SCIENCE CHINA
Life Sciences

•COVER ARTICLE•

May 2022 Vol.65 No.5: 851–860
https://doi.org/10.1007/s11427-021-2044-x

Exogenous artificial DNA forms chromatin structure with active transcription in yeast

Jianting Zhou[1,2†], Chao Zhang[3†], Ran Wei[3†], Mingzhe Han[1,2], Songduo Wang[4,5], Kaiguang Yang[4], Lihua Zhang[4], Weigang Chen[6], Mingzhang Wen[1,2], Cheng Li[3], Wei Tao[3*] & Ying-Jin Yuan[1,2*]

[1]*Frontier Science Center for Synthetic Biology and Key Laboratory of Systems Bioengineering (Ministry of Education), Tianjin University, Tianjin 300072, China;*
[2]*SynBio Research Platform, Collaborative Innovation Center of Chemical Science and Engineering (Tianjin), School of Chemical Engineering and Technology, Tianjin University, Tianjin 300072, China;*
[3]*The MOE Key Laboratory of Cell Proliferation and Differentiation, School of Life Sciences, Peking University, Beijing 100871, China;*
[4]*CAS Key Laboratory of Separation Science for Analytical Chemistry, National Chromatographic R. & A. Center, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China;*
[5]*University of Chinese Academy of Sciences, Beijing 100049, China;*
[6]*School of Microelectronics, Tianjin University, Tianjin 300072, China*

Received October 2, 2021; accepted December 10, 2021; published online December 27, 2021

Yeast artificial chromosomes (YACs) are important tools for sequencing, gene cloning, and transferring large quantities of genetic information. However, the structure and activity of YAC chromatin, as well as the unintended impacts of introducing foreign DNA sequences on DNA-associated biochemical events, have not been widely explored. Here, we showed that abundant genetic elements like TATA box and transcription factor-binding motifs occurred unintentionally in a previously reported data-carrying chromosome (dChr). In addition, we used state-of-the-art sequencing technologies to comprehensively profile the genetic, epigenetic, transcriptional, and proteomic characteristics of the exogenous dChr. We found that the data-carrying DNA formed active chromatin with high chromatin accessibility and H3K4 tri-methylation levels. The dChr also displayed highly pervasive transcriptional ability and transcribed hundreds of noncoding RNAs. The results demonstrated that exogenous artificial chromosomes formed chromatin structures and did not remain as naked or loose plasmids. A better understanding of the YAC chromatin nature will improve our ability to design better data-storage chromosomes.

yeast artificial chromosome, DNA storage, epigenetics, chromatin accessibility, histone H3K4 tri-methylation

Citation: Zhou, J., Zhang, C., Wei, R., Han, M., Wang, S., Yang, K., Zhang, L., Chen, W., Wen, M., Li, C., et al. (2022). Exogenous artificial DNA forms chromatin structure with active transcription in yeast. Sci China Life Sci 65, 851–860. https://doi.org/10.1007/s11427-021-2044-x

INTRODUCTION

Yeast artificial chromosomes (YACs) play an important role in genome sequencing (Sasaki et al., 2005), gene cloning (Kouprina and Larionov, 2008), and transgenesis (Lamb and Gearhart, 1995). Yeast cells can carry an additional YAC up to several megabase pairs in length (Marschall et al., 1999). YACs have chromosome elements with autonomously replicating sequences (ARS) and centromeric sequences as mini-chromosomes. YACs also persist in the host cell, with approximately one copy per haploid genome (Tschumper and Carbon, 1983). The presence of a centromere sequence

†Contributed equally to this work
*Corresponding authors (Wei Tao, email: weitao@pku.edu.cn; Ying-Jin Yuan, email: yjyuan@tju.edu.cn)

confers high segregational stability to YACs during mitosis (Hieter et al., 1985). YACs have also been considered to be yeast centromeric plasmids (YCps) (Gnügge and Rudolf, 2017). YAC copy number across individual cells varies, and many cells contain more than one. Asymmetric segregation of YACs during mitosis occurs at a frequency of about 10% per plasmid pair (Gnügge and Rudolf, 2017). *In vivo*, chromosomal DNA is wrapped around histones, which contributes to the formation of chromatin with three-dimensional structures (Keung et al., 2015). Chromatin accessibility and histone modification reflect the epigenetic state of the chromatin (Klemm et al., 2019; Lee et al., 2021) and differ significantly from that found in the plasmids. Studying the epigenetic features of exogenous DNA in yeast can provide a better understanding of YACs.

Due to the storage capacity of long DNA molecules and the high stability of replication, YACs have been a powerful tool for genome synthesis and pathway assembly in the rapidly progressing field of synthetic biology. In 2010, Gibson et al. assembled a 1.08 Mb artificial genome for *Mycoplasma mycoides* JCVI-syn1.0 into a YAC, allowing for the successful creation of the first microorganism controlled by a chemically synthesized genome (Gibson et al., 2010). Because of the high-efficiency of homologous recombination in *Saccharomyces cerevisiae*, subsequent genome synthesis has almost always used YACs as vectors for large DNA assembly; for example, the *M. mycoides* JCVI-syn3.0 minimal genome (Hutchison Iii et al., 2016), the codon recorded genomes of *Escherichia coli* (Fredens et al., 2019; Wang et al., 2016; Zhou et al., 2016) and the redesigned *S. cerevisiae* chromosomes (Kannan and Gibson, 2017; Shen et al., 2017; Wu et al., 2017; Xie et al., 2017; Zhang et al., 2017). Although yeast episomal plasmids (YEps) and yeast integrating plasmids (YIps) have been widely used in pathway construction for chemicals and pharmaceuticals production (Luo et al., 2019; Ro et al., 2006), constructing large pathways with YAC may also prove to be a valuable application of this tool (Srinivasan and Smolke, 2020). In 2021, Postma et al. constructed an orthogonal, supernumerary chromosome with YAC for modular pathway assembly in *S. cerevisiae* (Postma et al., 2021). The work paved the way for *de novo* designer YACs as a platform for rewiring native cellular processes.

Recently, our group constructed a 254 kb YAC for data storage (Chen et al., 2021; Lu and Ellis, 2021). This artificial chromosome contained information obtained from two pictures and one video clip and incorporated five replication sites and one centromere sequence. We used superposition with sparsified low-density parity-check (LDPC) codes and pseudo-random sequences to encode the images. During the procedure, no additional measures were used to avoid introducing genetic elements with potential biological activities, such as promoters, regulatory sequences, protein/RNA binding, and coding sequences. The artificial chromosome was constructed using transformation-associated recombination in *S. cerevisiae* and was stably replicated (Chen et al., 2021; Kouprina and Larionov, 2008). We systematically examined the stability and fidelity of a data-carrying chromosome (dChr) through a series of batch cultures in selective media. After 100 generations, no mutations were observed in the tested samples. Moreover, dChr in yeast was reported to have little effect on yeast growth (Chen et al., 2021). However, more in-depth exploration is required to determine how dChrs affect gene expression and chromatin formation in the host organism. As this was the first exogenous YAC designed independent of a DNA sequence in nature, its genetic and epigenetic activity in living cells is unclear.

