A new method for determining ribosomal DNA copy number shows differences between *Saccharomyces cerevisiae* populations

Diksha Sharma[1], Sylvie Hermann-Le Denmat[1,2] , Nicholas J. Matzke[1], Katherine Hannan[3,4], Ross D. Hannan[3,4,5,6,7], Justin M. O'Sullivan[8,9,10,11], Austen R. D. Ganley[1]

1. School of Biological Sciences, University of Auckland, Auckland, New Zealand
2. Ecole Normale Supérieure, PSL Research University, F-75005 Paris, France
3. ACRF Department of Cancer Biology and Therapeutics, The John Curtin School of Medical Research, ACT 2601, Australia
4. Department of Biochemistry and Molecular Biology, University of Melbourne, Parkville, Victoria 3010, Australia
5. Division of Research, Peter MacCallum Cancer Centre, Melbourne, Victoria, 3000, Australia
6. Sir Peter MacCallum Department of Oncology, University of Melbourne, Parkville, Victoria, 3010, Australia
7. Department of Biochemistry and Molecular Biology, Monash University, Clayton, Victoria 3168, Australia
8. Liggins Institute, University of Auckland, Auckland, New Zealand
9. Maurice Wilkins Center, University of Auckland, New Zealand
10. MRC Lifecourse Unit, University of Southampton, United Kingdom
11. Brain Research New Zealand, The University of Auckland, Auckland, New Zealand

23 Corresponding Author, email : a.ganley@auckland.ac.nz

Abstract

Ribosomal DNA genes (rDNA) encode the major ribosomal RNAs (rRNA) and in eukaryotic genomes are typically present as one or more arrays of tandem repeats. Species have characteristic rDNA copy numbers, ranging from tens to thousands of copies, with the number thought to be redundant for rRNA production. However, the tandem rDNA repeats are prone to recombination-mediated changes in copy number, resulting in substantial intra-species copy number variation. There is growing evidence that these copy number differences can have phenotypic consequences. However, we lack a comprehensive understanding of what determines rDNA copy number, how it evolves, and what the consequences are, in part because of difficulties in quantifying copy number. Here, we developed a genomic sequence read approach that estimates rDNA copy number from the modal coverage of the rDNA and whole genome to help overcome limitations in quantifying copy number with existing mean coverage-based approaches. We validated our method using strains of the yeast *Saccharomyces cerevisiae* with previously-determined rDNA copy numbers, and then applied our pipeline to investigate rDNA copy number in a global sample of 788 yeast isolates. We found that wild yeast have a mean copy number of 92, consistent with what is reported for other fungi but much lower than in laboratory strains. We show that different populations have different rDNA copy numbers. These differences can partially be explained by phylogeny, but other factors such as environment are also likely to contribute to population differences in copy number. Our results demonstrate the utility of the modal coverage method, and highlight the high level of rDNA copy number variation within and between populations.

Introduction

The ribosomal RNA gene repeats (rDNA) encode the major ribosomal RNA (rRNA) components of the ribosome, and thus are essential for ribosome biogenesis and protein translation. In most eukaryotes the rDNA forms large tandem repeat arrays on one or more chromosomes [1]. Each repeat unit comprises a coding region transcribed by RNA polymerase I (Pol-I) that encodes 18S, 5.8S and 28S rRNA [2], and an intergenic spacer region (IGS) that separates adjacent coding regions (**Fig 1**). The number of rDNA repeat copies varies widely between species, typically from tens to hundreds of thousands of copies [1, 3-5]. However, each species appears to have a 'set' or homeostatic (in the sense of [6]) rDNA copy number that is returned to if the number of copies deviates [7-10]. Deviation in rDNA copy number between individuals within a species is well documented and can be substantial [11-16]. This copy number variation is thought to be tolerated because of redundancy in rDNA copies [8, 17]. This redundancy can partly be explained the striking observation that only a subset of the repeats is transcribed at any one time [2]. Thus, cells can compensate for changes in rDNA copy number by activating or silencing repeats to maintain the same transcriptional output [18]. Variation in rDNA copy number is a consequence of unequal homologous recombination, which results in loss or gain of rDNA copies [8, 19-22]. This copy number variation is, somewhat counter-intuitively, what drives the high levels of sequence homogeneity observed between the rDNA copies within a genome, a pattern known as concerted evolution [23-25]. Recent results in *Saccharomyces cerevisiae* revealed an elegant mechanism through which homeostatic rDNA copy number is maintained in the face of rDNA copy number change via the abundance of the Pol-I transcription factor UAF (upstream activating factor) and the histone deacetylase Sir2 [26]. However, the selective pressure(s) that determines what the homeostatic rDNA copy

number is remains unknown. Nevertheless, there is growing evidence that rDNA copy number and the proportion of active/silent rDNA copies impact several aspects of cell biology beyond simply rRNA production [8, 12, 17, 22, 27-35].

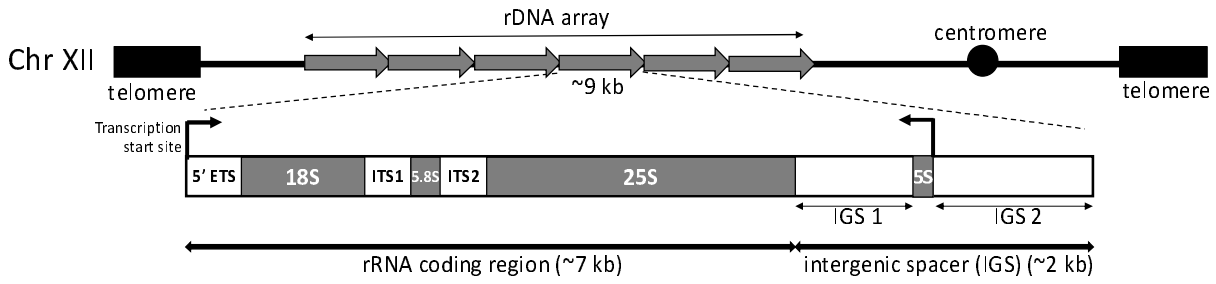

Figure 1. Organization of the rDNA repeats in *Saccharomyces cerevisiae*. Top shows a schematic of tandemly-repeated units in the rDNA array located on chromosome XII. Bottom shows the organization of an individual rDNA repeat including transcription start sites, the 5' external transcribed spacer (5'ETS), the rRNA (18S, 5.8S and 28S) coding genes, the two internal transcribed spacers (ITS1 and 2), and the intergenic spacer (IGS). The IGS is divided into two by a 5S rRNA gene. Schematic is not to scale.

Interest in the phenotypic consequences of rDNA copy number variation has led to a number of approaches being used to measure it. These include molecular biology approaches such as quantitative DNA hybridization [36-39], pulsed field gel electrophoresis (PFGE) [40, 41], quantitative real-time PCR (qPCR) [15, 42-46] and digital droplet PCR (ddPCR) [47, 48]. A major advance in the measurement of rDNA copy number has been the emergence of bioinformatic approaches that use whole genome (WG) next generation sequencing (NGS) reads to estimate copy number, based on the rationale that

90 sequence coverage of the rDNA correlates with copy number. This correlation is a
91 consequence of concerted evolution, with the high sequence identity between repeats
92 resulting in reads from all rDNA copies mapping to a single reference rDNA unit, thus
93 providing a high coverage signal that is proportional to copy number. Existing
94 bioinformatic approaches calculate the mean rDNA read coverage and normalize to the
95 mean WG coverage to estimate copy number [5, 12, 25, 34, 49], thus assuming that mean
96 coverage represents the "true coverage" for both the rDNA and the WG. However, there are
97 reasons to suspect this mean coverage approach assumption might not always hold.
98 Repetitive elements (e.g. microsatellites and transposons), PCR/sequencing bias (which is
99 particularly evident for the rDNA [50-54]; **Supplementary Figure 1**), and large-scale
100 mutations such as aneuploidies and segmental duplications may all cause the measured
101 mean coverage to differ from the real coverage. While efforts have been made to address
102 some of these potential confounders [12, 55, 56], estimated copy number varies depending
103 on which region of the rDNA is used [12, 34], thus the accuracy of this mean read coverage
104 approach has been called into question [5, 46].

105

106 Here we present a bioinformatics pipeline that measures rDNA copy number using modal
107 (most frequent) NGS read coverage as a way to overcome the limitations of the mean
108 coverage bioinformatics approach. We assessed the parameters important for performance
109 and validated the pipeline using *S. cerevisiae* strains with known rDNA copy numbers. We
110 then employed our pipeline to investigate whether *S. cerevisiae* populations maintain
111 different homeostatic rDNA copy numbers.

112 **Materials and Methods**

113

114 **Modified *Saccharomyces cerevisiae* genome**

115 Chromosome sequences for W303 were obtained from the NCBI (accession CM001806.1 -
116 CM001823.1) and concatenated. rDNA copies present within the W303 reference genome
117 were identified using BLAST and removed using Geneious (v. 11.0.3). The *S. cerevisiae*
118 W303 strain rDNA repeat unit from [23] was added as an extrachromosomal rDNA
119 reference, and this modified W303 yeast reference genome (W303-rDNA) was used in
120 subsequent analyses.

121

122 **Yeast strains/isolates and growth conditions**

123 Yeast strains/isolates that were cultured are listed in **Table 1**. Culturing was performed in
124 liquid or solid (2% agar) YPD (1% w/v yeast extract, 2% w/v peptone and 2 % w/v D+
125 glucose) medium at 30°C.

