Loss of inner kinetochore genes is associated with the transition to an unconventional point centromere in budding yeast

Nagarjun Vijay Corresp. 1

1 Computational Evolutionary Genomics Lab, Department of Biological Sciences, Indian Institute of Science Education and Research Bhopal, Bhopal, Madhya Pradesh, India

Corresponding Author: Nagarjun Vijay
Email address: nagarjun@iiserb.ac.in

Background: The genomic sequences of centromeres, as well as the set of proteins that recognize and interact with centromeres, are known to quickly diverge between lineages potentially contributing to post-zygotic reproductive isolation. However, the actual sequence of events and processes involved in the divergence of the kinetochore machinery is not known. The patterns of gene loss that occur during evolution concomitant with phenotypic changes have been used to understand the timing and order of molecular changes.

Methods: I screened the high-quality genomes of twenty budding yeast species for the presence of well-studied kinetochore genes. Based on the conserved gene order and complete genome assemblies, I identified gene loss events. Subsequently, I searched the intergenic regions to identify any un-annotated genes or gene remnants to obtain additional evidence of gene loss.

Results: My analysis identified the loss of four genes (NKP1, NKP2, CENPL/IML3, and CENPN/CHL4) of the inner kinetochore constitutive centromere-associated network (CCAN/also known as CTF19 complex in yeast) in both the Naumovozyma species for which genome assemblies are available. Surprisingly, this collective loss of four genes of the CCAN/CTF19 complex coincides with the emergence of unconventional centromeres in N.*castellii* and N.*dairenensis*. My study suggests a tentative link between the emergence of unconventional point centromeres and the turnover of kinetochore genes in budding yeast.

2 Loss of inner kinetochore genes is associated with the transition to an unconventional point

3 centromere in budding yeast

5 Nagarjun Vijay[1]

6 [1]Computational Evolutionary Genomics Lab, Department of Biological Sciences, Indian Institute

7 of Science Education and Research Bhopal, Bhopal, Madhya Pradesh, India

8 *Correspondence: nagarjun@iiserb.ac.in

10 **Keywords: point centromere, unconventional centromeres, kinetochore, interaction network, speciation**

12 **Abstract**

14 **Background**: The genomic sequences of centromeres, as well as the set of proteins that

15 recognize and interact with centromeres, are known to quickly diverge between lineages

16 potentially contributing to post-zygotic reproductive isolation. However, the actual sequence of

17 events and processes involved in the divergence of the kinetochore machinery is not known. The

18 patterns of gene loss that occur during evolution concomitant with phenotypic changes have been

19 used to understand the timing and order of molecular changes.

20 **Methods**: I screened the high-quality genomes of twenty budding yeast species for the presence

21 of well-studied kinetochore genes. Based on the conserved gene order and complete genome

22 assemblies, I identified gene loss events. Subsequently, I searched the intergenic regions to

23 identify any un-annotated genes or gene remnants to obtain additional evidence of gene loss.

24 **Results**: My analysis identified the loss of four genes (NKP1, NKP2, CENPL/IML3, and

25 CENPN/CHL4) of the inner kinetochore constitutive centromere-associated network

26 (CCAN/also known as CTF19 complex in yeast) in both the Naumovozyma species for which

27 genome assemblies are available. Surprisingly, this collective loss of four genes of the

28 CCAN/CTF19 complex coincides with the emergence of unconventional centromeres in

29 N.*castellii* and N.*dairenensis*. My study suggests a tentative link between the emergence of

30 unconventional point centromeres and the turnover of kinetochore genes in budding yeast.

31
32 **Introduction**

33
34 The increasing availability of genomic datasets across diverse species has allowed the use of

35 comparative genomic approaches to study kinetochore evolution. Such comparative approaches

36 have led to interesting evolutionary insights from species with unique kinetochores (Drinnenberg

37 & Akiyoshi, 2017). The kinetochore is a complex of interacting proteins that have undergone

38 frequent changes in gene content and exhibits strong signatures of adaptive evolution (Malik &

39 Henikoff, 2001; Talbert, Bryson & Henikoff, 2004; Schueler et al., 2010; Drinnenberg, Henikoff

40 & Malik, 2016). Despite changes in gene content, the kinetochore in diverse species is known to

41 consist of the inner kinetochore which is assembled close to the centromere DNA as a

42 specialised form of histone (CENPA) and the outer kinetochore which interacts with

43 microtubules. The inner and outer kinetochores are known to consist of several interacting

44 protein complexes (see van Hooff et al., 2017 for more details). These kinetochore protein

45 complexes show patterns of co-evolution with interacting components through correlated gene

46 loss events (Tromer, 2017). The CCAN/CTF19 complex interfaces with both the inner and outer

47 kinetochore acting as a foundation for kinetochore formation (Hara & Fukagawa, 2017; Hinshaw

48 & Harrison, 2019). Surprisingly, large-scale systematic screening for kinetochore genes in the

49 genomes of numerous eukaryotes has shown that the majority of the components of the

50 CCAN/CTF19 complex are lost in many lineages (van Hooff et al., 2017). Another unexpected

51 observation was that the CENPA (CENH3 & CSE4 homolog) gene, which performs a central

52 role in kinetochore function, was lost in many insect species. This potentially recurrent loss of

53 the CENPA/CSE4 gene in insect species coincides with their transition to holocentricity

54 (Drinnenberg et al., 2014). Interestingly, it has been shown that C.*elegans* holocentromeres are

55 actually point centromeres that are dispersed at transcription factor hotspots (Steiner & Henikoff,

56 2014). Hence, it has been speculated that changes in the gene content of kinetochores might be

57 functionally related to transitions in the centromere type (Drinnenberg & Akiyoshi, 2017).

58

59 Centromere sequences in the vast majority of species are repeat-rich regions that are thought to

60 be defined epigenetically and/or through recognition of dyad rich regions with non-B-form DNA

61 structures (Kasinathan & Henikoff, 2018). Despite advances in genome sequencing and

62 assembly methods, the high repeat content of these centromeres makes it harder to assemble and

63 study them. Hence, as a proxy to the study of centromere regions, comparison of tandem repeats

64 across large phylogenetic distances have been performed to understand the evolution of

65 centromere sequences (Melters et al., 2013).

66

67   One of the most interesting transitions during centromere evolution was the emergence of
68   genetically defined point centromeres in budding yeasts (Malik & Henikoff, 2009). These
69   ~150bp long centromeres found in Saccharomycetaceae have been easier to study due to the lack
70   of repetitive regions, availability of complete high-quality genome assemblies for multiple
71   closely related species, gene knockout libraries and ease of experimental manipulation. Hence,
72   budding yeast species are a popular system to study the evolution of centromeres and the set of
73   proteins that interact with them (Roy & Sanyal, 2011).

74

75   In addition to changing their sequences, centromeres are also known to change their genomic
76   positions without any change in the karyotype (Montefalcone et al., 1999). However, centromere
77   evolution can also be accompanied by changes in the karyotype (O'Neill, Eldridge & Metcalfe,
78   2004). Within mammals, centromeres are known to have undergone multiple repositioning
79   events (Rocchi et al., 2012). Centromere repositioning events seem to be fairly common and
80   could have a role in driving speciation or, at the very least, have a non-negligible role in
81   affecting the local genomic selection landscape.

