are suggested to be joined. (A) Contig 2 representing most of the cases that only parts of contigs were anchored by Bionano molecules. (B) One case indicating that two contigs (Contig 5 and Contig 30) were not anchored by Bionano but suggested to be combined. (C-E) Three cases displaying that contigs were anchored by Bionano and suggested to be combined. (E, F) Two cases showing that contigs can be fully covered by Bionano molecules. Map102 is a fragment trimmed from Contig 32. NGS, next generation sequencing, namely Psojae2019.1; BNG, Bionano *de novo* assembly; HYBRID, hybrid scaffold.
(TIF)

S10 Fig. Dot plot comparison of scaffolded assemblies against the original Psojae2019.1 assembly and the Sanger assembly. (A-D) Assemblies generated by the scaffolding programs npScarf, SSPACE and LINKS, and Bionano mapping were aligned respectively to the original SMARTdenovo assembly (Psojae2019.1), as well as the Sanger genome, and plotted utilizing the MUMmer package [92].
(TIF)

S1 Table. Metrics of ONT sequencing.
(DOCX)

S2 Table. Statistics of ChIP-seq samples.
(DOCX)

S3 Table. Metrics of scaffolded assemblies and their comparison to the Sanger and the Psojae2019.1 assemblies.
(DOCX)

S4 Table. Five incompletely assembled centromeres in the Psojae2019.1 assembly and their corresponding CENP-A regions mapped in the Sanger assembly.
(DOCX)

S5 Table. *P. sojae* strains used in the study.
(DOCX)

S6 Table. Primers used in this study.
(DOCX)

S1 Text. Nanopore sequencing and *de novo* assembly of the reference *P. sojae* genome.
(PDF)

S1 Movie. Time-lapse experiment showing cellular dynamics of CENP-A during *P. sojae* vegetative growth.
(AVI)

S1 File. Sequences of kinetochore orthologs identified in *P. sojae*.
(XLSX)

S2 File. Telomeres predicted in the Psojae2019.1 and the Sanger genome assemblies.
(XLSX)

S3 File. The 5 kb consensus sequence shared by *P. sojae* centromeres, and representative full-length CoLT homologs found in the *P. sojae*, *B. lactucae* and *P. citricola* genome assemblies.
(DOCX)

S4 File. Sequences used for constructing phylogenetic trees in Fig 6. (A) Sequences employed for the phylogenetic analysis in Fig 6A. (B) Genomic coordinates of the sequences used for constructing phylogenetic tree in Fig 6B.
(XLSX)

S5 File. Original names of the sorted *B. lactucae* scaffolds.
(XLSX)

S6 File. Bionano mapping report.
(TXT)

S7 File. CENP-A peaks called by MACS in the 10 contigs that harbor fully assembled centromeres.
(XLSX)

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.

Author Contributions

Conceptualization: Yufeng Fang, Marco A. Coelho, Minou Nowrousian, Joseph Heitman.

Data curation: Yufeng Fang, Marco A. Coelho, Haidong Shu, Minou Nowrousian, Joseph Heitman.

Formal analysis: Yufeng Fang, Marco A. Coelho, Haidong Shu, Klaas Schotanus, Bhagya C. Thimmappa, Vikas Yadav, Jeremy Wang, Minou Nowrousian, Joseph Heitman.

Funding acquisition: Suomeng Dong, Minou Nowrousian, Joseph Heitman.

Investigation: Yufeng Fang, Haidong Shu, Han Chen, Ewa P. Malc, Suomeng Dong, Minou Nowrousian, Joseph Heitman.

Methodology: Yufeng Fang, Minou Nowrousian.

Project administration: Joseph Heitman.

Resources: Brent Kronmiller, Brett M. Tyler, Joseph Heitman.

Supervision: Piotr A. Mieczkowski, Kaustuv Sanyal, Suomeng Dong, Joseph Heitman.

Validation: Yufeng Fang, Joseph Heitman.

Visualization: Yufeng Fang, Marco A. Coelho, Joseph Heitman.

Writing – original draft: Yufeng Fang, Marco A. Coelho, Minou Nowrousian.