126

127 **Table 1: *S. cerevisiae* strains/isolates cultured in this study**

Strain/isolate	Details	Source
Wild-type	*MATa ade2-1 ura3-1 his3-11 trp1-1 leu2-3, 112 can1-100 fob1Δ::HIS3*	NOY408-1bf; [17]
20-copy	*MATa ade2-1 ura3-1 his3-11 trp1-1 leu2-3, 112 can1-100 fob1Δ::HIS3*	[17]

40-copy	*MATa ade2-1 ura3-1 his3-11 trp1-1 leu2-3, 112 can1-100 fob1Δ::HIS3*	[17]
80-copy	*MATa ade2-1 ura3-1 his3-11 trp1-1 leu2-3, 112 can1-100 fob1Δ::HIS3*	[17]
YJM981	Human clinical isolate from Italy; *Mat* ***a***, *ho::HygMX, ura3::KanMX*-Barcode	[57]
DBVPG1373	Netherlands isolate from soil; *Mat* ***a***, *ho::HygMX, ura3::KanMX*-Barcode	[57]
UWOPS03-461-4	Malaysian isolate from nectar	[58]
UWOPS03-461-4 (Mat a)	Derivative of UWOPS03-461-4; *Mat* ***a***, *ho::HygMX, ura3::KanMX*-Barcode	[57]
UWOPS03-461-4 (Mat α)	Derivative of UWOPS03-461-4; *Mat α, ho::HygMX, ura3::KanMX*-Barcode	[57]
YPS128	US isolate from soil beneath *Quercus alba*	[58]
DBVPG1788	Finland isolate from vineyard soil	[58]

128

129 **Genomic DNA extraction**

130 High molecular weight genomic DNA (gDNA) was isolated as follows. Cell pellets from 3–5
131 mL liquid cultures were washed in 500 µL of 50 mM EDTA pH 8 and resuspended in 200 µL
132 of 50 mM EDTA pH 8 supplemented with zymolyase (3 mg/mL). After 1 hr at 37°C, the cell
133 lysate was mixed with 20 µL of 10% sodium dodecyl sulfate then with 150 µL of 3 M
134 potassium acetate (KAc) and incubated on ice for 1 hr. 100 µL of phenol-chloroform-

135 isoamyl alcohol was added to the SDS-KAc suspension, and following vortexing and
136 centrifugation, 600 µL of propanol-2 were added to the aqueous supernatant (≈ 300 µL).
137 The nucleic acid pellet was washed three times in 70% EtOH, dried and resuspended in
138 PCR grade water supplemented with RNase A (0.3 mg/mL). After 1 hr at 37°C, samples
139 were stored at -20°C.

140

141 **Whole genome sequence data**

142 gDNA extracted from four isogenic strains with different rDNA copy numbers (WT, 20-
143 copy, 40-copy and 80-copy; **Table 1**) was sequenced using Illumina MiSeq
144 (**Supplementary Table 1**). The raw sequence files are available through the NCBI SRA
145 (accession number SUB7882611).

146

147 **Read preparation**

148 Paired-end reads were combined and quality checked using SolexaQA [59]. Low-quality
149 ends of reads (score cutoff 13) were trimmed using DynamicTrim, and short reads were
150 removed using a length cutoff of 50 bp with LengthSort, both within SolexaQA, as follows:

151 **command:** ~/path/to/solexaQA/SolexaQA++ dynamictrim /fastq/file

152 **command:** ~/path/to/solexaQA/SolexaQA++ lengthsort -l 50 /trimmed/fastq/file

153

154 **Obtaining whole genome and rDNA coverages**

155 The W303-rDNA reference genome was indexed using the bowtie2 (v. 2.3.2) build
156 command as follows:

157 **command:** ~/bowtie2-2.3.2/bowtie2-build <reference_in> <bt2_base>

158 Coverage files for the whole genome and rDNA were obtained using a four step pipeline:

159 **Step-1:** Processed reads were mapped to the indexed W303-rDNA genome using bowtie2
160 (v. 2.3.2)

161 **command:** ~/bowtie2-2.3.2/bowtie2 -x /path/to/indexed/genome/ -U
162 /path/to/trimmed/reads/ -S /output SAM file/

163 **Step-2:** The subsequent SAM format alignment was converted to BAM format using the
164 SAMtools (v. 1.8) view command:

165 **command:** ~/samtools-1.8/samtools view -b -S -o <output_BAM> <input_SAM>

166 **Step-3:** Mapped reads in the BAM file were sorted according to the location they mapped to
167 in W303-rDNA using the SAMtools sort command:

168 **command:** ~/samtools-1.8/samtools sort <input_BAM> -o <output_sorted.bam>

169 **Step-4:** Per-base read coverages across the entire W303-rDNA genome and the rDNA were
170 obtained using BEDtools (v. 2.26.0):

171 **command:** ~/bedtools genomecov -ibam <aligned_sorted.bam> -g
172 <reference_genome.fasta> -d <bedtools_coverage_WG.txt>

173 **command:** grep "rDNA_BLAST" <bedtools_coverage_WG.txt>
174 <rDNA_bedtools_coverage.txt>

175

176 **Calculation of rDNA copy number using modal coverage**

177 Coverage frequency tables for the rDNA and whole genome (excluding mitochondrial DNA
178 and plasmids) were obtained from per-base read coverage files by computing the mean
179 coverage over a given sliding window size with a slide of 1 bp. The mean coverage for each
180 sliding window was then allocated into a coverage bin. The bin that includes read coverage
181 of zero was removed. The three highest frequency coverage bins from both the rDNA and
182 whole genome frequency tables were used to calculate rDNA copy number as follows:

$$\boldsymbol{rDNA\ copy\ number} = \frac{\boldsymbol{Peak\ rDNA\ coverage\ bin\ value}}{\boldsymbol{Peak\ whole\ genome\ coverage\ bin\ value}}$$

183 The rDNA copy number estimates were taken as the mean of all pairwise combinations of
184 these copy number values (**Supplementary Figure 2**).

185

186 **Pipeline availability**

187 The pipeline for modal calculation of rDNA copy number from an alignment of sequence
188 reads to a reference genome containing one rDNA copy is available through Github
189 (https://github.com/diksha1621/rDNA-copy-number-pipeline).

190

191 **Calculation of rDNA copy number using mean and median coverage**

192 The per-base read coverage across W303-rDNA from Bedtools was input into custom R-
193 scripts to obtain the mean and median coverage values for the whole genome and rDNA,
194 after removing the rDNA, 2-micron plasmid, and mitochondrial DNA coverage values from

195 the whole genome calculation. The rDNA copy number was then calculated for both mean
196 and median approaches as follows:

$$\boldsymbol{rDNA\ copy\ number = \frac{coverage\ across\ rDNA}{coverage\ across\ whole\ genome}}$$

197

198 **Subsampling**

199 To generate different coverage levels for copy number estimation, sequence reads were
200 randomly downsampled using the seqtk tool (https://github.com/lh3/seqtk) as follows:

201 **command:** ~/seqtk/seqtk sample –s$RANDOM <name of fastqfile> <number of reads
202 required> <outputfile>

203

204 **rDNA copy number measurement by ddPCR**

205 At least three independent cultures (biological replicates) were generated for each isolate
206 using one independent colony per culture. To evaluate rDNA copy number variation over
207 generations, cultures were propagated over four days (~60 generations) as follows:
208 individual colonies were initially grown in 3 mL YPD for 24 hr. 30 µL of this was used to
209 inoculate 3 mL YPD and this was grown for another 24 hr. This process was repeated for
210 four days. Cells were harvested after 24 hr (~15 generations) and four days, and cell pellets
211 frozen at -80˚C. gDNA was extracted as above, then linearized by *Xba*I in NEB2 buffer
212 following the manufacturer's instructions (NEB) to individualize rDNA repeats. gDNA
213 linearization was verified by separation on agarose gels and DNA concentration measured
214 on a Qubit Fluorometer using the Qubit dsDNA HS assay (Thermo Fisher). Linearized gDNA

215 was brought to 2 pg/μL by serial dilution. EvaGreen master mixes were prepared with an
216 rDNA primer pair (rDNAScSp_F2 5'- ATCTCTTGGTTCTCGCATCG-3', rDNAScSp_R2 5'-
217 GGAAATGACGCTCAAACAGG-3') or a single copy *RPS3* gene primer pair (RPS3ScSp_F2 5'-
218 CACTCCAACCAAGACCGAAG-3', RPS3ScSp_R2 5'-GACAAACCACGGTCTTGAAC-3'). *RPS3* and
219 rDNA ddPCR reactions were performed with 2 μL (4 pg) of the same linearized gDNA
220 dilution as template. Droplet generation and endpoint PCR were performed following the
221 manufacturer's instructions, and droplets were read using a QX200 droplet reader
222 (BioRad). Quantification was performed using QuantaSoft Analysis Pro (v. 1.0.596). rDNA
223 copy number was determined by the (rDNA copy/μL)/(*RPS3* copy/μL) ratio.

224

225 **Pulse field gel electrophoresis (PFGE)**

226 To make chromosome plugs [21], cells from overnight liquid cultures were resuspended in
227 50 mM EDTA pH 8.0 to 2.10^9 cells/mL, transferred to 45°C, and mixed with an equal
228 volume of 1.5% low melting point agarose in 50 mM EDTA prewarmed to 45°C. The
229 mixture was transferred by gentle pipetting to PFGE plug molds (BioRad) to set at 4°C for
230 15 min. Plugs were transferred to fresh spheroplasting solution (1 M Sorbitol, 20 mM EDTA
231 pH 8.0, 10 mM Tris-HCl pH 7.5, 14 mM 2-mercaptoethanol, 2 mg/mL zymolyase). After 6 hr
232 incubation at 37°C with occasional inversion, plugs were washed for 15 min in LDS buffer
233 (1% lithium dodecyl sulphate, 100 mM EDTA pH 8.0, 10 mM Tris-HCl pH 8.0), before
234 overnight incubation at 37°C in the same buffer with gentle shaking. Plugs were incubated
235 twice for 30 min each in NDS buffer (500 mM EDTA, 10 mM Tris-HCl, 1% sarkosyl, pH 9.5)
236 and at least three times for 30 min in TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0).
237 Plugs were stored at 4°C in fresh TE. For restriction digestion, half plugs were pre-washed

238 for two hours in TE, three times for 20 min each in TE, and three times for 20 min each in
239 300 µL restriction buffer supplemented with 100 µg/mL BSA, all at room temp. Restriction
240 digestion was performed overnight at the recommended temperature in 500 µL of
241 restriction buffer containing 100 U of restriction endonuclease. Digested plugs were
242 washed in 50 mM EDTA pH 8.0 and stored at 4˚C in 50 mM EDTA pH 8.0 before loading.
243 PFGE was performed using 1% agarose gel in 0.5X TBE (Thermo-Fisher) in a CHEF Master
244 XA 170-3670 system (BioRad) with the following parameters: auto algorithm separation
245 range 5 kb - 2 Mb (angle 120˚C, run 6 V/cm, initial switch time 0.22 s, final switch time 3
246 min 24 s, run time 916 min) at 14˚C. DNA was visualized by staining in ethidium bromide
247 (5 µg/mL) and imaging (Gel Doc XR+; BioRad).