82

83   Recently, it has been shown that in Naumovozyma, an unconventional centromere has come into
84   existence at a location that is largely distinct from that expected based on synteny with other
85   Saccharomycetaceae species (Kobayashi et al., 2015). The Naumovozyma genus has been
86   suggested as a model for comparative genomics and study of adaptive evolution due to the
87   various phenotypic differences compared to other yeast species (Karademir Andersson & Cohn,
88   2017).

89
90   In this study, I first screened the genome assemblies of eight Pre-WGD and twelve Post-WGD
91   yeast species for the presence/absence of homologs of 67 kinetochore genes. I find evidence for
92   the concurrent loss of multiple genes from the CCAN/CTF19 protein complex in Naumovozyma
93   species and corresponding sequence divergence of the N-terminus region of the CENPA/CSE4
94   gene that interacts with the CCAN/CTF19 complex. In contrast to this, I see high levels of
95   sequence conservation of the C-terminus region of the CENPA/CSE4 gene that mediates an
96   interaction between the kinetochore and centromere. My analysis finds an association between
97   gene loss events and the emergence of novel centromeres in Naumovozyma species.

98
99 **Materials and Methods**
100
101 **Gene presence/absence screening**
102
103 I compiled a list of 67 kinetochore associated genes in the yeast S. *cerevisiae* by downloading

104 genes annotated with the GO term kinetochore (GO: 0000776) in Ensembl release 91. The

105 orthologs of these genes in all twenty yeast species were identified from the Yeast Gene Order

106 Browser (YGOB) (Byrne & Wolfe, 2005). I screened the genomes of all twenty yeast species for

107 the presence of all 67 genes and identified eight genes that are (both copies lost in post-WGD)

108 lost in at least one species (see **Supplementary Material S1**). Kinetochore genes are known to

109 evolve at a very fast rate, making it hard to identify orthologs of these genes even in closely

110 related species (van Hooff et al., 2017). It is possible that these genes have evolved at a very fast

111 rate making it unfeasible to establish homology of these genes. Fortunately, the YGOB provides

112 not only the order of the genes but also the intergenic sequences between genes. Based on

113 flanking genes with conserved synteny in other species, I identified the intergenic regions that

114 correspond to the location of the missing genes (see **Supplementary Material S2**). I checked

115 these intergenic sequences for the presence of open reading frames (ORFs) using the NCBI ORF

116 finder program with default settings. The ORFs that were found in the intergenic regions did not

117 show any homology (inferred using blastn and blastp search against NCBI's nucleotide

118 collection and non-redundant protein sequences, respectively) to the genes that I have inferred to

119 be lost. Based on the evidence from the identification of syntenic regions using YGOB and

120 additional screening of intergenic regions I am confident of these gene loss events. Nonetheless,

121 the presence of highly diverged copies of these genes in non-syntenic regions and genome

122 assembly errors cannot be ruled out.

123

124 **Multiple sequence alignments of** CENPA/**CSE4 gene**

125 The complete open reading frame of the CENPA (CENH3 & CSE4 homolog) gene was used for

126 multiple sequence alignment. To ensure that the results I see are not the result of alignment

127 artifacts, I performed the multiple sequence alignment at the nucleotide sequence level and

128 amino acid residue level using four different programs (Clustal omega (default settings in the

129 webserver), M-coffee (command used and alignment scores are provided in **Supplementary**

130 **Material S3**), MUSCLE (default settings in MEGA) and Guidance with PRANK as the aligner

131   (command used is provided in **Supplementary Material S3**)). All the nucleotide and amino acid

132   sequence alignments are provided as **Supplementary Material S3**. I investigated the multiple

133   sequence alignments for evidence of lineage-specific patterns of selection in Naumovozyma

134   species using the programs (RELAX, MEME, FEL and BUSTED) available in the HYPHY

135   package (Kosakovsky Pond, Frost & Muse, 2005). The output files of these are also provided in

136   **Supplementary Material S3**. The presence of additional sequences in the genus Naumovozyma

137   was seen in all of the multiple sequence alignments that I have analysed. It is possible that longer

138   ORF's have been incorrectly annotated as the CDS for both Naumovozyma species. However,

139   the second methionine codon in the CDS occurs at the 226ᵗʰ and 40ᵗʰ residue from the currently

140   annotated start codon in N. *castellii* & N. *dairenensis,* respectively. If the second methionine is

141   the correct start codon, the N. *castellii* protein would be just 39 residues. This suggests that the

142   correct ORF is annotated in Naumovozyma species. To further rule out the possibility of

143   erroneous annotation of shorter ORF's in non-naumovozyma species I extracted the genomic

144   sequence found between genes flanking CENPA (CENH3 & CSE4 homolog) and searched them

145   for ORF's using the NCBI ORF finder program with default settings (see **Supplementary**

146   **Material S4**). I found that the longest ORF that could be identified in this sequence was the

147   currently annotated ORF itself. This further supports the validity of the annotation and multiple

148   sequence alignments generated by me. I used the amino acid alignment generated by MUSCLE

149   aligner to calculate the sequence conservation score using the al2co program (Pei & Grishin,

150   2001) with default settings along with the –a flag to calculate nine measures of sequence

151   conservation. Per base measures of sequence conservation are provided as **Supplementary**

152   **Material S5**.

153

154   **Calculation of dyad density**

155   Centromere sequences of the two Naumovozyma species were obtained from NCBI and YGOB

156   browser for the remaining 18 species. The intergenic regions in Naumovozyma species

157   orthologous to the old centromeres were extracted from YGOB. All these sequences and their

158   locations are provided in **Supplementary Material S6**. The program *palindrome* from the

159   EMBOSS package (Rice, Longden & Bleasby, 2000) was used to identify dyad symmetry (i.e.,

160   DNA sequences with base pairs that are inverted repeats of each other). The following settings

161   were used for running the program "-minpallen 5 -maxpallen 100 -gaplimit 20 –overlap". All the

162 palindromes identified for each of the sequences is provided as **Supplementary Material S7.**
163 The dyad density for each centromere sequence was calculated as the fraction of bases that are
164 part of a dyad. GC content for each sequence was calculated as the fraction of GC bases. GC
165 content and dyad density for each of the centromere sequences are provided in **Supplementary**
166 **Material S8.**

