Genome sequence assembly and annotation of *MATA* and *MATB* strains of *Yarrowia lipolytica*

Narges Zali[1,2], Osama El Damerash[1], Kapeel Chougule[1], Zhenyuan Lu [1],Doreen Ware[1,3] and Bruce Stillman[1]

[1] Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA
[2] Graduate Program in Genetics, Stony Brook University, Stony Brook, NY 11794, USA
[3] USDA-ARS Robert W. Holley Center for Agriculture and Health, Ithaca, NY, 14853, USA

Correspondence: Bruce Stillman: email stillman@cshl.edu, phone +1 (516) 367-8383

ABSTRACT

Yeast is commonly utilized in molecular and cell biology research, and *Yarrowia lipolytica* is favored by bio-engineers due to its ability to produce copious amounts of lipids, chemicals, and enzymes for industrial applications. *Y. lipolytica* is a dimorphic yeast that can proliferate in aerobic and hydrophobic environments conducive to industrial use. However, there is limited knowledge about the basic molecular biology of this yeast, including how the genome is duplicated and how gene silencing occurs. Genome sequences of *Y. lipolytica* strains have offered insights into this yeast species and have facilitated the development of new industrial applications. Although previous studies have reported the genome sequence of a few *Y. lipolytica* strains, it is of value to have more precise sequences and annotation, particularly for studies of the biology of this yeast. To further study and characterize the molecular biology of this microorganism, a high-quality reference genome assembly and annotation has been produced for two related *Y. lipolytica* strains of the opposite mating type, *MATA* (E122) and *MATB* (22301-5). The combination of short-read and long-read sequencing of genome DNA and short-read and long-read sequencing of transcript cDNAs allowed the genome assembly and a comparison with a distantly related *Yarrowia* strain.

INTRODUCTION

Y. lipolytica is an ascomycete yeast belonging to the class *Saccharomycetes* that proliferates in hydrophobic environments rich in lipids and proteins, in part due to its remarkable lipolytic and proteolytic abilities (1). *Yarrowia* therefore proliferates in environments rich in lipids and proteins, such as meat and dairy products, particularly fermented ones like cheeses and dry meats, as well as sewage or oil-polluted waters (2–4). *Yarrowia* is significantly different from other hemiascomycetous yeasts in terms of its genomic features. For instance, it is a heterothallic yeast with two distinct mating types, *MATA* and *MATB,* and most natural isolates of this yeast are predominantly haploid (5). Additionally, the G/C content in *Y. lipolytica* is notably high, averaging at 49% and reaching nearly 53% in genes, compared to other yeasts, particularly compared to the commonly used *Saccharomyces cerevisiae* with a genome containing 38.3% G/C (6). *Y. lipolytica* and *S. cerevisiae* are estimated to have a common ancestor that existed 300 million years ago (7), making this pair of yeasts attractive for comparing the evolution of fundamental biological processes such as genome function and DNA replication. *Yarrowia* also has an unusually high number of intron-containing genes compared to *S. cerevisiae* and its related species (8). The number of genes in *Yarrowia* is within the typical range for hemiascomycetous yeasts, but its genome size is 1.7 times larger than that of *S. cerevisiae,* which contains approximately the same number of genes.

Comparisons of the potential mechanisms of initiation of DNA replication in the budding yeasts, notably the predictions of DNA sequence-specific origins of DNA replication, have implicated major differences between *Y. lipolytica* and *S. cerevisiae* (9). *S. cerevisiae* and a small clade of highly related budding yeasts have origins of DNA replication that are composed of extensive DNA sequence-specific elements that constitute a functional origin of DNA replication (10, 11)). The origin DNA elements in *S. cerevisiae* are recognized by the Origin Recognition Complex (ORC) and Cdc6, and two essential initiator subunits of ORC confer base-specific interactions via an inserted alpha-helix in the Orc4 subunit (Orc4-IH) and a loop in the Orc2 subunit (Orc2-loop) (9, 12). In contrast, *Yarrowia*, like nearly all other fungi and all animals and plants, lack these DNA sequence-specific recognition domains in Orc2 and Orc4, suggesting that it may have a more relaxed DNA sequence-specificity at its origins of DNA replication. Moreover, *S. cerevisiae* has lost RNA interference (RNAi) mechanisms but has gained Silent Information Regulator (SIR) proteins (Sir1, Sir3, and Sir4) that function in gene silencing and suppression of recombination of repetitive DNA sequences such as ribosome DNA (rDNA) and telomeric DNA (13–15). Interestingly, *Y. lipolytica* lacks both RNAi and SIR-dependent gene-silencing proteins, including the RNAi silencing proteins such as Dicer and Argonaute (Ago) and SIR proteins, except for Sir2 (9)). All eukaryotes harbor the Sir2 gene that encodes an NAD-dependent histone deacetylase (16). Of relevance in *S. cerevisiae* and its related budding yeasts such as *Kluyveromyces lactis,* the interaction between ORC and SIR proteins and a role for ORC silencing the mating type genes (17–21)). It is, therefore, not known how *Y. lipolytica* silences gene expression or suppresses recombination of repetitive rDNA and telomeric DNA.