Here, we comprehensively mapped the genetic, epigenetics, and transcriptional characteristics of the dChr constructed previously. We found that the dChr forms active chromatin with high accessibility and high histone H3K4 trimethylation (H3K4me3) levels, factors that may facilitate the widespread transcription of the artificial chromosome. In addition, we observed that the circular dChr was organized in the Rabl configuration, which may contribute to dChr self-replication and maintenance of haploidy *in vivo* during yeast replication. Taken together, this work explores the chromatin state, transcriptional activity, and conformation of an exogenous artificial chromosome and provides extensive guidance for the design of *in vivo* DNA storage methods.

RESULTS

Abundant genetic elements in the data-carrying chromosome

We performed a local alignment search to compare the data-carrying sequence with native yeast genome sequences and found only four hits, all of which were ARS short sequences. Next, we analyzed the sequence and searched for potential promoter, TATA box, and transcription factor-binding motif sequences in the dChr. We found 54 putative promoters and 36 TATA boxes (Figure 1A). We also found transcription factor-binding motifs that were unintentionally generated in the dChr, including 100 STP4 and 98 UGA3 motifs (Figure 1B). These motifs play an important role in DNA replication and transcriptional activation (Sylvain et al., 2011; Tkach et al., 2012). The number of motifs per unit length of transcription factors in the dChr is comparable to that found in natural yeast chromosomes (Figure S1 in Supporting Information). These results suggest a high abundance of genetic elements were generated by chance in the artificial chromosome.

To better understand the molecular and conformational features of the dChr *in vivo*, we used state-of-the-art next-generation sequencing technologies to profile the

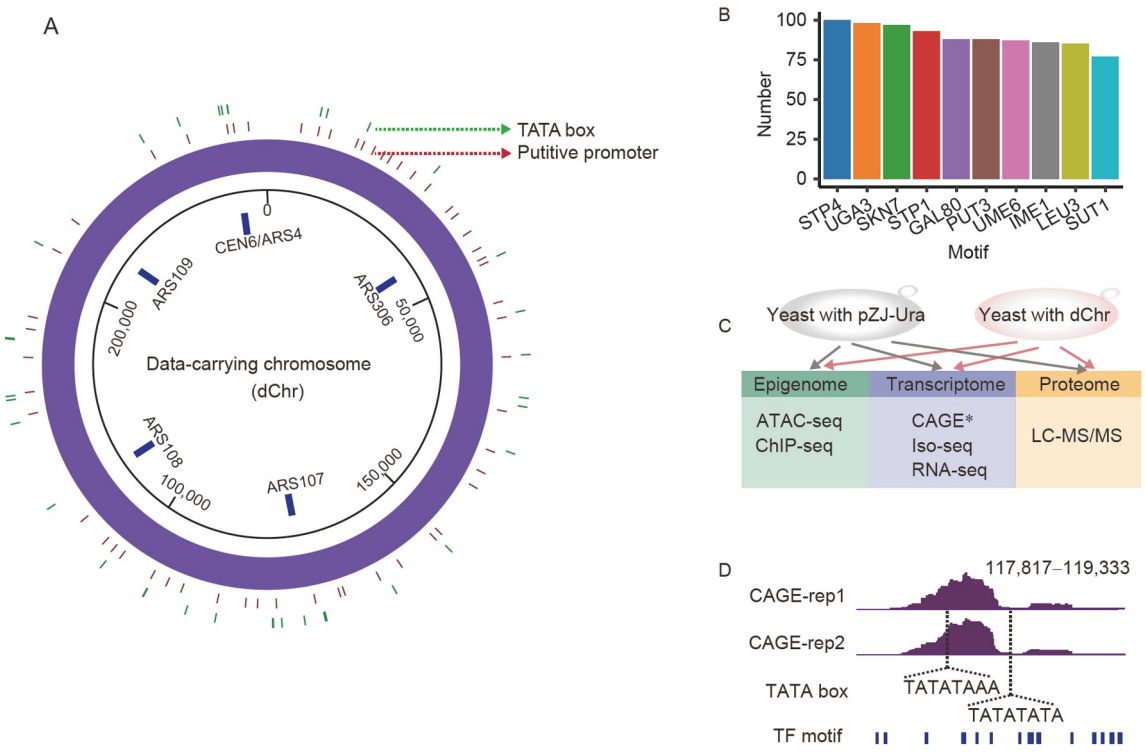

Figure 1 Genetic elements in the data-carrying DNA. A, Circle plot displaying the elements in the data-carrying chromosome. The locations of the five designed ARSs are marked in blue. B, Number of top 10 motifs in dChr. C, Schematic diagram of sequencing technologies and yeasts used in this study. Yeast with pZJ-Ura was used as a control to compare the effects of data-carrying DNA on the host. "*" CAGE was performed using only yeast with dChr. D, IGV visualization of CAGE signals in a dChr example region. TATA boxes and TF motifs are also shown.

epigenome, transcriptome, and proteome. We employed (i) the assay for transposase-accessible chromatin using sequencing (ATAC-seq) and chromatin immunoprecipitation with sequencing (ChIP-seq) to understand the chromatin state; (ii) cap-analysis of gene expression (CAGE), mRNA-seq, and long-read isoform sequencing (Iso-seq) to study the transcriptional characteristics; and (iii) in-depth proteome analysis using LC-MS/MS to identify the translated proteins. To compare the impact of the YACs between host cells with and without the data-carrying sequence, yeast with an empty vector, pZJ-Ura, was used as a control (Figure 1C). All data were generated with 2–3 biological replicates (Table S1 in Supporting Information). To validate the potential transcriptional start sites (TSSes) and promoters previously identified by computational sequence analysis in the dChr (Figure 1A), we used CAGE analysis. CAGE is an efficient method for genome-wide mapping of potential TSSes and is widely used to detect the locations of TSSes (Shiraki et al., 2003). We detected 1,606 TSSes across the yeast native genome, 97.6% (1,568 of 1,606) of which were previously reported as yeast gene TSSes. This result confirmed the high quality of our CAGE data. In addition, we found 20 *de novo* TSSes in the dChr as well as TATA boxes near these TSSes (Figure 1A and D), which demonstrated that TSSes were incorporated into the dChr DNA sequence. PolII may bind to these TSSes and further activate the transcription machine, producing stable RNAs.