167 **Results**
168
169 **Patterns of kinetochore gene loss**
170
171 My study system consists of eight Pre-WGD and twelve Post-WGD yeast species at varying
172 evolutionary distances (see **Fig. 1, phylogeny based on** (Gordon, Byrne & Wolfe, 2011; Feng et
173 al., 2017)). Based on screening of twenty yeast genomes for the presence of kinetochore genes I
174 identified that eight genes (NKP1, NKP2, CENPL/IML3, CENPN/CHL4, CENPT/CNN1,
175 CENPW/WIP1, FIN1 and BIR1) are lost at least once (see **Fig. 1** and **Supplementary Material**
176 **S1**). Intriguingly, five of the eight genes (i.e., NKP1, NKP2, CENPL/IML3, BIR1 and WIP1)
177 that have lost both their copies in the Post-WGD species are also the ones that are lost in the Pre-
178 WGD species (see **Fig. 1**). This hints at the dispensability of these genes over the course of
179 evolution. In this study, I focus on the set of four genes (NKP1, NKP2, CENPL/IML3, and
180 CENPN/CHL4) that are lost in Naumovozyma (a genus known to have novel point centromeres).
181 As the name suggests, the Non-essential Kinetochore Protein genes NKP1 & NKP2 produce
182 proteins that localise to kinetochores and when deleted produce viable single gene mutants
183 (Cheeseman et al., 2002). However, both single gene mutants are known to show elevated rates
184 of chromosome loss (Fernius & Marston, 2009). Knockout of both NKP1 and NKP2 only shows
185 a moderate increase in chromosome loss (Tirupataiah et al., 2014). Despite being non-essential
186 genes, at least one copy of both of these genes is found in 15 of the 20 yeast species screened in
187 this study. A broader phylogenetic search for homologs of these genes has shown the repeated
188 loss of these two genes in various taxa (Tromer, 2017). While it was known that NKP1 and
189 NKP2 bind to COMA complex (Hornung et al., 2014), recent work has shown that NKP1 and
190 NKP2 are positioned at the bottom of the CTF19 complex and form a four-chain helical coil
191 along with OKP1 and AME1 at their c-terminus (Hinshaw & Harrison, 2019). The main function
192 of the NKP1 and NKP2 heterodimer is thought to be the stabilisation of the COMA complex

193   (Schmitzberger et al., 2017). The COMA complex is known to interact directly with the
194   CENPA/CSE4 protein (Fischböck-Halwachs et al., 2019).
195
196   The CENPN/CHL4 (CHromosome Loss 4) single-gene knockouts are also viable but are known
197   to show increased levels of chromosome loss, miss-segregation, and abnormal kinetochores (Roy
198   et al., 1997). However, based on the chromatin state CENPN/CHL4 mutant cells are known to
199   show two distinct (high and low) levels of mitotic mobility (Roy & Sanyal, 2011). This suggests
200   that CENPN/CHL4 mutants can be compensated through changes in the chromatin state.
201   CENPL/IML3 (Increased Minichromosome Loss) protein forms a stable heterodimer with
202   CENPN/CHL4 protein (Hinshaw & Harrison, 2013). Although CENPL/IML3 and
203   CENPN/CHL4 are known interactors, the phylogenetic distribution of CENPL/IML3 seems to be
204   more restricted than that of CENPN/CHL4. The association of cohesin with the pericentromeric
205   regions is ensured by the action of CENPL/IML3 and CENPN/CHL4. The lack of these two
206   genes leads to reduced cohesin binding at the pericentromere that results in the miss-segregation
207   of chromosomes (Fernius & Marston, 2009).
208
209   The function of the genes lost in Naumovozyma may be performed by genes that are found in
210   Naumovozyma but are absent in S. *cerevisiae*. Naumovozyma species have 46 genes that are
211   absent in the other 18 species analysed in this study. To evaluate whether any of these genus
212   specific genes could have taken over the role performed by the four missing genes, I identified
213   homologs of each of these genes by performing blastp search against the non-redundant protein
214   database with an e-value cut-off of 10^{-3}. None of the identified homologs had a characterised role
215   related to the kinetochore machinery. However, profile-versus-profile search based comparisons
216   have identified that NKP1 and NKP2 are extremely similar to Mis12 and Nnf1 (Tromer et al.,
217   2019). Since, one copy (NCAS0H00450 & NDAI0K02740) of the Nnf1 gene is present in each
218   of the Naumovozyma species; it is possible that the functions of NKP1/2 are compensated.
219
220
221   **N-terminus divergence of CENPA/CSE4 gene in Naumovozyma**
222

223   The detailed study of the interactions between kinetochore proteins in S. *cerevisiae* has led to a

224   better understanding of their roles (Measday et al., 2005; Baetz, Measday & Andrews, 2006).

225   However, the evolution of the kinetochore network across Eukaryotes has been shown to be a

226   complex process that requires further investigation (van Hooff et al., 2017). In the current study,

227   I have focussed on well-studied genetic interactions of the selected genes that have lost both

228   copies in Post-WGD yeast species. Chromatin Immuno-Precipitation (ChIP) of CENPA/CSE4

229   protein followed by sequencing of DNA fragments was used by Kobayashi et al. (2015) to

230   identify all the CENPA/CSE4 binding sites in the N. *castellii* genome. Based on this

231   experimental data for CENPA/CSE4 along with similar ChIP-seq data for the NDC10, NDC80,

232   and CEP3 proteins the locations of the new centromeres in N. *castellii* have been validated

233   (Kobayashi et al., 2015). The multiple sequence alignment of amino-acid sequences across the

234   study species (see **Supplementary Material S3**) shows that the CENPA/CSE4 genes of N.

235   *castellii* and N. *dairenensis* have a stretch of approximately ten amino acid residues at the very

236   beginning of the CENPA/CSE4 gene that is absent in other species. It is known that the C-

237   terminus of the CENPA/CSE4 protein binds to the centromeric DNA sequence, and the N-

238   terminus interacts with the CCAN/CTF19 complex (Chen et al., 2000). The N-terminus region

239   has overall reduced sequence conservation across all species compared to the C-terminus region

240   (see **Fig. 2**). I also find evidence of significant relaxed selection in the N. *castellii* and N.

241   *dairenensis* lineages using the hypothesis testing framework available in the RELAX program

242   (version 2.1) on the multiple sequence alignments generated by MUSCLE. However, when the

243   alignments generated using guidance with PRANK as aligner was used, the relaxed selection in

244   the N. *castellii* lineage was not found to be statistically significant (see **Supplementary**

245   **Material S3**). These differences in the inferences based on the multiple sequence aligner used

246   are known to occur when the sequences are highly diverged (Blackburne & Whelan, 2013). I

247   have shown earlier that genes can experience relaxed selection when the c-terminal end is

248   increased in length due to change in the position of the stop codon (Shinde et al., 2019). Hence,

249   the relaxed selection detected in the CENPA/CSE4 gene could simply be a result of drastic

250   change in the length of the n-terminal sequence.

251

252   The N-terminus of the CENPA/CSE4 protein has been shown to play a role in the ubiquitin-

253   mediated proteolysis of the CENPA/CSE4 protein (Au et al., 2013). More recently, sumoylation

254 and ubiquitination of the N-terminus have been shown to be required to prevent the

255 mislocalization of CENPA/CSE4 to non-centromeric chromatin (Ohkuni et al., 2018). The lysine

256 65 residue (K65) in the CENPA/CSE4 gene has been identified as the residue important for

257 proper localisation. Changes in the N-terminus region of CENPA/CSE4 gene could have led to

258 changes in the post-translational modifications resulting in changes in the localisation patterns

259 and subsequent movement of centromeres in Naumovozyma.