One possible explanation for the occurrence DNA sequence-specific origins of DNA replication in some budding yeast species, such as *S. cerevisiae*, is that these organisms have lost much of the intergenic DNA and lack introns, therefore they possess a very gene-dense genome relative to their genome size. The presence of DNA sequence-defined origins in gene-rich organisms, such as *S. cerevisiae,* could provide an advantage in recruiting ORC to intergenic sites within these species, thereby avoiding conflicts between DNA transcription and the initiation of DNA replication, which can result in genome instability (9, 22)). As a result, organisms like *S. cerevisiae*, with high gene density and smaller intergenic regions, have evolved an efficient mechanism to ensure that the replication complex can find appropriate sites for initiation and avoid initiating DNA replication in a transcribed region. By gaining a deeper understanding of how DNA replication is initiated in a variety of species, including human cells and in diverse yeasts such as *Y. lipolytica*, further insights into how origins of DNA replication are located in the genome and the replication strategies in eukaryotic cells will become apparent (23)). For this reason, the complete genome sequences and assemblies of two *Y. lipolytica* strains of opposite mating type were performed to assist in subsequent studies of whole genome DNA replication and gene silencing mechanisms. Both long-read sequencing using PacBio and Oxford nanopore methods and short-read Illumina-based methods of both genomic DNA and cDNA were used to generate and assemble the genome of the two *Y. lipolytica* strains.

Previously, the main reference genome for *Y. lipolytica* was strain CLIB122 (also called E150), a derivative from a mating between a French isolate W29 and an American isolate YB423-12 (CBS 6124-

2) (see Figure 1), that was obtained through short read Sanger sequencing (6, 24)). A number of other strains have been shot-gun sequenced and their genomes compared, resulting in a recent summary of these genome comparisons (25)). The estimated size of the six-chromosome genome was ~21 Mb (24)). However, these assemblies contained gaps, and the telomeric ends as well as rDNA repeats could not be integrated into the genome assembly due to their repetitive nature. The CLIB122/E150 strain and its derivatives, including the commonly used PO1f strain (Figure 1) (26), are the main strains used for industrial purposes. A high-quality, near-contiguous genome assembly of a distantly related *Y. lipolytica* strain DSM 3286 (a German strain) was obtained using a combination of long-read and short-read genomic DNA sequencing (27). This allowed the characterization of the repetitive rDNA and telomeric regions and the observation that rDNA clusters are located near the telomeric regions. Additionally, the genetic and phenotypic diversity of 56 haploid strains of *Y. lipolytica* was investigated by sequencing of a diverse set of *Y. lipolytica* strains collected from various geographical and biological origins, and included revision of the version of the E150 *Y. lipolytica* strain genome sequence and annotation (25).

In this study, we have used both long and short-read sequencing of genomic DNA and cDNA copies of RNA transcripts to precisely compare the genomes of genetically related *MATA* and *MATB* strains that are distant from the DSM 3286 strain isolated in Germany and other geographically diverse strains that have been sequenced (25, 28)). Chromosomal rearrangements were observed comparing the multiple isolates. The sequences allowed annotation of the genome and revealed the presence of many repeated DNA sequences such as transposable elements, including LTR-retrotransposons, LINE elements, and DNA transposons from various families, which are distributed variably among strains. Moreover, the distribution of and number of rDNA repeats was analyzed.