Writing – review & editing: Yufeng Fang, Marco A. Coelho, Haidong Shu, Klaas Schotanus, Bhagya C. Thimmappa, Vikas Yadav, Han Chen, Ewa P. Malc, Jeremy Wang, Piotr A. Mieczkowski, Brent Kronmiller, Brett M. Tyler, Kaustuv Sanyal, Suomeng Dong, Minou Nowrousian, Joseph Heitman.

References

1. Kursel LE, Malik HS. Centromeres. Curr Biol. 2016; 26(12):R487–R90. Epub 2016/06/22. https://doi.org/10.1016/j.cub.2016.05.031 PMID: 27326706.
2. Stimpson KM, Sullivan BA. Epigenomics of centromere assembly and function. Curr Opin Cell Biol. 2010; 22(6):772–80. Epub 2010/08/03. https://doi.org/10.1016/j.ceb.2010.07.002 PMID: 20675111.
3. Buscaino A, Allshire R, Pidoux A. Building centromeres: home sweet home or a nomadic existence? Curr Opin Genet Dev. 2010; 20(2):118–26. Epub 2010/03/09. https://doi.org/10.1016/j.gde.2010.01.006 PMID: 20206496.
4. Wang N, Dawe RK. Centromere size and its relationship to haploid formation in plants. Mol Plant. 2018; 11(3):398–406. Epub 2017/12/27. https://doi.org/10.1016/j.molp.2017.12.009 PMID: 29277426.
5. Black BE, Foltz DR, Chakravarthy S, Luger K, Woods VL Jr, Cleveland DW. Structural determinants for generating centromeric chromatin. Nature. 2004; 430(6999):578–82. Epub 2004/07/30. https://doi.org/10.1038/nature02766 PMID: 15282608.
6. Guse A, Carroll CW, Moree B, Fuller CJ, Straight AF. In vitro centromere and kinetochore assembly on defined chromatin templates. Nature. 2011; 477(7364):354–8. Epub 2011/08/30. https://doi.org/10.1038/nature10379 PMID: 21874020.
7. Steiner FA, Henikoff S. Holocentromeres are dispersed point centromeres localized at transcription factor hotspots. Elife. 2014; 3:e02025. Epub 2014/04/10. https://doi.org/10.7554/eLife.02025 PMID: 24714495.
8. Comai L, Maheshwari S, Marimuthu MPA. Plant centromeres. Curr Opin Plant Biol. 2017; 36:158–67. Epub 2017/04/16. https://doi.org/10.1016/j.pbi.2017.03.003 PMID: 28411416.
9. McNulty SM, Sullivan BA. Alpha satellite DNA biology: finding function in the recesses of the genome. Chromosome Res. 2018; 26(3):115–38. Epub 2018/07/06. https://doi.org/10.1007/s10577-018-9582-3 PMID: 29974361.
10. Jin W, Melo JR, Nagaki K, Talbert PB, Henikoff S, Dawe RK, et al. Maize centromeres: organization and functional adaptation in the genetic background of oat. Plant Cell. 2004; 16(3):571–81. Epub 2004/02/20. https://doi.org/10.1105/tpc.018937 PMID: 14973167.
11. Smith KM, Galazka JM, Phatale PA, Connolly LR, Freitag M. Centromeres of filamentous fungi. Chromosome Res. 2012; 20(5):635–56. Epub 2012/07/04. https://doi.org/10.1007/s10577-012-9290-3 PMID: 22752455.
12. Yadav V, Sreekumar L, Guin K, Sanyal K. Five pillars of centromeric chromatin in fungal pathogens. PLoS Pathog. 2018; 14(8):e1007150. Epub 2018/08/24. https://doi.org/10.1371/journal.ppat.1007150 PMID: 30138484.
13. Yadav V, Sun S, Billmyre RB, Thimmappa BC, Shea T, Lintner R, et al. RNAi is a critical determinant of centromere evolution in closely related fungi. Proc Natl Acad Sci U S A. 2018; 115(12):3108–13. Epub 2018/03/07. https://doi.org/10.1073/pnas.1713725115 PMID: 29507212.
14. Smith KM, Phatale PA, Sullivan CM, Pomraning KR, Freitag M. Heterochromatin is required for normal distribution of *Neurospora crassa* CenH3. Mol Cell Biol. 2011; 31(12):2528–42. Epub 2011/04/21. https://doi.org/10.1128/MCB.01285-10 PMID: 21505064.
15. Schotanus K, Soyer JL, Connolly LR, Grandaubert J, Happel P, Smith KM, et al. Histone modifications rather than the novel regional centromeres of *Zymoseptoria tritici* distinguish core and accessory chromosomes. Epigenet Chromatin. 2015; 8(1):41. Epub 2015/10/03. https://doi.org/10.1186/s13072-015-0033-5 PMID: 26430472.

