The Host Adapted Fungal Pathogens of *Pneumocystis* Genus Utilize Genic Regional Centromeres

Authors:

Ousmane H. Cissé [1], Shelly Curran [1], H. Diego Folco [2], Yueqin Liu [1], Lisa Bishop [1], Honghui Wang [1], Elizabeth R. Fischer [3], A Sally Davis [4], Spenser Babb-Biernacki [5], Vinson P. Doyle [6], Jonathan K. Richards [6], Sergio A. Hassan [7], John P. Dekker [8,9], Pavel P. Khil [8,9], Jason M. Brenchley [10], Shiv Grewal [2], Melanie Cushion [11], Liang Ma [1] and Joseph A. Kovacs [1]

Affiliations:

[1] Critical Care Medicine Department, Clinical Center, National Institutes of Health, Bethesda, Maryland, USA.

[2] Laboratory of Biochemistry and Molecular Biology, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA.

[3] Microscopy Unit, Research Technologies Branch, Rocky Mountain Laboratories, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, Montana, USA.

[4] Diagnostic Medicine/Pathobiology, Kansas State University College of Veterinary Medicine, Manhattan, USA.

[5] Museum of Natural Science and Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana, USA.

[6] Department of Plant Pathology and Crop Physiology, Lousiana State University AgCenter, Baton Rouge, Louisiana, USA.

[7] Bioinformatics and Computational Biosciences Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, USA.

[8] Bacterial Pathogenesis and Antimicrobial Resistance Unit, National Institute of Allergy, and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, USA.

[9] Department of Laboratory Medicine, Clinical Center, National Institutes of Health, Bethesda, Maryland, USA.

[10] Laboratory of Viral Diseases, National Institute of Allergy, and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, USA.

[11] Department of Internal Medicine, College of Medicine, University of Cincinnati, Cincinnati, Ohio, USA.

Correspondence to: ousmane.cisse@nih.gov and/or jkovacs@mail.nih.gov

Abstract

Centromeres are genomic regions that coordinate accurate chromosomal segregation during mitosis and meiosis. Yet, despite their essential function, centromeres evolve rapidly across eukaryotes. Centromeres are often the sites of chromosomal breaks which contribute to genome shuffling and promote speciation by inhibiting gene flow. How centromeres form in strongly host-adapted fungal pathogens has yet to be investigated. Here, we characterized the centromere structures in closely related species of mammalian-specific pathogens of the fungal phylum of Ascomycota. Methods allowing reliable continuous culture of *Pneumocystis* species do not currently exist, precluding genetic manipulation. CENP-A, a variant of histone H3, is the epigenetic marker that defines centromeres in most eukaryotes. Using heterologous complementation, we show that the *Pneumocystis* CENP-A ortholog is functionally equivalent to CENP-A^{Cnp1} of *Schizosaccharomyces pombe*. Using organisms from a short-term *in vitro* culture or infected animal models and ChIP-seq, we identified centromeres in three *Pneumocystis* species that diverged ~100 million years ago. Each species has a unique short regional centromere (< 10kb) flanked by heterochromatin in 16-17 monocentric chromosomes. They span active genes and lack conserved DNA sequence motifs and repeats. CENP-C, a scaffold protein that links the inner centromere to the kinetochore appears dispensable in one species, suggesting a kinetochore rewiring. Despite the loss of DNA methyltransferases, 5-methylcytosine DNA methylation occurs in these species, though not related to centromere function. These features suggest an epigenetic specification of centromere function.

Short summary

Pneumocystis species offer a suitable genetic system to study centromere evolution in pathogens during host adaptation because of their unique specificity for mammals, and their phylogenetic proximity with the nonpathogenic yeast *Schizosaccharomyces pombe*, a popular model for cell biology. We used this system to explore how centromeres have evolved after divergence of the two clades ~460 million years ago. To address this question, we established a protocol combining short-term culture and ChIP-seq to characterize centromeres in multiple *Pneumocystis* species. We show that *Pneumocystis* have short epigenetic centromeres that function differently from those in *S. pombe* and exhibit similarities to centromeres in more distantly related host adapted fungal pathogens.

68 **Introduction**

69 Centromeres are genomic locations where spindles attach during chromosomal segregation. They are
70 essential for cell division. Errors in the chromosomal segregation can result in aneuploidy and cell
71 division defects with disastrous consequences such as cancer. In fungi, centromeres are fragile sites
72 that are often involved in karyotype variability (Sankaranarayanan, Ianiri et al. 2020, Seidl, Kramer et
73 al. 2020) and drug resistance (Selmecki, Forche et al. 2006). In most eukaryotes, centromeres are
74 defined by the presence of CENP-A nucleosomes, a centromere specific histone H3 variant that is
75 essential for the localization of known kinetochore components (McKinley and Cheeseman 2016).
76 There is a remarkable diversity in centromere structures ranging from sequence-dependent (point)
77 centromeres to epigenetically-regulated (regional) centromeres.

78 Point centromeres are exemplified by *Saccharomyces cerevisiae* with genetically defined 125 bp
79 long centromeres containing conserved DNA elements (I, II and III) that serve as anchors for the
80 recruitment of the centromeric DNA binding factor 3 (CBF3) complex (Clarke and Carbon 1980,
81 Gordon, Byrne et al. 2011). Regional centromeres are much longer, lack universally conserved DNA
82 patterns, and are often made up of repetitive DNA (Kursel and Malik 2016). Regional centromeres
83 are epigenetically regulated; however, which factors contribute to this epigenetic specification of
84 centromeres remains elusive.

85 *Pneumocystis* is a genus of pathogenic fungi that exclusively infect mammals and remain
86 unculturable. They belong to the Taphrinomycotina subphylum and form a monophyletic clade with
87 fission yeasts *Schizosaccharomyces*, with a separation time estimated at ~460 million years ago (Shen,

88 Steenwyk et al. 2020, Cisse, Ma et al. 2021). In addition to these two classes there are other classes, all
89 of which are plant or soil adapted organisms. Compared to *S. pombe*, *Pneumocystis* species have
90 streamlined genomes due to substantial gene losses during their transition to animal parasitism.
91 Recently, the genomes of multiple *Pneumocystis* species have been sequenced (Slaven, Meller et al.
92 2006, Cisse, Pagni et al. 2012, Ma, Chen et al. 2016, Cisse, Ma et al. 2021). However, centromeres
93 were not defined in these because they cannot be predicted bioinformatically. *Pneumocystis* have
94 retained CENP-A, which presumably binds to centromeres, and most of the kinetochore proteins.
95 Some data suggest that *Pneumocystis* centromeres differ structurally from those of *S. pombe*. For
96 example, missing in *Pneumocystis* genomes is the heterochromatin protein Swi6/HP1 (Almeida, Cisse
97 et al. 2015), which is required for replication origins at the centromeres flanking regions
98 (pericentromeres) (Hayashi, Takahashi et al. 2009). The RNAi pathway is lost in *Pneumocystis* (Cisse,
99 Pagni et al. 2014), further supporting mechanistic difference with *S. pombe*. Similar to *S. pombe*, DNA
100 methylation has been predicted to be absent in *Pneumocystis* based on the absence of DNA
101 methyltransferases (DMTs) (Freitag 2017).
102 In this work, we undertook to characterize *Pneumocystis* centromeres and compare them to
103 *Schizosaccharomyces pombe* centromeres. We found that *Pneumocystis* species have small regional
104 centromeres that are structurally different from those of *S. pombe*.
105
106
107

108 **Results**

109 **Survey of genes related to centromere function in Taphrinomycotina fungi.**

110 To understand the kinetochore evolution in Taphrinomycotina fungi, we screened the predicted
111 proteomes of 14 species spanning six orders, *Pneumocystis* (*n* = 7), *Schizosaccharomyces* (*n* = 3), *Taphrina*
112 (*n* = 1), *Saitoella* (*n* = 1), *Protomyces* (*n* = 1) and *Neolecta* (*n* = 1), as well as other fungi to have a broad
113 overview of kinetochore evolution.

114 We searched for homologs of CENP-A ^{Cnp1}, CENP-C^{Cnp3}, CCAN (Constitutive Centromere
115 Associated Network) and KMN (KNL-1/Mis12 complex/Ndc80 complex) proteins. We extended our
116 searches to pathways required for centromere function and chromosomal segregation
117 (heterochromatin, RNAi, and DNA methylation).

118 CENP-A, CENP-C, and outer kinetochore proteins are conserved across Taphrinomycotina
119 (**Supplementary figure 1**). Two genes of the inner kinetochore (CCAN) that are absent in
120 *Pneumocystis* are clearly innovations in *S. pombe* (*fta2* and *fta3* genes). The *fta2*, *fta3* and *fta4* gene
121 products associate with the central core (cnt) and innermost repeats (imr) region of the centromere
122 in *S. pombe* (Liu, McLeod et al. 2005). *Pneumocystis* also lack the *mis17* gene exclusively found in *S.*
123 *pombe*, which encodes a member of the Mis6-Mal2-Sim4 multiprotein required for CENP-A
124 recruitment in *S. pombe* (Shiroiwa, Hayashi et al. 2011). The *S. pombe* centromere-associated protein
125 B genes (*cbp1*, *cbh1* and *cbh2*), which are derived from the domestication of *pogo*-like transposases
126 (Casola, Hucks et al. 2008) have no identified homologs in other fungi.

127 The chromosomal passenger complex (CPC) is a heterotetrametric complex composed of the
128 Aurora B kinase Ark1 and three regulatory components Pic1 (inner centromere protein), Bir1p
129 (Survivin) and Nbl1 (Borealin) which mediates chromosome segregation and cytokinesis (Leverson,
130 Huang et al. 2002). This pathway is conserved in *Pneumocystis* and *S. pombe* (**Supplementary figure**
131 **1**).

132 CENP-A recruiting complex, which includes Mis18 and Mis16 (Hayashi, Fujita et al. 2004) as
133 well as the CENP-A histone chaperone Scm3 required for CENP-A loading at the centromere
134 (Williams, Hayashi et al. 2009) are conserved in *Pneumocystis*, as are the monopolin complex genes
135 that coordinate the kinetochore microtubules attached during mitosis (Corbett, Yip et al. 2010).

136 The network for heterochromatin formation in *Pneumocystis* seems intact with the presence
137 of the Clr4 methyltransferase complex (Clr4/SUV39H). The only notable loss is Swi6, which with
138 Chp2 are HP1 orthologs, while Chp2 is conserved. DNA methylases (DMTs) are not present in any
139 of the Taphrinomycotina fungi sequenced so far, which suggests that DNA methylation does not
140 occur in this subphylum. Overall, the *Pneumocystis* kinetochore protein catalog is similar to those
141 other Taphrinomycotina fungi except *S. pombe* which seems to be an outlier with a few gene
142 additions. We conclude that *Pneumocystis* species have retained an ancestral version of the
143 kinetochore.

144

145

146

147 ***Pneumocystis* CENP-A^{cnp1} localizes to the nucleus.**

148 Centromeres are identified by tracking CENP-A binding regions. There are currently seven
149 *Pneumocystis* species with fully sequenced genomes (Ma, Chen et al. 2016, Cisse, Ma et al. 2021). All
150 of them encode an ortholog of *S. pombe* CENP-A^{cnp1,} (**Figure 1a**). The CENP-A histones from these
151 species, henceforth referring to as *pn*CENP-A, display a high overall conservation (89% average
152 protein identity) with most divergence in the N-terminal tail.

153 Precise nuclear localization of CENP-A is required for accurate chromosomal segregation. To
154 determine if *pn*CENP-A localizes in the nucleus, we generated specific antibodies using synthetic
155 immunogenetic peptides representing sequences located in the N-terminus of the protein, to
156 minimize cross reactivity with host histones (**Supplementary figure 2; Supplementary methods**).
157 We used anti – *Pneumocystis* CENP-A antibodies in conjunction with DAPI to demonstrate that
158 *pn*CENP-A colocalizes with DAPI which support a nuclear localization within the *Pneumocystis*
159 nucleus (**Figure 1b**).

160

161 ***pn*CENP-A variants are functional in *S. pombe*, and support viability and centromere loading.**

162 To confirm that the putative *cenp-a* genes from *Pneumocystis* encode functional CENP-A
163 proteins, we expressed the *cenp-a* genes from *P. jirovecii* and *P. murina* in the related fission yeast *S.*
164 *pombe*, a tractable organism of the subdivision Taphrinomycotina widely used as model for
165 chromosome and cell biology. The transgenes were codon-optimized, expressed under the
166 endogenous *S. pombe cnp1*, the CENP-A ortholog, gene promoter and inserted in single copy at the

167 *lys1* locus. In addition, the gene products were tagged with GFP at their N-termini since it has been
168 shown that C-terminal tagging (*i.e.*, *cnp1-GFP*) displays growth retardation (Lando, Endesfelder et al.
169 2012). As *Pneumocystis* CENP-A proteins harbor an intervening sequence at the N-terminal tail that
170 may potentially interfere with function in *S. pombe*, we also expressed truncated versions missing ~25
171 aa from the N-terminal tail (**Figure 1c**). We first assessed the ability of *Pneumocystis* CENP-A to
172 rescue lethality of *cnp1-76* cells, a thermosensitive mutant, whose mutation T74M lies at the CATD
173 (CENP-A targeting domain) region (Castillo, Mellone et al. 2007). As previously reported (Folco,
174 Campbell et al. 2015), we found that the control transgene (SpCENP-A) restored growth of *cnp1-76*
175 cells at restrictive temperatures (i.e., 32°C and 35°C). Importantly, *Pneumocystis* counterparts were
176 also able to rescue the thermosensitive phenotype, albeit partially (**Figure 1c**), nevertheless proving
177 the functionality of these genes. Results were similar between constructs harboring either intact or
178 truncated N-termini (**Figure 1c**). Therefore, we continued our analyses with the full-length proteins.
179 It is well established that in interphase, *S. pombe* displays the Rabl configuration, in which
180 centromeres cluster to the spindle pole body while telomeres associate with each other near nuclear
181 periphery (Funabiki, Hagan et al. 1993, Mizuguchi, Fudenberg et al. 2014). As expected, expressing *S.*
182 *pombe* CENP-A tagged with GFP resulted in cells displaying single fluorescence foci that colocalized
183 with a *tetO* array inserted at *cen2* (NB: a TetR-tdTomato fusion is expressed to bind to the *tetO* array
184 and this locus is referred to as *cen2-tetO-tdTomato*) (**Figure 1d**). Consistent with the mentioned rescue
185 of *cnp1-ts* cells and further supporting their role as functional CENP-A proteins, *Pneumocystis* CENP-
186 A proteins also colocalize with *cen2-tetO-tdTomato* and remarkably, they exhibited single foci,

187 although the intensities of centromere foci were dimmer than that of the *S. pombe* counterpart

188 (**Figure 1e**). Finally, to confirm the specific localization of *Pneumocystis* CENP-A proteins to *S. pombe*

189 centromeres, we performed ChIP-qPCR (**Figure 1f**). Noteworthily, we observed that *Pneumocystis*

190 CENP-A localized to central core regions (i.e., *cc1&3, cnt2:ura4*$^{+}$) similarly to the *S. pombe*

191 counterpart, albeit the enrichment was lower, consistent with the results from live-cell imaging.

192 Additionally, *Pneumocystis* CENP-A localization seems to be specific as no detectable enrichment was

193 found at pericentromeric outer repeats coated with heterochromatin (i.e., *dg*) and euchromatin

194 locations (i.e., *fbp1*). Together, these results confirm that *Pneumocystis* CENP-A genes encode bona

195 fide CENP-A proteins.

196

197 ***pn*CENP-A binds to single genomic foci in *Pneumocystis* replicating cells.**

198 In *S. pombe*, CENP-A^{cnp1} is deposited at the centromeres during the G2 phase of the cell cycle (Lando,

199 Endesfelder et al. 2012). Although *Pneumocystis* cell replication is not fully understood, these

200 organisms take 5 to 8 days to replicate *in vivo* (Vestereng, Bishop et al. 2004). To determine *pn*CENP-

201 A bound genome regions, we established a short-term co-culture system for *P. murina* and *P. carinii*

202 (**Figure 2a**). We used quantitative PCR targeting the single copy gene dihydrofolate reductase (*dhfr*)

203 to assess organism growth; a typical growth curve shows a decline from day 0 to 7 followed by a

204 gradual increase (**Figures 2b and 2c**). Population growth is further supported by the presence of

205 mitotic or fusing cells 7 days post culture, though some cells do not appear to be dividing (**Figure**

206 **2d**). We assessed *Pn*CENP-A location on the *P. carinii* genome using ChIP-seq targeting chromatin

207 prepared from infected lung tissue lysates and cultured organisms harvested at three time points.

208 CENP-A binds to discrete areas, which is consistent with centromeres (**Figure 2e** and

209 **Supplementary figure 3**).

210 We mapped centromeres in all *P. carinii* and *P. murina* chromosomes by ChIP-seq using anti-

211 *pn*CENP-A and anti-*pn*CENP-C antibodies (**Supplementary figures 2a** and **b**) with cell cultures

212 collected from three timepoints: days 0 (normalized initial inoculum), 7 and 14; pre-immune sera

213 served as negative controls (Input). Five independent experiments were performed for each species.