MATERIAL AND METHODS

Strains: Two related *Yarrowia lipolytica* strains of opposite mating type were obtained from Richard A. Rachubunski, University of Alberta, Canada, and single clones were isolated and used for both genome and transcript sequencing. The strains were 22301-5 (*MATB*) and E122 (*MATA*, alternatively called CLIB120) (5, 29, 30) (Figure 1). E122 is *MATA*, *ura3-302*, *leu2-270*, *lys8-11* and is related to the *MATB* strain E150 (CLIB122 and is the current DNA sequence reference strain (6, 24)). Strain 22301-5 is *MATB, his-1, uras-302, leu2-270,* (Figure 1).

DNA and RNA Preparation and Sequencing

RNA: To isolate the high molecular weight RNA, the TRIzol Plus RNA Purification Kit from Thermo Fisher Scientifics was used. Briefly, 5ml liquid of yeast culture was grown to an OD^{600} of 2.0 and cells were pelleted and lysed with TRIzol™ reagent according to the user manual. Following lysis, the RNA present in the sample was bound to the PureLink RNA Mini Kit Spin Cartridge (12183018A, Invitrogen) where it was washed to remove contaminants. Lastly, eluted RNA was stored in 50-microliter aliquots at -80°C. To generate a short-read RNA, a Direct-zol RNA purification kit from Zymo Research (Cat # R2050) was used according to the user manual. Long-read sequencing of RNA transcripts was performed as follows. The RNA was prepared using the ONT SQK-PCS109 kit according to the manufacturer's instructions, and it was loaded onto a PromethION P24 system with a PROM-0002 flow cell. Base-calling was performed using the live hac base-calling guppy version 3.2.10. Two cells of *MATA* and *MATB* were run for each experiment.

DNA: For the generation of high molecular weight DNA and ultra-long nanopore sequencing reads, cells in a 100 ml culture grown overnight at 30°C in Yeast extract, Peptone and Dextrose (YPD) were harvested, washed in sterile distilled water and incubated for 2 hr at 37° C in 10 ml SEB buffer (0.9 M sorbitol, 0.1 M EDTA, 0.8% β -mercaptoethanol) containing 5 mg Zymolyase 20T (Sunrise Science products, CAT#N0766391). Protoplast formation was monitored by phase contrast microscopy. The protoplasts were then harvested, resuspended in 3 ml TE Buffer (Tris-EDTA, pH 8.0), then 300 µl 10% SDS was added, and the samples were incubated at 65°C for 30 minutes. 1 ml of 5 M potassium acetate was added, and the samples were kept on ice for 1 h. The supernatant was recovered after centrifugation, and DNA was precipitated by adding 0.1 volume of 3 M sodium acetate and 2.5 volumes of ethanol at −20°C for at least 1 h. The DNA was recovered by centrifugation and resuspended in 3 ml TE (31)). Then 100 µg/mL of proteinase K along with 50 µg/mL RNase A were added and the samples were incubated at 37°C for 3 hrs. After centrifugation for 45 min at 12,000 × g and 4°C, the supernatant was collected and transferred to a 2-ml Eppendorf tube. Samples were then extracted two more times with phenol/chloroform/isoamyl alcohol and one final time with chloroform. To precipitate DNA, 2-2.5 volumes of 100 % freeze-cold ethanol was added to the aqueous phase along with 1/10 volume of 3 M sodium acetate, mixing by inversion, and samples were incubated at −20°C for at least 1 h. The DNA was recovered by centrifugation for 20 min at 12,000 × g and 4°C, and the pellet was subsequently washed three times with 2 ml of 80% (vol/vol) ethanol. The pellet was then air dried and dissolved in 100 µl of Tris-EDTA.

DNA fragment length was assessed for molecular weight distributions of genomic DNA samples were evaluated using a Femto Pulse pulse-field capillary electrophoresis system (Agilent). >5ug of DNA was size selected via SRE XS (Circulomics). The full reaction was repaired and end prepped with NEBNext FFPE DNA Repair Buffer and Ultra II End prep kit (NEB). The reaction was cleaned up with 1X Ampure beads and precipitated with ethanol. DNA was bound to ONT adapter from the SQK-LSK109 kit (ONT) via NEBnext Quick T4 ligation module (NEB). DNA was resuspended in SQB buffer (ONT) and loading beads (ONT) and sequenced on one PromethION 24 cell PROM0002 with a three-day run time.