16. Sanyal K, Baum M, Carbon J. Centromeric DNA sequences in the pathogenic yeast *Candida albicans* are all different and unique. Proc Natl Acad Sci U S A. 2004; 101(31):11374–9. Epub 2004/07/24. https://doi.org/10.1073/pnas.0404318101 PMID: 15272074.

17. Hoeijmakers WA, Flueck C, Francoijs KJ, Smits AH, Wetzel J, Volz JC, et al. *Plasmodium falciparum* centromeres display a unique epigenetic makeup and cluster prior to and during schizogony. Cell Microbiol. 2012; 14(9):1391–401. Epub 2012/04/18. https://doi.org/10.1111/j.1462-5822.2012.01803.x PMID: 22507744.

18. Diner RE, Noddings CM, Lian NC, Kang AK, McQuaid JB, Jablanovic J, et al. Diatom centromeres suggest a mechanism for nuclear DNA acquisition. Proc Natl Acad Sci U S A. 2017; 114(29):E6015–E24. Epub 2017/07/05. https://doi.org/10.1073/pnas.1700764114 PMID: 28673987.

19. Brooks CF, Francia ME, Gissot M, Croken MM, Kim K, Striepen B. *Toxoplasma gondii* sequesters centromeres to a specific nuclear region throughout the cell cycle. Proc Natl Acad Sci U S A. 2011; 108(9):3767–72. Epub 2011/02/16. https://doi.org/10.1073/pnas.1006741108 PMID: 21321216.

20. Yadav V, Yang F, Reza MH, Liu S, Valent B, Sanyal K, et al. Cellular dynamics and genomic identity of centromeres in cereal blast fungus. mBio. 2019; 10(4):e01581–19. Epub 2019/08/01. https://doi.org/10.1128/mBio.01581-19 PMID: 31363034.

21. Navarro-Mendoza MI, Perez-Arques C, Panchal S, Nicolas FE, Mondo SJ, Ganguly P, et al. Early diverging fungus *Mucor circinelloides* lacks centromeric histone CENP-A and displays a mosaic of point and regional centromeres. Curr Biol. 2019; 29(22):3791–802 e6. Epub 2019/11/05. https://doi.org/10.1016/j.cub.2019.09.024 PMID: 31679929.

22. Chang CH, Chavan A, Palladino J, Wei X, Martins NMC, Santinello B, et al. Islands of retroelements are major components of *Drosophila* centromeres. PLoS Biol. 2019; 17(5):e3000241. Epub 2019/05/16. https://doi.org/10.1371/journal.pbio.3000241 PMID: 31086362.

23. Lan T, Renner T, Ibarra-Laclette E, Farr KM, Chang TH, Cervantes-Perez SA, et al. Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome. Proc Natl Acad Sci U S A. 2017; 114(22):E4435–E41. Epub 2017/05/17. https://doi.org/10.1073/pnas.1702072114 PMID: 28507139.

24. Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, Paten B, et al. Linear assembly of a human centromere on the Y chromosome. Nat Biotechnol. 2018; 36(4):321–3. Epub 2018/03/20. https://doi.org/10.1038/nbt.4109 PMID: 29553574.

25. Keeling PJ, Burki F. Progress towards the Tree of Eukaryotes. Curr Biol. 2019; 29(16):R808–R17. Epub 2019/08/21. https://doi.org/10.1016/j.cub.2019.07.031 PMID: 31430481.

26. Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RH, Aerts A, et al. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science. 2006; 313(5791):1261–6. Epub 2006/09/02. https://doi.org/10.1126/science.1128796 PMID: 16946064.