214 We removed host DNA contamination from raw sequence reads by mapping against the rat and

215 mouse genomes; filtered reads (0.3-5% of total reads) were aligned to the *Pneumocystis* genomes. We

216 the genome-wide ratios of immunoprecipitated DNA sequences and matching controls (input

217 sequences) to identify significantly enriched genomic loci using a false discovery rate [FDR] < 0.05

218 and a minimum fold enrichment of 1.2. Enriched regions following immunoprecipitation (IP) were

219 inspected using IGV genome browser. DeepTools fingerprint plots show that most reads are mapped

220 within a small fraction of the genome (~3% of the total 7-8 Mb), which indicates that the antibodies

221 demonstrated a restricted binding (**Supplementary figure 2c**). Using ChIP-qPCR, we confirmed

222 *pn*CENP-A enrichment over regions flagged by ChIP-seq compared to a control region taken 200

223 kilobases away from centromere 1 of each species) for 16 of 17 *P. murina* chromosomes and all 17 *P.*

224 *carinii* chromosomes (**Supplementary figure 4**).

225 In *P. murina*, each of the 17 chromosome-level scaffolds display a single region with

226 overlapping peaks of CENP-A and CENP-C (**Figures 3a and 3b**). The *P. carinii* genome also displays

227 monocentric chromosomes enriched with CENP-A but they lack any CENP-C enrichment (**Figures**
228 **3c and 3d**).

229 Enriched regions in both species span 4.8 to 8.0 kb, with an average length of 6.7 kb, and
230 have an average of 3.9% lower GC content than the rest of the genome (**Figures 4a and 4b;**
231 **Supplementary Table 1**). *P. murina* and *P. carinii* centromeres encode 74 and 58 genes respectively;
232 all of them are conserved in the genomes of both species except a prefoldin gene that is lost in *P.*
233 *carinii* (**Supplementary Table 2**). Of the 74 *P. murina* centromeric genes, 73 were expressed based on
234 RNA-seq and four were further detected by protein mass spectrometry mapping (LC-MS); of the 58
235 *P. carinii* genes, 56 are expressed and 53 were detected by LC-MS. Analysis of the predicted function
236 of these genes revealed housekeeping functions without major differences compared to randomly
237 sampled genes.

238 To map *P. macacae* centromeres, we performed ChIP-seq on infected lung tissues from two
239 Simian immunodeficiency virus (SIV)-infected macaques with untreated *Pneumocystis* pneumonia (cf.
240 **Supplementary Methods**). We prepared chromatin directly from lung tissues without culture.
241 Because anti-CENP-A and anti-CENP-C antibodies specific to *P. macacae* (**Supplementary figures 2a**
242 **and 2b**) yielded nonspecific peaks in ChIP-seq experiments, we utilized two antibodies targeting *P.*
243 *macacae* Mis12 (**Supplementary figure 2d**), a member of the outer kinetochore that has been used as
244 replacement of CENP-A for centromeres identification in the fungal pathogen *Mucor circinelloides*
245 (Navarro-Mendoza, Perez-Arques et al. 2019). Mis12 ChIP-seq identified -4-6 peaks per chromosome
246 (**Figure 4c; Supplementary figure 5**). Mis12 peaks tend to occur in GC rich regions. It is unclear if

247 the multiple peaks correspond to ChIP-seq artifacts or if Mis12 truly binds to multiple locations per
248 chromosome. However, there is only one peak flanked by heterochromatin (H3K9me2) per
249 chromosome, which also flanked *P. carinii* and *P. murina* centromeres (**Figures 4a and 4b,**
250 **Supplementary figures 6 and 7**). DNA sequence similarity searches showed that this peak is
251 homologous to *P. carinii* and *P. murina* centromeres. We consider this peak to be a putative
252 centromere.
253 We were unable to map centromeres in *P. jirovecii* despite utilizing anti-CENP-A and anti-
254 CENP-C antibodies specific for this species (inconclusive ChIP-seq results), possibly because the only
255 samples available for Chip-seq studies were from patients previously treated with anti-*Pneumocystis*
256 drugs.
257
258 **Centromeric nucleosomes assemble differently among species.**
259 CENP-C binds to specific motifs in CENP-A in the C-terminal cupin domain involved in
260 dimerization (Kato, Jiang et al. 2013). To confirm the absence of CENP-C ChIP-seq peaks at *P. carinii*
261 centromeres, we analyzed chromatin by co-immunoprecipitation (Co-IP) with anti-CENP-A, anti-
262 CENP-C and anti-H4 antibodies followed by LC-MS analysis. While CENP-A was detected by
263 peptide mapping with high confidence (peptide: "WQSTAILCLQEATEAFLVHLFEDTNLCAIHAK",
264 Qvality q-value = 0.00553986), none of the CENP-C peptides were identified in either of the two
265 runs. Of note, we detected CENP-C in total proteins from both species by LC-MS and Western
266 blotting (**Supplementary figures 2e and f**). These results are consistent with ChIP-seq results and

267 suggest that *P. carinii* CENP-C does not localize at the *P. carinii* centromeres. *pn*CENP-C may be
268 dispensable in *P. carinii*, similar to *S. pombe* (Sridhar and Fukagawa 2022).
269 To gain further molecular insight into the CENP-A nucleosome differences, we performed
270 molecular dynamics (MD) simulations of centromeric complex models of *P. carinii*, *P, murina*, and *S.*
271 *pombe* (cf. **Supplementary Methods**). The analysis suggests that the observed differences can be
272 explained by the different histones H2A:H2B dimers recruited by each species at the centromeres
273 (**Figure 5a**). Comparative analysis shows that variation in the terminal regions of these proteins can
274 induce structural and dynamic changes in the region surrounding the central crevice formed by the
275 spatial confluence of four histones (cf. **Figure 5b and Supplementary animation D2**). This region
276 of the octameric complex is solvent-exposed and poised for binding to inner-kinetochore proteins
277 such as CENP-C. Slight variations in its average size or dynamics can affect protein selection and
278 binding process, hence the early-stage assembly mechanism, even when its amino acid composition is
279 identical, as is the case of *P. carinii* and *P. murina*. Compared to these species, *S. pombe* presents
280 several mutations lining the crevice (cf. **Figure 5a**), which adds physicochemical features to consider
281 when comparing the two organisms. For the *Pneumocystis* sequences modeled here, the structural and
282 dynamic differences between the two species stem from the longer C-terminal domain of H2A in *P.*
283 *murina*, which adopts a helical conformation with several nonpolar moieties. This helix is quite
284 mobile and tend to interact hydrophobically with the disordered N-terminus of CENP-A
285 (**Supplementary Table 3**). However, when free from these interactions, the helix can transiently
286 gravitate towards the core center, affecting the exposure of the crevice to the solvent, hence to other

287 proteins (cf. **Figure 5b**). Transient helices in intrinsically disordered segments can have functional
288 implications (Chen, Yang et al. 2021, Zhao, Nguyen et al. 2022), and hydrophobic helixes, in
289 particular, can play a role in molecular recognition and drive protein association and aggregation
290 (Zhao, Nguyen et al. 2022).
291
292 ***Pneumocystis* CENP-A replaces the canonical histone H3 at the centromeres.**
293 In *S. pombe*, CENP-A^{cnp1} enrichment at the centromeres correlates with a depletion of histone H3
294 (Thakur, Talbert et al. 2015). To determine if this feature is present in *Pneumocystis*, we quantified the
295 relative abundance of canonical histones H3 and H4 using ChIP-seq. We examined the 16-17
296 centromeres per species, their flanking regions (30 kb) and control regions selected 25 kb away from
297 centromeres. H3 occupancy (expressed as ratio between H3 and H4) was found to be significantly
298 reduced at centromeres relative to controls in *P. carinii* and *P. murina* but not *P. macacae* (Wilcoxon
299 test; $P < .0001$; **Supplementary figure 8**). In parallel to ChIP-seq, we found by ChIP-qPCR a
300 reduction of histone H3 in 15 of the 17 *P. carinii* centromeres, and in 2 of the 17 *P. murina*
301 centromeres compared to the controls (**Supplementary figure 9**). The discrepancy between *P.*
302 *murina* ChIP-seq and ChIP-qPCR is potentially due to variability in culture experiments.
303
304 **Centromeres sequences are unique and lack repetitive sequences.**
305 Centromeres are located in non-repetitive sequences within the *Pneumocystis* genomes
306 (**Figures 3e and 3f**). Motif searches also did not reveal any conserved motif either in the centromeric

307 regions as a whole or when focused on *pn*CENP-A peaks. To determine if repeats (DNA transposons
308 and retrotransposons) are associated to centromeres, we searched for known signatures of repeats. In
309 *P. murina*, we detected long terminal repeats (LTRs) (*copia* and *gypsy*) in the vicinity of 5 of the 17
310 centromeres but they are not directly flanking the CENs (**Supplementary figure 7**). In *P. carinii*, of
311 the 17 centromeres, the CEN2 contains a duplicated copy of *copia*-retrotransposons and CEN16 has a
312 *gypsy* repeat in its upstream region (**Figure 4b; Supplementary figure 6**). The repeated DNA
313 elements are likely nonfunctional for further transposition because they lack essential domains
314 required for transposition activity (LTRs, reverse transcriptase, integrase) and they are integrated into
315 genes. No repeat is present in the vicinity of *P. macacae* centromeres (**Supplementary figure 5**).
316
317 **Centromeres are flanked by heterochromatin and lack DNA methylation marks.**
318 In *S. pombe*, H3K9me2 marking of chromatin is associated with the repression of centromeres,
319 subtelomeres, ribosomal rDNA and the mating locus (Cam, Sugiyama et al. 2005). There are two
320 types of chromatins: euchromatin, the lightly packed form of the chromatin enriched with H3K4me2
321 (di-methylation of the 4th lysine residue of the histone H3), which is associated with active
322 transcription, and heterochromatin enriched with H3K9me2 and H3K9me3 (di and tri-methylation
323 of the 9th lysine residue of the histone H3 protein). Heterochromatin regulates mating and
324 sporulation gene expression in *S. pombe*, in addition to controlling mating-type switching (Allshire
325 and Ekwall 2015). To test if this feature is shared in *Pneumocystis*, we performed ChIP-seq with
326 antibodies targeting histones H3 modifications (H3K9me2/3 and H3K4me2). In *Pneumocystis*, peaks

327 of H3K9me2 and H3K9me3 are only present at the pericentromeres (Supplementary figures 5,6 and
328 7). In all three species analyzed (*P. carinii*, *P. murina* and *P. macacae)*, H3K9me2 enrichment is
329 significantly more pronounced in pericentromeric regions compared to centromeres (**Figure 4;**
330 Wilcoxon test, $P < .00001$; **Supplementary figure 8**). H3K9me3 enrichment at pericentromeric
331 regions compared to centromeres is only significant in *P. carinii*. We observed a reduced
332 transcriptional activity by RNA-seq analysis only in *P. carinii* pericentromeres (**Supplementary**
333 **figure 8**). H3K4me2 is correlated with H3K9me2 in the centromeres (Pearson rho = .72) but not at
334 the whole genome level (r = 0.05). However, research has shown that the histone modifications
335 H3K4me2 and H3K9me2 decorate heterochromatin and euchromatin regions, respectively (Yasuhara
336 and Wakimoto 2008). Therefore, the correlation observed between H3K4me2 and H3K9me2 may be
337 due to the *Pneumocystis* population of cells used in the ChIP analysis. Furthermore, similar to *S.*
338 *pombe* (Cam, Sugiyama et al. 2005), a small amount of H3K4me2 is present at the CENP-A
339 containing of centromeres in *Pneumocystis* (**Supplementary figure 8**). These results suggest that
340 *Pneumocystis* centromeres are flanked by heterochromatin.
341 DNA methylation is frequently associated with centromeres in fungi, where they silence
342 repeats (Miura, Yonebayashi et al. 2001, Smith, Galazka et al. 2012). To determine whether DNA
343 methylation plays a role in *Pneumocystis* centromeres, we performed bisulfite sequencing (5
344 methylcytosine) in *P. carinii* and *P. macacae*. The overall level of 5mC DNA methylation as measured
345 by the average weighted methylation percentage is 0.6% for *P. carinii* and 2% for *P. macacae* at the CG
346 dinucleotides (**Supplementary figure 10**). These levels are in range with reported levels for other

347 fungi e.g., *Verticillium* (0.4%) (Cook, Kramer et al. 2020) and *Neurospora* (2.5%) (Hosseini, Meunier et
348 al. 2020). To assess the potential role of DNA methylation implication in centromere function, we
349 analyzed the DNA methylation patterns over different genomic features (genes, intergenic spacers,
350 and centromeres). DNA methylation levels are higher in genes compared to centromeres (Mann-
351 Whitney U-test $P = .005$; **Supplementary figure 10**). However, there is no significant difference
352 among centromeric or pericentromeric regions (defined as 30 kb flanking the centromeres: P-value >
353 0.3), and the randomly selected genomic regions (genomic background). These results suggest that 5-
354 mC DNA methylation is not required for centromere function in *Pneumocystis*, consistent with the
355 absence of repeats in the centromeres.
356
357 **Centromere conservation during speciation**
358 *Pneumocystis* species diverged about 100 Mya (**Figure 2a**). Yet, homologous regions act as
359 centromeres in *P. carinii*, *P. murina* and *P. macacae* (**Figure 6a, Supplementary table 4**). Centromeres
360 in these three species have similar DNA sequences and in *P. jirovecii* as well as *P. wakefieldiae*, *P.*
361 *oryctolagi* and *P. canis* (**Supplementary figure 11**), which indicates that sequences are vertically
362 transmitted.
363 To investigate footprints of selection acting on centromeres, we computed conservation
364 scores (PhastCons) from whole genome alignments. PhastCons conservation scores range from 0 to 1
365 and represent probabilities of negative selection. Centromeric regions are more conserved than
366 genomic background (Wilcoxon test $P < 2.2e-16$**; Figures 6b and 6c**). This suggests that centromeres
367 are maintained by negative selection. This is consistent with the hypothesis that centromere

368 positioning tends to be conserved in obligate sexual fungi due to their role in meiosis (Schotanus,
369 Soyer et al. 2015).
370 Centromeres may contribute to karyotypic diversity in fungi. However, we found little
371 support for this hypothesis here because most centromeres do not overlap with chromosomal breaks
372 (only 3 of 17 in the *P. carinii* versus *P. murina* pairwise comparison) (**Figure 6a**).
373
374 **Discussion**
375 In the current study we have identified putative centromeres of multiple *Pneumocystis* species and
376 demonstrated that they are short regional centromeres with pericentromeric heterochromatin and
377 span active genes (**Figure 7**). Cell replication and chromosomal segregation pathways are unexplored
378 in these species. Central to these pathways are the centromeres.
379 As centromeres cannot be predicted bioinformatically, the locations and characteristics of
380 *Pneumocystis* centromeres were unknown. Moreover, there are several roadblocks for characterizing
381 centromeres in *Pneumocystis*, with the most significant one being the lack of continuous culture and
382 transfection tools. Here we utilized two species of organisms in a short-term culture system to
383 determine where *pn*CENP-A, the epigenetic marker for centromeres, binds in the genome. We
384 demonstrated that *Pn*CENP-A functions similarly to CENP-A^{Cnp1} in *S. pombe*, though less efficiently,
385 validating its use in characterizing *Pneumocystis* centromeres. Unlike the phylogenetically related *S.*
386 *pombe*, in *Pneumocystis* these regions overlap with active genes, but like *S. pombe*, these centromeres
387 are flanked by heterochromatin. Our work provides the first experimental evidence that DNA

388 methylation occurs in these species, although it may not be involved in centromere function. Taken
389 together, our results suggest that *Pneumocystis* has small epigenetically regulated centromeres.
390 Individual centromeres are conserved among different species, though not across chromosomes
391 within species.

392 Differences between *P. carinii* and other species analyzed here (*P. murina* and *P. macacae*)
393 need to be clarified. *P. carinii* and *P. murina* centromeres are homologous, yet there seems to be
394 heterochromatin silencing only in *P. carinii*. Also, CENP-C is not linked to CENP-A or centromeres
395 in *P. carinii*, which may instead be using another histone linker (e.g., CENP-T) to connect the
396 centromere to the kinetochore (Sridhar and Fukagawa 2022).

397 The presence of genes within centromeres is rare and has only been described in rice (Nagaki,
398 Cheng et al. 2004) and the plant fungal pathogen *Zymoseptoria tritici* (Schotanus, Soyer et al. 2015).
399 Centromeres bearing genes are interpreted as young centromeres (neocentromeres), in which the
400 genes are progressively inactivated. This hypothesis lacks support here because there is no sign of
401 pseudogenization in these genes: nearly all are expressed and translated, they are conserved in all
402 species, and many are involved in housekeeping cellular pathways that are presumably critical for
403 organism survival.

404 Findings in *Pneumocystis* cannot be generalized to other Taphrinomycotina species. The
405 determination of ancestral traits between *Pneumocystis* and fission yeast centromeres will require
406 characterizing the centromeres in additional Taphrinomycotina species. The loss of RNAi is often

407 associated with shortening of centromeres (Yadav, Sun et al. 2018); consistent with this, *Pneumocystis*
408 centromeres are much smaller than *Schizosaccharomyces* centromeres.

409 Our study will benefit from a microscopy validation of CENP-A loading kinetics when a
410 reliable long term culture system becomes available. Live imaging would help to determine at which
411 cell cycle stage CENP-A is loaded.