For short-read DNA sequencing, the YeaStar Genomic DNA Kit from Zymo Research (CAT# D2002) was used according to the user manual. DNA sequencing libraries were prepared per the manufacturer's instructions with a Kapa DNA hyperprep kit (Roche CAT #KK8504). It was loaded on an Illumina MiSeq with a PE150 v2 format.

DNA Sequence Assembly

Long-Read processing: The unprocessed long-reads were produced using Guppy v.5 base-caller from Oxford Nanopore Technologies (https://github.com/nanoporetech). To assemble the reads, the long-read assembly pipeline Flye v. 2.9-b1774 (https://github.com/fenderglass/Flye) (32) was used in nano-hq mode, which is intended for high-quality reads (<5% error rate). The minimum overlap between reads was set to 7KB. The pipeline was run with five iterations of polishing.

Short-Read processing: The paired-end reads were trimmed using Cutadapt v.3.7 (33). Cutadapt removes adapter sequences from high-throughput sequencing reads (33)). BWA v.0.7.17-r1188 (34) was used to index the long-read assembly and align the trimmed short-reads to the assembly. The alignments were sorted and indexed using Samtools v.1.14 (35). Pilon v1.24 (https://github.com/broadinstitute/pilon) (36) was used for polishing the long-read assembly with the aligned short-reads. We obtained exactly one contig per chromosome and mitochondria for *MATB* and an extra contig for *MATA*. To scaffold the extra contig, we used RaGOO (37) and the assembled *MATB* as the reference genome sequence.

Transcript assembly

For the transcriptome assembly, three different transcriptome assemblies were combined: one from the short-read sequencing, a second one from the Nanopore long cDNA reads, and a third combining both. The short-reads were first trimmed using Trimmomatic v.0.38 (38)). The trimmed reads were aligned to the genome using hisat2 (v.2.2.1). A short-read only transcriptome was then assembled using Stringtie v.1.3.6 (39)). The long cDNA reads were *de novo* assembled using Oxford Nanopore's Workflow Transcriptomes (wf-transcriptomes) pipeline (v1.1.1). A third transcriptome that used long- and short-reads was assembled using TASSEL (40) (https://github.com/kainth-amoldeep/TASSEL). These three transcriptomes were then combined using gff compare (v.0.12.2) (41) and the output was used as input to the annotation pipeline.

Gene annotations

The MAKER-P (v.3.0) (42) pipeline was used to annotate protein-coding genes in the two strains 22301-5 (*MATB*) and E122 (*MATA*). As evidence, we used all annotated proteins from *Yarrowia lipolytica* (budding yeasts) downloaded from the NCBI protein database. These protein sequences were clustered using CDHit-est (v4.6) (43) with parameters (-c 0.95 -n 10 -d 0 -M 3000 -t 1). For transcript evidence, the combined transcriptome from gff compare was used and transcript assembly.The assembled transcripts were checked and filtered for intron retention using Suppa (v2) (44)). For gene prediction, we used Augustus (v3.3) (45, 46) trained on *Y. lipolytica* and FGENESH (http://www.softberry.com) trained on *S. pombe*, respectively. Repeat masking was done using repeatmasker (RepeatMasker Open-4.0) with the ensembl repeat annotation pipeline using parameters (-nolow -gccalc -species "Fungi" -engine ncbi). The repeat masked genomes were used in annotation with MAKER-P evidence. Additional improvements to structural annotations were done using PASA (v2.3.3) (47) using the assembled transcriptome and fungal EST from NCBI using query (EST[Keyword]) AND fungi [Organism]. Gene identifiers were assigned using existing nomenclature schema established for *Yarrowia* for each strain (6, 24)). Functional domain identification was completed with InterProScan (v5.38-76.0)_(48)). TRaCE (49)_was used to assign canonical transcripts based on domain coverage, protein length, and similarity to transcripts assembled by StringTie (v1.3.4a) (39)). Finally, the gene annotations were imported to ensembl core databases, verified, and validated for translation using the ensembl API (50)).

Genome comparative analysis was done using 5 *Yarrowia lipolytica* strains including 15 closely related species and outgroups providing the foundation for building protein-based gene trees based on the EnsemblCompara pipeline (51)).