Data-carrying chromosome forms highly accessible chromatin with active histone modifications

As there was no expectation that YAC would not form chromatin, we hypothesized that the data-carrying DNA might also form chromatin. To test this hypothesis, we profiled dChr chromatin accessibility using ATAC-seq and then profiled chromatin activity using histone H3K4me3 ChIP-seq (Figure S2A in Supporting Information). Data-carrying DNA chromatin accessibility was significantly lower than that of mitochondrial (mt) DNA, which does not possess histones (Figure 2A). This suggested that data-carrying DNA is not naked but is instead wrapped around histones. In addition, data-carrying DNA chromatin accessibility was distinctly higher than that of wild-type yeast DNA (Figure 2B), demonstrating that the dChr was more accessible than the wild-type yeast chromosome.

Like chromatin accessibility, H3K4me3 histone modification has been reported as an active epigenetic marker for promoter elements (Chong et al., 2020). We performed H3K4me3 ChIP-seq and found that H3K4me3 was enriched in wild-type yeast gene promoters (Figure S2B in Supporting Information). In addition, we found that there were 44 ChIP-seq peaks in the dChr after immunoprecipitation of the

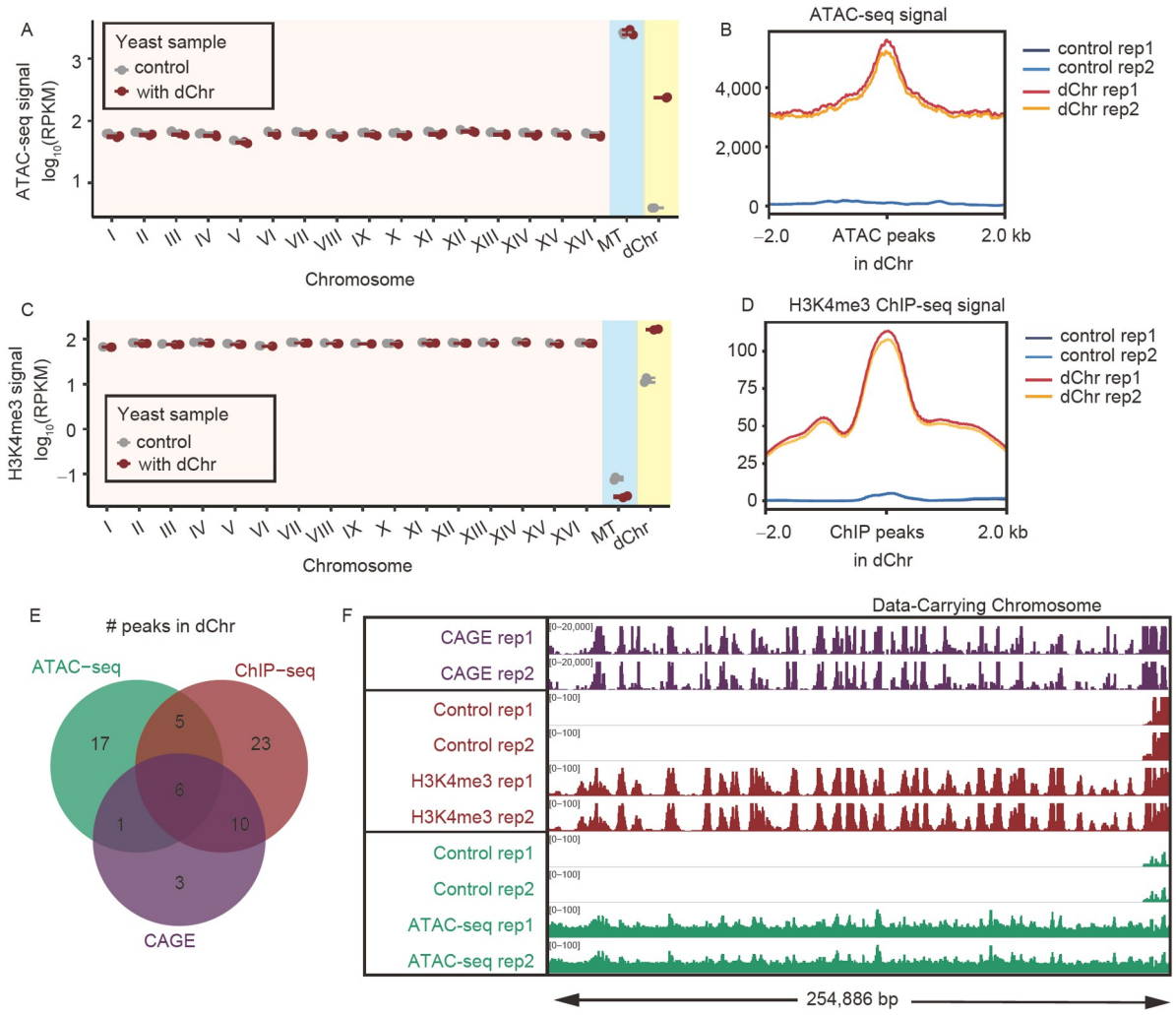

Figure 2 Data-carrying chromosome forms highly accessible chromatin with active histone modifications. A, Average normalized ATAC-seq signals among the chromosomes. The normalized signal is calculated by RPKM (reads per kilobase per million mapped reads for each chromosome). The gray dots represent the control yeast without dChr; the red dots represent the yeast with dChr. Two biological replicates are shown. B, Distribution of normalized ATAC-seq signals around ATAC-seq peaks in dChr. Two biological replicates are shown. C, Average normalized H3K4me3 ChIP-seq signals among all chromosomes. The normalized signal method is the same as the ATAC-seq signal. D, Distribution of normalized H3K4me3 ChIP-seq signals around H3K4me3 peaks located in dChr. E, Venn diagram showing the overlap between ATAC-seq, ChIP-seq, and CAGE peak regions in dChr. F, IGV visualization of CAGE, H3K4me3 ChIP-seq, and ATAC-seq signals for dChr. Peaks are visualized only on the pZJ-Ura sequence on the right side of the plots in the control.

H3K4me3-marked chromatin. The average H3K4me3 level was higher in the dChr than in the wild-type yeast chromosome (Figure 2C and D). This finding demonstrated that the data-carrying artificial DNA could form chromatin with H3K4me3 modifications. Moreover, no global changes of H3K4me3 level in the wild-type yeast chromosomes were found when comparing yeasts with and without the data-carrying DNA (Figure 2C; Figure S2C in Supporting Information), indicating that the data-carrying DNA did not affect the chromatin state of wild-type yeast chromosomes significantly.