260

261 **Dyad density at point centromeres**

262

263 Dyad symmetry refers to the presence of inverted repeats or palindromes in the sequence. The

264 inverted repeats identified on the centromere of S. *cerevisiae* are provided in **Fig. 3** as an

265 example. It has been shown that centromere regions are defined by the presence of Non-B-form

266 DNA structures resulting from the presence of dyad symmetry in the nucleotide sequence

267 (Kasinathan & Henikoff, 2018). While the new Naumovozyma centromeres are enriched for

268 dyad symmetries and non-B-form DNA, the dyad symmetry was less and the SIST DNA melting

269 and cruciform extrusion scores were lower in Naumovozyma (sensu lato, demarcated by a red

270 box in **Fig. 1**) compared to sensu strictu species (demarcated by a blue box in **Fig. 1** (Pulvirenti

271 et al., 2000)). However, they utilised the contrast of sensu stricto vs sensu lato and ignore other

272 species that have been phylogenetically placed closer to the sensu strictu species. Using a larger

273 sample size (centromeres from various budding yeast species) I show that the GC content shows

274 a significant negative correlation (Kendall's rank correlation coefficient tau of -0.59, p-value <

275 2.2e-16) with dyad density (see **Fig. 4A**). This correlation is reflective of the genome-wide

276 pattern seen in S. *cerevisiae* (Lisnić et al., 2005).

277

278 The unconventional centromeres in the yeast species N. *castellii* and N. *dairenensis* are in most

279 cases located at a different genomic locus compared to the point centromeres found in other

280 budding yeast species such as S. *cerevisiae* (Kobayashi et al., 2015). This movement of

281 centromeres has partly been attributed to chromosomal re-arrangements. I show that the GC

282 content of the intergenic regions corresponding to the older (S. *cerevisiae* like) centromeres is

283 very high and dyad density is low (see **Fig. 4A**; red coloured circles). On the other hand the new

284 unconventional centromeres that have been identified in Naumovozyma (see **Fig. 4A**; blue

285 coloured circles) are having higher dyad densities and lower GC content. I note that the pattern is

286   the same for both Naumovozyma species (see **Fig. 4A**; filled circles for N. *castellii* and unfilled

287   circles for N. *dairenensis*). The new (mean: 25.80, median: 24.77, min: 20.00 & max: 35.16)

288   centromeres in N. *dairenensis* have a lower GC content than the older regions (mean: 31.58,

289   median: 32.90, min: 22.03 & max: 37.32). Similarly, the new (mean: 21.36, median: 20.91, min:

290   18.18 & max: 24.55) centromeres in N. *castellii* have a lower GC content than the older regions

291   (mean: 30.15, median: 30.35, min: 21.82 & max: 36.47). I used a pair-wise wilcoxon test with

292   holm multiple testing correction (also see **Fig. 4B**) to compare the GC content of the old and new

293   centromere regions and find a difference in both N. *castellii* (q-value: 0.0067) and N. *dairenensis*

294   (q-value: 0.0778). This change in dyad density and GC content probably reflects a divergence

295   that occurred after the change in the location of the centromere.

296   **Discussion**

297

298

299   Nucleotide sequence divergence at the centromeres themselves as well as in the coding sequence

300   of proteins that recognise and bind to centromeres have been proposed as potential mechanisms

301   for the build-up of post-zygotic reproductive isolation (Borodin et al., 2001). The paradoxical

302   behaviour of centromeres to evolve rapidly while still being inherited stably makes them good

303   candidates for loci that contribute to the process of speciation (Henikoff, Ahmad & Malik, 2001).

304   However, the observation that Drosophila species produce fertile offspring despite the extensive

305   divergence of the sequence of centromere binding proteins seems to contradict this idea (Sainz et

306   al. 2003; but see Thomae et al. 2013). Computational identification and functional

307   characterisation of the effects of sequence divergence, loss, and duplication of the kinetochore

308   genes in diverse species will help clarify the role of the centromere in facilitating reproductive

309   isolation.

310

311   In this study, I provide evidence that suggests that both copies of four genes of the

312   CCAN/CTF19 complex are lost in Naumovozyma budding yeast species that have transitioned to

313   unconventional centromeres (Kobayashi et al., 2015). The loss of these CCAN/CTF19 complex

314   genes is potentially mediated by changes in the N-terminus region of the CENPA/CSE4 gene

315   (see **Fig. 3B**). Although I have no experimental data to suggest that the changes in the N-

316   terminus region of the CENPA/CSE4 gene are a consequence of changes in the sequence of the

317   centromere being bound by the C-terminus region, it is one potential scenario that could explain

318   the observed association between gene loss and transition to unconventional centromeres. In
319   contrast to the well-studied case of loss of CENH3 (CENPA/CSE4 homolog) gene in multiple
320   insect lineages (Drinnenberg et al., 2014) and the multiple cases of exon gain and loss (Fan et al.,
321   2013), I show that transition to a new centromere sequence might be sufficient for the loss of
322   CCAN/CTF19 complex genes. Understanding the sequence of events involved in the loss of the
323   CCAN/CTF19 complex genes and the emergence of the unconventional centromere in budding
324   yeast species would help understand the loss of CCAN/CTF19 complex genes in various
325   eukaryotic lineages (van Hooff et al., 2017). The CENPA/CSE4 gene has additional sequence in
326   Naumovozyma compared to other species suggesting that the unconventional centromeres might
327   potentially be remnants of the ancestral state of point centromeres. I discuss a few other potential
328   scenarios for the sequence of events that could have led to the association observed.

329   The dispensability of the genes that are lost in Naumovozyma is supported by the viability of S.
330   *cerevisiae* knockouts and independent loss in the two Pre-WGD Eremothecium species. It is
331   possible that initially, these genes were sequentially lost, and the emergence of novel
332   centromeres was an adaptive response to compensate for the lost genes. Yet, the loss of three out
333   of these four genes in Eremothecium species (E. *gossypii* and E. *cymbalariae*) does not seem to
334   be associated with any noticeable changes in centromere type. The evolution of novel
335   centromeres could just be one possible solution to the loss of these genes from the CCAN/CTF19
336   complex, and Eremothecium species might have come up with a different solution. Further
337   phenotypic characterisation of Eremothecium species might shed light on this.

338
339   It has been shown that loss of RNAi leads to a shortening of the centromeres and is an important
340   determinant of centromere evolution in fungi (Yadav et al., 2018). The N. *castellii* species in
341   addition to having unconventional point centromeres, have also been shown to have a functional
342   RNA interference pathway (Drinnenberg et al., 2009). Despite having a role in fission yeast
343   heterochromatin specification at the centromeres, siRNA's have not been found to have any
344   centromere-specific role in N. *castellii* (Kobayashi et al., 2015). Nonetheless, it would be
345   interesting to know what role the presence of RNAi machinery in Naumovozyma might have had
346   in the loss of CCAN/CTF19 complex genes. The Drinnenberg et al. study (2009) reports the
347   gene expression fold change of ago1 knockout and dcr1 knockout in N. *castellii* compared to the

348   wild type. However, none of the genes showing a fold-change greater than two are part of the

349   kinetochore gene set.