412 In summary, we have identified atypical short regional centromeres in genetically intractable
413 micro-organisms. Our results provide insights into the formation of centromeres in host-adapted
414 fungal pathogens. Our ultimate goal is to use *Pneumocystis* centromeres to stabilize plasmids for
415 future genetic manipulation. This is the first step along this path, which should lead to better
416 understanding of the biology of *Pneumocystis* and facilitate the discovery of novel interventions to
417 effectively control and prevent the disease caused by this pathogen.

418

419 **MATERIAL AND METHODS**

420 **Ethics and organism source**

421 Studies involving human and mouse samples were approved by the NIH Institutional Review Board
422 (IRB) (IRB protocols 99-I-0084 and CCM 19-05, respectively), rat *Pneumocystis* studies by the Animal
423 Care and Use Committees of the Cincinnati VA Medical Center (protocol 20-11-08-01) and
424 macaque *Pneumocystis* studies by the National Institute of Allergy and Infectious Diseases (NIAID)
425 Division of Intramural Research (protocol LVD 26).

426 *P. carinii* organisms were collected from heavily infected lungs of corticosteroid-treated
427 Sprague-Dawley male rats; *P. murina* from heavily infected lungs of CD40 ligand knock out mouse;
428 *P. macacae* from heavily infected lungs of two Simian Immunodeficiency Virus (SIV) infected
429 macaques (P2C and P3C); and *P. jirovecii* organisms were isolated from a single bronchoalveolar
430 lavage from one patient and 3 heavily infected autopsy lung tissues from three different patients. The
431 list of samples is presented in the key resource file.
432
433 ***Pneumocystis* culture and growth quantification**
434 *P. murina* and *P. carinii* were partially purified by Ficoll-Hypaque density gradient centrifugation
435 (Kovacs, Halpern et al. 1988) and frozen at -80ºC in cell recovery media (Gibco). For short term
436 *Pneumocystis* cultures, A549 (ATCC) and LET1 (a gift from Dr. Paul Thomas, St Jude Children's
437 Research Hospital) cells were cultured in culture medium F12 media with 2.5% or 5% heat
438 inactivated fetal bovine serum (GIBCO) and Penicillin-Streptomycin (GIBCO), plated to
439 approximately 60 – 80% confluency and incubated for 24h at 37°C. The next day, frozen *Pneumocystis*
440 vials were thawed, washed in 50ml 1x PBS, and centrifuged at 2,000g for 20 minutes. *Pneumocystis*
441 cell pellets were resuspended in culture medium and added to the plated cells. Media were partially
442 changed every 3 or 4 days. At set time points, wells were scraped for collection. Cell/organism
443 suspensions were centrifuged at 10,000g for 3 min, the supernatant was removed, and the pellets
444 were frozen at -80°C. Genomic DNA was extracted using QiAmp DNA Extraction Kit (Qiagen).

445 Quantitative PCR targeting the single copy *dhfr* gene was performed as described previously (Liu,
446 Davis et al. 2020).
447
448 **_Pn_CENP-A complementation and ChIP-qPCR in _S. pombe_**
449 Standard procedures were used for fission yeast growth, genetics, and manipulations (Sabatinos and
450 Forsburg 2010). *P. jirovecii* and *P. murina* CENP-A full-length and N-terminal truncated DNA
451 sequences were codon optimized for *S. pombe* codon bias (**Supplementary dataset 1**), synthesized
452 commercially (GenScript, NJ) and subcloned into pS2 vector (Takayama, Sato et al. 2008).
453 Recombinant plasmids were isolated from *E. coli* strain *DH5α* colonies (Invitrogen) using QIAprep
454 Spin Miniprep Kit (Qiagen), linearized with *Nsi*I and introduced by transformation at the *lys1+* locus
455 of *cnp1-76* (Castillo, Mellone et al. 2007) and *cen2-tetO-tdTomato* (Sakuno and Watanabe 2009). For
456 ChIP-qPCR experiments, *S. pombe* cells were grown overnight at 30°C in rich medium YEA and 2 µl
457 of anti-GFP (ab290, Abcam) was used for each immunoprecipitation. PCR oligonucleotides detecting
458 central core (*cc1&3; ura4*), heterochromatic *dg* and euchromatin control *fbp1* were previously
459 reported (Folco, McCue et al. 2019).
460
461 **Ortholog identification**
462 Fungal proteomes were retrieved from NCBI. Ortholog identification was performed using
463 OrthoFinder (Emms and Kelly 2019). Remote homology detection was performed using PSI-BLASTp
464 (Altschul, Madden et al. 1997) and HMMER v.3.x.x (http://www.hmmer.org) (E-value < 10^{-5}) using *S.*

465 *pombe* and *S. cerevisiae* proteins as seeds. Ortholog results were further validated by BLAST searches
466 on NCBI website, and searches in eggNOG 6.0 (Hernandez-Plaza, Szklarczyk et al. 2023). Gene
467 reannotations were performed using protein-to-genome alignment with Exonerate (Slater and Birney
468 2005) followed by manual adjustments if necessary (cf. **Supplementary Methods**). *P. macacae* strain
469 P3C was sequenced using Oxford nanopore sequencing (cf. **Supplementary Methods**). Protein
470 multiple sequence alignments were generated using Clustal Omega (Sievers, Wilm et al. 2011).
471 Maximum likelihood phylogeny was performed using RAxML-ng (Kozlov, Darriba et al. 2019).
472 Predicted histones phylogenetic classification was performed with reference protein sequences
473 identified by (Postberg, Forcob et al. 2010). Immunogenetic peptides were selected for *Pneumocystis*
474 CENP-A, CENP-C and mis12 orthologs, used for antibody production and quality control
475 (**Supplementary Methods**).
476
477 **Light, electron and immunofluorescence microscopy**
478 *P. murina* and A549-LET1 cells described above were grown on Thermanox™ coverslips (Ted Pella,
479 Redding, CA, USA), harvested at day 7, fixed with 2% paraformaldehyde/2.5% glutaraldehyde in 0.1
480 M Sorenson's phosphate buffer and then post-fixed with 1.0% osmium tetroxide/0.8% potassium
481 ferricyanide in 0.1 M sodium cacodylate buffer for 1 hour, washed with buffer then stained with
482 1% tannic acid in dH_2O for 1 hr. After additional buffer washes, the samples were further
483 osmicated with 2% osmium tetroxide in 0.1M sodium cacodylate, then washed with dH_2O.
484 Specimens were then stained overnight at 4C with 1% aqueous uranyl acetate. The cells were
485 then washed with dH_2O and dehydrated with a graded ethanol series, prior to embedding in

486 Spurr's resin. Thin sections were cut with a Leica UC7 ultramicrotome (Buffalo Grove, IL) prior
487 to viewing at 120 kV on a FEI BT Tecnai transmission electron microscope (ThermoFisher/FEI,
488 Hillsboro, OR). Digital images were acquired with a Gatan Rio camera (Gatan, Pleasanton,
489 CA).

490 Fluorescence labelling of *Pneumocystis*-infected tissues was performed by Histoserv, Inc
491 (Germantown, Maryland) and visualized using Leica DMi8 microscope (DMI Medical Inc).

492 *S. pombe* cells were grown overnight at 32°C in minimal medium PMG until logarithmic
493 phase and then mounted in PMG 2% agarose as described (Tran, Paoletti et al. 2004). Images were
494 acquired on a Delta Vision Elite microscope (Applied Precision) with a 100X 1.35 NA oil lens
495 (Olympus). Twenty 0.35-µm z sections were acquired generating maximum intensity projections.
496 Further image processing including background-subtracted centromere foci intensity measurements,
497 was performed using ImageJ (National Institutes of Health).

498

499 **ChIP-seq**

500 Purified cells (~5 x 10^6 cells) and tissue preparations were fixed by 37% formaldehyde treatment.
501 Chromatin preparation, immunoprecipitation and Illumina sequencing libraires were prepared
502 using the Low Cell ChiP-Seq kit (Active Motif, Carlsbad, CA) according to the manufacturer
503 instructions. ChIP-seq libraries were sequenced commercially (Psomagen Rockville, Maryland, USA)
504 using Illumina NovaSeq 6000 (150-bp paired end reads).

505

506

507 **RNA-Seq**

508 Total RNA was extracted using RNeasy Mini kit (Qiagen). RNA quality and integrity were
509 estimated using Bioanalyzer RNA 6000 Pico Assay (Agilent) and sequenced commercially using
510 Illumina HiSeq 4000 (150-base paired end libraries, Novogene Inc, USA). Reads were mapped using
511 STAR v.2.7.9a (Dobin, Davis et al. 2013), and normalized to bins per million mapped reads (BPM).
512 Transcripts counts were quantified using kallisto 0.44.0 (Bray, Pimentel et al. 2016).

513

514 **ChIP-seq peak calling and sequence analysis**

515 Raw reads were quality-checked with multi-QC (Ewels, Magnusson et al. 2016). Host reads (human,
516 macaque, rat, mice) were removed by mapping using Bowtie2 (Langmead and Salzberg 2012) with
517 default parameters. Unmapped reads were retrieved and mapped to *Pneumocystis* reference genomes
518 (*P. murina* (GenBank accession number GCA_000349005.2), *P. carinii* (GCA_001477545.1), *P.*
519 *macacae* (GCA_018127085.1) and *P. jirovecii* (GCA_001477535.1)) using the Burrows-Wheeler
520 Aligner BWA-MEM (Li 2013). Alignments were filtered and sorted using SAMTools (Li, Handsaker
521 et al. 2009) and duplicates marked using PICARD MarkDuplicates
522 http://broadinstitute.github.io/picard). Quality controls were performed using deepTools
523 plotFingerprint (Ramirez, Ryan et al. 2016) and phantompeakqualtools (Landt, Marinov et al. 2012),
524 which estimate the normalized strand coefficient (NSC) and the relative strand correlation (RSC).
525 Chip-Seq coverage data were calculated as 50-bp bin and normalized per genomic content (1x

526 normalization). Chromosome ideograms were generated using chromoMap v0.3 (Lakshay Anand
527 2020).

528 ChIP enrichment peaks were detected using Model-Based Analysis for ChIP-Sequencing
529 (Zhang, Liu et al. 2008)(MACS3 v3.0.0a5) callpeak function with default parameters and estimated
530 genome sizes and sorted by fold-enrichment and FDR values (broad cutoff 0.1; FDR < 0.05).
531 Enriched peaks were further inspected using deepTools bamCompare versus shuffle non centromeric
532 genomic regions. Peaks were further examined in conjunction with other data (e.g. conservation
533 score, RNA-seq, methylation data, GC content, repeats) with the Integrative Genomics Viewer
534 (Robinson, Thorvaldsdottir et al. 2011), ensuring that peaks are present in all IP samples and absent
535 in controls. *De novo* DNA motif searches were performed using MEME v5.1.0 (Bailey, Boden et al.
536 2009) and Homer v4.9 (Heinz, Benner et al. 2010) with a p-value cut off of 10^{-10}. GC content across
537 windows (bin) was computed using BEDTools (Quinlan and Hall 2010). Normalized coverage and
538 annotations across chromosomes were visualized using pyGenomeTracks (Lopez-Delisle, Rabbani et
539 al. 2021).

540 Repeated elements (DNA transposons, retrotransposons and low complexity repeats) were
541 identified using RepeatMasker (Smit AFA. 2013-2015) and RepBase (Bao, Kojima et al. 2015).
542

543 **Genome synteny and conservation**

544 DNA alignments were generated using Satsuma2 (Grabherr, Russell et al. 2010) and visualized with
545 Circos (Krzywinski, Schein et al. 2009). To estimate the sequence conservation scores, we performed

546 one-to-one pairwise whole genome alignments using LAST (Kielbasa, Wan et al. 2011) with the
547 MAM4 seeding scheme (Frith and Noe 2014). We consider two trios (1) *P. jirovecii*/*P. macacae*/*P.*
548 *oryctolagi* and (2) *P. carinii*/*P. murina*/*P. wakefieldiae.* We generated a neutral model using phyloFit
549 (Siepel and Haussler 2004) from PHAST which fit the tree model to multiple sequence alignments by
550 maximum likelihood using the REV substitution model. Sequence conservation scores were
551 estimated using phastCons (Siepel, Bejerano et al. 2005) (--target-coverage 0.25 --expected-length 20 --
552 estimate-trees). Wig files and Bedgraph files were converted using 'wigToBigWig' and
553 'bigWigToBedGraph' tools (Kent, Zweig et al. 2010).
554
555 **Bisulfite sequencing**
556 *Pneumocystis* DNA (5 μg) was extracted by a *Pneumocystis* DNA enrichment protocol (Ma, Chen et al.
557 2016). Bisulfite conversion, library preparation and sequencing (Illumina NovaSeq 150 bp paired end
558 reads) were performed commercially (Novogene). Bisulfite conversion of non-methylated DNA was
559 performed using EZ DNA Methylation kit. Data analysis was performed using BSMAP and
560 methratio script (Xi and Li 2009). Only cytosine positions with more than 5-fold coverage were
561 considered. We used weighted methylation percentage (Cook, Kramer et al. 2020), which is
562 calculated as the number of reads supporting methylation over the number of cytosines sequenced to
563 quantify methylation levels. Methylation data were partitioned over different genomic compartments
564 using BEDTools (Quinlan 2014). Statistical comparisons were performed using R (https://www.R-

565 project.org) to compute non-parametric Mann-Whitney U test and Bonferroni correction to adjust *P*-
566 values for multiple comparisons.
567
568 **ChIP-qPCR**
569 CENP-A enrichment and H3 depletion relative to H4 within the centromeres were evaluated using
570 primers specific to centromeric and non-centromeric loci (see **Supplementary Table 5** for primers).
571 Primers were designed using primer3 (Koressaar, Lepamets et al. 2018) from regions visualized using
572 IGV genome browser. Primers with potential matches against mammal genomes were excluded
573 using NCBI BLASTn. Dilutions of 1:10 were used based on preliminary test runs. Each reaction
574 contains 5 µl of iTaq universal Sybr green supermix (Biorad), 1.25 µl of primers (500 nM)) and 2.5 µl
575 of DNA. All assays were run on CFX96 thermocycler (Biorad). Three technical replicates were taken
576 for each assay, and the standard errors of the mean were calculated. The PCR program was as follow:
577 initial denaturation for 2 min at 95°C, followed by 30 cycles of 30 seconds at 95°C, 60°C for 30s, and
578 72° for 30s. Locus ΔCt values were normalized using the formula: ($^{\Delta\Delta Ct}$ ChIP – $^{\Delta Ct}$ Input;
579 https://www.sigmaaldrich.com/US/en/technical-documents/protocol/genomics/qpcr/chip-qpcr-data-
580 analysis) and the fold enrichment of the ChIP DNA over the input was computed as log2 ($2^{\Delta\Delta Ct}$). The
581 plots and statistical analyses were performed with GraphPad Prism 8.
582
583
584

585 **ChIP-LC-MS**

586 Formalin-fixed purified chromatin preparations were separately co-immunoprecipitated with anti
587 CENP-A, CENP-C, H4 and pre immune serum (control) using Pierce™ Direct Magnetic IP/Co-IP
588 Kit (ThermoFisher). Each IP reaction was performed separately on aliquots from the same chromatin
589 preparation. Elution was performed using 1% formic acid. Extracts were frozen in dry ice, lyophilized
590 for 1 hour and digested using trypsin, diluted, and injected to a Thermo Orbitrap Fusion LC-MS/MS
591 to identify unique peptides. LC-MS/MS data were searched against the proteomes using Proteome
592 Discoverer 2.4 (Thermo Fisher Scientific, Waltham, MA). Label-free quantification was analyzed
593 based on the peak intensity of the precursor ion.

594

595 **Data availability**

596 ChIP-seq datasets and associated MACS2 outputs generated in this study (CENP-A, CENP-C, Mis12,
597 H3, H4, H3K9me2/3, H3K4me2) are deposited at NCBI GEO (accession numbers GSE226536,
598 GSE230598).

599

600 **Code availability**

601 Codes generated for this project are available at GitHub: https://github.com/ocisse/pneumo-cens
602 Multi panel figures were assembled in Inkscape using original plots.

603

604 **Acknowledgements**

605 We thank Rene Costello for animal studies, Drs. Keytam Awad, Salina Gairhe and Shuibang Wang
606 for discussion about ChIP-seq. This work has been funded in whole or in part with federal funds
607 from the Intramural Research Program of the US National Institutes of Health (NIH) Clinical Center
608 and the National Institute of Allergy and Infectious Diseases (NIAID). The views expressed in this
609 work are those of the authors and do not necessarily represent the official positions of the NIH or
610 U.S. Government. This study used the Office of Cyber Infrastructure and Computational Biology
611 (OCICB) High Performance Computing (HPC) cluster at the National Institute of Allergy and
612 Infectious Diseases (NIAID), Bethesda, MD. This study also utilized the high-performance
613 computational capabilities of the Biowulf Linux cluster at the NIH, Bethesda, MD
614 (http://biowulf.nih.gov).
615

616 **Figure legends**

617 **Figure 1. *Pneumocystis* CENP-A variants localize to the nucleus, and support viability and**
618 **centromere loading in a heterologous system.**

619 (**a**) Protein sequence alignment of full-length CENP-A^{cnp-1} homologs in *Schizosaccharomyces pombe*
620 and seven *Pneumocystis* species. Each *Pneumocystis* species only infects a single mammalian host: *P.*
621 *jirovecii* (infecting humans), *P. macacae* (macaques), *P. oryctolagi* (rabbits), *P. canis* (dogs), *P.*
622 *murina* (mice), *P. carinii* (rats), and *P. wakefieldiae* (rats). The ~25 amino acid insert in
623 *Pneumocystis* CENP-A N-terminal regions is unique to this genus.