ATAC-seq and H3K4me3 ChIP-seq signals were also enriched in promoters (Chong et al., 2020; Klemm et al., 2019). We found that the peaks in the CAGE, ATAC-seq, and H3K4me3 ChIP-seq signals overlapped significantly (Figure 2E). Visualization of the signals in the dChr with Integrative Genomics Viewer (IGV) showed the same results (Figure 2F) and also indicated that functional promoters were present in the dChr. Taken together, data-carrying DNA formed chromatin with histones *in vivo*, and the artificial chromatin displayed an active epigenetic state with high chromatin accessibility and H3K4me3 modification levels as well as active promoters.

The circular data-carrying chromosome is arranged in the Rabl configuration in the nucleus

The dChr was designed with five ARS and one centromere (Figure 1A), which are responsible for maintaining the haploid copy during yeast replication (Chen et al., 2021). A

previous study demonstrated that the yeast chromosome was organized in a Rabl configuration (Kim et al., 2017; Pouokam et al., 2019; Wang et al., 2015) and featured centromeres from different chromosomes grouped together spatially. However, how the dChr maintains haploidy and chromatin conformation *in vivo* is unclear.

We used Hi-C (Kim et al., 2017; Pouokam et al., 2019) to investigate the organization of the dChr in the host nucleus. First, we focused on the dChr cis-interaction map. The contact probability of wild-type yeast chromosomes decayed with genome distance, consistent with previous studies (Kim et al., 2017; Lieberman-Aiden et al., 2009; Zhang et al., 2020). However, the contact probability of the dChr increased substantially after the initial decrease (Figure S3A in Supporting Information). This result was confirmed by visualizing the dChr interaction heatmap (Figure 3A). There was a clear, high contact frequency from the beginning to the end of the dChr sequence, indicating a circular conformation *in vivo*. Next, we inspected the trans-interaction (inter-chromosome) matrix and confirmed the yeast chromosomes exhibited a Rabl configuration (Figure 3B and C). The dChr uniformly interacted with wild-type yeast chromosomes and showed no preference for interacting with specific chromosomes (Figure S3B and C in Supporting Information). In addition, the centromere of the dChr co-organized with those of the wild-type yeast chromosome (Figure 3B–D), contributing to the Rabl conformation. These findings suggested that the synthetic artificial centromere DNA sequence plays an important role *in vivo*. The organization of the dChr in the nucleus suggested that the artificial centromere could function in sister chromatid segregation during yeast replication and thereby maintain the haploidy of the artificial chromosome.

Highly pervasive transcription of the data-carrying chromosome

Both the ATAC-seq and H3K4me3 ChIP-seq results suggest that the dChr can be actively transcribed. To confirm this observation, we used RNA-seq to determine the transcriptome profile of the dChr. Indeed, the dChr showed highly pervasive transcription (Figure 4A and B), and its average expression level was higher than in the wild-type yeast chromosome (Figure S4A in Supporting Information). This result was consistent with the finding that the data-carrying DNA formed more highly accessible chromatin and had higher H3K4me3 levels than wild-type yeast chromosomes.

We next wondered what transcript isoforms were transcribed by the dChr. To answer this question, we performed Iso-seq using PacBio single-molecule real-time sequencing. Full-length Iso-seq can capture the whole transcript isoform in a single sequencing read without breaking the mRNA fragment (Sharon et al., 2013), making it more precise for identifying novel isoforms (Treutlein et al., 2014). We found 488 *de novo* isoforms in the dChr, most of which contained only one exon (473 of 488) (Figure S4B in Supporting Information). The average length of the dChr isoforms was ~3.6 kb (but ranged from 159 to 10,132 bp), which was longer than the wild-type yeast isoforms (~1.3 kb) (Figure 4C). The Iso-seq data confirmed that many long isoforms were transcribed from the dChr.

While systematically analyzing the transcriptional profile, we also examined whether any transcripts from the dChr were translated into polypeptides. As genome-encoded functional short peptides naturally exist, and long noncoding RNAs (ncRNAs) can also contain internal sequences that code for small proteins (Choi et al., 2019), we set the minimum length cutoff to 10 codons in open reading frames (ORF) Finder (http://www.bioinformatics.org/sms2/orf_find.html). The entire dChr sequence includes 9,202 predicted proteins (Supplementary Materials), none of which overlapped significantly (>7 a.a.) with genome-encoded proteins. This thus rationalizes the feasibility of LC-MS/MS analysis of the data-carrying yeast, which revealed a total of 13,261 peptides that mapped to 1,999 proteins. Despite the sufficient coverage of our proteomic analysis, indicated by a broad range of protein abundance (6 orders of magnitude) (Figure S5 in Supporting Information), none of the 9,202 predicted proteins were detected. The large number of transcripts from dChr were indeed ncRNAs.

DISCUSSION

In this study, we described the transcriptional landscape and epigenetic modifications of a fully exogenous YAC containing an audio-visual data sequence *in vivo* for the first time. Our results are useful for informing the design of many such data-storage chromosomes in the future. A previous study showed that the data-carrying yeast had a phenotype similar to that of wild-type yeast (Chen et al., 2021). The dChr was highly and pervasively transcribed, consistent with the pervasive nature of transcription in eukaryotes (Lu and Lin, 2019; Tudek et al., 2015).

ncRNAs are derived from sense or antisense coding regions and intergenic regions in yeast (Tudek et al., 2015). They also possess regulatory functions, such as regulating yeast colonies and biofilms (Wilkinson et al., 2018). We asked how the expression of wild-type yeast genes was affected by the inserted dChr. RNA-seq data showed that many genes involved in amino acid biosynthesis were down-regulated, and the up-regulated yeast native host proteins involved in protein folding and unfolded protein binding included *HSP30*, *HSP42*, *HSP104*, *SSA4* (Figure S4C and D in Supporting Information). A previous study showed that