RESULTS

A hybrid assembly approach was employed by pairing Illumina short-read DNA sequences and high-quality long-read DNA sequences with much-improved base-calling using NANOPORE Technologies' Guppy 5 base caller with very high mean coverage, as indicated in Figure 2A and Table 1. Longer contiguous and quasi-contiguous genome assemblies for *MATB* and *MATA* were obtained, compared to previous assemblies of the related French isolates. The improvements over the reference assembly for strain CLIB122-E150 included the incorporation of telomeric repeats on each chromosome, as well as a decrease in the estimated number of missing essential gene markers according to BUSCO assessment. We assessed the quality of the assemblies using BUSCO version 4.1.1 (52) , which employs AUGUSTUS as the gene predictor in genome mode on the *Saccharomycetes* lineage set of 2137 essential genes**.** The results are shown in Figure 2B and indicate a completeness index of 96.8%. *Yarrowia* appears to be missing genes that are present in other *Saccharomycetes,* or the genes are diverged enough so that BUSCO does not detect them.

A dot plot produced with chromeister (release 1.5.a.) (53) compared the E122 *MATA* and 22301-5 *MATB* genomes, revealing very high similarity between these two related genomes, as expected (Figure 3A). The similarity score indicated a divergence of just 0.003. In contrast, a similar comparison between 22301-5 *MATB* and the German isolate DSM 3286 showed considerable genome rearrangements and a reduced similarity score of 0.0031 (Figure 3B), as noted previously (27)). When the genes that were expressed in E122 *MATA* but not 22301-5 *MATB* were analyzed, they were in and surrounding the *MATA* locus and included *Sla2,* encoding an actin binding protein and *Apn2* encoding a DNA-(apurinic or apyrimidinic site) lyase, both genes that flank the *MATA* locus (54, 55)**,** as well as the *MATA1* and *MATA2* mating type genes (55)).

Gene annotations and comparative analysis

The structural gene annotation pipeline identified 7,728 and 7,769 genes in *Yarrowia lipolytica* strains E122 (*MATA*) and 22301-5 (*MATB*), respectively (Table 2). This gene count surpasses that of previously reported *Yarrowia* strains DSM 3286 (27) with 6,467 protein-coding genes and the reference strain CLIB-122 (56) with 6,448 protein-coding genes. To further evaluate annotation quality, we utilized the Annotation Edit Distance (AED) score generated by MAKER-P (42)). An AED score of 0 indicates genes supported by evidence, while a score of 1 indicates a lack of evidence. Employing mRNA and homology evidence, as described in the methods, to calculate AED scores yielded 6,297 and 6,212 genes with some evidence (AED score < 1) in strains E122 *MATA* and 22301-5 *MATB*, respectively. In contrast, strains DSM 3286 and CLIB-122 (E150) exhibited 6,236 and 6,387 protein-coding genes with AED scores <1 (Table 2).

When comparing annotation features between the strains E122 *MATA* and 22301-5 *MATB* using genes filtered by AED <1, it was observed that 78% of the protein-coding genes contained a single exon. The shortest introns were ~40 base pairs and the longest intron was 6782 base pairs (Figure 4C and D). Moreover, we found more genes in E122 *MATA* and 22301-5 *MATB* with multiple introns compared to the previously analyzed strains (Figure 4A). The median gene lengths in E122 *MATA* and 22301-5 *MATB* were higher than previously estimated, reflecting their increased intron counts compared to the other two strains (Table 2). *Yarrowia lipolytica* genomes are known to be intron-rich, with previous estimates of 15% of genes containing introns, which is 4 times that of *S. cerevisiae* (8)). Intron-containing genes in E122 *MATA*, 22301-5 *MATB* and the DSM 3286 strain represented ~20% of the protein-coding genes, highlighting the improvement in genome assembly using long-read technology compared to CLIB-122. 80% of these intron-containing genes were mono-intronic, compared to 20% that were multi-intronic (with up to five introns). The internal exons of the multi-intronic genes were mostly short compared to 1st intron (Figure 4B and D).