350

351   The current study does not perform any experimental characterisation of phenotypes across the

352   twenty yeast species that would be caused by knocking out the genes that I identify as lost.

353   However, kinetochore genes have been studied extensively in S. *cerevisiae*. Single and double

354   gene knockouts have been phenotypically characterised in great detail (Measday et al., 2005).

355   This functional data from S. *cerevisiae* can be extrapolated to other species by supplementing it

356   with computational predictions. Such extrapolation would, of course, have to be experimentally

357   verified by subsequent studies in the focal species. Hence, our study is merely focussed on

358   demonstrating strong patterns of associations that suggest kinetochore network rewiring in

359   Naumovozyma.

360

361   Stable separation of chromosomes into daughter cells requires proper functioning of the

362   kinetochore machinery. Reduced efficiency in such segregation of chromosomes would result in

363   reduced fertility. The high levels of gene turnover seen in kinetochore genes suggest frequent

364   changes in the interactome of kinetochore. Such frequent changes in gene content could lead to

365   differences in the interactome of closely related species. These differences in the interactome can

366   have a prominent role in mediating speciation through reproductive isolation. Reconstruction of

367   the sequence of events leading to the turnover of kinetochore machinery and centromere

368   sequence at the molecular level in budding yeast species might serve as a good test case for

369   understanding its role in speciation.

370

371   **Conclusions**

372   Rapid changes in the genomic sequence of centromeres and associated kinetochore proteins

373   between closely related species are thought to have an important role in speciation. Yet, it is not

374   conclusively known whether the centromere sequence divergence occurs first or kinetochore

375   proteins evolve to use a different centromere sequence. I look at patterns of kinetochore gene

376   loss in twenty species of yeast to identify major lineage-specific events. Interestingly, the loss of

377   four genes of the CCAN/CTF19 complex coincides with the emergence of unconventional

378   centromeres in N.*castellii* and N.*dairenensis*. I speculate that loss of these genes might have

379 occurred subsequent to the divergence of the centromere sequences as their role might have been

380 taken over by other genes. While our results cannot establish the sequence of events, the

381 identified lineage-specific loss of kinetochores genes that are known to functionally interact

382 serves as a molecular footprint of genetic changes that contribute to reproductive isolation

383 between species.

384

385 **References**

386

387 Au WC, Dawson AR, Rawson DW, Taylor SB, Baker RE, Basrai MA. 2013. A novel role of the
388   N terminus of budding yeast histone H3 variant Cse4 in ubiquitin-mediated proteolysis.
389   *Genetics* 194:513–518. DOI: 10.1534/genetics.113.149898.
390 Baetz K, Measday V, Andrews B. 2006. Revealing hidden relationships among yeast genes
391   involved in chromosome segregation using systematic synthetic lethal and synthetic dosage
392   lethal screens. *Cell Cycle* 5:592–595. DOI: 10.4161/cc.5.6.2583.
393 Blackburne BP, Whelan S. 2013. Class of multiple sequence alignment algorithm affects
394   genomic analysis. *Molecular Biology and Evolution*. DOI: 10.1093/molbev/mss256.
395 Borodin PM, Henikoff S, Ahmad K, Malik HS. 2001. Speciation and centromere evolution [3]
396   (multiple letters). *Science* 294:2478–2480. DOI: 10.1126/science.294.5551.2478.
397 Byrne KP, Wolfe KH. 2005. The Yeast Gene Order Browser: Combining curated homology and
398   syntenic context reveals gene fate in polyploid species. *Genome Research* 15:1456–1461.
399   DOI: 10.1101/gr.3672305.
400 Cheeseman IM, Anderson S, Jwa M, Green EM, Kang J seog, Yates JR, Chan CSM, Drubin DG,
401   Barnes G. 2002. Phospho-regulation of kinetochore-microtubule attachments by the Aurora
402   kinase Ipl1p. *Cell* 111:163–172. DOI: 10.1016/S0092-8674(02)00973-X.
403 Chen Y, Baker RE, Keith KC, Harris K, Stoler S, Fitzgerald-Hayes M. 2000. The N terminus of
404   the centromere H3-like protein Cse4p performs an essential function distinct from that of
405   the histone fold domain. *Molecular and Cellular Biology* 20:7037–7048. DOI:
406   10.1128/mcb.20.18.7037-7048.2000.
407 Drinnenberg IA, Akiyoshi B. 2017. Evolutionary Lessons from Species with Unique
408   Kinetochores. In: *Centromeres and Kinetochores*. 111–138. DOI: 10.1007/978-3-319-
409   58592-5_5.
410 Drinnenberg IA, DeYoung D, Henikoff S, Malik HS ingh. 2014. Recurrent loss of CenH3 is
411   associated with independent transitions to holocentricity in insects. *eLife* 3:e03676. DOI:
412   10.7554/eLife.03676.
413 Drinnenberg IA, Henikoff S, Malik HS. 2016. Evolutionary Turnover of Kinetochore Proteins: A
414   Ship of Theseus? *Trends in Cell Biology* 26:498–510. DOI: 10.1016/j.tcb.2016.01.005.
415 Drinnenberg IA, Weinberg DE, Xie KT, Mower JP, Wolfe KH, Fink GR, Bartel DP. 2009.
416   RNAi in budding yeast. *Science* 326:544–550. DOI: 10.1126/science.1176945.
417 Fan XY, Yu L, Xu HL, Li Y. 2013. Multiple intron gain and loss events occurred during the
418   evolution of Cenp-A gene. *Chinese Science Bulletin* 58:2174–2178. DOI: 10.1007/s11434-
419   012-5623-z.
420 Feng B, Lin Y, Zhou L, Guo Y, Friedman R, Xia R, Hu F, Liu C, Tang J. 2017. Reconstructing
421   yeasts phylogenies and ancestors from whole genome data. *Scientific Reports* 7:1–12. DOI:

422   10.1038/s41598-017-15484-5.

423   Fernius J, Marston AL. 2009. Establishment of cohesion at the pericentromere by the Ctf19
424   kinetochore subcomplex and the replication fork-associated factor, Csm3. *PLoS Genetics*
425   5:e1000629. DOI: 10.1371/journal.pgen.1000629.

426   Fischböck-Halwachs J, Singh S, Potocnjak M, Hagemann G, Solis-Mezarino V, Woike S,
427   Ghodgaonkar-Steger M, Weissmann F, Gallego LD, Rojas J, Andreani J, Köhler A, Herzog
428   F. 2019. The COMA complex interacts with Cse4 and positions Sli15/ipl1 at the budding
429   yeast inner kinetochore. *eLife* 8:e42879. DOI: 10.7554/eLife.42879.

430   Gordon JL, Byrne KP, Wolfe KH. 2011. Mechanisms of chromosome number evolution in yeast.
431   *PLoS Genetics* 7:e1002190. DOI: 10.1371/journal.pgen.1002190.

432   Hara M, Fukagawa T. 2017. Critical Foundation of the Kinetochore: The Constitutive
433   Centromere-Associated Network (CCAN). *Progress in molecular and subcellular biology*
434   56:29–57. DOI: 10.1007/978-3-319-58592-5_2.