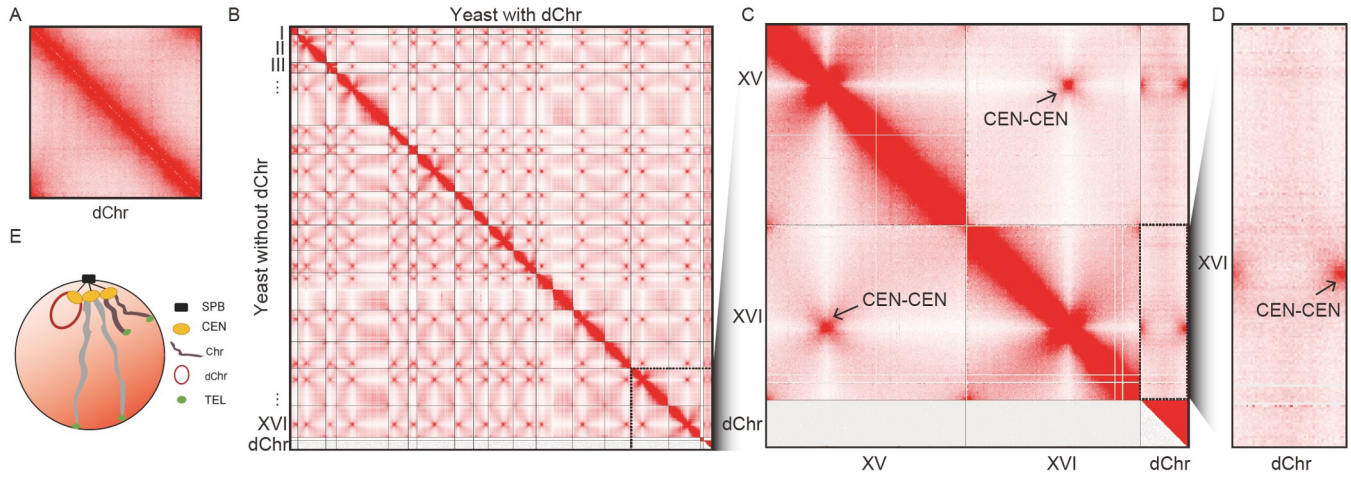

Figure 3 The circular data-carrying chromosome is organized into a RabI configuration in the nucleus. A, Hi-C interaction heatmap of dChr (1 kb resolution). B, Genome-wide interaction heatmap of the yeast with (top right corner) and without dChr (bottom left corner). C, Magnification of the boxed regions in (B), showing interaction heatmap for XV, XVI, and dChr chromosomes. Arrows indicate centromere-centromere interactions and where XV centromere region frequently interacts with XVI centromere regions. D, Further magnification of the boxed regions in (C) showing interactions between XVI and dChr. The arrows indicate high frequency interaction loci between the XVI centromere region and dChr centromere regions. E, Schematic diagram of chromatin conformation for yeast genome and dChr.

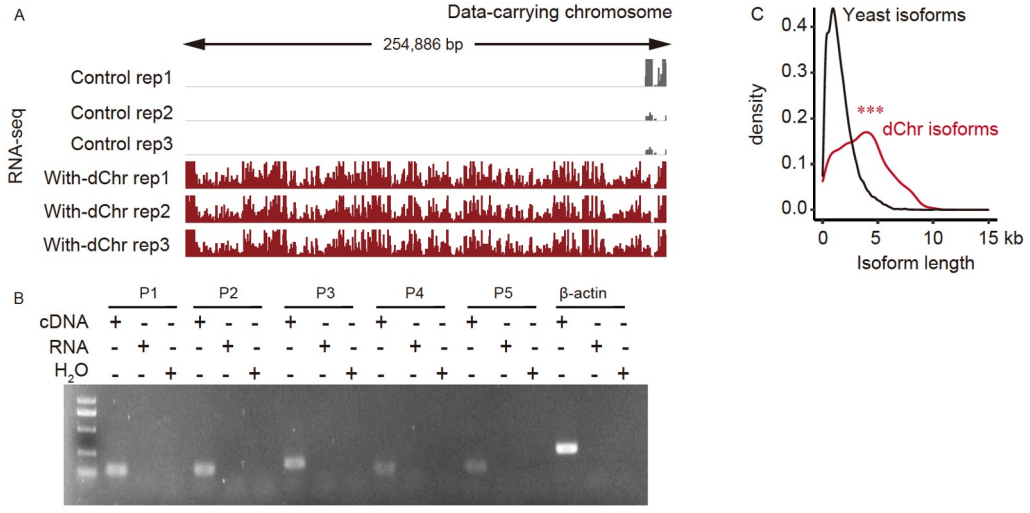

Figure 4 Highly pervasive transcription of the data-carrying chromosome. A, IGV visualization of RNA-seq signals for dChr. Control and data-carrying yeasts were both sequenced with three biological replicates. B, The five regions were randomly selected for RT-PCR analysis; all regions could be reverse-transcribed, and β-actin expression was used as a positive control. C, The black curve represents isoform lengths from the wild-type yeast genome, and the red curve represents isoform lengths transcribed in dChr. The Wilcoxon test was performed. ***, $P<0.001$.

the abundance of HSP42 and HSP104 increased in response to DNA replication stress (Tkach et al., 2012). Our results suggest that the dChr might increase replication pressure in its host yeast, although only 5.7% of total yeast genes experienced significant differences in expression: 81 genes were up-regulated and 202 were down-regulated. Moreover, one may ask whether the transcriptome and proteome were influenced by transcripts originating from dChr. However, determining whether there is uncharacterized interactions between transcripts and the native genome or host RNAs requires further in-depth study.

The LDPC coding rule apparently was not designed to avoid genetic elements. As a result, TATA boxes and transcription factor-binding motifs were unintentionally incorporated into our dChr sequence. Despite the production of a large number of long ncRNAs, many of which were of high yield, dChr had only a slight impact on the internal transcriptional and proteomic activities of yeast. This could be because the heterologous protein was expressed below the detection limit. That yeast fitness can be maintained in the presence of a data-carrying chromosome is encouraging, but as this is only the first case report, more data with additional cases of dChr use is needed to confirm this finding.

MATERIALS AND METHODS

Strains and cell culture before library preparation

Yeast strains yMH007 (containing dChr) and yMH104 (Chen et al., 2021) were cultured in tubes containing 4 mL SC-Ura liquid medium in a shaker incubator overnight at 30°C. yMH104 contains pZJ-Ura with basic elements including CEN6/ARS4 and a marker gene *URA3*, as well as elements of a bacterial artificial chromosome. After incubation, liquid culture was diluted with 50 mL fresh SC-Ura liquid medium to an optical density at 600 nm (A_{600}) of 0.2. After culturing at 30°C until the A_{600} reached 1.0–1.2, sufficient yeast cells for downstream applications were harvested: RNA extraction, ChIP-seq, ATAC-seq, Hi-C, and CAGE-seq.

Local alignment search and genetic elements prediction

The blastn algorithm (designed for cross-species comparison with shorter queries) of the BLAST search local alignment tool was used to search for sequences analogous to those in our dChr. The entire dChr data-carrying sequence (not including the vector sequence) was used as a query against the *S. cerevisiae* S288C genomic sequence.

Putative promoters were predicted using a promoter prediction webserver (Knudsen, 1999). TATA boxes were identified in the dChr sequence by searching for the consensus sequence TATA(A/T)A(A/T)(A/G) (Basehoar et al., 2004). Transcription factor-binding motifs were found using FIMO (Grant et al., 2011) in MEME Suite version 5.4.1 (Bailey et al., 2015).