248

249 **1002 Yeast Genome project rDNA copy number estimation**

250 Illumina reads from the 1002 Yeast Genome project were obtained from the European
251 Nucleotide Archive (www.ebi.ac.uk/) under accession number ERP014555. We omitted
252 clades with few members, mosaic clades, and unclustered isolates, giving a total of 788
253 isolates. Reads were downsampled to 10-fold-coverage using seqtk() and rDNA copy
254 number for each isolate was calculated using W303-rDNA as the reference. Bin sizes of
255 1/200th of the mean coverage for rDNA and 1/50th for the whole genome, and a window
256 size of 600 bp for both estimates, were used. Violin plots were plotted using the ggplot()
257 package in R.

258

259 **Phylogenetic analyses**

260 To create a neighbour-joining phylogeny based on rDNA copy number values, rDNA copy
261 number for each isolate (after removing 30 isolates for which SNP data were not available)
262 was normalized on a 0-1 scale. Normalized values were used to calculate pairwise
263 Euclidean distances between each pair of isolates to generate a distance matrix that was
264 applied to construct a phylogeny via neighbour-joining using MEGA X [60].

265

266 Phylocorrelograms of copy number and the SNP phylogeny were generated using
267 phylosignal v.1.3 [61] (https://cran.r-project.org/web/packages/phylosignal/index.html).
268 Phylocorrelograms representing a no phylogenetic signal dataset (a "white noise" random
269 distribution) and a high phylogenetic signal dataset (a character evolving on the SNP tree
270 according to a Brownian motion model) were also generated. For the white noise
271 distribution, data were simulated from a normal distribution with mean and standard
272 deviation matching those of the observed copy number data (mean=92.5, sd=30.8). For the
273 Brownian motion model, we first estimated the ancestral mean (z0=83.2) and the rate
274 parameter (σ2=72557.2) from the observed copy number data using the fitContinuous
275 function from geiger [62] (https://cran.r-project.org/package=geiger). Then, we simulated
276 from these parameters on the SNP tree using fastBM from phytools 0.7 (https://cran.r-
277 project.org/package=phytools). Phylocorrelograms were generated for the observed and
278 the two simulated datasets, estimating correlations at a series of 100 phylogenetic
279 distances using 100 bootstrap replicates.

280

281 **Comparing intra-species variation in rDNA copy number**

282 Copy number estimates for twelve isolates from the 1002 Yeast Genome data were
283 randomly drawn 1000 times using a custom bash-script to obtain rDNA copy number
284 ranges.

285

286 **Statistical analyses**

287 All statistical analyses to evaluate differences in rDNA copy number between clades were
288 performed in R. Significance was calculated using the Welch *t*-test (*t*-test), the non-
289 parametric Wilcoxon-Mann-Whitney test (wilcox test) and ANOVA, with *p*-values
290 considered statistically significant at $p < 0.05$.

291 **Results and Discussion**

292 **Establishment of a modal coverage bioinformatics pipeline for estimating rDNA copy**
293 **number**

294 The abundance of data generated from NGS platforms has led a number of studies to use
295 mean read depth to estimate rDNA copy number [5, 12, 25, 34, 49, 55, 56]. However, repeat
296 elements, sequence biases and large-scale changes like aneuploidies can potentially result
297 in non-normal read coverage distributions where the mean coverage does not accurately
298 represent the true coverage. To overcome these limitations, we developed a novel
299 sequence read-based rDNA copy number calculation approach based on the most frequent
300 (modal) coverage. The rationale for this approach is that modal coverage will provide an
301 estimate of the relative coverage representation of a given region in a genome that is more
302 robust to biases away from normality than the mean or median, The approach allocates
303 coverage across a reference genome into coverage bins. The ratio of the most frequently
304 occurring coverage bins for the rDNA and the WG is then used to calculate rDNA copy
305 number (per haploid genome). We implemented this modal coverage approach as a simple
306 pipeline to calculate rDNA copy number from mapped sequence reads (**Fig 2**). To help
307 smooth across positions that stochastically vary in coverage, an issue that is particularly
308 prevalent with very low coverage datasets, we used a sliding window approach to calculate
309 coverage. Our straightforward pipeline uses a sorted BAM file of reads aligned to a
310 reference genome for which the position of the rDNA is known (either embedded in the
311 genome or as a separate contig) to calculate copy number

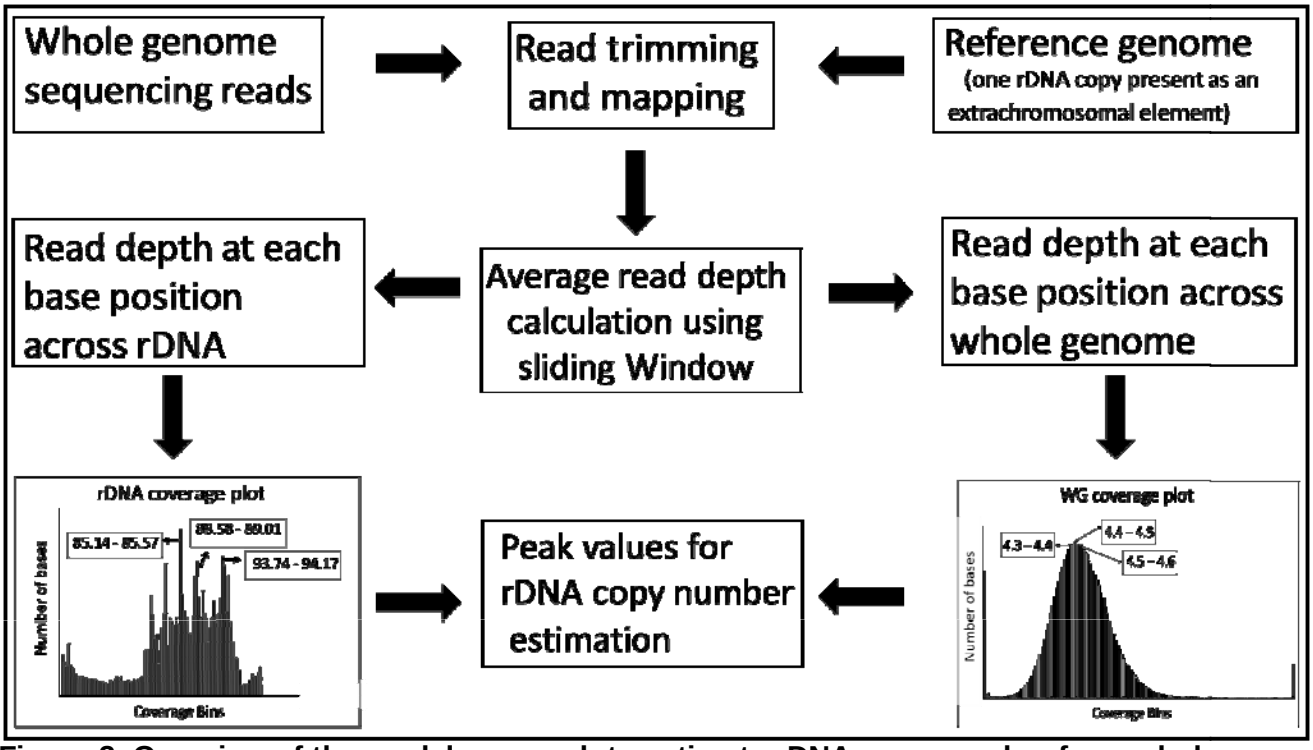

Figure 2. Overview of the modal approach to estimate rDNA copy number from whole genome sequence data. Whole genome (WG) sequence reads are mapped against a reference genome containing a single rDNA copy. Mean read depth for each postion is calculated across the rDNA and the WG using a sliding window, then allocated into coverage bins (shown as histograms). To calculate modal rDNA copy number, the highest frequency coverage bins for both the rDNA and WG are used to compute ratios that represent the rDNA copy number range. The histograms shown were plotted using a 20-copy yeast strain at 5-fold WG coverage with bin sizes of 1/200th of mean coverage for rDNA and 1/50th for WG, and a sliding window of 600 bp for both. The coverage ranges for the three most frequent bins for each are indicated in boxes.

To implement our modal coverage approach, we generated test datasets by performing WG Illumina sequencing of a haploid wild-type laboratory *S. cerevisiae* strain reported to have 150-200 rDNA copies, and three isogenic derivatives where the rDNA has been artificially

327 reduced to 20, 40 and 80 copies, and "frozen" in place through disruption of a gene (*FOB1*)
328 that promotes rDNA copy number change [17] (**Table 1**). Initially, we investigated which
329 parameters provide the most accurate results by applying our pipeline to the WG sequence
330 data obtained from a strain with 20 rDNA copies (20-copy strain). We obtained a genome-
331 wide read coverage of 13.1-fold (**Supplementary Table 1**) and mapped these reads to the
332 W303-rDNA yeast reference genome that has a single rDNA copy. The mapping output was
333 used to determine per-base coverage values, which were placed into coverage bins using a
334 sliding window. We investigated a range of sliding window sizes, from 100 bp (previously
335 reported to have an approximately normal distribution of WG sequence read coverage
336 [63]) to 1,000 bp (large sliding window sizes, whilst smoothing stochastic coverage
337 variation, converge on the mean coverage as the window size approaches the rDNA unit
338 length). We also assessed the impact of coverage on copy number estimation by
339 downsampling the sequence reads. We ran analyses with 100 technical replicates and
340 computed the rDNA copy number means and ranges. We found that, as expected, the
341 accuracy and precision (defined here as <u>similarity to known copy number</u> and <u>copy</u>
342 <u>number range</u>, respectively) of the pipeline was poorer at lower coverage levels, while
343 larger sliding window sizes could compensate for a lack of reads to improve both measures
344 (**Fig 3**). Coverage levels above 10-fold with a sliding window size between 500-800 bp
345 produced accurate rDNA estimates. However, our method also demonstrated adequate
346 performance even with a coverage level of 5-fold, when the sliding window was 600-700
347 bp (**Fig 3**). We found that the method works similarly when just using the rRNA coding
348 region (**Supplementary Figure 3**) rather than the full repeat, which is important as the full
349 rDNA unit sequence is often not available. We also examined the performance of median
350 coverage, but found that while it had greater precision compared to the modal coverage

approach, the accuracy was poorer (**Supplementary Figure 4**). Given the rapid rate at which copy number changes even during vegetative growth [21], the lower precision of our method may more accurately represent the range of copy numbers likely to be present in samples that consist of multiple cells.