435   Henikoff S, Ahmad K, Malik HS. 2001. The centromere paradox: Stable inheritance with rapidly
436   evolving DNA. *Science* 293:1098–1102. DOI: 10.1126/science.1062939.

437   Hinshaw SM, Harrison SC. 2013. An Iml3-Chl4 Heterodimer Links the Core Centromere to
438   Factors Required for Accurate Chromosome Segregation. *Cell Reports* 5:29–36. DOI:
439   10.1016/j.celrep.2013.08.036.

440   Hinshaw SM, Harrison SC. 2019. The structure of the ctf19c/ccan from budding yeast. *eLife*
441   8:e48215. DOI: 10.7554/eLife.44239.

442   van Hooff JJ, Tromer E, van Wijk LM, Snel B, Kops GJ. 2017. Evolutionary dynamics of the
443   kinetochore network in eukaryotes as revealed by comparative genomics. *EMBO reports*
444   18:1559–1571. DOI: 10.15252/embr.201744102.

445   Hornung P, Troc P, Malvezzi F, Maier M, Demianova Z, Zimniak T, Litos G, Lampert F,
446   Schleiffer A, Brunner M, Mechtler K, Herzog F, Marlovits TC, Westermann S. 2014. A
447   cooperative mechanism drives budding yeast kinetochore assembly downstream of CENP-
448   A. *Journal of Cell Biology* 206:509–524. DOI: 10.1083/jcb.201403081.

449   Karademir Andersson A, Cohn M. 2017. Naumovozyma castellii: an alternative model for
450   budding yeast molecular biology. *Yeast* 34:95–109. DOI: 10.1002/yea.3218.

451   Kasinathan S, Henikoff S. 2018. Non-B-Form DNA Is Enriched at Centromeres. *Molecular*
452   *Biology and Evolution*.

453   Kobayashi N, Suzuki Y, Schoenfeld LW, Müller CA, Nieduszynski C, Wolfe KH, Tanaka TU.
454   2015. Discovery of an Unconventional Centromere in Budding Yeast Redefines Evolution
455   of Point Centromeres. *Current Biology* 25:2026–2033. DOI: 10.1016/j.cub.2015.06.023.

456   Kosakovsky Pond SL, Frost SDW, Muse S V. 2005. HyPhy: Hypothesis testing using
457   phylogenies. *Bioinformatics*. DOI: 10.1093/bioinformatics/bti079.

458   Lisnić B, Svetec IK, Šarić H, Nikolić I, Zgaga Z. 2005. Palindrome content of the yeast
459   Saccharomyces cerevisiae genome. *Current Genetics* 47:289–297. DOI: 10.1007/s00294-
460   005-0573-5.

461   Malik HS, Henikoff S. 2001. Adaptive evolution of Cid, a centromere-specific histone in
462   Drosophila. *Genetics* 157:1293–1298.

463   Malik HS, Henikoff S. 2009. Major Evolutionary Transitions in Centromere Complexity. *Cell*
464   138:1067–1082. DOI: 10.1016/j.cell.2009.08.036.

465   Measday V, Baetz K, Guzzo J, Yuen K, Kwok T, Sheikh B, Ding H, Ueta R, Hoac T, Cheng B,
466   Pot I, Tong A, Yamaguchi-Iwai Y, Boone C, Hieter P, Andrews B. 2005. Systematic yeast
467   synthetic lethal and synthetic dosage lethal screens identify genes required for chromosome

468   segregation. *Proceedings of the National Academy of Sciences* 102:13956–13961. DOI:
469   10.1073/pnas.0503504102.
470   Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby J, Sebra R, Peluso P, Eid J,
471   Rank D, Garcia J, DeRisi JL, Smith T, Tobias C, Ross-Ibarra J, Korf I, Chan SW. 2013.
472   Comparative analysis of tandem repeats from hundreds of species reveals unique insights
473   into centromere evolution. *Genome Biology* 14:R10. DOI: 10.1186/gb-2013-14-1-r10.
474   Montefalcone G, Tempesta S, Rocchi M, Archidiacono N. 1999. Centromere repositioning.
475   *Genome Research* 9:1184–1188. DOI: 10.1101/gr.9.12.1184.
476   O'Neill RJ, Eldridge MDB, Metcalfe CJ. 2004. Centromere dynamics and chromosome
477   evolution in marsupials. In: *Journal of Heredity*. 375–381. DOI: 10.1093/jhered/esh063.
478   Ohkuni K, Levy-Myers R, Warren J, Au W-C, Takahashi Y, Baker RE, Basrai MA. 2018. N-
479   terminal Sumoylation of Centromeric Histone H3 Variant Cse4 Regulates Its Proteolysis To
480   Prevent Mislocalization to Non-centromeric Chromatin. *G3 (Bethesda, Md.)* 8:1215–1223.
481   DOI: 10.1534/g3.117.300419.
482   Pei J, Grishin N V. 2001. AL2CO: Calculation of positional conservation in a protein sequence
483   alignment. *Bioinformatics* 17:700–712. DOI: 10.1093/bioinformatics/17.8.700.
484   Pulvirenti A, Nguyen H-V, Caggia C, Giudici P, Rainieri S, Zambonelli C. 2000. *Saccharomyces*
485   *uvarum* , a proper species within *Saccharomyces* sensu stricto. *FEMS Microbiology Letters*
486   192:191–196. DOI: 10.1111/j.1574-6968.2000.tb09381.x.
487   Rice P, Longden L, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open
488   Software Suite. *Trends in Genetics* 16:276–277.
489   Rocchi M, Archidiacono N, Schempp W, Capozzi O, Stanyon R. 2012. Centromere repositioning
490   in mammals. *Heredity* 108:59–67. DOI: 10.1038/hdy.2011.101.
491   Roy N, Poddar A, Lohia A, Sinha P. 1997. The mcm17 mutation of yeast shows a size dependent
492   segregational defect of a mini-chromosome. *Current Genetics* 32:182–189. DOI:
493   10.1007/s002940050264.
494   Roy B, Sanyal K. 2011. Diversity in requirement of genetic and epigenetic factors for
495   centromere function in fungi. *Eukaryotic Cell* 10:1384–1395. DOI: 10.1128/EC.05165-11.
496   Sainz A, Wilder JA, Wolf M, Hollocher H. 2003. Drosophila melanogaster and D. simulans
497   rescue strains produce fit offspring, despite divergent centromere-specific histone alleles.
498   *Heredity* 91:28–35. DOI: 10.1038/sj.hdy.6800275.
499   Schmitzberger F, Richter MM, Gordiyenko Y, Robinson C V, Dadlez M, Westermann S. 2017.
500   Molecular basis for inner kinetochore configuration through <scp>RWD</scp> domain–
501   peptide interactions. *The EMBO Journal* 36:3458–3482. DOI: 10.15252/embj.201796636.
502   Schueler MG, Swanson W, Thomas PJ, Green ED. 2010. Adaptive evolution of foundation
503   kinetochore proteins in primates. *Molecular Biology and Evolution* 27:1585–1597. DOI:
504   10.1093/molbev/msq043.
505   Shinde SS, Teekas L, Sharma S, Vijay N. 2019. Signatures of Relaxed Selection in the CYP8B1
506   Gene of Birds and Mammals. *Journal of Molecular Evolution* 87:209–220. DOI:
507   https://doi.org/10.1007/s00239-019-09903-6.
508   Steiner FA, Henikoff S. 2014. Holocentromeres are dispersed point centromeres localized at
509   transcription factor hotspots. *eLife* 2014:e02025. DOI: 10.7554/eLife.02025.001.
510   Talbert PB, Bryson TD, Henikoff S. 2004. Adaptive evolution of centromere proteins in plants
511   and animals. *Journal of biology* 3:18. DOI: 10.1186/jbiol11.
512   Thomae AW, Schade GOM, Padeken J, Borath M, Vetter I, Kremmer E, Heun P, Imhof A. 2013.
513   A Pair of Centromeric Proteins Mediates Reproductive Isolation in Drosophila Species.