RNA extraction and reverse transcription polymerase chain reaction (RT-PCR)

Total RNA from yeast was purified according to the instructions for the Qiagen RNeasy Mini Kit (QIAGEN, Germany). Reverse transcription was performed using the First Strand cDNA Synthesis Kit (GeneCopoeia, USA). Equivalent cDNA, as well as RNA and ddH_2O, was used to carry out PCR using the primers listed in Table S2 in Supporting Information. The cycling parameters were 95°C for 5 min, 30 cycles of 95°C for 15 s, 54°C for 15 s and 72°C for 30 s, followed by a 5 min incubation at 72°C. After amplification, 5 μL of each PCR product was loaded onto 1% agarose gels for electrophoresis.

RNA-seq and Iso-seq

ToFU (Transcript isOforms: Full-length and Unassembled; see Gordon et al. (2015)) was used for Iso-seq library preparation and sequencing to generate a *de novo* transcriptome. RNA-seq library preparation followed standard procedures (for Illumina and PacBio, respectively). The HiSeq PE150 mode on the Illumina NovaSeq™ 6000 was used for RNA-seq, and the PacBio single-molecule long-read platform was used for Iso-seq. RNA-seq raw reads were cleaned by using Cutadapt (version 2.6) to trim the sequencing adapters. Clean reads were then mapped to the yeast reference genome (we merged the sacCer3 genome and the dChr DNA sequence to obtain our reference genome, and the reference genome was indexed for mapping) using hisat2 (version 2.0.5), after which Stringtie (version 2.0) was used to quantify the gene expression levels. The mapped bam files were converted to the Bigwig format using bamCoverage from Deeptools (version 3.3.1). Afterward, the Bigwig files were visualized in the IGV. The differentially expressed genes were called using DESeq2 (R package, version 1.26.0) in R (version 3.6.0).

Hi-C library preparation and sequencing

Crosslinking and MboI enzyme (NEB, USA) digestion were performed by following a previous study (Lieberman-Aiden et al., 2009). Briefly, cohesive ends were blunted and labeled by biotin-14-dCTP (TriLINK). Blunt-end ligation was carried out using T4 DNA ligase (Thermo Fisher Scientific, USA). After purification, the DNA was sheared to a length of ~400 bp. Dynabeads MyOne Streptavidin C1 (Thermo Fisher Scientific) was used to pull down point ligation junctions. The Hi-C library was prepared with the NEBNext Ultra II DNA Library Prep Kit according to the manufacturers' instructions. Sequencing was carried out on an Illumina HiSeq X Ten platform (USA) using the PE150 mode. Two replicates were generated for each group of materials. HiC-Pro (version 2.7.9) (Servant et al., 2015) was used to analyze the Hi-C data. Briefly, paired-end reads were first aligned to the reference genome. Unmapped reads were cut at restriction sites and re-mapped. After removing PCR duplicates and reads without restriction sites (MobI), the contact matrixes were used to generate the final valid contacts. Finally, the contact matrix was normalized using ICE. Hicpro2juicebox (a script from HiC-Pro) was used to convert the valid contacts to the ".hic" format, and these files were imported into Juicebox (Durand et al., 2016) for visualization.

CAGE-seq library preparation and sequencing

A previous study was used as a reference for CAGE-seq library preparation (Adiconis et al., 2018). For high-throughput sequencing, the libraries were prepared following the manufacturer's instructions and sequenced on the Illumina NovaSeq 6000 system using the 150 nt paired-end sequencing mode. Cwutadapt (version 2.6) was used to find and remove adapter sequences with the parameters "-a GATCGGAAGAGCACACGTCTGAACTCCAGTCAC -A

AGATCGGAAGAGCGTCGTGTAGGGAAA -m 20 -q 15". Bowtie2 was used to map clean reads to reference genome, and Picard was used to remove duplicates. MACS2 was used to call significant peaks with the parameters "-f BAMPE -g 1.2e7 - -p 0.05". bamCoverage was used to convert bam files to Bigwig files for visualization.

ATAC-seq library preparation and sequencing

Yeast cells cultured from a previous log growth phase culture were harvested by centrifuge and then washed three times with phosphate-buffered saline (PBS). Cell pellets were lysed via incubation for 10 min at 4°C on the rotation mixer. DNA fragmentation by Tn5 was carried out according to standard protocols (Corces et al., 2017). The final library was sequenced on the Illumina HiSeq X Ten platform using the PE150 mode. ATAC-seq raw reads were first trimmed with sequencing adapters, similar to the process for RNA-seq. Afterward, clean reads were aligned to a reference genome using Bowtie2 (version 2.3.5) with the following parameters: "-L 25 -X 2000 -t -q -N 1 –no-mixed –no-discordant". PCR duplicates were removed using Picard (version 1.118), and MACS2 (version 2.2.5) was used to call significant peaks with "-F BAM -g 1.2e7 -w - -nomodel - -shift -100 - -extsize 200". The final mapped bam files were converted to the Bigwig format using bamCoverage for visualization in IGV.

ChIP-seq library preparation and sequencing

Forty milliliters of yeast culture were combined with 1.11 mL of 37% formaldehyde (1% final concentration) in a 50 mL Falcon tube. The lid was tightened to prevent leakage, and the tube was placed in a silent mixer at room temperature for 25 min. Afterward, 2.74 mL of 2 mol L^{-1} glycine was added and mixed at room temperature for 10 min to stop crosslink reactions. The mixture was washed three times with 20 mL of cold (4°C) PBS and resuspended with 3 mL of PBS that contained protease inhibitor. A standard ChIP-seq protocol (Kim and Dekker, 2018) was carried out using an H3K4me3 antibody (Cell Signaling Technology, USA) and then sequenced on the Illumina NovaSeq6000, using the PE150 mode. The ChIP-seq analysis was similar to the ATAC-seq analysis. The ChIPed and input sample pairs were mapped to the reference genome using Bowtie2, and enriched peaks were called using MACS2. The mapped bam files were finally converted to Bigwig format via bamCoverage using RPKM normalization for visualization.