27. Grattepanche JD, Walker LM, Ott BM, Paim Pinto DL, Delwiche CF, Lane CE, et al. Microbial diversity in the eukaryotic SAR clade: Illuminating the darkness between morphology and molecular data. Bioessays. 2018; 40(4):e1700198. Epub 2018/03/08. https://doi.org/10.1002/bies.201700198 PMID: 29512175.

28. Erwin DC, Ribeiro OK. Phytophthora diseases worldwide: American Phytopathological Society (APS Press); 1996.

29. Jiang RH, Tyler BM. Mechanisms and evolution of virulence in oomycetes. Annu Rev Phytopathol. 2012; 50:295–318. Epub 2012/08/28. https://doi.org/10.1146/annurev-phyto-081211-172912 PMID: 22920560.

30. Savary S, Willocquet L, Pethybridge SJ, Esker P, McRoberts N, Nelson A. The global burden of pathogens and pests on major food crops. Nat Ecol Evol. 2019; 3(3):430–9. Epub 2019/02/06. https://doi.org/10.1038/s41559-018-0793-y PMID: 30718852.

31. Fang Y, Cui L, Gu B, Arredondo F, Tyler BM. Efficient genome editing in the oomycete *Phytophthora sojae* using CRISPR/Cas9. Curr Protoc Microbiol. 2017; 44:21A 1.1–A.1 6. Epub 2017/02/07. https://doi.org/10.1002/cpmc.25 PMID: 28166383.

32. Judelson HS, Coffey MD, Arredondo FR, Tyler BM. Transformation of the oomycete pathogen *Phytophthora-megasperma* f.sp. *glycinea* occurs by DNA integration into single or multiple chromosomes. Curr Genet. 1993; 23(3):211–8. Epub 1993/03/01. https://doi.org/10.1007/bf00351498 PMID: 8382110.

33. Fang Y, Tyler BM. Efficient disruption and replacement of an effector gene in the oomycete *Phytophthora sojae* using CRISPR/Cas9. Mol Plant Pathol. 2016; 17(1):127–39. Epub 2015/10/29. https://doi.org/10.1111/mpp.12318 PMID: 26507366.

34. McGowan J, Fitzpatrick DA. Genomic, network, and phylogenetic analysis of the oomycete effector arsenal. mSphere. 2017; 2(6):e00408–17. Epub 2017/12/05. https://doi.org/10.1128/mSphere.00408-17 PMID: 29202039.

35. Tyler BM, Gijzen M. The *Phytophthora sojae* genome sequence: foundation for a revolution. Genomics of plant-associated fungi and oomycetes: dicot pathogens: Springer; 2014. p. 133–57.

36. Malar CM, Yuzon JD, Das S, Das A, Panda A, Ghosh S, 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. 2019; 32(8):1047–60. Epub 2019/02/23. https://doi.org/10.1094/MPMI-08-18-0222-R PMID: 30794480.

37. Fletcher K, Gil J, Bertier LD, Kenefick A, Wood KJ, Zhang L, et al. Genomic signatures of heterokaryosis in the oomycete pathogen *Bremia lactucae*. Nat Commun. 2019; 10(1):2645. Epub 2019/06/16. https://doi.org/10.1038/s41467-019-10550-0 PMID: 31201315.

38. van Hooff JJ, Tromer E, van Wijk LM, Snel B, Kops GJ. Evolutionary dynamics of the kinetochore network in eukaryotes as revealed by comparative genomics. EMBO Rep. 2017; 18(9):1559–71. Epub 2017/06/24. https://doi.org/10.15252/embr.201744102 PMID: 28642229.

39. Fulneckova J, Sevcikova T, Fajkus J, Lukesova A, Lukes M, Vlcek C, et al. A broad phylogenetic survey unveils the diversity and evolution of telomeres in eukaryotes. Genome Biol Evol. 2013; 5(3):468–83. Epub 2013/02/12. https://doi.org/10.1093/gbe/evt019 PMID: 23395982.