514   *Developmental Cell* 27:412–424. DOI: 10.1016/j.devcel.2013.10.001.

515 Tirupataiah S, Jamir I, Srividya I, Mishra K. 2014. Yeast Nkp2 is required for accurate

516   chromosome segregation and interacts with several components of the central kinetochore.

517   *Molecular Biology Reports* 41:787–797. DOI: 10.1007/s11033-013-2918-3.

518 Tromer EC. 2017. Evolution of the Kinetochore Network in Eukaryotes.

519 Tromer EC, van Hooff JJE, Kops GJPL, Snel B. 2019. Mosaic origin of the eukaryotic

520   kinetochore. *Proceedings of the National Academy of Sciences of the United States of*

521   *America* 116:12873–12882. DOI: 10.1073/pnas.1821945116.

522 Yadav V, Sun S, Billmyre RB, Thimmappa BC, Shea T, Lintner R, Bakkeren G, Cuomo CA,

523   Heitman J, Sanyal K. 2018. RNAi is a critical determinant of centromere evolution in

524   closely related fungi. *Proceedings of the National Academy of Sciences* 115:3108–3113.

525   DOI: 10.1073/pnas.1713725115.

526

527

528

529 **Figure legends**

530 **Figure 1: Observed loss of kinetochore genes in budding yeast species.** Gene loss events have

531   been mapped onto the widely accepted tree topology obtained from (Gordon, Byrne &

532   Wolfe, 2011). The whole genome duplication event separating the Pre-WGD and Post-

533   WGD species is denoted by a star. Point centromeres are common to all twenty budding

534   yeast species and their emergence is depicted by a dark blue dot. Emergence of

535   unconventional point centromere in Naumovozyma is depicted by a light blue dot. Sensu

536   strictu yeast species are demarcated by a blue box and sensu lato species with neo-

537   centromeres are demarcated by a red box.

538 **Figure 2: Conserved c-terminus of CENPA/CSE4 contrasts with the diverse n-terminus.**

539   Sequence conservation (unweighted entropy measure) along the length of the

540   CENPA/CSE4 gene calculated using the al2co program is shown for each amino acid

541   position. While the c-terminus is highly conserved, the n-terminus has very little sequence

542   conservation. Refer to **Supplementary Material 3** for actual multiple sequence alignments.

543 **Figure 3: Dyads identified in the S. *cerevisiae* centromere CEN1.** Inverted repeats were

544   identified by the program palindrome from the emboss package. Each of the lines has the

545   118 bp sequence of CEN1 and the dyad region is highlighted in red colour. The last line in

546   the figure shows the regions (coloured green) that are covered by at least one dyad. Dyad

547   density is calculated as the fraction of bases that are covered by a dyad.

548

549 **Figure 4: Dyad density and GC content at the centromere regions. (A) Negative correlation**

550   **between GC content and dyad density at the point centromeres across budding yeast**

551   **species**. The filled circles represent N. *castellii* and unfilled circles show N. *dairenensis*.

552   Red colour is used for the intergenic regions corresponding to the older (S. cerevisiae like)

553   centromeres. Blue colour is used for the new unconventional centromeres that have been

554   identified in Naumovozyma. Centromeres of the remaining budding species have the *

555   shape and are coloured black. (B) **Boxplots comparing the GC content of old and new**

556   **centromeres in N. *castellii*and N. *dairenensis***. The new (mean: 25.80, median: 24.77, min:

557   20.00 & max: 35.16) centromeres in N. *dairenensis* have a lower GC content than the older

558   regions (mean: 31.58, median: 32.90, min: 22.03 & max: 37.32). Similarly, the new (mean:

559   21.36, median: 20.91, min: 18.18 & max: 24.55) centromeres in **N. *castellii*** have a lower

560   GC content than the older regions (mean: 30.15, median: 30.35, min: 21.82 & max: 36.47).

561   Pair-wise Wilcoxon test is used to compare the GC content of the old and new centromeres.

562   The q-values are obtained based on holm multiple testing correction.

563

Figure 1

Observed loss of kinetochore genes in budding yeast species.