In-depth proteome analysis

Sample preparation for proteome analysis followed the ionic liquid-based filter-aided sample preparation (i-FASP) procedure (Fang et al., 2020; Zhao et al., 2017). Briefly, yeast proteins were extracted using 10% (m/v) C12Im-Cl solution and reduced with 100 mmol L^{-1} 1,4-dithio-D-threitol at 95°C for 5 min. Subsequently, the cooled samples were transferred to 10 kD filter devices and washed with a 50 mmol L^{-1} ammonium bicarbonate (NH_4CO_3) buffer by centrifugation. Then, iodoacetamide was added to the final concentration of 150 mmol L^{-1} and incubated for 30 min in the dark. After being washed three times with 50 mmol L^{-1} NH_4CO_3 buffer, the samples were digested by trypsin with an enzyme/protein ratio (m/m) of 1:30 at 37°C for 12 h. The protein digests were analyzed using a Q-Exactive mass spectrometer equipped with an ultra-HPLC EASY-nLC 1000 system (Thermo Fisher Scientific). Peptides were separated by a 15 cm C18 capillary analytical column (150 μm i.d., packed with ReproSil-Pur C18-AQ 1.9 μm beads (Dr. Maisch GmbH, Germany)) over an 85 min elution gradient. The MS analysis was performed in a data-dependent mode with full scans (m/z 300–1,800, resolution 70,000@ m/z 200) obtained in an Orbitrap mass analyzer. After peptides were fragmented by higher energy collisional dissociation, the fragment ions were transferred into the Orbitrap and acquired at a resolution of 17,500@ m/z 200.

MS data were processed using the MaxQuant software (Version 1.6.5.0) against the *S. cerevisiae* (strain ATCC 204508/S288c) database (Uniprot Proteome, 2020_11, 6,049 entries), combined with 9,202 protein-coding sequences constructed from the dChr gene sequence, and translated to ORFs. All other parameters were set to default.

Gene ontology analysis

Gene ontology enrichment analysis for differentially expressed genes between data-carrying yeast and control yeast was performed with goseq (R package, version 1.38.0; (Young et al., 2010)). Only biological process terms were focused on exclusively, and the top five enriched terms were selected for visualization. All yeast genes were selected as background genes.

Data availability

Related sequencing data have been uploaded to NCBI's Gene Expression Omnibus and are accessible through the GEO Series accession number GSE183492. Other data supporting the findings of the present study are available from the corresponding author upon reasonable request.

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

Acknowledgements *This work was supported by the National Key Research and Development Program of China (2121YFA0909300), the*

National Natural Science Foundation of China (31861143017, 21621004, and 31901019) and the China Postdoctoral Science Foundation (2021M692389). We thank Prof. Yan Zhang at Tianjin University for the manuscript revision.

References

Adiconis, X., Haber, A.L., Simmons, S.K., Levy Moonshine, A., Ji, Z., Busby, M.A., Shi, X., Jacques, J., Lancaster, M.A., Pan, J.Q., et al. (2018). Comprehensive comparative analysis of 5′-end RNA-sequencing methods. Nat Methods 15, 505–511.

Bailey, T.L., Johnson, J., Grant, C.E., and Noble, W.S. (2015). The MEME Suite. Nucleic Acids Res 43, W39–W49.

Basehoar, A.D., Zanton, S.J., and Pugh, B.F. (2004). Identification and distinct regulation of yeast TATA box-containing genes. Cell 116, 699–709.

Chen, W., Han, M., Zhou, J., Ge, Q., Wang, P., Zhang, X., Zhu, S., Song, L., and Yuan, Y. (2021). An artificial chromosome for data storage. Natl Sci Rev 8, nwab028.

Choi, S.W., Kim, H.W., and Nam, J.W. (2019). The small peptide world in long noncoding RNAs. Brief Bioinform 20, 1853–1864.

Chong, S.Y., Cutler, S., Lin, J.J., Tsai, C.H., Tsai, H.K., Biggins, S., Tsukiyama, T., Lo, Y.C., and Kao, C.F. (2020). H3K4 methylation at active genes mitigates transcription-replication conflicts during replication stress. Nat Commun 11, 809.

Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna, S., Satpathy, A.T., Rubin, A.J., Montine, K. S., Wu, B., et al. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat Methods 14, 959–962.

Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S., and Aiden, E.L. (2016). Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst 3, 99–101.

Fang, F., Zhao, Q., Chu, H., Liu, M., Zhao, B., Liang, Z., Zhang, L., Li, G., Wang, L., Qin, J., et al. (2020). Molecular dynamics simulation-assisted ionic liquid screening for deep coverage proteome analysis. Mol Cell Proteomics 19, 1724–1737.

Fredens, J., Wang, K., de la Torre, D., Funke, L.F.H., Robertson, W.E., Christova, Y., Chia, T., Schmied, W.H., Dunkelmann, D.L., Beránek, V., et al. (2019). Total synthesis of *Escherichia coli* with a recoded genome. Nature 569, 514–518.

Gibson, D.G., Glass, J.I., Lartigue, C., Noskov, V.N., Chuang, R.Y., Algire, M.A., Benders, G.A., Montague, M.G., Ma, L., Moodie, M.M., et al. (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. Science 329, 52–56.

Gnügge, R., and Rudolf, F. (2017). *Saccharomyces cerevisiae* Shuttle vectors. Yeast 34, 205–221.

Gordon, S.P., Tseng, E., Salamov, A., Zhang, J., Meng, X., Zhao, Z., Kang, D., Underwood, J., Grigoriev, I.V., Figueroa, M., et al. (2015). Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. PLoS ONE 10, e0132628.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics 27, 1017–1018.

Hieter, P., Mann, C., Snyder, M., and Davis, R.W. (1985). Mitotic stability of yeast chromosomes: a colony color assay that measures nondisjunction and chromosome loss. Cell 40, 381–392.

Hutchison Iii, C.A., Chuang, R.Y., Noskov, V.N., Assad-Garcia, N., Deerinck, T.J., Ellisman, M.H., Gill, J., Kannan, K., Karas, B.J., Ma, L., et al. (2016). Design and synthesis of a minimal bacterial genome. Science 351, aad6253.

Kannan, K., and Gibson, D.G. (2017). Yeast genome, by design. Science 355, 1024–1025.

Keung, A.J., Joung, J.K., Khalil, A.S., and Collins, J.J. (2015). Chromatin regulation at the frontier of synthetic biology. Nat Rev Genet 16, 159–171.

Kim, S., Liachko, I., Brickner, D.G., Cook, K., Noble, W.S., Brickner, J.H., Shendure, J., and Dunham, M.J. (2017). The dynamic three-dimensional organization of the diploid yeast genome. eLife 6, e23623.

Kim, T.H., and Dekker, J. (2018). ChIP-quantitative polymerase chain reaction (ChIP-qPCR). Cold Spring Harb Protoc 2018(5), pdb. prot082628.

Klemm, S.L., Shipony, Z., and Greenleaf, W.J. (2019). Chromatin accessibility and the regulatory epigenome. Nat Rev Genet 20, 207–220.

Knudsen, S. (1999). Promoter2.0: for the recognition of PolII promoter sequences. Bioinformatics 15, 356–361.