40. Tooley PW, Carras MM. Separation of chromosomes of *Phytophthora* species using CHEF gel electrophoresis. Experimental Mycology. 1992; 16(3):188–96. https://doi.org/10.1016/0147-5975(92)90027-O

41. Sansome E, Brasier C. Polyploidy associated with varietal differentiation in the megasperma complex of *Phytophthora*. Transactions of the British Mycological Society. 1974; 63(3):461–IN11.

42. Sullivan BA, Karpen GH. Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. Nature Structural & Molecular Biology. 2004; 11 (11):1076–83. https://doi.org/10.1038/nsmb845 PMID: 15475964

43. Jiang RHY, Dawe AL, Weide R, van Staveren M, Peters S, Nuss DL, et al. Elicitin genes in *Phytophthora infestans* are clustered and interspersed with various transposon-like elements. Molecular Genetics and Genomics. 2005; 273(1):20–32. https://doi.org/10.1007/s00438-005-1114-0 PMID: 15702346

44. Basnayake S, Maclean DJ, Whisson SC, Drenth A. Identification and occurrence of the LTR-Copia-like retrotransposon, PSCR and other Copia-like elements in the genome of *Phytophthora sojae*. Curr Genet. 2009; 55(5):521–36. Epub 2009/07/31. https://doi.org/10.1007/s00294-009-0263-9 PMID: 19641921.

45. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015; 6(1):11. Epub 2015/06/06. https://doi.org/10.1186/s13100-015-0041-9 PMID: 26045719.

46. Henikoff S, Ahmad K, Malik HS. The centromere paradox: stable inheritance with rapidly evolving DNA. Science. 2001; 293(5532):1098–102. Epub 2001/08/11. https://doi.org/10.1126/science.1062939 PMID: 11498581.

47. Volpe TA, Kidner C, Hall IM, Teng G, Grewal SI, Martienssen RA. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. Science. 2002; 297(5588):1833–7. Epub 2002/08/24. https://doi.org/10.1126/science.1074973 PMID: 12193640.

48. Li XY, Wang XF, He K, Ma YQ, Su N, He H, et al. High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression. Plant Cell. 2008; 20(2):259–76. https://doi.org/10.1105/tpc.107.056879 PMID: 18263775

49. Allshire RC, Karpen GH. Epigenetic regulation of centromeric chromatin: old dogs, new tricks? Nat Rev Genet. 2008; 9(12):923–37. Epub 2008/11/13. https://doi.org/10.1038/nrg2466 PMID: 19002142.

50. Blower MD, Sullivan BA, Karpen GH. Conserved organization of centromeric chromatin in flies and humans. Dev Cell. 2002; 2(3):319–30. Epub 2002/03/07. https://doi.org/10.1016/s1534-5807(02)00135-1 PMID: 11879637.

51. Scott KC, Merrett SL, Willard HF. A heterochromatin barrier partitions the fission yeast centromere into discrete chromatin domains. Curr Biol. 2006; 16(2):119–29. Epub 2006/01/25. https://doi.org/10.1016/j.cub.2005.11.065 PMID: 16431364.

52. Allshire RC, Madhani HD. Ten principles of heterochromatin formation and function. Nat Rev Mol Cell Biol. 2018; 19(4):229–44. Epub 2017/12/14. https://doi.org/10.1038/nrm.2017.119 PMID: 29235574.

53. Greaves IK, Rangasamy D, Ridgway P, Tremethick DJ. H2A.Z contributes to the unique 3D structure of the centromere. Proc Natl Acad Sci U S A. 2007; 104(2):525–30. Epub 2006/12/30. https://doi.org/10.1073/pnas.0607870104 PMID: 17194760.

54. Friedman S, Freitag M. Evolving centromeres and kinetochores. Advances in genetics. 98: Elsevier; 2017. p. 1–41. https://doi.org/10.1016/bs.adgen.2017.07.001 PMID: 28942791

55. Brown JD, O'Neill RJ. The evolution of centromeric DNA sequences. In: John Wiley & Sons Ltd, editor. eLS. John Wiley & Sons, Ltd: Chichester, UK; 2014. https://doi.org/10.1002/9780470015902.a0020827.pub2

56. Yin L, Wang Q, Ning F, Zhu X, Zuo Y, Shan W. Identification of a repetitive sequence element for DNA fingerprinting in *Phytophthora sojae*. Wei Sheng Wu Xue Bao. 2010; 50(4):524–9. Epub 2010/06/22. PMID: 20560357.