Gene loss events have been mapped onto the widely accepted tree topology obtained from (Gordon, Byrne & Wolfe, 2011) . The whole genome duplication event separating the Pre-WGD and Post-WGD species is denoted by a star. Point centromeres are common to all twenty budding yeast species and their emergence is depicted by a dark blue dot. Emergence of unconventional point centromere in Naumovozyma is depicted by a light blue dot. Sensu strictu yeast species are demarcated by a blue box and sensu lato species with neo-centromeres are demarcated by a red box.
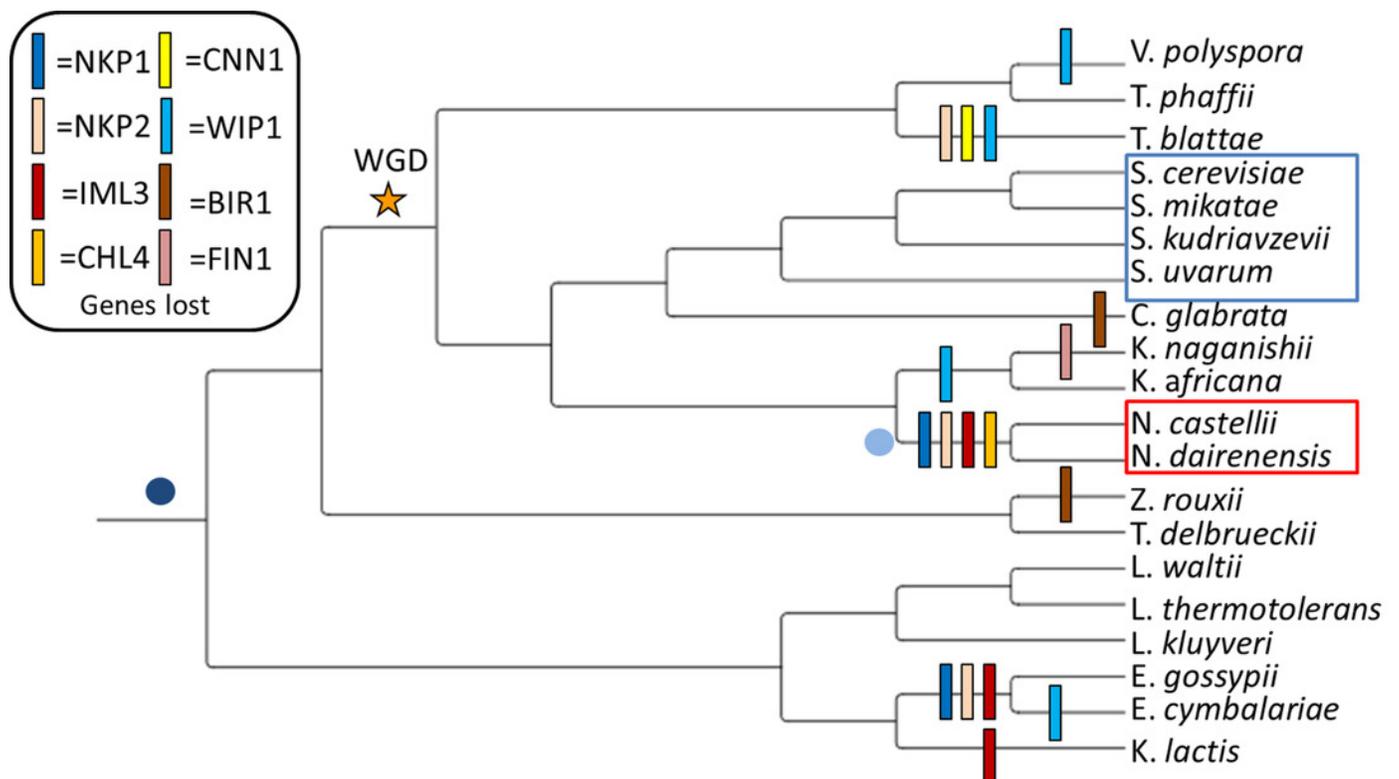
Figure 2

Conserved c-terminus of CENPA/CSE4 contrasts with the diverse n-terminus

Sequence conservation (unweighted entropy measure) along the length of the CENPA/CSE4 gene calculated using the al2co program is shown for each amino acid position. While the c-terminus is highly conserved, the n-terminus has very little sequence conservation. Refer to **Supplementary Material 3** for actual multiple sequence alignments.

Figure 3

Dyads identified in the S. *cerevisiae* centromere CEN1

Inverted repeats were identified by the program palindrome from the emboss package. Each of the lines has the 118 bp sequence of CEN1 and the dyad region is highlighted in red colour. The last line in the figure shows the regions (coloured green) that are covered by at least one dyad. Dyad density is calculated as the fraction of bases that are covered by a dyad.
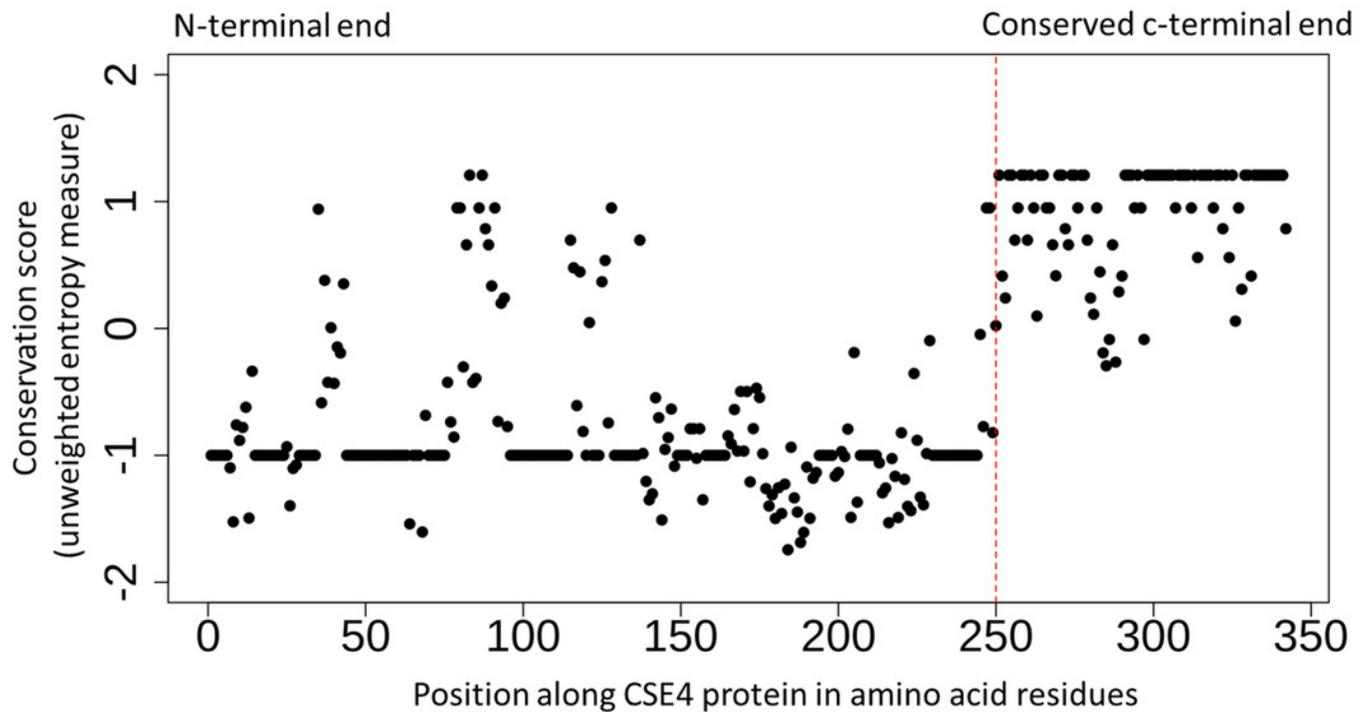
Figure 4

Dyad density and GC content at the centromere regions.

(A) Negative correlation between GC content and dyad density at the point centromeres across budding yeast species. The filled circles represent N. *castellii* and unfilled circles show N. *dairenensis*. Red colour is used for the intergenic regions corresponding to the older (S. cerevisiae like) centromeres. Blue colour is used for the new unconventional centromeres that have been identified in Naumovozyma. Centromeres of the remaining budding species have the * shape and are coloured black. **(B) Boxplots comparing the GC content of old and new centromeres in N. *castellii* and N. *dairenensis***. The new (mean: 25.80, median: 24.77, min: 20.00 & max: 35.16) centromeres in N. *dairenensis* have a lower GC content than the older regions (mean: 31.58, median: 32.90, min: 22.03 & max: 37.32). Similarly, the new (mean: 21.36, median: 20.91, min: 18.18 & max: 24.55) centromeres in N. *castellii* have a lower GC content than the older regions (mean: 30.15, median: 30.35, min: 21.82 & max: 36.47). Pair-wise Wilcoxon test is used to compare the GC content of the old and new centromeres. The q-values are obtained based on holm multiple testing correction.