Kouprina, N., and Larionov, V. (2008). Selective isolation of genomic loci from complex genomes by transformation-associated recombination cloning in the yeast *Saccharomyces cerevisiae*. Nat Protoc 3, 371–377.

Lamb, B.T., and Gearhart, J.D. (1995). YAC transgenics and the study of genetics and human disease. Curr Opin Genet Dev 5, 342–348.

Lee, C.S.K., Cheung, M.F., Li, J., Zhao, Y., Lam, W.H., Ho, V., Rohs, R., Zhai, Y., Leung, D., and Tye, B.K. (2021). Humanizing the yeast origin recognition complex. Nat Commun 12, 33.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289–293.

Lu, X., and Ellis, T. (2021). Self-replicating digital data storage with synthetic chromosomes. Natl Sci Rev 8.

Lu, Z., and Lin, Z. (2019). Pervasive and dynamic transcription initiation in *Saccharomyces cerevisiae*. Genome Res 29, 1198–1210.

Luo, X., Reiter, M.A., d'Espaux, L., Wong, J., Denby, C.M., Lechner, A., Zhang, Y., Grzybowski, A.T., Harth, S., Lin, W., et al. (2019). Complete biosynthesis of cannabinoids and their unnatural analogues in yeast. Nature 567, 123–126.

Marschall, P., Malik, N., and Larin, Z. (1999). Transfer of YACs up to 2.3 Mb intact into human cells with polyethylenimine. Gene Ther 6, 1634–1637.

Postma, E.D., Dashko, S., van Breemen, L., Taylor Parkins, S.K., van den Broek, M., Daran, J.M., and Daran-Lapujade, P. (2021). A supernumerary designer chromosome for modular *in vivo* pathway assembly in *Saccharomyces cerevisiae*. Nucleic Acids Res 49, 1769–1783.

Pouokam, M., Cruz, B., Burgess, S., Segal, M.R., Vazquez, M., and Arsuaga, J. (2019). The Rabl configuration limits topological entanglement of chromosomes in budding yeast. Sci Rep 9, 6795.

Ro, D.K., Paradise, E.M., Ouellet, M., Fisher, K.J., Newman, K.L., Ndungu, J.M., Ho, K.A., Eachus, R.A., Ham, T.S., Kirby, J., et al. (2006). Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature 440, 940–943.

Sasaki, T., Matsumoto, T., Antonio, B.A., and Nagamura, Y. (2005). From mapping to sequencing, post-sequencing and beyond. Plant Cell Physiol 46, 3–13.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol 16, 259.

Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. Nat Biotechnol 31, 1009–1014.
Shen, Y., Wang, Y., Chen, T., Gao, F., Gong, J., Abramczyk, D., Walker, R., Zhao, H., Chen, S., Liu, W., et al. (2017). Deep functional analysis of synII, a 770-kilobase synthetic yeast chromosome. Science 355, aaf4791.
Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., et al. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci USA 100, 15776–15781.
Srinivasan, P., and Smolke, C.D. (2020). Biosynthesis of medicinal tropane alkaloids in yeast. Nature 585, 614–619.
Sylvain, M.A., Liang, X.B., Hellauer, K., and Turcotte, B. (2011). Yeast zinc cluster proteins Dal81 and Uga3 cooperate by targeting common coactivators for transcriptional activation of γ-aminobutyrate responsive genes. Genetics 188, 523–534.
Tkach, J.M., Yimit, A., Lee, A.Y., Riffle, M., Costanzo, M., Jaschob, D., Hendry, J.A., Ou, J., Moffat, J., Boone, C., et al. (2012). Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. Nat Cell Biol 14, 966–976.
Treutlein, B., Gokce, O., Quake, S.R., and Südhof, T.C. (2014). Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. Proc Natl Acad Sci USA 111, E1291–E1299.
Tschumper, G., and Carbon, J. (1983). Copy number control by a yeast centromere. Gene 23, 221–232.
Tudek, A., Candelli, T., and Libri, D. (2015). Non-coding transcription by RNA polymerase II in yeast: Hasard or nécessité? Biochimie 117, 28–36.
Wang, K., Fredens, J., Brunner, S.F., Kim, S.H., Chia, T., and Chin, J.W. (2016). Defining synonymous codon compression schemes by genome recoding. Nature 539, 59–64.
Wang, R., Mozziconacci, J., Bancaud, A., and Gadal, O. (2015). Principles of chromatin organization in yeast: relevance of polymer models to describe nuclear organization and dynamics. Curr Opin Cell Biol 34, 54–60.
Wilkinson, D., Váchová, L., Hlaváček, O., Maršíková, J., Gilfillan, G.D., and Palková, Z. (2018). Long noncoding RNAs in yeast cells and differentiated subpopulations of yeast colonies and biofilms. Oxid Med Cell Longev 2018, 1–12.
Wu, Y., Li, B.Z., Zhao, M., Mitchell, L.A., Xie, Z.X., Lin, Q.H., Wang, X., Xiao, W.H., Wang, Y., Zhou, X., et al. (2017). Bug mapping and fitness testing of chemically synthesized chromosome X. Science 355, eaaf4706.
Xie, Z.X., Li, B.Z., Mitchell, L.A., Wu, Y., Qi, X., Jin, Z., Jia, B., Wang, X., Zeng, B.X., Liu, H.M., et al. (2017). "Perfect" designer chromosome V and behavior of a ring derivative. Science 355, aaf4704.
Young, M.D., Wakefield, M.J., Smyth, G.K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol 11, R14.
Zhang, C., Xu, Z., Yang, S., Sun, G., Jia, L., Zheng, Z., Gu, Q., Tao, W., Cheng, T., Li, C., et al. (2020). tagHi-C reveals 3D chromatin architecture dynamics during mouse hematopoiesis. Cell Rep 32, 108206.
Zhang, W., Zhao, G., Luo, Z., Lin, Y., Wang, L., Guo, Y., Wang, A., Jiang, S., Jiang, Q., Gong, J., et al. (2017). Engineering the ribosomal DNA in a megabase synthetic chromosome. Science 355, eaaf3981.
Zhao, Q., Fang, F., Shan, Y., Sui, Z., Zhao, B., Liang, Z., Zhang, L., and Zhang, Y. (2017). In-depth proteome coverage by improving efficiency for membrane proteome analysis. Anal Chem 89, 5179–5185.
Zhou, J., Wu, R., Xue, X., and Qin, Z. (2016). CasHRA (Cas9-facilitated Homologous Recombination Assembly) method of constructing megabase-sized DNA. Nucleic Acids Res 44, e124.

SUPPORTING INFORMATION

The supporting information is available online at https://doi.org/10.1007/s11427-021-2044-x. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.