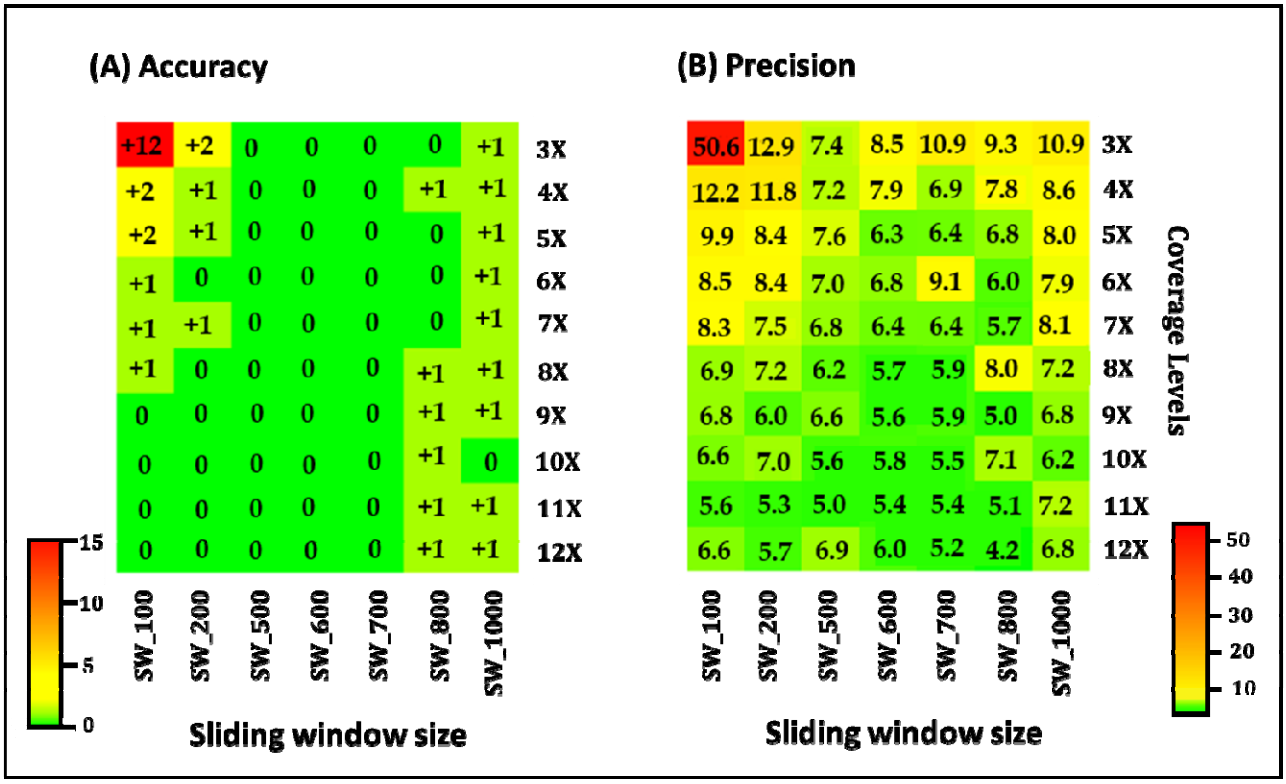

Figure 3. Assessing parameters for rDNA copy number estimation accuracy and precision. Cells represent the (**A**) deviation of the calculated modal rDNA copy number from 20, and (**B**) maximum variation of rDNA copy number calculated from the 100 technical replicates for each coverage level and sliding window (SW) size combination. The heatmap scales used are indicated. In (**A**), rDNA copy number was rounded to the nearest integer.

364 We then assessed the performance of our pipeline with the 40-copy, 80-copy, and WT *S.*
365 *cerevisiae* strain data. Illumina WG sequence reads (**Supplementary Table 1**) obtained
366 from these strains were downsampled to generate 100 technical replicates at 10-fold
367 coverage for each strain, and rDNA copy numbers were calculated using our modal
368 coverage pipeline with a sliding window of 600 bp. The resultant rDNA copy numbers
369 were: 32-40 (x□ = 36 copies) for the 40-copy strain; 57-72 (x□ = 64 copies) for the 80-
370 copy strain; 129-177 (x□ = 157 copies) for the WT strain. These values, while similar to the
371 reported copy numbers for these strains, are not identical. Therefore, to check the actual
372 copy numbers of these strains, and to provide a direct validation of our modal pipeline
373 method, we next experimentally determined the rDNA copy numbers of these strains.

374

375 We chose ddPCR to experimentally determine rDNA copy number because it is less
376 sensitive than qPCR to biases in secondary structure regions that are common in the rDNA
377 coding region [22]. The ddPCR data showed that the rDNA copy numbers of our strains are
378 similar to those calculated by our modal coverage method, with both methods suggesting
379 that the "80-copy" strain actually has substantially fewer copies than reported (**Table 2**;
380 **Supplementary Figure 5A**), perhaps due to a stochastic change in copy number that has
381 occurred in our version of this strain. We also compared our modal coverage approach
382 with the mean coverage calculated from the same datasets. We used a simple mean
383 calculation to match the implementation of our modal approach, using the same down-
384 sampled 10-fold WG coverage datasets. The copy number estimates made using the mean
385 coverage approach were uniformly lower than the other estimates (**Supplementary Table**
386 **2**), which we suggest is the result of sequencing biases against regions in the rRNA coding

387 region. Importantly, correlating read coverage and ddPCR copy number estimates showed
388 the modal coverage slope was closer to the expected value of 1 than the mean coverage
389 slope (**Fig 4**). We also estimated the copy number using pulsed field gel electrophoresis
390 based on the size of the restriction fragment encompassing the entire rDNA array divided
391 by the rDNA unit size (accounting for the sizes of the flanking regions), again with
392 consistent results (**Supplementary Figure 5B,C**). Together, these results suggest the
393 modal coverage approach is an accurate way to estimate rDNA copy number.

394

395

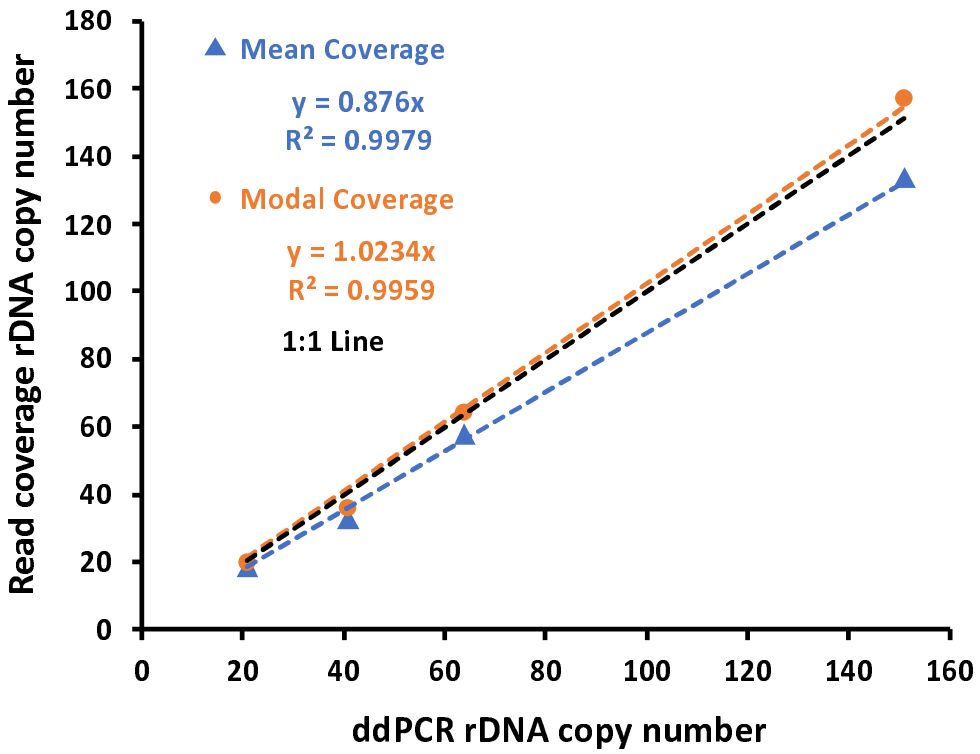

Figure 4. Comparison of modal and mean coverage copy number estimation methods. Plot of rDNA copy number for the 20, 40, 80 and WT *S. cerevisiae* strains (10-fold coverage) calculated using modal (orange line) and mean (blue line) coverage methods versus the copy number determined by ddPCR. The expected 1:1 correlation between read coverage and ddPCR methods is shown in black. Note that while the mean coverage method gives a higher R^2, the modal coverage results are a closer fit to the expected 1:1 line.

Our results suggest that the modal coverage pipeline provides robust estimates of rDNA copy number even when coverage is less than 5-fold. This reliability may partly be a consequence of the larger sliding window size we used compared to that commonly applied for mean coverage methods. It was previously reported that coverage below ~65X

409 results in precision issues when estimating rDNA copy number [5]. However, we did not
410 find this, either for our method or using mean coverage, suggesting that the issues might be
411 specific to the approach or dataset used in that study. The simple implementation of our
412 modal approach coupled with its good performance make it an attractive method for
413 estimating rDNA copy number from sequence read data. Furthermore, a modal approach is
414 expected to be more robust to features that can perturb mean coverage approaches by
415 skewing coverage distributions, such as repeat elements, large duplications and deletions,
416 regions exhibiting sequencing biases, modest sequence divergence from the reference
417 sequence, and aneuploidies [46]. Although we have developed our pipeline for measuring
418 rDNA copy number, in principle it can be used to calculate copy number for any repeated
419 sequence where all reads map to a single repeat copy and the sequence is known, such as
420 mitochondrial and chloroplast genome copy numbers. Given its strong performance, we
421 applied our method to characterize the inter-population distributions of rDNA copy
422 number in *S. cerevisiae*.