624 (**b**) Immunofluorescence microscopy showing that anti-*Pneumocystis murina* CENP-A (*pn*CENP-A)
625 reacts with *P. murina* organisms. *P. murina*-infected CD40 ligand knockout lung tissue sections
626 were dual labeled using an anti CENP-A antibody(magenta) and 4′,6-diamidino-2-phenylindole
627 (DAPI, light blue). The image is a maximum projection of a 5 µm thick z-stack. The images were
628 minimally adjusted to enhance the contrast. The green box is magnified in the inset showing
629 *pn*CENP-A inside *Pneumocystis* organisms (small dots). *pn*CENP-A signals overlap or are close in
630 proximity with DAPI (represented by a white color produced by the overlap between magenta
631 and light blue colors).

632 (**c**) Left, rescue experiment of fission yeast *cnp1-76* thermosensitive cells using the GFP-tagged
633 CENP-A transgenes. Right, protein sequence alignment of *Pneumocystis* CENP-A N-terminal
634 regions (*P. jirovecii* and *P. murina*) with conserved amino acids highlighted (grey areas). Plasmids
635 containing full length and N-terminal region truncated (**) versions of *Pneumocystis* CENP-A
636 were constructed and integrated as single copy. Growth was assayed at indicated temperatures
637 using 10-fold serial dilutions plated on rich YEA medium. The transgene encoding *S. pombe*
638 CENP-A serves as a non-temperature-sensitive control.

639 (**d**) Representative images of *cen2-tetO-TdTomato* strains expressing the indicated GFP-tagged CENP-
640 A transgenes. On the bottom, GFP and tdTomato fluorescence as well as brightfield images are
641 merged.

642 (**e**) Integrated fluorescence intensity of GFP foci was measured and plotted for the strains used in
643 panel e. Mean (bars) and SD (error bars) are shown. n>176. AU, arbitrary units.

644 (**f**) Anti-GFP ChIP-qPCR analysis of indicated loci for strains used in panel e. %IP represents the
645 percentage of input that was immunoprecipitated. Error bars denote the SD (n = 3).
646 Abbreviations: cc1 &3, *S. pombe* centromere central core 1 & 3; cnt2:ura4^{+}, ura4^{+} insertion at
647 central core 2; *dg*, a class of heterochromatic outer repeats within pericentromeric regions
648 (control); *fbp1*, euchromatic locus (control)

649

650

651

652

653 **Figure 2. *Pneumocystis* CENP-A binds to single genomic foci in replicating cells**

654 **a**) Study workflow and sample preparation. Four *Pneumocystis* species infecting mice, rats,
655 macaques, and humans respectively were analyzed. *P. murina* and *P. carinii* were obtained
656 from CD40 ligand knock-out female mice and immunosuppressed Sprague-Dawley male rats,
657 respectively, and cultured on a co-culture of human lung adenocarcinoma cells (A549) and
658 immortalized murine lung epithelial type 1 cells (Let-1) for 14 days. *P. macacae* organisms
659 were obtained from a Simian Immunodeficiency Virus infected macaque (P2C), and *P.*
660 *jirovecii* organisms were obtained from a single bronchoalveolar lavage from a non-infected
661 HIV patient and three autopsy samples from patients with HIV infection. Species phylogeny
662 and estimated speciation timing are presented. Animal icons were obtained
663 from http://phylopic.org under creative commons
664 licenses https://creativecommons.org/licenses/by/3.0/: mouse (Anthony Caravaggi; license CC
665 BY-NC-SA 3.0); and rat (by Rebecca Groom; license CC BY-NC-SA 3.0).

666 **b**) *P. murina* population growth in duplicate wells were measured by quantitative PCR targeting
667 a single copy dihydrofolate reductase (*dhfr*) gene over 14 days. Error bars represent the
668 standard deviation (n=2).

669 **c**) *P. carinii* growth measurement by qPCR targeting *dhfr* gene. Error bars represent the standard
670 deviation (n=2).

671 **d**) Electron micrograph showing a possibly dividing *P. murina* trophic form seven days post
672 culture (black arrow). The field also include a non-dividing trophozoite (white arrow). 2 μm
673 scale bar.

674 **e**) *pn*CENP-A ChIP-seq study in *P. carinii* organisms from infected lungs and cultures at days 0,
675 7 and 14. CENP-A remains strongly enriched at a single genomic location up to 7 days. Only
676 one representative chromosome is presented. The remaining 16 chromosomes have similar
677 ChIP-seq profiles and presented in supplementary figure 4.

678

679 **Figure 3. *Pneumocystis* displays seventeen centromeres.**

680 **a**) Scaled ideogram of 17 chromosomal level scaffolds (grey) of *P. murina* genome showing
681 CENP-A binding regions (constricted areas) and telomeric repeat motif (red).

682 **b**) CENP-A and CENP-C binding regions delineate putative centromeres (CENs) in *P. murina*
683 genome. Color coded peaks represent enrichment of immunoprecipitated DNA (IP DNA)
684 relative to controls (Input DNA). The 1-kb flanking sequences from the center of the
685 enrichment peak are presented. Each scaffold displays two peaks labelled M (Major) and m
686 (minor) according to the enrichment level.

687 **c**) Scaled ideogram of 17 chromosomal level scaffolds (grey) of *P. carinii* genome showing
688 CENP-A binding regions (constricted areas) and telomeric repeat motif (red).

689 **d**) CENP-A binding regions delineate putative centromeres (CENs) in *P. carinii* genome. Note
690 that no peak was observed with anti-CENP-C antibody.

691 **e**) A DNA dot plot of CENP-A binding regions in *P. murina* showing regional self-similarity.
692 The main diagonal represents the sequence alignment with itself. Lines off the main diagonal
693 which are repetitive patterns within the sequences are not observed. The plot shows that each
694 CEN is unique and repeat-free.

695 **f**) DNA dot plot of CENP-A binding regions in *P. carinii* showing that each CEN is unique and
696 repeat-free.

697

698

699

700 **Figure 4. Centromeres are flanked by heterochromatin and contain active genes.**
701 All three scaffolds are homologous.
702 **a**) Genomic view of chromosome 2 of *P. murina* genome subsequently showing annotated genes
703 (directed grey boxes), DNA repeats (not present), percent GC content (blue), ChIP-seq read
704 coverage distribution (BPM normalized over bins of 50 bp; input subtracted) of CENP-A,
705 CENP-C, histones H3 and H4 ratio, heterochromatin-associated modifications (H3K9me2
706 and H3K9me3), euchromatin (H3K4me2) and gene expression (RNA-seq) in relation with
707 centromeres. Genome views of remaining 16 chromosomes are presented in Supplementary
708 dataset figure 7.
709 **b**) Genomic view of chromosome 2 of *P. carinii* genome with the same features presented as *P.*
710 *murina*. A duplicated copy of *copia*-retrotransposon is presented (red arrows), which is present
711 in syntenic regions in other *Pneumocystis* genomes. Genome views of all remaining 16
712 chromosomes are presented in supplementary figure 6.
713 **c**) Genomic view of chromosome 14 of *P. macacae* genome showing the distribution of gene
714 density, repeats (DNA transposons), homology region with GC content, mis12 ChIP-seq read
715 coverage, chromatin-associated modifications (H3K9me2 and H3K9me3 and H3K4me2) and
716 RNA-seq in relation with centromeres. The white box indicates a DNA sequence homologous
717 to *P. carinii* centromere 2. A single copy of the DNA transposon EnSPM (CACTA family) and
718 the class 2DNA transposon Helitron are present upstream the centromere (forward red
719 arrows). These elements are not present in syntenic regions in other *Pneumocystis* genomes.
720 Genome views of all remaining 15 chromosomes are presented in Supplementary figure 5.
721
722

723 **Figure 5. Modeling of Molecular dynamics of CENP-A nucleosomes in *P.* carinii, *P. murina* and**
724 ***S. pombe.***

725 **a**) Upper panel: Sequence alignment of the centromeric histones used in modeling and
726 simulations. Differences between *P. carinii* and *P. murina* colored green and red; differences
727 between *S. pombe* and *pneumocystis* colored blue. Lower panel: surface representation of the
728 centromeric octamers of *P. carinii* (left), *P. murina* (middle), and *S. pombe* (right), showing the
729 spatial location of the mutations (snapshots from the simulations); DNA (not simulated)
730 shown only as a reference based on the crystal structure of the human centromere nucleosome,
731 PDB: 3AN2). The white circles show the central crevices, the behaviors of which are expected
732 to dictate the assembly of the inner kinetochore in each species.

733 **b**) Aggregation of snapshots along the dynamic trajectory illustrating the extent of species-specific
734 conformational variations of the central crevices (see Animation in Supplementary dataset 2);
735 histones color code is presented at the bottom). The C-terminal tip of CENP-A is located at the
736 center of the core (white ribbon; residues 146-151 in *Pneumocystis* and 120 in *S. pombe*; see
737 alignment in panel a). The observed differences between *P. carinii* and *P. murina* are due to
738 structural and dynamic changes induced by the different (H2A:H2B) dimers (cf. alignment in
739 panel a); the differences between these species and *S. pombe* are likely the result of more
740 extensive sequence variations in the four histones, including several solvent-exposed mutations
741 at the core center (cf. panel a). Early-stage assembly of the inner kinetochore is thus expected
742 to be strongly determined by features of this crevice, including its size, dynamics, and other
743 physicochemical features.
744
745
746

747 **Figure 6. Centromere locations, not sequences, are evolutionary constrained by negative**
748 **selection.**

749 **a**) Genome organization and synteny in *Pneumocystis*. Circos plots depicting pairwise *P. carinii*,
750 *P. murina* and *P. macacae* genome synteny. Rodent infecting Pneumocystis (*P. carinii* and *P.*
751 *murina*) and macaques infecting *Pneumocystis* (*P. macacae*) have 17 and 16 chromosomal level
752 scaffolds, respectively. Colored connectors indicate regions of synteny between species.
753 Centromeres that overlap with recent chromosomal breakpoints are indicated (*). The square
754 highlights *P. carinii* centromere 4 displayed in the panel b.

755 **b**) Genome view of centromere 4 in the *P. carinii* genome. Genes are represented by directed
756 boxes (grey for protein coding genes and cyan for polymorphic major surface glycoprotein
757 genes), *pn*CENP-A binding region (centromere) and sequence conservation scores which
758 were calculated from whole genome alignments of *P. carinii*. *P. murina* and *P. wakefieldiae*
759 (PhasCons). The phastCons scores represent probabilities of negative selection and range
760 between 0 (no conservation) and 1 (total conservation).

761 **c**) Boxplot of conservation scores per genomic context summarized over all *P. carinii*
762 centromeres (CEN), 30-kb regions flanking the centromeres (Cenflk), major surface
763 glycoproteins (Msg) and random genomic background (Bckg). Statistical differences for the
764 indicated comparisons were obtained using one-sided nonparametric Mann-Whitney test; *P*-
765 values < .0001, ****; $P < .001$, ***; $P < .01$, **; $P < .05$, *; ns, not significant.
766
767

768
769 **Figure 7. Model of regional centromere structures in *Pneumocystis* and other representative**
770 **fungi.**
771 On the left is presented the phylogeny of selected fungi inferred from maximum likelihood
772 phylogenetic analysis of shared core protein orthologs. The overall centromere structures for
773 *Pneumocystis carinii* (details supporting our model are provided in the text), *Schizosaccharomyces pombe*
774 (Pidoux and Allshire 2004), *Candida albicans* (Sanyal, Baum et al. 2004), *Zymoseptoria tritici*
775 (Schotanus, Soyer et al. 2015) and *Cryptococcus neoformans* (Janbon, Ormerod et al. 2014) are used as
776 representative to showcase the diversity of regional centromeres in fungi. Animal pathogens are
777 highlighted in pale olive and the plant fungal pathogen *Zymoseptoria tritici* is in pale green. For sake
778 of brevity, only *Pneumocystis carinii* is presented for brevity, although there are differences with other
779 *Pneumocystis* species. In the middle are presented DNA structures of centromeres. *P. carinii* has 17
780 centromeres (one for each of its 17 chromosomes) that share the same overall architecture.
781 Centromeres are delineated by a localized enrichment of the centromeric histone CENP-A, which
782 overlap with a reduction of the canonical histone H3 (inverted grey triangle). Centromeres are
783 flanked by heterochromatin H3K9me (stands here stands for both H3K9me2 and H3K9me3). All 17
784 *Pneumocystis carinii* centromeres span active genes (dark grey boxes). Each centromere sequence is
785 different and lacks shared DNA sequence motif. *S. pombe* has three centromeres that share the same
786 overall structure in which a central core (*cnt*) domain is surrounded by innermost repeats (*imr*) and
787 outer repeats (*otr*). The *imr* repeats incorporate clusters of transfer RNAs (tRNAs) that play a role in
788 restricting CENP-A spread. *S. pombe* centromeres are flanked by heterochromatin (H3K9me). Genes
789 are found 0.75-1.5 kb beyond the limits of the centromeres. *C. albicans* has 8 unique and different
790 centromeres that are gene free and lack shared sequence motifs. *Z. tritici* has 21 centromeres ranging
791 from 6 to 14 kb in size, that partially overlap with genes. *C. neoformans* has 14 centromeres that are
792 gene free and enriched with Tcn transposons.
793
794

795 **List of Supplementary files:**
796 **Supplementary Methods.**
797 **Supplementary figure 1.** Conservation of kinetochore, chromatin formation, chromatin modifying
798 methyltransferases, DNA methylation and RNAi components in Taphrinomycotina and other
799 representative fungi.
800 **Supplementary figure 2.** Antibody selection and testing
801 **Supplementary figure 3.** *Pneumocystis* CENP-A binds to single genomic foci in replicating cells.
802 **Supplementary figure 4.** CENP-A enrichment at *Pneumocystis* centromeres by ChIP-qPCR.
803 **Supplementary figure 5.** Genomic view of *P. macacae* genome showing pericentromeric
804 heterochromatin.
805 **Supplementary figure 6**. Genomic view of *P. carinii* genome showing pericentromeric
806 heterochromatin.
807 **Supplementary figure 7**. Genomic view of *P. murina* genome showing pericentromeric
808 heterochromatin.
809 **Supplementary figure 8.** Centromeres are flanked by heterochromatin and contain active genes.
810 **Supplementary figure 9.** Analysis of canonical histone H3 at *Pneumocystis* centromeres by ChIP-
811 qPCR.
812 **Supplementary figure 10.** 5-methylcytosine (5mC) DNA methylation in *Pneumocystis*
813 **Supplementary figure 11.** Genome synteny and centromere locations in *Pneumocystis species.*
814 **Supplementary table 1.** Coordinates, length, and GC content (in %) of Pneumocystis centromeres
815 identified by direct ChIP.
816 **Supplementary table 2.** Expression of genes located in homologous *P. murina* and *P. carinii*
817 centromeres.
818 **Supplementary table 3.** Persistent hydrophobic/nonpolar interactions between the tetramer
819 (H4:CENP-A)2 and dimers (H2A:H2B)
820 **Supplementary table 4.** Centromeres sequence similarity across *Pneumocystis* species.
821 **Supplementary table 4.** *Pneumocystis* ChIP-qPCR oligo nucleotides
822 **Key Resources table.**
823
824 **Supplementary dataset 1.** *P. jirovecii* and *P. murina* CENP-A full-length and N-terminal truncated
825 DNA sequences codon optimized for *S. pombe* codon bias.
826 **Supplementary dataset 2.** Dynamic animations of the molecular dynamics trajectories showing
827 both sides of each octameric complex in *P. carinii*, *P. murina* and *S. pombe*.
828

829 **Literature Cited**

830

831 Allshire, R. C. and K. Ekwall (2015). "Epigenetic Regulation of Chromatin States in
832 Schizosaccharomyces pombe." Cold Spring Harb Perspect Biol **7**(7): a018770.
833 Almeida, J. M., O. H. Cisse, A. Fonseca, M. Pagni and P. M. Hauser (2015). "Comparative genomics
834 suggests primary homothallism of Pneumocystis species." MBio **6**(1).
835 Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997).
836 "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic
837 Acids Res **25**(17): 3389-3402.
838 Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li and W. S.
839 Noble (2009). "MEME SUITE: tools for motif discovery and searching." Nucleic Acids Res **37**(Web
840 Server issue): W202-208.
841 Bao, W., K. K. Kojima and O. Kohany (2015). "Repbase Update, a database of repetitive elements in
842 eukaryotic genomes." Mob DNA **6**: 11.
843 Bray, N. L., H. Pimentel, P. Melsted and L. Pachter (2016). "Near-optimal probabilistic RNA-seq
844 quantification." Nat Biotechnol **34**(5): 525-527.
845 Cam, H. P., T. Sugiyama, E. S. Chen, X. Chen, P. C. FitzGerald and S. I. Grewal (2005).
846 "Comprehensive analysis of heterochromatin- and RNAi-mediated epigenetic control of the fission
847 yeast genome." Nat Genet **37**(8): 809-819.
848 Casola, C., D. Hucks and C. Feschotte (2008). "Convergent domestication of pogo-like transposases
849 into centromere-binding proteins in fission yeast and mammals." Mol Biol Evol **25**(1): 29-41.
850 Castillo, A. G., B. G. Mellone, J. F. Partridge, W. Richardson, G. L. Hamilton, R. C. Allshire and A.
851 L. Pidoux (2007). "Plasticity of fission yeast CENP-A chromatin driven by relative levels of histone
852 H3 and H4." PLoS Genet **3**(7): e121.
853 Chen, J. W., L. Yang, C. Santos, S. A. Hassan, P. L. Collins, U. J. Buchholz and C. Le Nouen (2021).
854 "Reversion mutations in phosphoprotein P of a codon-pair-deoptimized human respiratory syncytial
855 virus confer increased transcription, immunogenicity, and genetic stability without loss of
856 attenuation." PLoS Pathog **17**(12): e1010191.
857 Cisse, O. H., L. Ma, J. P. Dekker, P. P. Khil, J. H. Youn, J. M. Brenchley, R. Blair, B. Pahar, M. Chabe,
858 K. K. A. Van Rompay, R. Keesler, A. Sukura, V. Hirsch, G. Kutty, Y. Q. Liu, L. Peng, J. Chen, J. Song,
859 C. Weissenbacher-Lang, J. Xu, N. S. Upham, J. E. Stajich, C. A. Cuomo, M. T. Cushion and J. A.
860 Kovacs (2021). "Genomic insights into the host specific adaptation of the Pneumocystis genus."
861 Communications Biology **4**(1).
862 Cisse, O. H., M. Pagni and P. M. Hauser (2012). "De novo assembly of the Pneumocystis jirovecii
863 genome from a single bronchoalveolar lavage fluid specimen from a patient." MBio **4**(1): e00428-
864 00412.
865 Cisse, O. H., M. Pagni and P. M. Hauser (2014). "Comparative Genomics Suggests That the Human
866 Pathogenic Fungus Pneumocystis jirovecii Acquired Obligate Biotrophy through Gene Loss."
867 Genome Biology and Evolution **6**(8): 1938-1948.