57. Xiao J, Miao M, Gao K, Xiang G, Yang S, Dong S, et al. Genetic diversity of *Phytophthora sojae* in China based on RFLP. Scientia Agricultura Sinica. 2011; 44(20):4190–8.

58. Cao MD, Nguyen SH, Ganesamoorthy D, Elliott AG, Cooper MA, Coin LJ. Scaffolding and completing genome assemblies in real-time with nanopore sequencing. Nat Commun. 2017; 8:14515. Epub 2017/02/22. https://doi.org/10.1038/ncomms14515 PMID: 28218240.

59. Boetzer M, Pirovano W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics. 2014; 15(1):211. https://doi.org/10.1186/1471-2105-15-211 PMID: 24950923

60. Warren RL, Yang C, Vandervalk BP, Behsaz B, Lagman A, Jones SJ, et al. LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. Gigascience. 2015; 4(1):35. Epub 2015/08/06. https://doi.org/10.1186/s13742-015-0076-3 PMID: 26244089.

61. Compton DA. Mechanisms of aneuploidy. Curr Opin Cell Biol. 2011; 23(1):109–13. Epub 2010/09/03. https://doi.org/10.1016/j.ceb.2010.08.007 PMID: 20810265.

62. Todd RT, Forche A, Selmecki A. Ploidy variation in fungi: polyploidy, aneuploidy, and genome evolution. Microbiol Spectr. 2017; 5(4):599–618. Epub 2017/07/29. https://doi.org/10.1128/microbiolspec.FUNK-0051-2016 PMID: 28752816.

63. Elliott M, Yuzon J, C MM, Tripathy S, Bui M, Chastagner GA, et al. Characterization of phenotypic variation and genome aberrations observed among *Phytophthora ramorum* isolates from diverse hosts. BMC Genomics. 2018; 19(1):320. Epub 2018/05/04. https://doi.org/10.1186/s12864-018-4709-7 PMID: 29720102.

64. Dobrowolski MP, Tommerup IC, Blakeman HD, O'Brien PA. Non-Mendelian inheritance revealed in a genetic analysis of sexual progeny of *Phytophthora cinnamomi* with microsatellite markers. Fungal Genet Biol. 2002; 35(3):197–212. Epub 2002/04/04. https://doi.org/10.1006/fgbi.2001.1319 PMID: 11929210.

65. Kasuga T, Bui M, Bernhardt E, Swiecki T, Aram K, Cano LM, et al. Host-induced aneuploidy and phenotypic diversification in the sudden oak death pathogen *Phytophthora ramorum*. BMC Genomics. 2016; 17(1):385. Epub 2016/05/22. https://doi.org/10.1186/s12864-016-2717-z PMID: 27206972.

66. van der Lee T, Testa A, Robold A, van't Klooster J, Govers F. High-density genetic linkage maps of *Phytophthora infestans* reveal trisomic progeny and chromosomal rearrangements. Genetics. 2004; 167 (4):1643–61. Epub 2004/09/03. https://doi.org/10.1534/genetics.104.029652 PMID: 15342505.

67. Hu J, Shrestha S, Zhou Y, Mudge J, Liu X, Lamour K. Dynamic Extreme Aneuploidy (DEA) in the vegetable pathogen *Phytophthora capsici* and the potential for rapid asexual evolution. PLoS One. 2020; 15(1):e0227250. Epub 2020/01/08. https://doi.org/10.1371/journal.pone.0227250 PMID: 31910244.

68. Lin L, Ye W, Wu J, Xuan M, Li Y, Gao J, et al. The MADS-box transcription factor PsMAD1 is involved in zoosporogenesis and pathogenesis of *Phytophthora sojae*. Front Microbiol. 2018; 9(2259):2259. Epub 2018/10/16. https://doi.org/10.3389/fmicb.2018.02259 PMID: 30319576.