423

424 **Within-species evolutionary dynamics of rDNA copy number**

425 Studies in model organisms have provided evidence that each species has a homeostatic
426 copy number which is returned to following copy number perturbations [7-10]. This
427 homeostatic copy number appears to have a genetic basis [5, 26], which suggests it might
428 vary between populations, as well as between species. However, few studies have
429 addressed this question. Given that variation in rDNA copy number has been associated
430 with altered phenotypes [8, 12, 17, 22, 27-35], we decided to undertake a comprehensive

431 assessment of *S. cerevisiae* rDNA copy number at the population level using the global wild
432 yeast dataset from the 1002 Yeast Genome project [64].

433

434 We obtained WG sequence data for 788 isolates from the 1002 Yeast Genome project.
435 Reads for each isolate were downsampled to 10X genome coverage, mapped to our W303-
436 rDNA reference genome, and rDNA copy numbers estimated using our modal coverage
437 pipeline. The rDNA copy numbers ranged between 22-227 ($\bar{x}$ = 92) across the 788
438 isolates (**Supplementary Table 3**). The copy numbers of 11 wild *S. cerevisiae* isolates
439 included in our dataset had previously been estimated [14, 25], and our results are largely
440 consistent with these (**Supplementary Table 4**). However, the copy number we estimate
441 are, in general, much lower than those (~150-200) measured for most laboratory strains
442 (e.g. [17, 38, 41]). We looked to see whether ploidy affects rDNA copy number, given that
443 laboratory strains are predominantly haploid, while the wild *S. cerevisiae* isolates we
444 analyzed are mostly diploid. We observed a small difference in copy number between
445 haploid and diploid isolates (104 vs 91 copies, respectively; **Supplementary Figure 6** and
446 **Supplementary Information**), but overall do not find a strong effect of ploidy on copy
447 number. Thus, the copy number differences between lab and most wild *S. cerevisiae*
448 isolates seem to be a property of these isolates.

449

450 The difference in copy number between lab and wild *S. cerevisiae* isolates suggests that *S.*
451 *cerevisiae* populations may harbor different rDNA copy numbers. To test this, we used the
452 23 phylogenetic clades defined by [64] as proxies for *S. cerevisiae* populations and looked
453 at the distributions of rDNA copy number number within and between these populations

(**Fig 5**). ANOVA analysis rejects homogeneity of rDNA copy number between these populations ($p = 4.37e^{-15}$), suggesting there are population-level differences in copy number within *S. cerevisiae*.

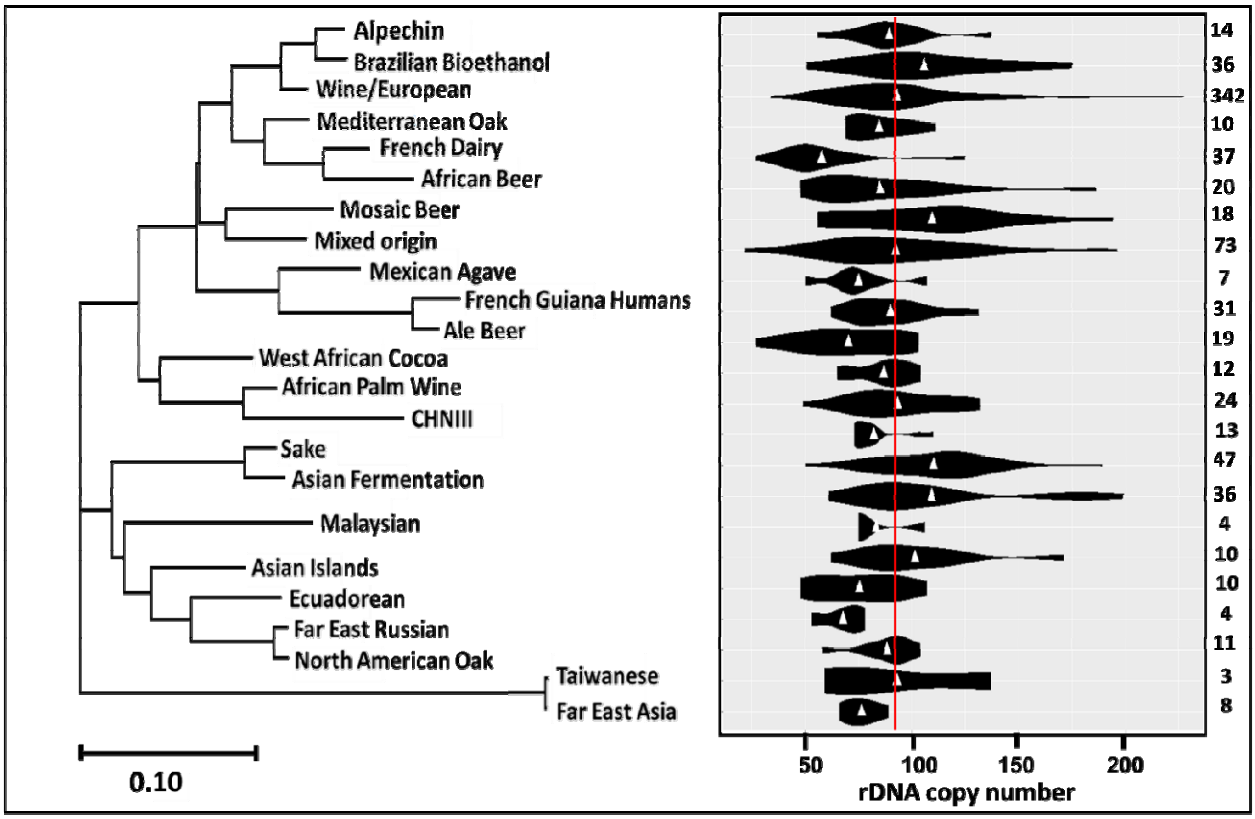

Fig 5. rDNA copy number in *S. cerevisiae* populations. To the left is the phylogeny of the 23 *S. cerevisiae* clades from [64] that encompass the 788 isolates included in this study. The scale represents substitutions per site. On the right, rDNA copy number calculated using the modal coverage method is displayed as a violin plot for each clade with mean population copy numbers indicated by white triangles. Numbers to the right represent the number of isolates in each clade. The red vertical line represents the overall mean rDNA copy number (92 copies). Copy number estimations were determined using 10-fold coverage and a 600 bp sliding window.

467

468 We next wanted to look for complementary evidence that *S. cerevisiae* populations have
469 different rDNA copy numbers, as an alternative explanation for our results is different
470 populations happened to have different copy numbers simply due to stochastic variation
471 [21]. If the stochastic variation explanation is correct, we would expect divergent copy
472 numbers to return to the homeostatic value over time. To test this, we used ddPCR to
473 measure the rDNA copy numbers of six of the 1002 Yeast Genome project isolates that
474 represent the range of copy numbers observed, including one that we had three different
475 isolates of. We grew three biological replicates of each isolate for ~60 generations to allow
476 any fluctuation in rDNA copy number to return to the homeostatic level [7]. The rDNA copy
477 numbers, both before and after the ~60 generations, resemble the copy numbers we
478 estimated from the sequence data and show no tendency to converge on the overall *S.*
479 *cerevisiae* mean copy number (**Table 3; Supplementary Table 5**). These results strongly
480 suggest that our method of estimating rDNA copy number is robust and that the copy
481 numbers of isolates are not recovering towards a common copy number value. From this
482 we conclude that different *S. cerevisiae* populations have different homeostatic rDNA copy
483 numbers.

484

485

Table 3. *S. cerevisiae* rDNA copy number does not recover to a common value following ~60 generations of growth

Isolates	rDNA CN at start[a]	rDNA CN after ~60 generations[a]		Original modal CN estimation[b]
***S. cerevisiae* wild-type** rep1[c]	213	130	185[d]	157
rep2		217		
rep3		208		
YJM981 rep1	174	120	175	171
rep2		183		
rep3		221		
DBVPG1373 rep1	69	77	85	78
rep2		72		
rep3		107		
UWOPS03-461-4[e] rep1	85	113	95	106
rep2		88		
rep3		83		
UWOPS03-461-4[e] **(Mat**a) rep1	244	164	146	
rep2		167		
rep3		106		
UWOPS03-461-4[e] **(Matα)** rep1	ND[f]	108	109	
rep2		115		
rep3		105		
YPS128 rep1	89	87	79	89
rep2		73		
rep3		77		
DBVPG1788 rep1	95	126	108	87
rep2		100		
rep3		97		

[a] Measured using ddPCR

[b] Measured using our modal coverage pipeline

[c] rep: biological replicate

[d] Mean of the three replicates to the nearest integer

[e] UWOPS03-461-4 is the parent isolate of UWOPS03-461-4 Mat**a** and UWOPS03-461-4 Matα derivatives

[f] Not determined

498 Copy number has previously been shown to correlate with phylogeny for species across the
499 fungal kingdom [5]. Given the differences in rDNA copy number we observe, we wondered
500 whether a similar correlation exists between *S. cerevisiae* populations. To test this, we
501 constructed a neighbour-joining phylogeny using rDNA copy number as the phylogenetic
502 character for 758 isolates (30 were removed as SNP data were not available) and
503 compared this to the reported *S. cerevisiae* phylogeny created from genomic SNP data [64].
504 To assess how well the two phylogenies correlate, we used Moran's Index of spatial
505 autocorrelation *I*, which quantifies the correlation between two traits. Moran's *I* indicated a
506 modest positive correlation between rDNA copy number and phylogeny at short
507 phylogenetic distances (**Fig 6**), but not a significant negative correlation at greater
508 phylogenetic distances like that previously observed above the species level [5]. These
509 results suggest that phylogeny only partially explains the distribution of rDNA copy
510 numbers amongst *S. cerevisiae* populations.