868 Clarke, L. and J. Carbon (1980). "Isolation of a yeast centromere and construction of functional small
869 circular chromosomes." Nature **287**(5782): 504-509.
870 Cook, D. E., H. M. Kramer, D. E. Torres, M. F. Seidl and B. Thomma (2020). "A unique chromatin
871 profile defines adaptive genomic regions in a fungal plant pathogen." Elife **9**.
872 Corbett, K. D., C. K. Yip, L. S. Ee, T. Walz, A. Amon and S. C. Harrison (2010). "The Monopolin
873 Complex Crosslinks Kinetochore Components to Regulate Chromosome-Microtubule Attachments."
874 Cell **142**(4): 556-567.
875 Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson and T. R.
876 Gingeras (2013). "STAR: ultrafast universal RNA-seq aligner." Bioinformatics **29**(1): 15-21.
877 Emms, D. M. and S. Kelly (2019). "OrthoFinder: phylogenetic orthology inference for comparative
878 genomics." Genome Biol **20**(1): 238.
879 Ewels, P., M. Magnusson, S. Lundin and M. Kaller (2016). "MultiQC: summarize analysis results for
880 multiple tools and samples in a single report." Bioinformatics **32**(19): 3047-3048.
881 Folco, H. D., C. S. Campbell, K. M. May, C. A. Espinoza, K. Oegema, K. G. Hardwick, S. I. S. Grewal
882 and A. Desai (2015). "The CENP-A N-tail confers epigenetic stability to centromeres via the CENP-T
883 branch of the CCAN in fission yeast." Curr Biol **25**(3): 348-356.
884 Folco, H. D., A. McCue, V. Balachandran and S. I. S. Grewal (2019). "Cohesin Impedes
885 Heterochromatin Assembly in Fission Yeast Cells Lacking Pds5." Genetics **213**(1): 127-141.
886 Freitag, M. (2017). "Histone Methylation by SET Domain Proteins in Fungi." Annu Rev Microbiol
887 **71**: 413-439.
888 Frith, M. C. and L. Noe (2014). "Improved search heuristics find 20,000 new alignments between
889 human and mouse genomes." Nucleic Acids Res **42**(7): e59.
890 Funabiki, H., I. Hagan, S. Uzawa and M. Yanagida (1993). "Cell cycle-dependent specific positioning
891 and clustering of centromeres and telomeres in fission yeast." J Cell Biol **121**(5): 961-976.
892 Gordon, J. L., K. P. Byrne and K. H. Wolfe (2011). "Mechanisms of chromosome number evolution
893 in yeast." PLoS Genet **7**(7): e1002190.
894 Grabherr, M. G., P. Russell, M. Meyer, E. Mauceli, J. Alfoldi, F. Di Palma and K. Lindblad-Toh
895 (2010). "Genome-wide synteny through highly sensitive sequence alignment: Satsuma."
896 Bioinformatics **26**(9): 1145-1151.
897 Hayashi, M. T., T. S. Takahashi, T. Nakagawa, J. Nakayama and H. Masukata (2009). "The
898 heterochromatin protein Swi6/HP1 activates replication origins at the pericentromeric region and
899 silent mating-type locus." Nature Cell Biology **11**(3): 357-U284.
900 Hayashi, T., Y. Fujita, O. Iwasaki, Y. Adachi, K. Takahashi and M. Yanagida (2004). "Mis16 and
901 Mis18 are required for CENP-A loading and histone deacetylation at centromeres." Cell **118**(6): 715-
902 729.
903 Heinz, S., C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh and
904 C. K. Glass (2010). "Simple Combinations of Lineage-Determining Transcription Factors Prime cis-
905 Regulatory Elements Required for Macrophage and B Cell Identities." Molecular Cell **38**(4): 576-589.
906 Hernandez-Plaza, A., D. Szklarczyk, J. Botas, C. P. Cantalapiedra, J. Giner-Lamia, D. R. Mende, R.
907 Kirsch, T. Rattei, I. Letunic, L. J. Jensen, P. Bork, C. von Mering and J. Huerta-Cepas (2023).

908 "eggNOG 6.0: enabling comparative genomics across 12 535 organisms." Nucleic Acids Res **51**(D1):
909 D389-D394.
910 Hosseini, S., C. Meunier, D. Nguyen, J. Reimegard and H. Johannesson (2020). "Comparative
911 analysis of genome-wide DNA methylation in Neurospora." Epigenetics **15**(9): 972-987.
912 Janbon, G., K. L. Ormerod, D. Paulet, E. J. Byrnes, 3rd, V. Yadav, G. Chatterjee, N. Mullapudi, C. C.
913 Hon, R. B. Billmyre, F. Brunel, Y. S. Bahn, W. Chen, Y. Chen, E. W. Chow, J. Y. Coppee, A. Floyd-
914 Averette, C. Gaillardin, K. J. Gerik, J. Goldberg, S. Gonzalez-Hilarion, S. Gujja, J. L. Hamlin, Y. P.
915 Hsueh, G. Ianiri, S. Jones, C. D. Kodira, L. Kozubowski, W. Lam, M. Marra, L. D. Mesner, P. A.
916 Mieczkowski, F. Moyrand, K. Nielsen, C. Proux, T. Rossignol, J. E. Schein, S. Sun, C. Wollschlaeger,
917 I. A. Wood, Q. Zeng, C. Neuveglise, C. S. Newlon, J. R. Perfect, J. K. Lodge, A. Idnurm, J. E. Stajich,
918 J. W. Kronstad, K. Sanyal, J. Heitman, J. A. Fraser, C. A. Cuomo and F. S. Dietrich (2014). "Analysis
919 of the genome and transcriptome of Cryptococcus neoformans var. grubii reveals complex RNA
920 expression and microevolution leading to virulence attenuation." PLoS Genet **10**(4): e1004261.
921 Kato, H., J. Jiang, B. R. Zhou, M. Rozendaal, H. Feng, R. Ghirlando, T. S. Xiao, A. F. Straight and Y.
922 Bai (2013). "A conserved mechanism for centromeric nucleosome recognition by centromere protein
923 CENP-C." Science **340**(6136): 1110-1113.
924 Kent, W. J., A. S. Zweig, G. Barber, A. S. Hinrichs and D. Karolchik (2010). "BigWig and BigBed:
925 enabling browsing of large distributed datasets." Bioinformatics **26**(17): 2204-2207.
926 Kielbasa, S. M., R. Wan, K. Sato, P. Horton and M. C. Frith (2011). "Adaptive seeds tame genomic
927 sequence comparison." Genome Res **21**(3): 487-493.
928 Koressaar, T., M. Lepamets, L. Kaplinski, K. Raime, R. Andreson and M. Remm (2018).
929 "Primer3_masker: integrating masking of template sequence with primer design software."
930 Bioinformatics **34**(11): 1937-1938.
931 Kovacs, J. A., J. L. Halpern, J. C. Swan, J. Moss, J. E. Parrillo and H. Masur (1988). "Identification of
932 antigens and antibodies specific for Pneumocystis carinii." J Immunol **140**(6): 2023-2031.
933 Kozlov, A. M., D. Darriba, T. Flouri, B. Morel and A. Stamatakis (2019). "RAxML-NG: a fast, scalable
934 and user-friendly tool for maximum likelihood phylogenetic inference." Bioinformatics **35**(21): 4453-
935 4455.
936 Krzywinski, M., J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones and M. A. Marra
937 (2009). "Circos: an information aesthetic for comparative genomics." Genome Res **19**(9): 1639-1645.
938 Kursel, L. E. and H. S. Malik (2016). "Centromeres." Curr Biol **26**(12): R487-R490.
939 Lakshay Anand, C. M. R. L. (2020). "chromoMap: An R package for Interactive Visualization and
940 Annotation of Chromosomes." bioRxiv.
941 Lando, D., U. Endesfelder, H. Berger, L. Subramanian, P. D. Dunne, J. McColl, D. Klenerman, A. M.
942 Carr, M. Sauer, R. C. Allshire, M. Heilemann and E. D. Laue (2012). "Quantitative single-molecule
943 microscopy reveals that CENP-A(Cnp1) deposition occurs during G2 in fission yeast." Open Biology
944 **2**.
945 Landt, S. G., G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B. E. Bernstein, P.
946 Bickel, J. B. Brown, P. Cayting, Y. Chen, G. DeSalvo, C. Epstein, K. I. Fisher-Aylor, G. Euskirchen,
947 M. Gerstein, J. Gertz, A. J. Hartemink, M. M. Hoffman, V. R. Iyer, Y. L. Jung, S. Karmakar, M. Kellis,

948 P. V. Kharchenko, Q. Li, T. Liu, X. S. Liu, L. Ma, A. Milosavljevic, R. M. Myers, P. J. Park, M. J.
949 Pazin, M. D. Perry, D. Raha, T. E. Reddy, J. Rozowsky, N. Shoresh, A. Sidow, M. Slattery, J. A.
950 Stamatoyannopoulos, M. Y. Tolstorukov, K. P. White, S. Xi, P. J. Farnham, J. D. Lieb, B. J. Wold and
951 M. Snyder (2012). "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia."
952 Genome Res **22**(9): 1813-1831.
953 Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nat Methods
954 **9**(4): 357-359.
955 Leverson, J. D., H. K. Huang, S. L. Forsburg and T. Hunter (2002). "The Schizosaccharomyces pombe
956 aurora-related kinase Ark1 interacts with the inner centromere protein Pic1 and mediates
957 chromosome segregation and cytokinesis." Mol Biol Cell **13**(4): 1132-1143.
958 Li, H. (2013). "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM."
959 arXiv: 1303.3997.
960 Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin
961 and S. Genome Project Data Processing (2009). "The Sequence Alignment/Map format and
962 SAMtools." Bioinformatics **25**(16): 2078-2079.
963 Liu, X., I. McLeod, S. Anderson, J. R. Yates, 3rd and X. He (2005). "Molecular analysis of kinetochore
964 architecture in fission yeast." EMBO J **24**(16): 2919-2930.
965 Liu, Y., A. S. Davis, L. Ma, L. Bishop, O. H. Cisse, G. Kutty and J. A. Kovacs (2020). "MUC1 mediates
966 Pneumocystis murina binding to airway epithelial cells." Cell Microbiol **22**(6): e13182.
967 Lopez-Delisle, L., L. Rabbani, J. Wolff, V. Bhardwaj, R. Backofen, B. Gruning, F. Ramirez and T.
968 Manke (2021). "pyGenomeTracks: reproducible plots for multivariate genomic datasets."
969 Bioinformatics **37**(3): 422-423.
970 Ma, L., Z. Chen, W. Huang da, G. Kutty, M. Ishihara, H. Wang, A. Abouelleil, L. Bishop, E. Davey,
971 R. Deng, X. Deng, L. Fan, G. Fantoni, M. Fitzgerald, E. Gogineni, J. M. Goldberg, G. Handley, X.
972 Hu, C. Huber, X. Jiao, K. Jones, J. Z. Levin, Y. Liu, P. Macdonald, A. Melnikov, C. Raley, M. Sassi, B.
973 T. Sherman, X. Song, S. Sykes, B. Tran, L. Walsh, Y. Xia, J. Yang, S. Young, Q. Zeng, X. Zheng, R.
974 Stephens, C. Nusbaum, B. W. Birren, P. Azadi, R. A. Lempicki, C. A. Cuomo and J. A. Kovacs (2016).
975 "Genome analysis of three Pneumocystis species reveals adaptation mechanisms to life exclusively in
976 mammalian hosts." Nat Commun **7**: 10740.
977 McKinley, K. L. and I. M. Cheeseman (2016). "The molecular basis for centromere identity and
978 function." Nat Rev Mol Cell Biol **17**(1): 16-29.
979 Miura, A., S. Yonebayashi, K. Watanabe, T. Toyama, H. Shimada and T. Kakutani (2001).
980 "Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis."
981 Nature **411**(6834): 212-214.
982 Mizuguchi, T., G. Fudenberg, S. Mehta, J. M. Belton, N. Taneja, H. D. Folco, P. FitzGerald, J.
983 Dekker, L. Mirny, J. Barrowman and S. I. S. Grewal (2014). "Cohesin-dependent globules and
984 heterochromatin shape 3D genome architecture in S. pombe." Nature **516**(7531): 432-435.
985 Nagaki, K., Z. Cheng, S. Ouyang, P. B. Talbert, M. Kim, K. M. Jones, S. Henikoff, C. R. Buell and J.
986 Jiang (2004). "Sequencing of a rice centromere uncovers active genes." Nat Genet **36**(2): 138-145.

987 Navarro-Mendoza, M. I., C. Perez-Arques, S. Panchal, F. E. Nicolas, S. J. Mondo, P. Ganguly, J.
988 Pangilinan, I. V. Grigoriev, J. Heitman, K. Sanyal and V. Garre (2019). "Early Diverging Fungus
989 Mucor circinelloides Lacks Centromeric Histone CENP-A and Displays a Mosaic of Point and
990 Regional Centromeres." Curr Biol **29**(22): 3791-3802 e3796.
991 Pidoux, A. L. and R. C. Allshire (2004). "Kinetochore and heterochromatin domains of the fission
992 yeast centromere." Chromosome Res **12**(6): 521-534.
993 Postberg, J., S. Forcob, W. J. Chang and H. J. Lipps (2010). "The evolutionary history of histone H3
994 suggests a deep eukaryotic root of chromatin modifying mechanisms." BMC Evol Biol **10**: 259.
995 Quinlan, A. R. (2014). "BEDTools: The Swiss-Army Tool for Genome Feature Analysis." Curr Protoc
996 Bioinformatics **47**: 11 12 11-34.
997 Quinlan, A. R. and I. M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic
998 features." Bioinformatics **26**(6): 841-842.
999 Ramirez, F., D. P. Ryan, B. Gruning, V. Bhardwaj, F. Kilpert, A. S. Richter, S. Heyne, F. Dundar and
1000 T. Manke (2016). "deepTools2: a next generation web server for deep-sequencing data analysis."
1001 Nucleic Acids Res **44**(W1): W160-165.
1002 Robinson, J. T., H. Thorvaldsdottir, W. Winckler, M. Guttman, E. S. Lander, G. Getz and J. P.
1003 Mesirov (2011). "Integrative genomics viewer." Nat Biotechnol **29**(1): 24-26.
1004 Sabatinos, S. A. and S. L. Forsburg (2010). "Molecular genetics of Schizosaccharomyces pombe."
1005 Methods Enzymol **470**: 759-795.
1006 Sakuno, T. and Y. Watanabe (2009). "[Kinetochore geometry defined by cohesin within the
1007 centromere]." Tanpakushitsu Kakusan Koso **54**(14): 1842-1849.
1008 Sankaranarayanan, S. R., G. Ianiri, M. A. Coelho, M. H. Reza, B. C. Thimmappa, P. Ganguly, R. N.
1009 Vadnala, S. Sun, R. Siddharthan, C. Tellgren-Roth, T. L. J. Dawson, J. Heitman and K. Sanyal (2020).
1010 "Loss of centromere function drives karyotype evolution in closely related Malassezia species." Elife
1011 **9**.
1012 Sanyal, K., M. Baum and J. Carbon (2004). "Centromeric DNA sequences in the pathogenic yeast
1013 Candida albicans are all different and unique." Proc Natl Acad Sci U S A **101**(31): 11374-11379.
1014 Schotanus, K., J. L. Soyer, L. R. Connolly, J. Grandaubert, P. Happel, K. M. Smith, M. Freitag and E.
1015 H. Stukenbrock (2015). "Histone modifications rather than the novel regional centromeres of
1016 Zymoseptoria tritici distinguish core and accessory chromosomes." Epigenetics Chromatin **8**: 41.
1017 Seidl, M. F., H. M. Kramer, D. E. Cook, G. L. Fiorin, G. C. M. van den Berg, L. Faino and B.
1018 Thomma (2020). "Repetitive Elements Contribute to the Diversity and Evolution of Centromeres in
1019 the Fungal Genus Verticillium." mBio **11**(5).
1020 Selmecki, A., A. Forche and J. Berman (2006). "Aneuploidy and isochromosome formation in drug-
1021 resistant Candida albicans." Science **313**(5785): 367-370.
1022 Shen, X. X., J. L. Steenwyk, A. L. LaBella, D. A. Opulente, X. Zhou, J. Kominek, Y. Li, M.
1023 Groenewald, C. T. Hittinger and A. Rokas (2020). "Genome-scale phylogeny and contrasting modes
1024 of genome evolution in the fungal phylum Ascomycota." Sci Adv **6**(45).