A total of 9453 unfiltered orthologous genes were found between E122 *MATA*, 22301-5 *MATB*, the DSM 3286 and the CLIB-122 strains (Figure 5A), of which 5,975 were core genes (that were found in all four strain genome sequences (Figure 5B, left hand set). These core genes were closer to the number of core genes (6,042) detected from 7 *Yarrowia lipolytica* strains and slightly lower than the pan-genome genes (6,528) detected with 54 strains (25). Prior studies also suggest that *Yarrowia lipolytica* exhibited lower genetic diversity since the core genes were barely different than the pan-genome (25). The *MATA* and *MATB* strains had 1,204 unique gene ortholog groups present in both strains but not present in DSM 3286 or CLIB122 (Figure 5B, second set from left). Subsequent Gene Ontology (GO) analysis of this subset identified only 14 genes with associated GO terms and GO enrichment analysis highlighted significant enrichment in molecular processes including 2 iron, 2 sulfur cluster binding (GO:0051537), methylmalonate-semialdehyde dehydrogenase (acylating) activity (GO:0004491), oxidoreductase activity (GO:0016491) and TBP-class protein binding (GO:0017025) (Figure 5C) (57), as well as biological processes (GO:0006352), for DNA-templated transcription initiation (Figure 5D).

Analysis of Repeat Sequences

Using RepeatMasker software, a comparison of the repeated DNA sequences in the E122 *MATA* and 22301-5 *MATB* strains, as well as the reference CLIB122 strain and the DSM 3286 strain, showed many more repeats in the two genomes sequenced here (Figure 6A AND 6B). In particular, there is an increase in the number of RNA repeats, LTRs, LINE and SINE repeats in the E122 *MATA* and 22301-5 *MATB* strains compared to the DSM 3286 strain. It is not clear whether this difference is due to biological variation or to technical issues with genome sequencing and analysis, but it is likely the latter. The distribution of repeat elements in the E122 *MATA* and 22301-5 *MATB* strains showed similar profiles, with RNA repeats making up about 30%, simple repeats around 35-40%, and LTRs at 15-18%. Both have smaller proportions of Type I Transposons (LINE and SINE) and Type II Transposons, each accounting for about 3-7% of the total repeats. DSM 3286, on the other hand, has a higher percentage of simple repeats (~55%) and a lower proportion of RNA repeats (~10%), with LTRs making up 20% of its repeat content. CLIB122 is distinct with 30% of its repeats being LTRs and about 10% RNA repeats. Across all strains, low complexity regions and unknown repeats remain minimal, each contributing around 1-2% of the total repeats. In summary, while simple and RNA repeats dominate the repeat landscape in these yeast strains, there is significant variability in the proportion of LTRs and other transposon types, particularly between DSM 3286, CLIB122, whereas the repeats are more similar in the E122 *MATA* and 22301-5 *MATB* strains. This variability in repeat element distribution probably reflects a combination of technical differences as well as biological variation in genomic evolution among the strains.

As previously observed (27) , the rDNA repeats consisting of the 18S and 28S genes were located at the ends of chromosomes B (right end), C (both ends), E (right end) and F (both ends), and lie adjacent to the telomeres (Figure 7A). The 5S rDNA genes are scattered throughout the genome on every chromosome. For the E122 *MATA* and 22301-5 MATB strains, the calculated size of these repeats in kilobase pairs (yellow bar) and the number of rDNA repeats (blue bar) for each region of the genome is shown in Figure 7B.

DISCUSSION

Heterothallic yeast, like *Yarrowia lipolytica*, typically engage in outcrossing, where genetic material is exchanged between individuals of different mating types. This may lead to greater genetic diversity and adaptability to changing environments and contribute to divergence and speciation (58)). In contrast, homothallic yeast, such as *S. cerevisiae*, primarily engage in selfing since they can switch their mating type through a gene conversion process initiated by the HO endonuclease (59) , where recombination occurs within the same individual. This process may result in the fixation of beneficial alleles or the accumulation of deleterious mutations, potentially leading to lower genetic diversity (60)). These differences in reproductive strategies lead to distinct recombination pathways in the two types of yeast (61)). However, a recent comparison of 56 shot-gun sequenced strains showed a very low level of genetic diversity, indicating that *Y. lipolytica* may be a species that has recently emerged (25)).

Y. lipolytica exhibits a remarkably low rate of mating and spore viability between different lineages due to chromosomal rearrangements, which may contribute to its poor fertility. Chromosomal rearrangements in *Y. lipolytica* could have been caused by crossing-over events facilitated by the different types of transposable elements present in the organism (27)). Yeast genomes contain mobile genetic elements, such as transposons and retrotransposons, which can translocate within the genome. These elements can be inserted into new locations within the genome or cause chromosomal reorganization by combining with various regions. The prevalence of repeated sequences found in the E122 *MATA* and 22301-5 *MATB* strains may also play a role in rearranging and evolving the genome of this yeast species, consistent with the notion that transposable elements and other repetitive elements can be significant contributors to genome evolution (25, 27)).