511

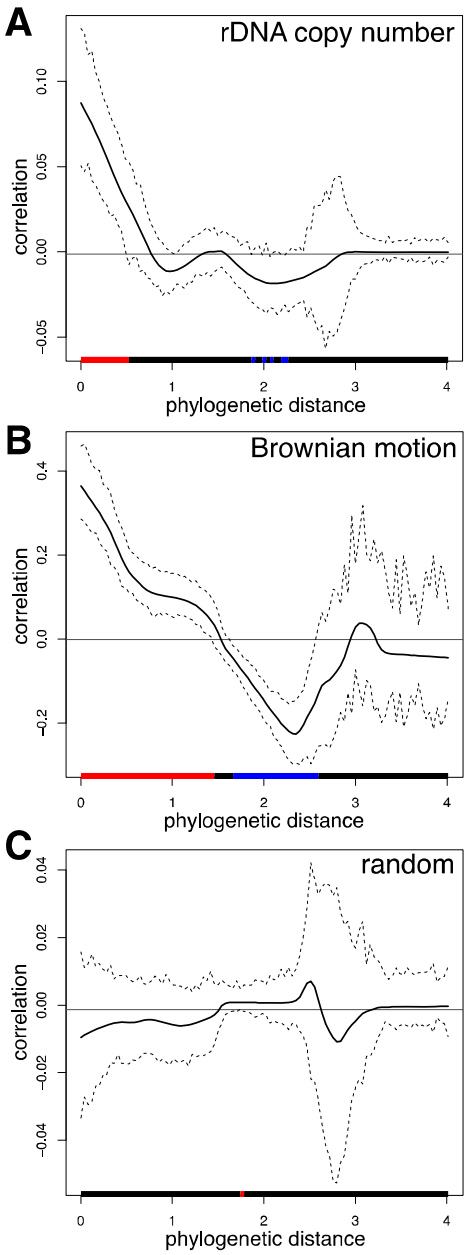

Figure 6. **Phylocorrelograms of autocorrelation based on Moran's *I***. Phylogenetic distance spatial autocorrelations between the SNP-based *S. cerevisiae* phylogeny and the rDNA copy number phylogeny (**A**), a Brownian motion phylogeny (**B**), and random data (**C**) are plotted. Red segments beneath each phylocorrelogram indicate significant positive autocorrelation; black no significant autocorrelation, and blue significant negative

autocorrelation. Dotted lines indicate autocorrelation 95% confidence intervals. Significance is based on comparison to zero phylogenetic autocorrelation (horizontal black line at 0).

Another feature that might explain the distribution of rDNA copy numbers between *S. cerevisiae* populations is the environment, given that nutritional conditions have been proposed to influence copy number [65, 66]. To investigate this, we compared the rDNA copy numbers from two phylogenetically-diverged *S. cerevisiae* populations that are associated with oak trees, which we took as a proxy for similar environments. We found the oak populations did not show significantly different copy numbers (p-value = 0.52), as expected if environment is contributing to copy number. Thus, rDNA copy number might be partially determined by the environmental conditions the population has evolved in. However, we found no consistent pattern of similarities or differences with the copy numbers of the nearest phylogenetic neighbours of these oak clades (**Supplementary Information**), thus we cannot rule out these results simply representing stochastic variation. We suggest that a better understanding of what environmental factors modulate rDNA copy number is necessary before we can properly evaluate the impact of the environment on patterns of rDNA copy number variation.

Finally, we wondered whether large range in estimated *S. cerevisiae* rDNA copy number (22-227 copies) might reflect an unusually large variance in copy number in this species, given this range is almost the same as that reported across 91 different fungal species from three different fungal phyla (11-251 copies, excluding one outlier of 1442 copies) [5]. However, comparing the *S. cerevisiae* copy number range generated by drawing twelve *S.*

cerevisiae isolates at random from our data 1,000 times to that previously measured for twelve isolates of one fungal species (*Suillus brevipes*; [5]) shows that the *S. brevipes* range falls in the middle of the *S. cerevisiae* distribution of copy number ranges (**Fig 7**). These results suggest that *S. cerevisiae* rDNA copy number is no more variable than that of *S. brevipes* at least, and illustrate the tremendous variation in rDNA copy number that is likely to be present in many eukaryotic species.

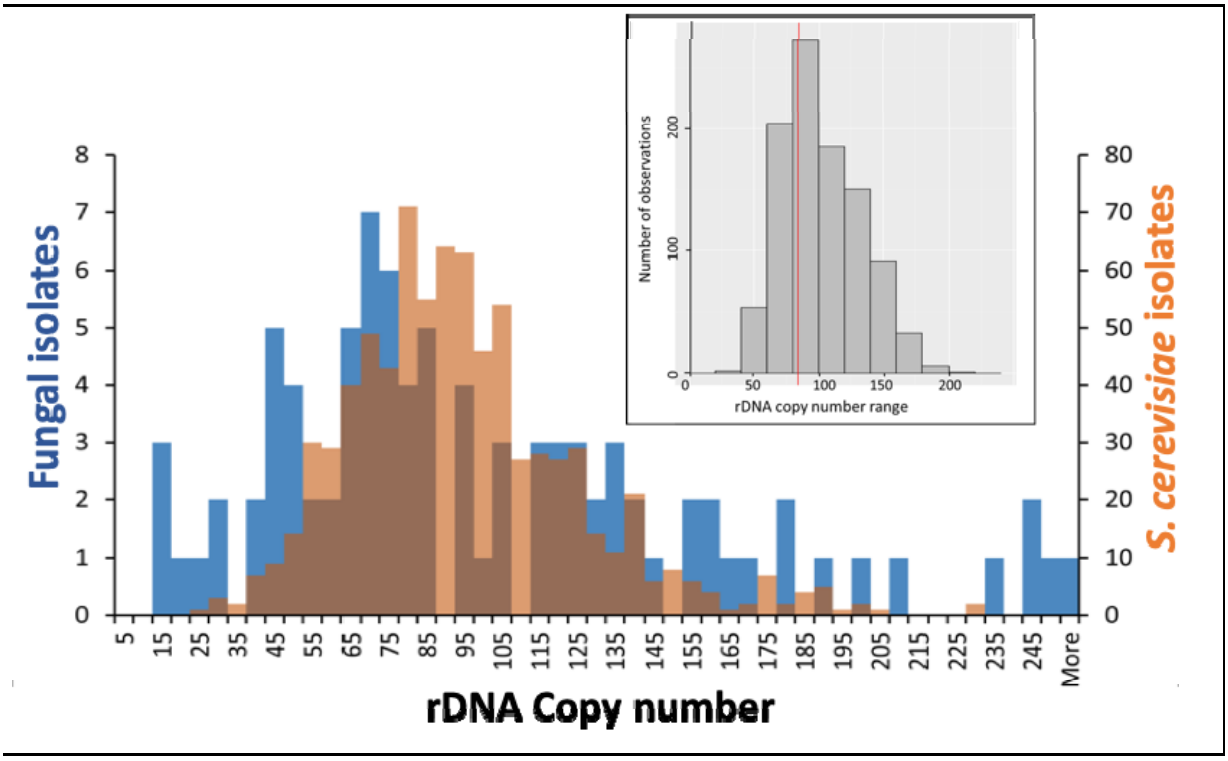

Figure 7. Distribution of rDNA copy number for fungal and *S. cerevisiae* isolates. The main histogram represents rDNA copy number (x-axis) for 91 previously published fungal taxa (blue bars, y-axis on left; [5]) and the 788 *S. cerevisiae* isolates (orange bars; y-axis on right) from this study. Brown represents overlaps. Inset histogram shows the distribution of total rDNA copy number ranges from 1,000 randomly drawn sets of twelve *S. cerevisiae*

isolates. The red vertical line represents the total copy number range (84) observed amongst twelve *Suillus brevipes* isolates [5].

Conclusions

Our results demonstrate that modal coverage can be used to robustly determine rDNA copy number from NGS data. Using our novel approach, we demonstrate that the mean rDNA copy number across all wild *S. cerevisiae* populations is 92. This is substantially lower than the copy numbers documented for lab *S. cerevisiae* strains, but overlaps the 'typical' rDNA copy numbers reported for fungi [5]. We show that *S. cerevisiae* populations have different homeostatic rDNA copy numbers, consistent with a previous study [14]. We found some correlation between rDNA copy number and phylogeny, but not enough to suggest that homeostatic copy number is simply drifting apart with increasing phylogenetic distance. We provide circumstantial evidence that environmental factors might help drive the homeostatic rDNA copy number differences. This is consistent with demonstrations that nutritional factors can induce physiological rDNA copy number changes [65, 66] and that such differences have phenotypic consequences [8, 12, 17, 22, 27-35]. However, it has been shown that rDNA copy number does not correlate with trophic mode in fungi [5]. Therefore, more work is required to determine what really drives copy number dynamics between populations. One caveat to our conclusions is that while studies from a variety of organisms have demonstrated that copy number recovers from perturbation [7-10], presumably as a result of mechanisms maintaining homeostatic copy number [26], some recent studies in *S. cerevisiae* and *Drosophila* have reported the persistence of stochastic copy number changes without recovery [65, 67]. It will be important to reconcile these

577 conflicting results and to determine to what extent the population-level differences we
578 observe are the result of copy number homeostasis (as we interpret them) versus copy
579 number inertia.

580

581 Our results showing population-level differences in rDNA copy number suggest that such
582 differences can arise relatively quickly in evolutionary time, although the very high level of
583 copy number variation between individuals obscures this pattern. Therefore, it is
584 important to take the large variances and rapid copy number dynamics of the rDNA into
585 account when interpreting the impact of copy number variation on phenotype.
586 Bioinformatics pipelines, such as the one we have developed here, in conjunction with the
587 increasing availability of appropriate NGS datasets provide a way to establish baseline data
588 on rDNA copy number variation between cells, individuals, populations, and species, as
589 well as to investigate the phenotypic consequences of this variation. Finally, while we
590 report population-level differences in rDNA copy number in *S. cerevisiae*, diverse human
591 populations have been reported to not differ in rDNA copy number [12, 46]. Whether this
592 reflects a difference in biology (such as differences in the level of genetic divergence
593 between populations) or an incomplete understanding of human population rDNA copy
594 number will require further clarification.