1025 Shiroiwa, Y., T. Hayashi, Y. Fujita, A. Villar-Briones, N. Ikai, K. Takeda, M. Ebe and M. Yanagida
1026 (2011). "Mis17 Is a Regulatory Module of the Mis6-Mal2-Sim4 Centromere Complex That Is
1027 Required for the Recruitment of CenH3/CENP-A in Fission Yeast." Plos One **6**(3).
1028 Siepel, A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J.
1029 Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller
1030 and D. Haussler (2005). "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast
1031 genomes." Genome Res **15**(8): 1034-1050.
1032 Siepel, A. and D. Haussler (2004). "Phylogenetic estimation of context-dependent substitution rates
1033 by maximum likelihood." Mol Biol Evol **21**(3): 468-488.
1034 Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M.
1035 Remmert, J. Soding, J. D. Thompson and D. G. Higgins (2011). "Fast, scalable generation of high-
1036 quality protein multiple sequence alignments using Clustal Omega." Mol Syst Biol **7**: 539.
1037 Slater, G. S. and E. Birney (2005). "Automated generation of heuristics for biological sequence
1038 comparison." BMC Bioinformatics **6**: 31.
1039 Slaven, B. E., J. Meller, A. Porollo, T. Sesterhenn, A. G. Smulian and M. T. Cushion (2006). "Draft
1040 assembly and annotation of the Pneumocystis carinii genome." J Eukaryot Microbiol **53 Suppl 1**:
1041 S89-91.
1042 Smit AFA., H. R., Green P. (2013-2015). "RepeatMasker Open-4.0.".
1043 Smith, K. M., J. M. Galazka, P. A. Phatale, L. R. Connolly and M. Freitag (2012). "Centromeres of
1044 filamentous fungi." Chromosome Res **20**(5): 635-656.
1045 Sridhar, S. and T. Fukagawa (2022). "Kinetochore Architecture Employs Diverse Linker Strategies
1046 Across Evolution." Front Cell Dev Biol **10**: 862637.
1047 Takayama, Y., H. Sato, S. Saitoh, Y. Ogiyama, F. Masuda and K. Takahashi (2008). "Biphasic
1048 incorporation of centromeric histone CENP-A in fission yeast." Mol Biol Cell **19**(2): 682-690.
1049 Thakur, J., P. B. Talbert and S. Henikoff (2015). "Inner Kinetochore Protein Interactions with
1050 Regional Centromeres of Fission Yeast." Genetics **201**(2): 543-561.
1051 Tran, P. T., A. Paoletti and F. Chang (2004). "Imaging green fluorescent protein fusions in living
1052 fission yeast cells." Methods **33**(3): 220-225.
1053 Vestereng, V. H., L. R. Bishop, B. Hernandez, G. Kutty, H. H. Larsen and J. A. Kovacs (2004).
1054 "Quantitative real-time polymerase chain-reaction assay allows characterization of Pneumocystis
1055 infection in immunocompetent mice." Journal of Infectious Diseases **189**(8): 1540-1544.
1056 Williams, J. S., T. Hayashi, M. Yanagida and P. Russell (2009). "Fission yeast Scm3 mediates stable
1057 assembly of Cnp1/CENP-A into centromeric chromatin." Mol Cell **33**(3): 287-298.
1058 Xi, Y. and W. Li (2009). "BSMAP: whole genome bisulfite sequence MAPping program." BMC
1059 Bioinformatics **10**: 232.
1060 Yadav, V., S. Sun, R. B. Billmyre, B. C. Thimmappa, T. Shea, R. Lintner, G. Bakkeren, C. A. Cuomo,
1061 J. Heitman and K. Sanyal (2018). "RNAi is a critical determinant of centromere evolution in closely
1062 related fungi." Proc Natl Acad Sci U S A **115**(12): 3108-3113.
1063 Yasuhara, J. C. and B. T. Wakimoto (2008). "Molecular landscape of modified histones in Drosophila
1064 heterochromatic genes and euchromatin-heterochromatin transition zones." PLoS Genet **4**(1): e16.

1065 Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M.
1066 Myers, M. Brown, W. Li and X. S. Liu (2008). "Model-based analysis of ChIP-Seq (MACS)." Genome
1067 Biol **9**(9): R137.
1068 Zhao, H., A. Nguyen, D. Wu, Y. Li, S. A. Hassan, J. Chen, H. Shroff, G. Piszczek and P. Schuck
1069 (2022). "Plasticity in structure and assembly of SARS-CoV-2 nucleocapsid protein." PNAS Nexus
1070 **1**(2): pgac049.
1071 Zhao, H., A. Nguyen, D. Wu, Y. Li, S. A. Hassan, J. Chen, H. Shroff, G. Piszczek and P. Schuck
1072 (2022). "Plasticity in structure and assembly of SARS-CoV-2 nucleocapsid protein." bioRxiv.
1073

Figure 1. *Pneumocystis* CENP-A variants localize to the nucleus, and support viability and centromere loading in a heterologous system.

(**a**) Protein sequence alignment of full-length CENP-A^{cnp-1} homologs in *Schizosaccharomyces pombe* and seven *Pneumocystis* species. Each *Pneumocystis* species only infects a single mammalian host: *P. jirovecii* (infecting humans), *P. macacae* (macaques), *P. oryctolagi* (rabbits), *P. canis* (dogs), *P. murina* (mice), *P. carinii* (rats), and *P. wakefieldiae* (rats). The ~25 amino acid insert in *Pneumocystis* CENP-A N-terminal regions is unique to this genus.

(**b**) Immunofluorescence microscopy showing that anti-*Pneumocystis murina* CENP-A (*pn*CENP-A) reacts with *P. murina* organisms. *P. murina*-infected CD40 ligand knockout lung tissue sections were dual labeled using an anti CENP-A antibody(magenta) and 4′,6-diamidino-2-phenylindole (DAPI, light blue). The image is a maximum projection of a 5 µm thick z-stack. The images were minimally adjusted to enhance the contrast. The green box is magnified in the inset showing *pn*CENP-A inside *Pneumocystis* organisms (small dots). *pn*CENP-A signals overlap or are close in proximity with DAPI (represented by a white color produced by the overlap between magenta and light blue colors).

(**c**) Left, rescue experiment of fission yeast *cnp1-76* thermosensitive cells using the GFP-tagged CENP-A transgenes. Right, protein sequence alignment of *Pneumocystis* CENP-A N-terminal regions (*P. jirovecii* and *P. murina*) with conserved amino acids highlighted (grey areas). Plasmids containing full length and N-terminal region truncated (**) versions of *Pneumocystis* CENP-A were constructed and integrated as single copy. Growth was assayed at indicated temperatures using 10-fold serial dilutions plated on rich YEA medium. The transgene encoding *S. pombe* CENP-A serves as a non-temperature-sensitive control.

(**d**) Representative images of *cen2-tetO-TdTomato* strains expressing the indicated GFP-tagged CENP-A transgenes. On the bottom, GFP and tdTomato fluorescence as well as brightfield images are merged.

(**e**) Integrated fluorescence intensity of GFP foci was measured and plotted for the strains used in panel e. Mean (bars) and SD (error bars) are shown. n>176. AU, arbitrary units.

(**f**) Anti-GFP ChIP-qPCR analysis of indicated loci for strains used in panel e. %IP represents the percentage of input that was immunoprecipitated. Error bars denote the SD (n = 3). Abbreviations: cc1 &3, *S. pombe* centromere central core 1 & 3; cnt2:ura4^{+}, ura4^{+} insertion at central core 2; *dg*, a class of heterochromatic outer repeats within pericentromeric regions (control); *fbp1*, euchromatic locus (control).

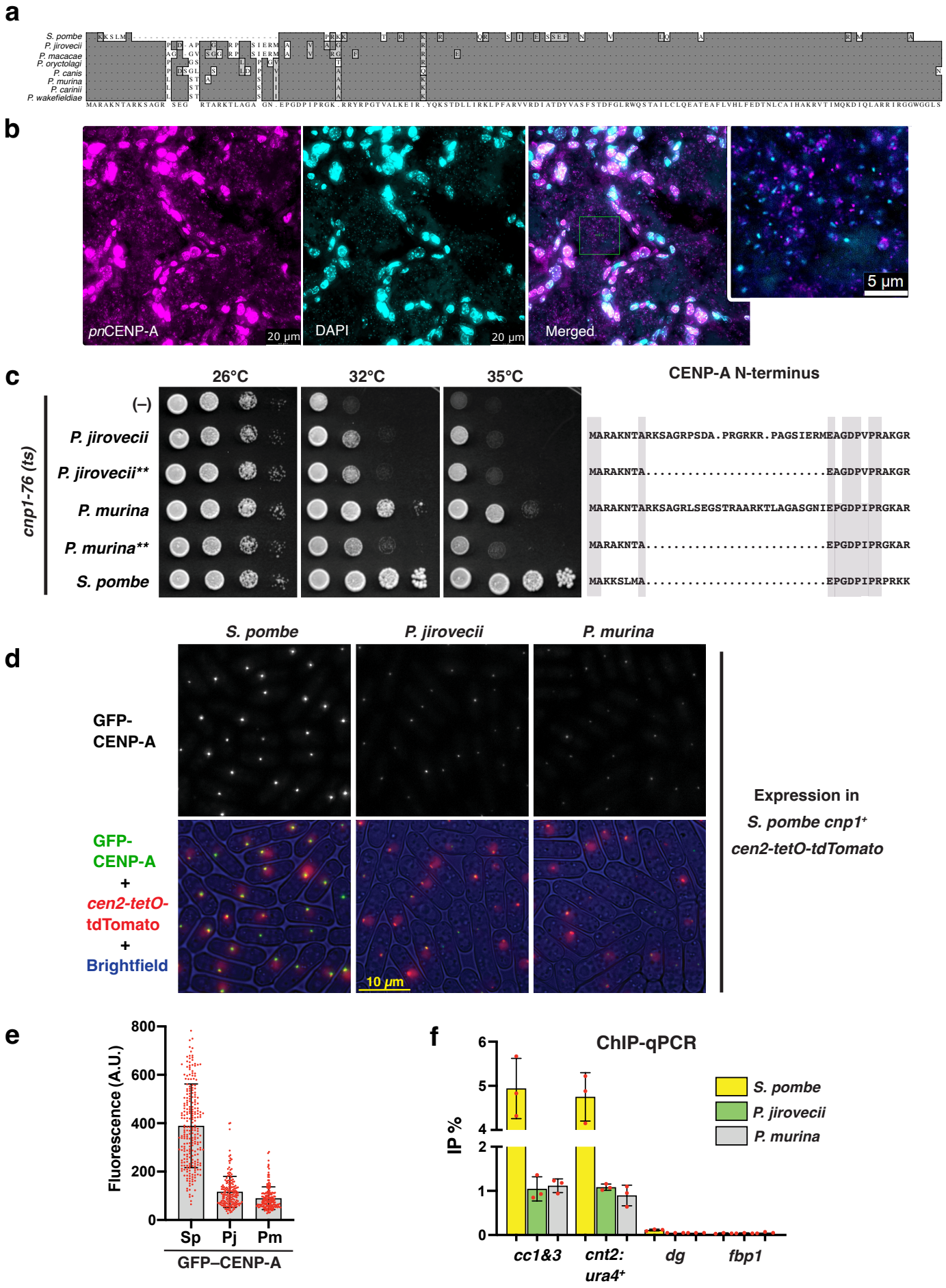

a
S. pombe
P. jirovecii
P. macacae
P. oryctolagi
P. canis
P. murina
P. carinii
P. wakefieldiae
b
pnCENP-A
20 µm
DAPI
20 µm
Merged
5 µm
c
26°C
32°C
35°C
CENP-A N-terminus
cnp1-76 (ts)
(–)
P. jirovecii
P. jirovecii**
P. murina
P. murina**
S. pombe
MARAKNTARKSAGRPSDA.PRGRKR.PAGSIERMEAGDPVPRAKGR
MARAKNTA.....EAGDPVPRAKGR
MARAKNTARKSAGRLSEGSTRAARKTLAGASGNIEPGDPIPRGKAR
MARAKNTA.....EPGDPIPRGKAR
MAKKSLMA.....EPGDPIPRPRKK
d
S. pombe
P. jirovecii
P. murina
GFP-CENP-A
GFP-CENP-A + cen2-tetO-tdTomato + Brightfield
10 µm
Expression in S. pombe cnp1+ cen2-tetO-tdTomato
e
Fluorescence (A.U.)
800
600
400
200
0
Sp
Pj
Pm
GFP–CENP-A
f
ChIP-qPCR
IP %
6
5
4
2
1
0
S. pombe
P. jirovecii
P. murina
cc1&3
cnt2: ura4+
dg
fbp1

Figure 2. *Pneumocystis* CENP-A binds to single genomic foci in replicating cells

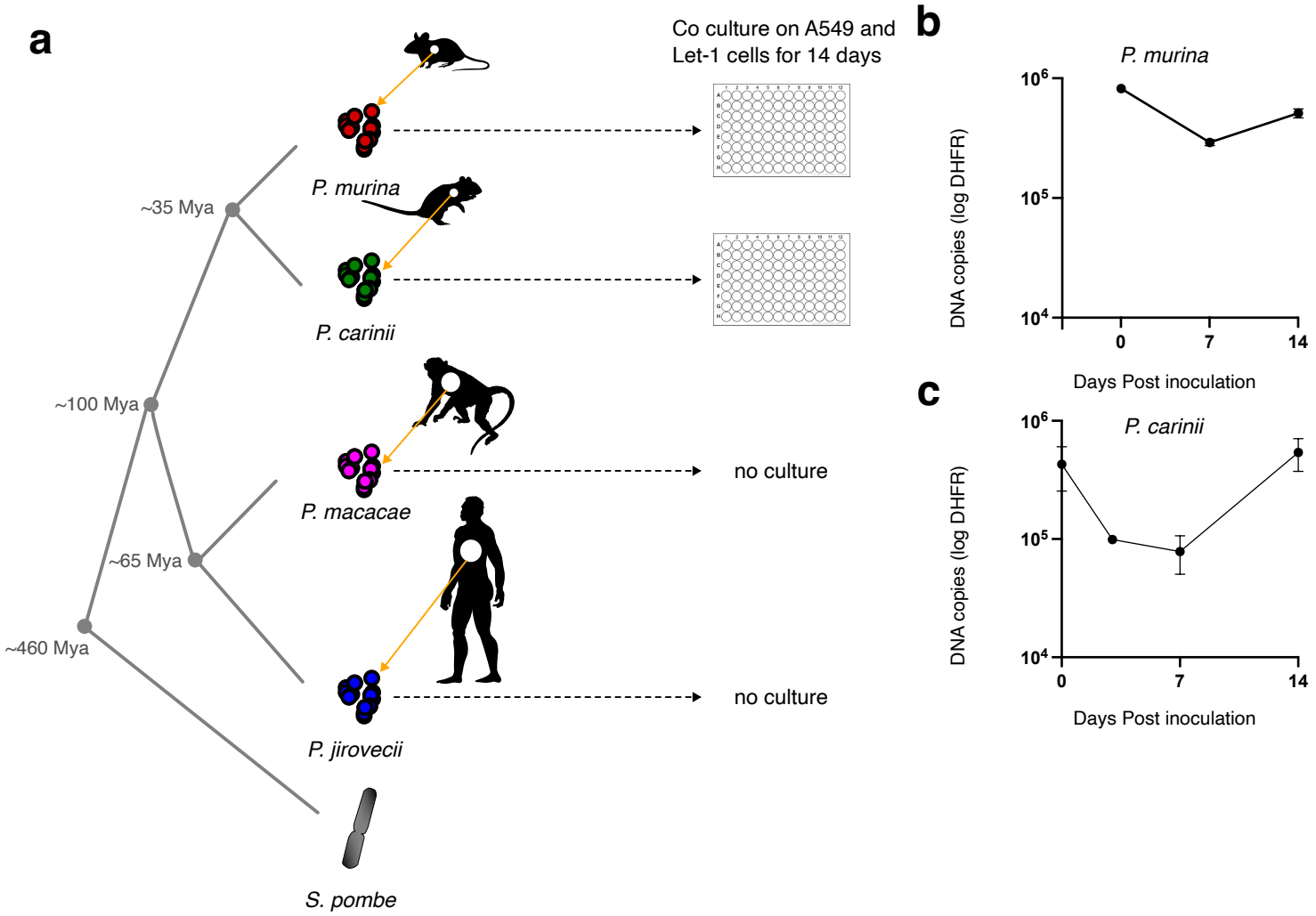
a) Study workflow and sample preparation. Four *Pneumocystis* species infecting mice, rats, macaques, and humans respectively were analyzed. *P. murina* and *P. carinii* were obtained from CD40 ligand knock-out female mice and immunosuppressed Sprague-Dawley male rats, respectively, and cultured on a co-culture of human lung adenocarcinoma cells (A549) and immortalized murine lung epithelial type 1 cells (Let-1) for 14 days. *P. macacae* organisms were obtained from a Simian Immunodeficiency Virus infected macaque (P2C), and *P. jirovecii* organisms were obtained from a single bronchoalveolar lavage from a non-infected HIV patient and three autopsy samples from patients with HIV infection. Species phylogeny and estimated speciation timing are presented. Animal icons were obtained from http://phylopic.org under creative commons licenses https://creativecommons.org/licenses/by/3.0/: mouse (Anthony Caravaggi; license CC BY-NC-SA 3.0); and rat (by Rebecca Groom; license CC BY-NC-SA 3.0).