We found a higher number of introns in the E122 *MATA* intron (Figure 4). and 22301-5 *MATB* strains compared to the DSM 3286 and CLIB122 strains, with most genes having a single In particular, we observed a higher proportion of genes with more than one intron. This intron-rich genome may enable the production of multiple protein isoforms from a single gene, offering *Yarrowia* the ability to rapidly adapt to changing environments or industrial processes. This could prove especially valuable for *Yarrowia*, as it frequently operates in diverse and challenging growth conditions.

The enhanced genome assembly and annotation of *Yarrowia lipolytica* strain E122 *MATA* and 22301-5 *MATB*, was made possible by employing a hybrid sequencing approach that combined the precision of

Illumina short reads with the depth of Oxford Nanopore long reads and advanced base calling with Guppy 5. This high-quality assembly allowed us to capture telomeric regions, rDNA repeats and improve the completeness of essential gene markers, achieving a BUSCO score of 96.8%. These results establish a strong foundation for further functional and comparative studies on *Y. lipolytica* and its applications.

Gene Annotation and Genetic Diversity

Our comparative gene analysis revealed significant differences between the sequenced E122 *MATA* and 22301-5 *MATB* strains and other previously studied *Yarrowia* strains, such as DSM 3286 and CLIB-122. The increased gene count in E122 *MATA* and 22301-5 *MATB* and the presence of alternative isoforms highlight a potentially broader genetic repertoire and greater regulatory complexity in these strains. This added complexity may reflect adaptive mechanisms developed in response to specific environmental and industrial conditions. For instance, the increased median gene length, attributed to a higher intron count, suggests unique gene structures that could enhance regulatory flexibility, supporting more intricate metabolic or stress-response pathways.

The core gene analysis indicates that E122 *MATA* and 22301-5 *MATB* share 5,975 core genes with other *Yarrowia* strains, consistent with prior findings of limited genetic diversity within *Yarrowia lipolytica* (25)). However, identifying 1,204 unique ortholog groups in *MATA* and *MATB* suggests subtle genomic differences that could contribute to strain-specific phenotypes. Gene Ontology (GO) enrichment analysis of these unique genes emphasizes metabolic processes and transcription initiation. These functions are advantageous in environmental settings where efficient resource utilization and adaptability to stressful conditions are beneficial.

Repeat Elements and Genomic Evolution

The investigation into repetitive DNA elements highlights more repeats in E122 *MATA* and 22301-5 *MATB*, particularly in RNA, LTRs, LINE, and SINE elements, compared to DSM 3286 and CLIB122 (Figure 6B). The distinct repeat profiles observed in E122 *MATA* and 22301-5 *MATB*—with a balance of RNA and simple repeats making up 30-40% of the genome—suggest unique genomic architectures that may influence adaptation. The higher proportions of certain repeat types, particularly simple and RNA repeats, could facilitate rapid genomic changes, enhancing adaptability in dynamic environments like industrial fermentation.

These repeat variations among strains may indicate different genomic stability and plasticity strategies. For example, the high LTR content in DSM 3286 may signify historical transposon activity, promoting genomic rearrangements. In contrast, the more balanced and stable repeat landscape in E122 *MATA* and 22301-5 *MATB* suggests a refined evolutionary adaptation that could confer resilience in industrial contexts. The conserved nature of certain repeat types across strains, such as LINE elements, suggests shared functional roles across *Yarrowia* lineages, whereas the unique repeat profiles of E122 *MATA* and 22301-5 *MATB* reflect strain-specific evolutionary pressures.

rDNA Repeats and Size

Analysis of the distribution of rDNA repeats in the E122 *MATA* strain revealed distinct patterns in repeat length and counts across multiple chromosome regions, adjacent to the telomeres as shown in Figure 7A. Regions chrC-R, chrE-R, and chrF-R exhibit the most extended total rDNA lengths (~11 KB) with moderate counts, suggesting these regions contain larger rDNA repeats or a higher density of sizeable elements. In contrast, chrC-L has a shorter total length (~8 KB) and a lower count, indicating fewer and potentially smaller rDNA repeats. This heterogeneity could indicate region-specific roles or stability requirements for rDNA within the MAT-A strain. We note that the location of the rDNA repeats in the two strains analyzed herein is essentially the same as in DSM 3286, but the estimated number of rDNA repeats differs. We suggest that this difference reflects both technical and biological variation.