69. Fang Y, Jang HS, Watson GW, Wellappili DP, Tyler BM. Distinctive nuclear localization signals in the oomycete *Phytophthora sojae*. Front Microbiol. 2017; 8:10. Epub 2017/02/18. https://doi.org/10.3389/fmicb.2017.00010 PMID: 28210240.

70. Gent JI, Wang N, Dawe RK. Stable centromere positioning in diverse sequence contexts of complex and satellite centromeres of maize and wild relatives. Genome Biol. 2017; 18(1):121. Epub 2017/06/24. https://doi.org/10.1186/s13059-017-1249-4 PMID: 28637491.

71. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26(6):841–2. Epub 2010/01/30. https://doi.org/10.1093/bioinformatics/btq033 PMID: 20110278.

72. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013–2015. 2015.

73. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008; 18:188–96. https://doi.org/10.1101/gr.6743907 PMID: 18025269

74. Haas BJ, Kamoun S, Zody MC, Jiang RHY, Handsaker RE, Cano LM, et al. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. Nature. 2009; 461:393–8. https://doi.org/10.1038/nature08358 PMID: 19741609

75. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007; 35(9):3100–8. Epub 2007/04/25. https://doi.org/10.1093/nar/gkm160 PMID: 17452365.

76. Lowe TM, Chan PP. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res. 2016; 44(W1):W54–7. Epub 2016/05/14. https://doi.org/10.1093/nar/gkw413 PMID: 27174935.

77. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009; 19(9):1639–45. Epub 2009/06/23. https://doi.org/10.1101/gr.092759.109 PMID: 19541911.

78. Jain C, Koren S, Dilthey A, Phillippy AM, Aluru S. A fast adaptive algorithm for computing whole-genome homology maps. Bioinformatics. 2018; 34(17):i748–i56. Epub 2018/11/14. https://doi.org/10.1093/bioinformatics/bty597 PMID: 30423094.

79. Okonechnikov K, Golosova O, Fursov M, team U. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics. 2012; 28(8):1166–7. Epub 2012/03/01. https://doi.org/10.1093/bioinformatics/bts091 PMID: 22368248.

80. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30(4):772–80. Epub 2013/01/19. https://doi.org/10.1093/molbev/mst010 PMID: 23329690.

81. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009; 25(15):1972–3. Epub 2009/06/10. https://doi.org/10.1093/bioinformatics/btp348 PMID: 19505945.

82. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015; 32(1):268–74. Epub 2014/11/06. https://doi.org/10.1093/molbev/msu300 PMID: 25371430.

83. Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. Molecular Biology and Evolution. 2017; 35(2):518–22.

84. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010; 59(3):307–21. Epub 2010/06/09. https://doi.org/10.1093/sysbio/syq010 PMID: 20525638.

85. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019; 47(W1):W256–W9. Epub 2019/04/02. https://doi.org/10.1093/nar/gkz239 PMID: 30931475.

86. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007; 8(12):973–82. Epub 2007/11/07. https://doi.org/10.1038/nrg2165 PMID: 17984973.

87. Xiong Y, Eickbush TH. Origin and evolution of retroelements based upon their reverse-transcriptase sequences. Embo J. 1990; 9(10):3353–62. https://doi.org/10.1002/j.1460-2075.1990.tb07536.x PMID: 1698615

88. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A resource for timelines, timetrees, and divergence times. Mol Biol Evol. 2017; 34(7):1812–9. Epub 2017/04/08. https://doi.org/10.1093/molbev/msx116 PMID: 28387841.

89. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018; 34(13):i142–i50. Epub 2018/06/29. https://doi.org/10.1093/bioinformatics/bty266 PMID: 29949969.

90. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018; 34(18):3094–100. Epub 2018/05/12. https://doi.org/10.1093/bioinformatics/bty191 PMID: 29750242.

91. Sović I, Šikić M, Wilm A, Fenlon SN, Chen S, Nagarajan N. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. Nat Commun. 2016; 7:11307. https://doi.org/10.1038/ncomms11307 PMID: 27079541

92. Marcais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol. 2018; 14(1):e1005944. Epub 2018/01/27. https://doi.org/10.1371/journal.pcbi.1005944 PMID: 29373581.