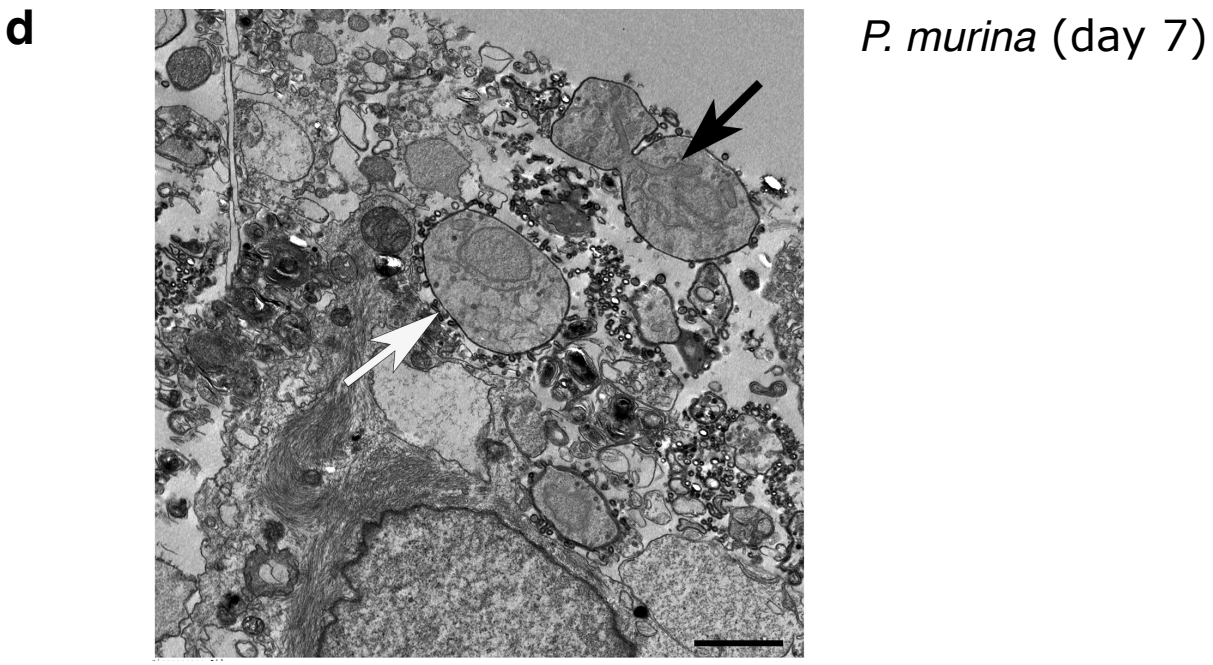
b) *P. murina* population growth in duplicate wells were measured by quantitative PCR targeting a single copy dihydrofolate reductase (*dhfr*) gene over 14 days. Error bars represent the standard deviation (n=2).

c) *P. carinii* growth measurement by qPCR targeting *dhfr* gene. Error bars represent the standard deviation (n=2).

d) Electron micrograph showing a possibly dividing *P. murina* trophic form seven days post culture (black arrow). The field also include a non-dividing trophozoite (white arrow). 2 μm scale bar.

e) *pn*CENP-A ChIP-seq study in *P. carinii* organisms from infected lungs and cultures at days 0, 7 and 14. CENP-A remains strongly enriched at a single genomic location up to 7 days. Only one representative chromosome is presented. The remaining 16 chromosomes have similar ChIP-seq profiles and presented in supplementary figure 4.


a
Co culture on A549 and
Let-1 cells for 14 days
P. murina
P. carinii
P. macacae
P. jirovecii
S. pombe
~35 Mya
~100 Mya
~65 Mya
~460 Mya
no culture
no culture
b
P. murina
DNA copies (log DHFR)
Days Post inoculation
c
P. carinii
DNA copies (log DHFR)
Days Post inoculation


d
P. murina (day 7)

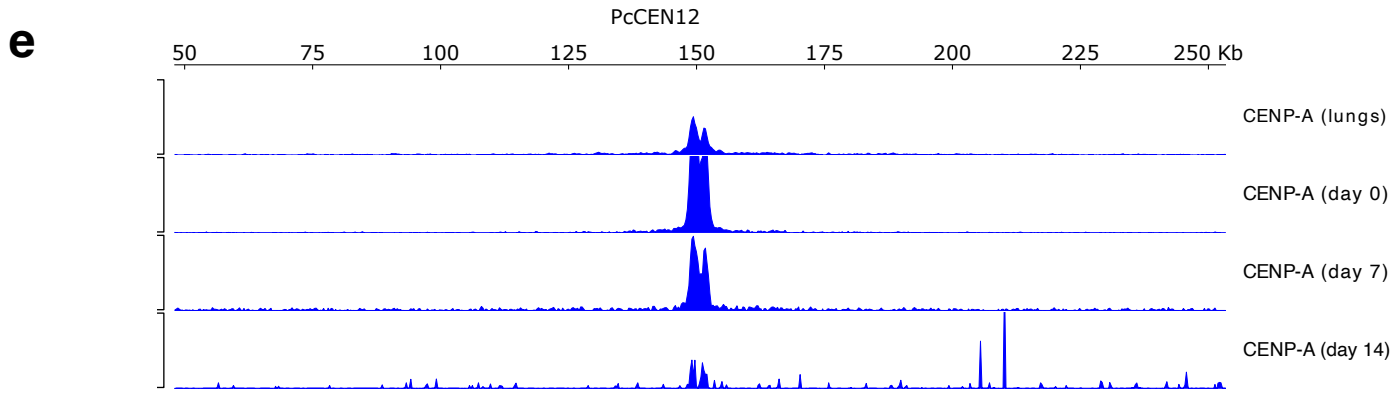

e
PcCEN12
50
75
100
125
150
175
200
225
250 Kb
CENP-A (lungs)
CENP-A (day 0)
CENP-A (day 7)
CENP-A (day 14)

Figure 3. *Pneumocystis* displays seventeen centromeres.

a) Scaled ideogram of 17 chromosomal level scaffolds (grey) of *P. murina* genome showing CENP-A binding regions (constricted areas) and telomeric repeat motif (red).

b) CENP-A and CENP-C binding regions delineate putative centromeres (CENs) in *P. murina* genome. Color coded peaks represent enrichment of immunoprecipitated DNA (IP DNA) relative to controls (Input DNA). The 1-kb flanking sequences from the center of the enrichment peak are presented. Each scaffold displays two peaks labelled M (Major) and m (minor) according to the enrichment level.

c) Scaled ideogram of 17 chromosomal level scaffolds (grey) of *P. carinii* genome showing CENP-A binding regions (constricted areas) and telomeric repeat motif (red).

d) CENP-A binding regions delineate putative centromeres (CENs) in *P. carinii* genome. Note that no peak was observed with anti-CENP-C antibody.

e) A DNA dot plot of CENP-A binding regions in *P. murina* showing regional self-similarity. The main diagonal represents the sequence alignment with itself. Lines off the main diagonal which are repetitive patterns within the sequences are not observed. The plot shows that each CEN is unique and repeat-free.

f) DNA dot plot of CENP-A binding regions in *P. carinii* showing that each CEN is unique and repeat-free.

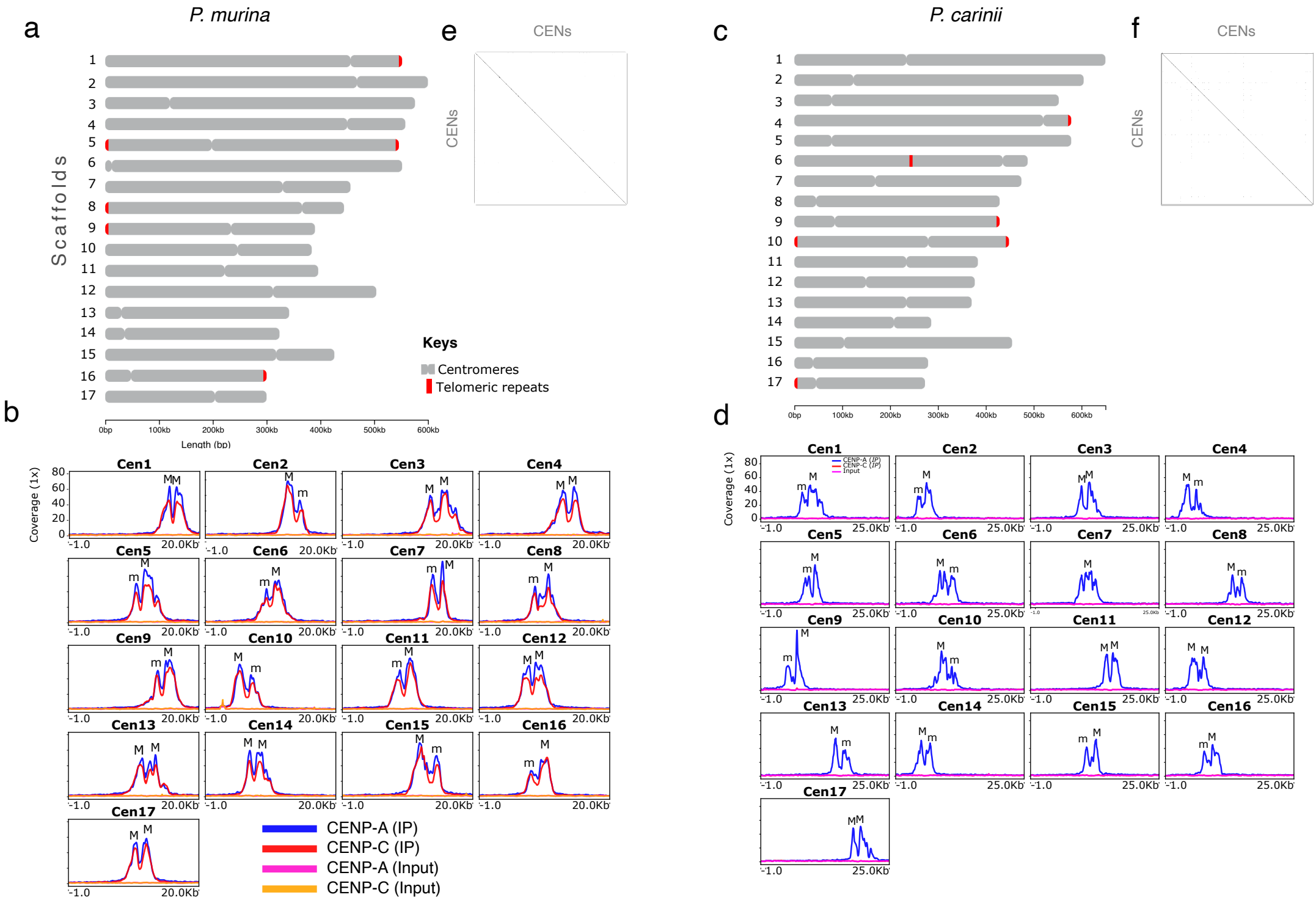

a
P. murina
Scaffolds
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
0bp
100kb
200kb
300kb
400kb
500kb
600kb
Length (bp)
Keys
Centromeres
Telomeric repeats
e
CENs
CENs
b
Coverage (1x)
Cen1
Cen2
Cen3
Cen4
Cen5
Cen6
Cen7
Cen8
Cen9
Cen10
Cen11
Cen12
Cen13
Cen14
Cen15
Cen16
Cen17
-1.0
20.0Kb
CENP-A (IP)
CENP-C (IP)
CENP-A (Input)
CENP-C (Input)
c
P. carinii
f
CENs
CENs
d
Coverage (1x)
25.0Kb

Figure 4. Centromeres are flanked by heterochromatin and contain active genes.

All three scaffolds are homologous.

a) Genomic view of chromosome 2 of *P. murina* genome subsequently showing annotated genes (directed grey boxes), DNA repeats (not present), percent GC content (blue), ChIP-seq read coverage distribution (BPM normalized over bins of 50 bp; input subtracted) of CENP-A, CENP-C, histones H3 and H4 ratio, heterochromatin-associated modifications (H3K9me2 and H3K9me3), euchromatin (H3K4me2) and gene expression (RNA-seq) in relation with centromeres. Genome views of remaining 16 chromosomes are presented in Supplementary dataset figure 7.

b) Genomic view of chromosome 2 of *P. carinii* genome with the same features presented as *P. murina*. A duplicated copy of *copia*-retrotransposon is presented (red arrows), which is present in syntenic regions in other *Pneumocystis* genomes. Genome views of all remaining 16 chromosomes are presented in supplementary figure 6.

c) Genomic view of chromosome 14 of *P. macacae* genome showing the distribution of gene density, repeats (DNA transposons), homology region with GC content, mis12 ChIP-seq read coverage, chromatin-associated modifications (H3K9me2 and H3K9me3 and H3K4me2) and RNA-seq in relation with centromeres. The white box indicates a DNA sequence homologous to *P. carinii* centromere 2. A single copy of the DNA transposon EnSPM (CACTA family) and the class 2DNA transposon Helitron are present upstream the centromere (forward red arrows). These elements are not present in syntenic regions in other *Pneumocystis* genomes. Genome views of all remaining 15 chromosomes are presented in Supplementary figure 5.

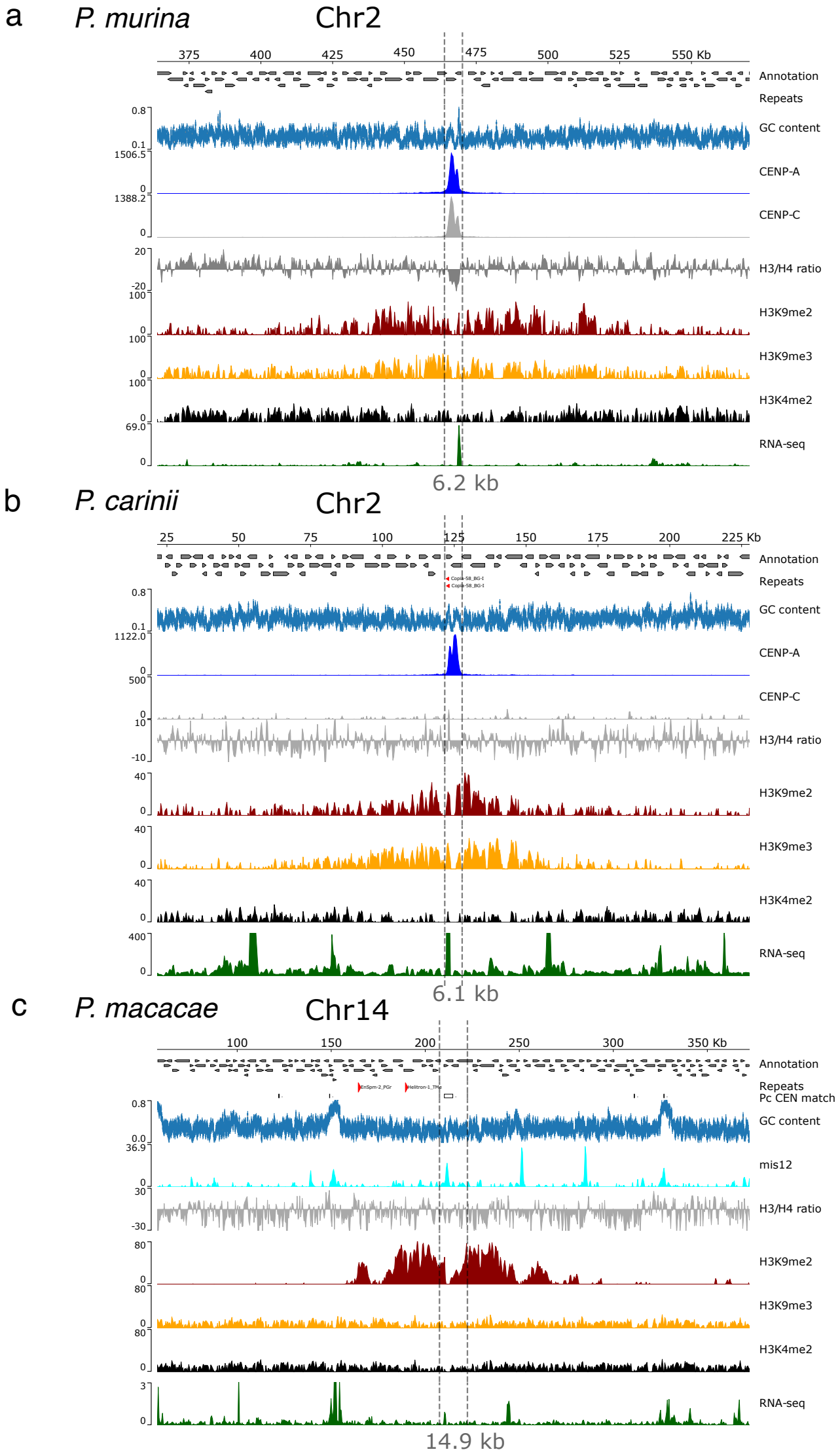
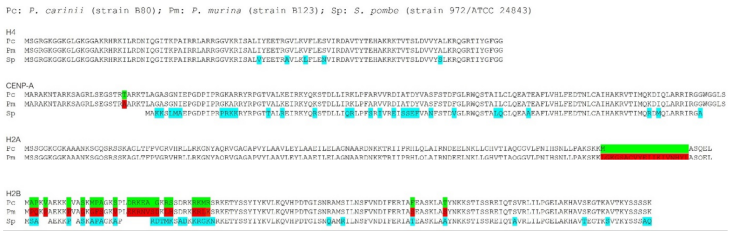

a
P. murina
Chr2
375
400
425
450
475
500
525
550 Kb
Annotation
Repeats
GC content
CENP-A
CENP-C
H3/H4 ratio
H3K9me2
H3K9me3
H3K4me2
RNA-seq
6.2 kb
b
P. carinii
Chr2
25
50
75
100
125
150
175
200
225 Kb
Annotation
Repeats
GC content
CENP-A
CENP-C
H3/H4 ratio
H3K9me2
H3K9me3
H3K4me2
RNA-seq
6.1 kb
c
P. macacae
Chr14
100
150
200
250
300
350 Kb
Annotation
Repeats
Pc CEN match
GC content
mis12
H3/H4 ratio
H3K9me2
H3K9me3
H3K4me2
RNA-seq
14.9 kb