595

596 **Acknowledgements**

597 We thank Gianni Liti (IRCAN, CNRS, INSERM, Université Côte d'Azur, University of Nice) for
598 kindly providing strains. We acknowledge the Centre for eResearch (CER) and

599 New Zealand eScience Infrastructure (NeSI) for providing high performance computing
600 facilities, consulting support, and training services. We thank Auckland Genomics for
601 advice and whole genome sequencing, and Kevin Chang for help with statistical analyses.
602 We thank Alastair Harris for helpful suggestions, and the Ganley Lab for helpful comments
603 and feedback on the manuscript. This work was supported by a grant from the New
604 Zealand Marsden Fund (14-MAU-053) and a University of Auckland Faculty Research
605 Development Grant (3712288), both to ARDG.

606 **References**

607 1. Long EO, Dawid IB. Repeated genes in eukaryotes. Annu Rev Biochem. 1980;49:727-
608 64.

609 2. McStay B, Grummt I. The epigenetics of rRNA genes: From molecular to chromosome
610 biology. Annu Rev Cell Dev Biol. 2008;24:131-57.

611 3. Prokopowich CD, Gregory TR, Crease TJ. The correlation between rDNA copy
612 number and genome size in eukaryotes. Genome. 2003;46:48-50.

613 4. Torres-Machorro AL, Hernández R, Cevallos AM, López-Villasenor I. Ribosomal RNA
614 genes in eukaryotic microorganisms: witnesses of phylogeny? FEMS Microbiol Rev.
615 2010;34:59–86.

616 5. Lofgren LA, Uehling JK, Branco S, Bruns TD, Martin F, Kennedy PG. Genome-based
617 estimates of fungal rDNA copy number variation across phylogenetic scales and ecological
618 lifestyles. Mol Ecol. 2019;28(4):721-30. Epub 2018/12/26. doi: 10.1111/mec.14995.
619 PubMed PMID: 30582650.

620 6. Iida T, Kobayashi T. How do cells count multi-copy genes?: "Musical Chair" model for
621 preserving the number of rDNA copies. Curr Genet. 2019;65(4):883-5. Epub 2019/03/25.
622 doi: 10.1007/s00294-019-00956-0. PubMed PMID: 30904990.

623 7. Kobayashi T, Heck DJ, Nomura M, Horiuchi T. Expansion and contraction of
624 ribosomal DNA repeats in *Saccharomyces cerevisiae*: requirement of replication fork
625 blocking (Fob1) protein and the role of RNA polymerase I. Genes and Development.
626 1998;12:3821-30.

627 8. Hawley RS, Marcus CH. Recombinational controls of rDNA redundancy in *Drosophila*.
628 Annu Rev Genet. 1989;23:87-120.

629 9. Russell PJ, Rodland KD. Magnification of rRNA gene number in a *Neurospora crassa*
630 strain with a partial deletion of the nucleolus organizer. Chromosoma. 1986;93:337-40.

631 10. Rodland KD, Russell PJ. Regulation of ribosomal RNA cistron number in a strain of
632 *Neurospora crassa* with a duplication of the nucleolus organizer region. Biochimica et
633 Biophysica Acta. 1982;697:162-9.

634 11. Lyckegaard EM, Clark AG. Ribosomal DNA and Stellate gene copy number variation
635 on the Y chromosome of *Drosophila melanogaster*. PNAS. 1989;86(6):1944-8. Epub
636 1989/03/01. doi: 10.1073/pnas.86.6.1944. PubMed PMID: 2494656; PubMed Central
637 PMCID: PMCPMC286821.

638 12. Gibbons JG, Branco AT, Yu S, Lemos B. Ribosomal DNA copy number is coupled with
639 gene expression variation and mitochondrial abundance in humans. Nature
640 Communications. 2014;5:4850. Epub 2014/09/12. doi: 10.1038/ncomms5850. PubMed
641 PMID: 25209200.

642 13. Cowen LE, Sanglard D, Calabrese D, Sirjusingh C, Anderson JB, Kohn LM. Evolution of
643 drug resistance in experimental populations of *Candida albicans*. J Bacteriol.
644 2000;182:1515-22.

645 14. West C, James SA, Davey RP, Dicks J, Roberts IN. Ribosomal DNA sequence
646 heterogeneity reflects intraspecies phylogenies and predicts genome structure in two
647 contrasting yeast species. Syst Biol. 2014;63(4):543-54. Epub 2014/04/01. doi:

648 10.1093/sysbio/syu019. PubMed PMID: 24682414; PubMed Central PMCID:
649 PMCPMC4055870.

650 15. Herrera ML, Vallor AC, Gelfond JA, Patterson TF, Wickes BL. Strain-dependent
651 variation in 18S ribosomal DNA Copy numbers in *Aspergillus fumigatus*. J Clin Microbiol.
652 2009;47(5):1325-32. Epub 2009/03/06. doi: 10.1128/JCM.02073-08. PubMed PMID:
653 19261786; PubMed Central PMCID: PMCPMC2681831.

654 16. Stults DM, Killen MW, Pierce HH, Pierce AJ. Genomic architecture and inheritance of
655 human ribosomal RNA gene clusters. Genome Res. 2008;18:13-8.

656 17. Ide S, Miyazaki T, Maki H, Kobayashi T. Abundance of ribosomal RNA gene copies
657 maintains genome integrity. Science. 2010;327:693-6.

658 18. French SL, Osheim YN, Cioci F, Nomura M, Beyer AL. In exponentially growing
659 *Saccharomyces cerevisiae* cells, rRNA synthesis is determined by the summed RNA
660 polymerase I loading rate rather than the number of active genes. Mol Cell Biol.
661 2003;23:1558-68.

662 19. Kobayashi T, Ganley ARD. Recombination regulation by transcription-induced
663 cohesin dissociation in rDNA repeats. Science. 2005;309:1581-4.

664 20. Szostak JW, Wu R. Unequal crossing over in the ribosomal DNA of *Saccharomyces*
665 *cerevisiae*. Nature. 1980;284(3 Apr):426-30.

666 21. Ganley ARD, Kobayashi T. Monitoring the rate and dynamics of concerted evolution
667 in the ribosomal DNA repeats of *Saccharomyces cerevisiae* using experimental evolution.
668 Mol Biol Evol. 2011;28:2883-91. Epub 2011/05/07. PubMed PMID: 21546356.

669 22. Salim D, Gerton JL. Ribosomal DNA instability and genome adaptability.
670 Chromosome Research. 2019;27(1-2):73-87. Epub 2019/01/04. doi: 10.1007/s10577-018-
671 9599-7. PubMed PMID: 30604343.

672 23. Ganley ARD, Kobayashi T. Highly efficient concerted evolution in the ribosomal DNA
673 repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data.
674 Genome Res. 2007;17:184-91.

675 24. Eickbush TH, Eickbush DG. Finely orchestrated movements: evolution of the
676 ribosomal RNA genes. Genetics. 2007;175:477-85.

677 25. James SA, O'Kelly MJT, Carter DM, Davey RP, van Oudenaarden A, Roberts IN.
678 Repetitive sequence variation and dynamics in the ribosomal DNA array of *Saccharomyces*
679 *cerevisiae* as revealed by whole-genome resequencing. Genome Res. 2009;19:626-35.

680 26. Iida T, Kobayashi T. RNA polymerase I activators count and adjust ribosomal RNA
681 gene copy number. Mol Cell. 2019;73(4):645-54. Epub 2019/01/08. doi:
682 10.1016/j.molcel.2018.11.029. PubMed PMID: 30612878.

683 27. Delany ME, Muscarella DE, Bloom SE. Effects of rRNA gene copy number and
684 nucleolar variation on early development: inhibition of gastrulation in rDNA-deficient chick
685 embryos. J Hered. 1994;85(3):211-7. Epub 1994/05/01. doi:
686 10.1093/oxfordjournals.jhered.a111437. PubMed PMID: 8014461.

687 28. Kobayashi T. Regulation of ribosomal RNA gene copy number and its role in
688 modulating genome integrity and evolutionary adaptibility in yeast. Cell Mol Life Sci.
689 2011;68:1395-403.

690 29. Paredes S, Maggert KA. Ribosomal DNA contributes to global chromatin regulation.
691 PNAS. 2009;106:17829-34.

692 30. Paredes S, Branco AT, Hartl DL, Maggert KA, Lemos B. Ribosomal DNA deletions
693 modulate genome-wide gene expression: "rDNA-sensitive" genes and natural variation.
694 PLoS Genet. 2011;7:e1001376.

695 31. Michel AH, Kornmann B, Dubrana K, Shore D. Spontaneous rDNA copy number
696 variation modulates Sir2 levels and epigenetic gene silencing. Genes and Development.
697 2005;19:1199-210.

698 32. Bughio F, Maggert KA. The peculiar genetics of the ribosomal DNA blurs the
699 boundaries of transgenerational epigenetic inheritance. Chromosome Research.
700 2019;27(1-2):19-30. doi: 10.1007/s10577-018-9591-2. PubMed PMID: 30511202; PubMed
701 Central PMCID: PMCPMC6393165.

702 33. Cullis CA. Quantitative variation of ribosomal RNA genes in flax genotrophs.
703 Heredity. 1979;42:237-46.

704 34. Xu B, Li H, Perry JM, Singh VP, Unruh J, Yu Z, et al. Ribosomal DNA copy number loss
705 and sequence variation in cancer. PLoS Genet. 2017;13(6):e1006771. Epub 2017/06/24.
706 doi: 10.1371/journal.pgen.1006771. PubMed PMID: 28640831; PubMed Central PMCID:
707 PMCPMC5480814.

708 35. Zhou J, Sackton TB, Martinsen L, Lemos B, Eickbush TH, Hartl DL. Y chromosome
709 mediates ribosomal DNA silencing and modulates the chromatin state in *Drosophila*. PNAS.

710 2012;109(25):9941-6. Epub 2012/06/06. doi: 10.1073/pnas.1207367109. PubMed PMID:
711 22665801; PubMed Central PMCID: PMCPMC3382510.

712 36. Ritossa FM, Spiegelman S. Localization of DNA complementary to ribosomal RNA in
713 the nucleolus organizer region of *Drosophila melanogaster*. PNAS. 1965;53:737-45. Epub
714 1965/04/01. doi: 10.1073/pnas.53.4.737. PubMed PMID: 14324529; PubMed Central
715 PMCID: PMCPMC221060.