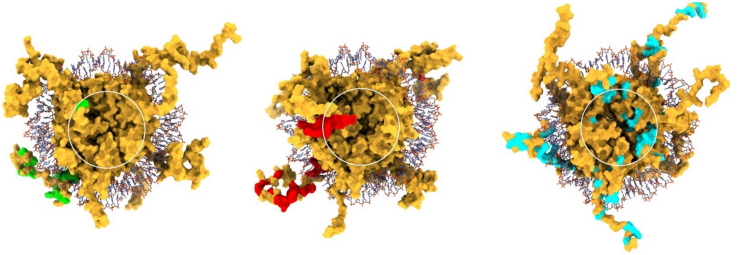
Figure 5. Modeling of Molecular dynamics of CENP-A nucleosomes in *P.* carinii, *P. murina* and *S. pombe.*

a) Upper panel: Sequence alignment of the centromeric histones used in modeling and simulations. Differences between *P. carinii* and *P. murina* colored green and red; differences between *S. pombe* and *pneumocystis* colored blue. Lower panel: surface representation of the centromeric octamers of *P. carinii* (left), *P. murina* (middle), and *S. pombe* (right), showing the spatial location of the mutations (snapshots from the simulations); DNA (not simulated) shown only as a reference based on the crystal structure of the human centromere nucleosome, PDB: 3AN2). The white circles show the central crevices, the behaviors of which are expected to dictate the assembly of the inner kinetochore in each species.

b) Aggregation of snapshots along the dynamic trajectory illustrating the extent of species-specific conformational variations of the central crevices (see Animation in Supplementary dataset 2); histones color code is presented at the bottom). The C-terminal tip of CENP-A is located at the center of the core (white ribbon; residues 146-151 in *Pneumocystis* and 120 in *S. pombe*; see alignment in panel a). The observed differences between *P. carinii* and *P. murina* are due to structural and dynamic changes induced by the different (H2A:H2B) dimers (cf. alignment in panel a); the differences between these species and *S. pombe* are likely the result of more extensive sequence variations in the four histones, including several solvent-exposed mutations at the core center (cf. panel a). Early-stage assembly of the inner kinetochore is thus expected to be strongly determined by features of this crevice, including its size, dynamics, and other physicochemical features.

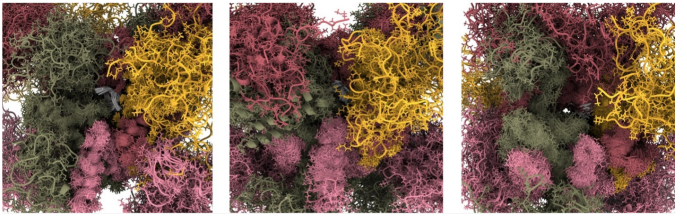
a





b

P. carinii *P. murina* *S. pombe*



CENP-A H4 H2A H2B

Figure 6. Centromere locations, not sequences, are evolutionary constrained by negative selection.

a) Genome organization and synteny in *Pneumocystis*. Circos plots depicting pairwise *P. carinii*, *P. murina* and *P. macacae* genome synteny. Rodent infecting Pneumocystis (*P. carinii* and *P. murina*) and macaques infecting *Pneumocystis* (*P. macacae*) have 17 and 16 chromosomal level scaffolds, respectively. Colored connectors indicate regions of synteny between species. Centromeres that overlap with recent chromosomal breakpoints are indicated (*). The square highlights *P. carinii* centromere 4 displayed in the panel b.

b) Genome view of centromere 4 in the *P. carinii* genome. Genes are represented by directed boxes (grey for protein coding genes and cyan for polymorphic major surface glycoprotein genes), *pn*CENP-A binding region (centromere) and sequence conservation scores which were calculated from whole genome alignments of *P. carinii*. *P. murina* and *P. wakefieldiae* (PhasCons). The phastCons scores represent probabilities of negative selection and range between 0 (no conservation) and 1 (total conservation).

c) Boxplot of conservation scores per genomic context summarized over all *P. carinii* centromeres (CEN), 30-kb regions flanking the centromeres (Cenflk), major surface glycoproteins (Msg) and random genomic background (Bckg). Statistical differences for the indicated comparisons were obtained using one-sided nonparametric Mann-Whitney test; *P*-values < .0001, ****; $P < .001$, ***; $P < .01$, **; $P < .05$, *; ns, not significant.

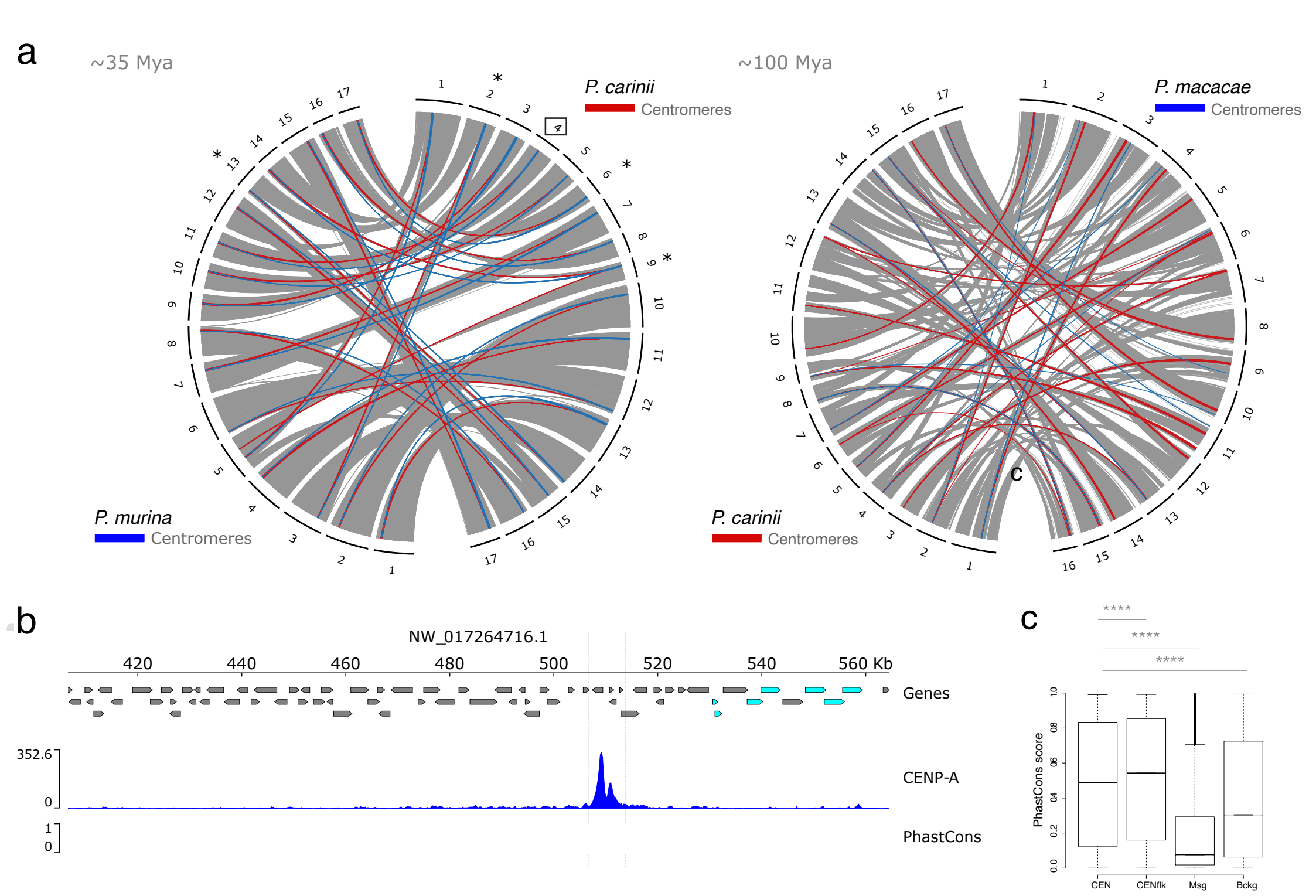
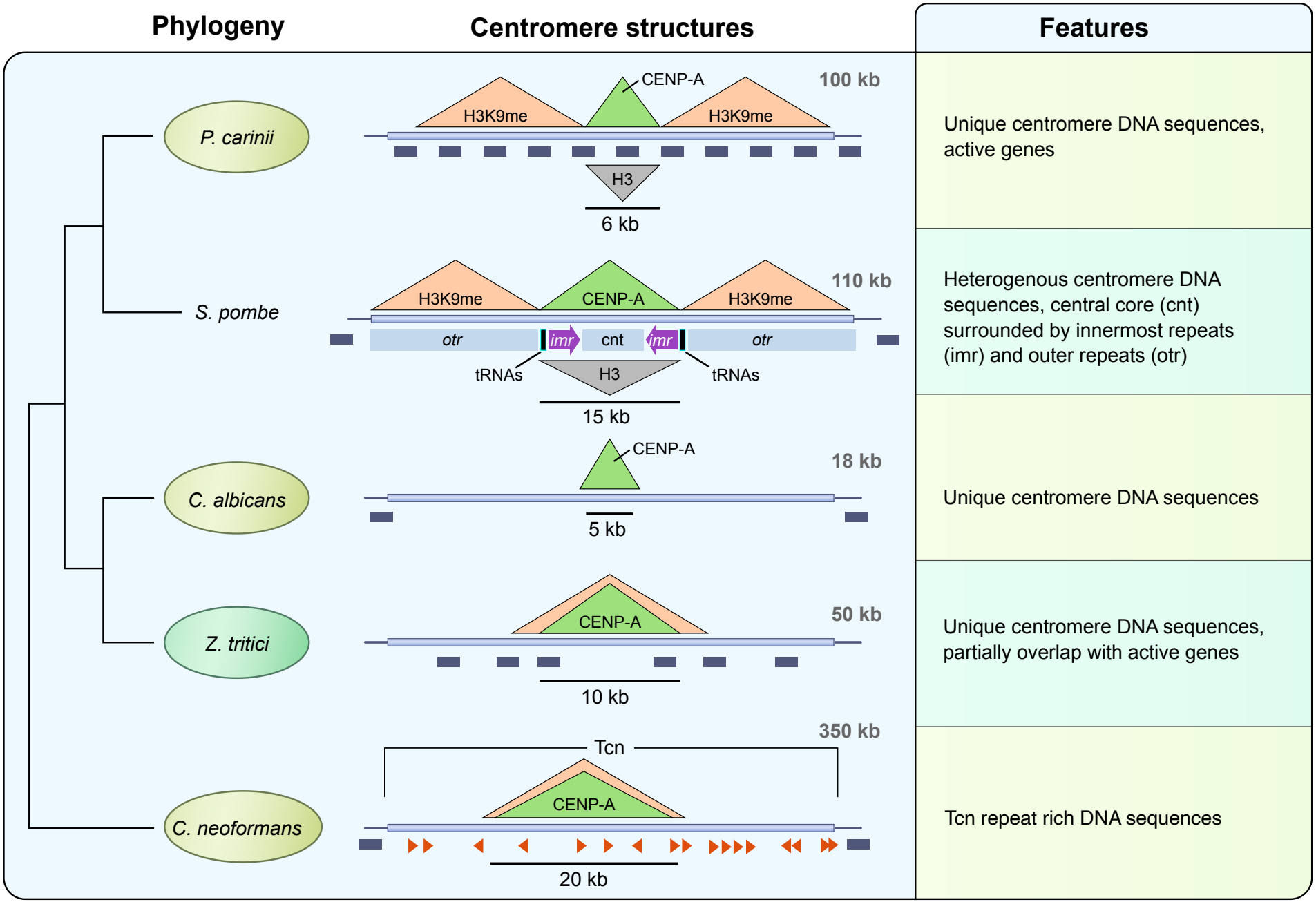

a
~35 Mya
P. carinii
Centromeres
P. murina
Centromeres
~100 Mya
P. macacae
Centromeres
P. carinii
Centromeres
b
NW_017264716.1
420
440
460
480
500
520
540
560 Kb
Genes
352.6
0
CENP-A
1
0
PhastCons
c
****
****
****
PhastCons score
CEN
CENflk
Msg
Bckg

Figure 7. Model of regional centromere structures in *Pneumocystis* and other representative fungi.

On the left is presented the phylogeny of selected fungi inferred from maximum likelihood phylogenetic analysis of shared core protein orthologs. The overall centromere structures for *Pneumocystis carinii* (details supporting our model are provided in the text), *Schizosaccharomyces pombe* (Pidoux and Allshire 2004), *Candida albicans* (Sanyal, Baum et al. 2004), *Zymoseptoria tritici* (Schotanus, Soyer et al. 2015) and *Cryptococcus neoformans* (Janbon, Ormerod et al. 2014) are used as representative to showcase the diversity of regional centromeres in fungi. Animal pathogens are highlighted in pale olive and the plant fungal pathogen *Zymoseptoria tritici* is in pale green. For sake of brevity, only *Pneumocystis carinii* is presented for brevity, although there are differences with other *Pneumocystis* species. In the middle are presented DNA structures of centromeres. *P. carinii* has 17 centromeres (one for each of its 17 chromosomes) that share the same overall architecture. Centromeres are delineated by a localized enrichment of the centromeric histone CENP-A, which overlap with a reduction of the canonical histone H3 (inverted grey triangle). Centromeres are flanked by heterochromatin H3K9me (stands here stands for both H3K9me2 and H3K9me3). All 17 *Pneumocystis carinii* centromeres span active genes (dark grey boxes). Each centromere sequence is different and lacks shared DNA sequence motif. *S. pombe* has three centromeres that share the same overall structure in which a central core (*cnt*) domain is surrounded by innermost repeats (*imr*) and outer repeats (*otr*). The *imr* repeats incorporate clusters of transfer RNAs (tRNAs) that play a role in restricting CENP-A spread. *S. pombe* centromeres are flanked by heterochromatin (H3K9me). Genes are found 0.75-1.5 kb beyond the limits of the centromeres. *C. albicans* has 8 unique and different centromeres that are gene free and lack shared sequence motifs. *Z. tritici* has 21 centromeres ranging from 6 to 14 kb in size, that partially overlap with genes. *C. neoformans* has 14 centromeres that are gene free and enriched with Tcn transposons.


Phylogeny
Centromere structures
Features
P. carinii
S. pombe
C. albicans
Z. tritici
C. neoformans
CENP-A
H3K9me
H3K9me
100 kb
H3
6 kb
Unique centromere DNA sequences, active genes
H3K9me
CENP-A
H3K9me
110 kb
otr
imr
cnt
imr
otr
tRNAs
H3
tRNAs
15 kb
Heterogenous centromere DNA sequences, central core (cnt) surrounded by innermost repeats (imr) and outer repeats (otr)
CENP-A
18 kb
5 kb
Unique centromere DNA sequences
CENP-A
50 kb
10 kb
Unique centromere DNA sequences, partially overlap with active genes
350 kb
Tcn
CENP-A
20 kb
Tcn repeat rich DNA sequences