716 37. Wallace H, Birnstiel ML. Ribosomal cistrons and the nucleolar organizer. Biochimica
717 et Biophysica Acta. 1966;114(2):296-310. Epub 1966/02/21. doi: 10.1016/0005-
718 2787(66)90311-x. PubMed PMID: 5943882.

719 38. Schweizer E, MacKechnie C, Halvorson HO. The redundancy of ribosomal and
720 transfer RNA genes in *Saccharomyces cerevisiae*. J Mol Biol. 1969;40:261-77.

721 39. Matsuda K, Siegel A. Hybridization of plant ribosomal RNA to DNA: the isolation of a
722 DNA component rich in ribosomal RNA cistrons. PNAS. 1967;58(2):673-80. Epub
723 1967/08/01. doi: 10.1073/pnas.58.2.673. PubMed PMID: 5234327; PubMed Central
724 PMCID: PMCPMC335687.

725 40. Maleszka R, Clark-Walker GD. Yeasts have a four-fold variation in ribosomal DNA
726 copy number. Yeast. 1993;9:53-8.

727 41. Saka K, Takahashi A, Sasaki M, Kobayashi T. More than 10% of yeast genes are
728 related to genome stability and influence cellular senescence via rDNA maintenance.
729 Nucleic Acids Res. 2016;44(9):4211-21. Epub 2016/02/26. doi: 10.1093/nar/gkw110.
730 PubMed PMID: 26912831; PubMed Central PMCID: PMCPMC4872092.

731 42. Paredes S, Maggert KA. Expression of *I-Cre*I endonuclease generates deletions within
732 the rDNA of Drosophila. Genetics. 2009;181:1661-71.

733 43. Chestkov IV, Jestkova EM, Ershova ES, Golimbet VE, Lezheiko TV, Kolesina NY, et al.
734 Abundance of ribosomal RNA gene copies in the genomes of schizophrenia patients.
735 Schizophrenia Research. 2018;197:305-14. Epub 2018/01/18. doi:
736 10.1016/j.schres.2018.01.001. PubMed PMID: 29336872.

737 44. LeRiche K, Eagle SH, Crease TJ. Copy number of the transposon, *Pokey*, in rDNA is
738 positively correlated with rDNA copy number in *Daphnia obtuse*. PLoS One.
739 2014;9(12):e114773. Epub 2014/12/10. doi: 10.1371/journal.pone.0114773. PubMed
740 PMID: 25490398; PubMed Central PMCID: PMCPMC4260951.

741 45. Son J, Hannan KM, Poortinga G, Hein N, Cameron DP, Ganley ARD, et al. rDNA
742 chromatin activity status as a biomarker of sensitivity to the RNA polymerase I
743 transcription inhibitor CX-5461. Frontiers in Cell and Developmental Biology. 2020;8:568.

744 46. Valori V, Tus K, Laukaitis C, Harris DT, LeBeau L, Maggert KA. Human rDNA copy
745 number is unstable in metastatic breast cancers. Epigenetics. 2020;15(1-2):85-106. Epub
746 2019/07/30. doi: 10.1080/15592294.2019.1649930. PubMed PMID: 31352858; PubMed
747 Central PMCID: PMCPMC6961696.

748 47. Alanio A, Sturny-Leclere A, Benabou M, Guigue N, Bretagne S. Variation in copy
749 number of the 28S rDNA of *Aspergillus fumigatus* measured by droplet digital PCR and
750 analog quantitative real-time PCR. J Microbiol Methods. 2016;127:160-3. Epub
751 2016/06/19. doi: 10.1016/j.mimet.2016.06.015. PubMed PMID: 27316653.

752 48. Salim D, Bradford WD, Freeland A, Cady G, Wang J, Pruitt SC, et al. DNA replication
753 stress restricts ribosomal DNA copy number. PLoS Genet. 2017;13(9):e1007006. Epub
754 2017/09/16. doi: 10.1371/journal.pgen.1007006. PubMed PMID: 28915237; PubMed
755 Central PMCID: PMCPMC5617229.

756 49. Rosato M, Kovarik A, Garilleti R, Rossello JA. Conserved organisation of 45S rDNA
757 sites and rDNA gene copy number among major clades of early land plants. PLoS One.
758 2016;11(9):e0162544. Epub 2016/09/14. doi: 10.1371/journal.pone.0162544. PubMed
759 PMID: 27622766; PubMed Central PMCID: PMCPMC5021289.

760 50. Xu J, Xu Y, Yonezawa T, Li L, Hasegawa M, Lu F, et al. Polymorphism and evolution of
761 ribosomal DNA in tea (*Camellia sinensis*, Theaceae). Mol Phylogen Evol. 2015;89:63-72.
762 Epub 2015/04/15. doi: 10.1016/j.ympev.2015.03.020. PubMed PMID: 25871774.

763 51. Xu J, Zhang Q, Xu X, Wang Z, Qi J. Intragenomic variability and pseudogenes of
764 ribosomal DNA in stone flounder *Kareius bicoloratus*. Mol Phylogen Evol. 2009;52(1):157-
765 66. Epub 2009/04/08. doi: 10.1016/j.ympev.2009.03.031. PubMed PMID: 19348952.

766 52. Agrawal S, Ganley ARD. Complete sequence construction of the highly repetitive
767 ribosomal RNA gene repeats in eukaryotes using whole genome sequence data. Methods in
768 Molecular Biology. 2016;1455:161-81. Epub 2016/09/01. doi: 10.1007/978-1-4939-3792-
769 9_13. PubMed PMID: 27576718.

770 53. Buckler ES, Ippolito A, Holtsford TP. The evolution of ribosomal DNA: Divergent
771 paralogues and phylogenetic implications. Genetics. 1997;145(March):821-32.

772 54. Mayol M, Rosselló JA. Why nuclear ribosomal DNA spacers (ITS) tell different stories
773 in *Quercus*. Mol Phylogen Evol. 2001;19:167-76.

774 55. Wang M, Lemos B. Ribosomal DNA copy number amplification and loss in human
775 cancers is linked to tumor genetic context, nucleolus activity, and proliferation. PLoS Genet.
776 2017;13(9):e1006994. Epub 2017/09/08. doi: 10.1371/journal.pgen.1006994. PubMed
777 PMID: 28880866; PubMed Central PMCID: PMCPMC5605086.

778 56. Gong W, Marchetti A. Estimation of 18S gene copy number in marine eukaryotic
779 plankton using a next-generation sequencing approach. Frontiers in Marine Science.
780 2019;6:219.

781 57. Cubillos FA, Louis EJ, Liti G. Generation of a large set of genetically tractable haploid
782 and diploid *Saccharomyces* strains. FEMS Yeast Res. 2009;9(8):1217-25. Epub 2009/10/21.
783 doi: 10.1111/j.1567-1364.2009.00583.x. PubMed PMID: 19840116.

784 58. Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, et al. Population
785 genomics of domestic and wild yeasts. Nature. 2009;458(7236):337-41. Epub 2009/02/13.
786 doi: 10.1038/nature07743. PubMed PMID: 19212322; PubMed Central PMCID:
787 PMCPMC2659681.

788 59. Cox MP, Peterson DA, Biggs PJ. SolexaQA: At-a-glance quality assessment of Illumina
789 second-generation sequencing data. BMC Bioinformatics. 2010;11:485. Epub 2010/09/30.
790 doi: 10.1186/1471-2105-11-485. PubMed PMID: 20875133; PubMed Central PMCID:
791 PMCPMC2956736.

792 60. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular evolutionary
793 genetics analysis across computing platforms. Mol Biol Evol. 2018;35(6):1547-9. Epub
794 2018/05/04. doi: 10.1093/molbev/msy096. PubMed PMID: 29722887; PubMed Central
795 PMCID: PMCPMC5967553.

796 61. Keck F, Rimet F, Bouchez A, Franc A. phylosignal: an R package to measure, test, and
797 explore the phylogenetic signal. Ecology and Evolution. 2016;6(9):2774-80. Epub
798 2016/04/12. doi: 10.1002/ece3.2051. PubMed PMID: 27066252; PubMed Central PMCID:
799 PMCPMC4799788.

800 62. Pennell MW, Eastman JM, Slater GJ, Brown JW, Uyeda JC, FitzJohn RG, et al. geiger
801 v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic
802 trees. Bioinformatics. 2014;30(15):2216-8. Epub 2014/04/15. doi:
803 10.1093/bioinformatics/btu181. PubMed PMID: 24728855.

804 63. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy
805 number variants using read depth of coverage. Genome Res. 2009;19(9):1586-92. Epub
806 2009/08/07. doi: 10.1101/gr.092981.109. PubMed PMID: 19657104; PubMed Central
807 PMCID: PMCPMC2752127.

808 64. Peter J, De Chiara M, Friedrich A, Yue JX, Pflieger D, Bergstrom A, et al. Genome
809 evolution across 1,011 *Saccharomyces cerevisiae* isolates. Nature. 2018;556(7701):339-44.
810 Epub 2018/04/13. doi: 10.1038/s41586-018-0030-5. PubMed PMID: 29643504; PubMed
811 Central PMCID: PMCPMC6784862.

812 65. Aldrich JC, Maggert KA. Transgenerational inheritance of diet-induced genome
813 rearrangements in Drosophila. PLoS Genet. 2015;11(4):e1005148. Epub 2015/04/18. doi:

814 10.1371/journal.pgen.1005148. PubMed PMID: 25885886; PubMed Central PMCID:
815 PMCPMC4401788.

816 66. Jack CV, Cruz C, Hull RM, Keller MA, Ralser M, Houseley J. Regulation of ribosomal
817 DNA amplification by the TOR pathway. PNAS. 2015;112(31):9674-9. Epub 2015/07/22.
818 doi: 10.1073/pnas.1505015112. PubMed PMID: 26195783; PubMed Central PMCID:
819 PMCPMC4534215.

820 67. Mansisidor A, Molinar T, Srivastava P, Dartis DD, Pino Delgado A, Blitzblau HG, et al.
821 Genomic copy-number loss is rescued by self-limiting production of DNA circles. Mol Cell.
822 2018;72(3):583-93. Epub 2018/10/09. doi: 10.1016/j.molcel.2018.08.036. PubMed PMID:
823 30293780; PubMed Central PMCID: PMCPMC6214758.

824