The stability of the rDNA and telomeric repeats needs explanation. In *S. cerevisiae*, the SIR proteins play an important role in the maintenance of the repeats by preventing recombination (14, 62). *Y. lipolytica* lacks the SIR proteins, except for *SIR*2, which is present in all eukaryotes, and it also lacks genes encoding RNAi components that in other eukaryotes suppress gene expression in heterochromatin (16, 63)). This raises the interesting issue of how the rDNA and telomeric repeats resist recombination and thus maintain stability.

GC Content and Evolutionary Implications. The overall GC content in *Y. lipolytica* is higher than in *S. cerevisiae*, which aligns with the significant evolutionary divergence between these species, estimated at around 300 million years. The consistent GC content across E122 *MATA*, 22301-5 *MATB*, and DSM 3286 (48.9%) compared to the low GC content of the mitochondrial chromosome (22.59%) suggests differences in selective pressures and genome organization between nuclear and mitochondrial genomes. This higher GC content may have implications for DNA stability, transcription efficiency, and DNA replication dynamics, offering insights into the evolutionary and functional constraints on the *Yarrowia* genome. For example, the well-characterized origins of DNA replication in *S. cerevisiae* are AT-rich. We are analyzing the genome replication and origins of DNA replication in *Yarrowia* to determine if the genome has GC-rich origins of DNA replication that are more akin to the GC-rich origins in human cells. The more complete genome sequences of the E122 *MATA* and 22301-5 *MATB* strains should facilitate the analysis of genome replication patterns and mechanisms.

Conclusions and Future Directions

Our findings provide a comprehensive view of the genomic landscape and diversity within *Yarrowia lipolytica* strains *MATA* and *MATB*, laying the groundwork for further research in functional genomics and strain optimization. The variability in repeat elements, the distinct genomic organization, and the elevated gene complexity observed in *MATA* and *MATB* highlight the evolutionary and functional divergence within *Y. lipolytica*. The insights gained from future studies of basic molecular biology in *Y. lipolytica* will contribute to our understanding of the molecular underpinnings that enable *Yarrowia lipolytica* to thrive in highly varied environments, ultimately advancing strain development for biotechnology.

DATA AVAILABILITY

The DNA sequence and annotation data are available at Dryad https://datadryad.org/share/cvheElifiD1l5ooEar-cRwTXDDAmKF8pr3ayrUHA8xg The data are:

Files and variables

File: matB_final.fasta Description**:** Strain 22301-5 Genome assembly

File: matA_final.fasta Description**:** Strain E122 Genome assembly

File: matA_AEDcln.gff Description: Strain E122 Transcriptome assembly and annotation

File: matB_AEDcln.gff Description: Strain 22301-5 Transcriptome assembly and annotation

AUTHOR CONTRIBUTIONS

N.Z. did the experiments and the DNA sequencing using the Cold Spring Harbor Laboratory Sequencing and Technology Core facility. O. E D., K.P., Z.L. and D.W. performed data analysis. B.S. conceived the project and oversaw all aspects of the research. N.Z. K.P. D.W. and B.S. wrote the paper.

ACKNOWLEDGEMENTS

The authors thank Dr. Richard A. Rachubunski, University of Alberta, Canada for providing the Yarrowi strains used in this project. We thank members of the Cold Spring Harbor Laboratory Sequencing and Technology Core facility for DNA sequence analysis.

FUNDING

Funding of this research was supported by a grant (GM45436) to B.S. from the National Institute of General Medical Sciences at the National Institutes of Health and a grant (8062-21000-051-000D) to D.W. from the United States Department of Agriculture (USDA), Agricultural Research Service (ARS). USDA ARS. The Cold Spring Harbor Laboratory Genome Sequencing and Technologies Core was supported in part by the Cold Spring Harbor Laboratory Cancer Center grant (CA13106) from the National Cancer Institute at the National Institutes of Health.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

Figures and Tables

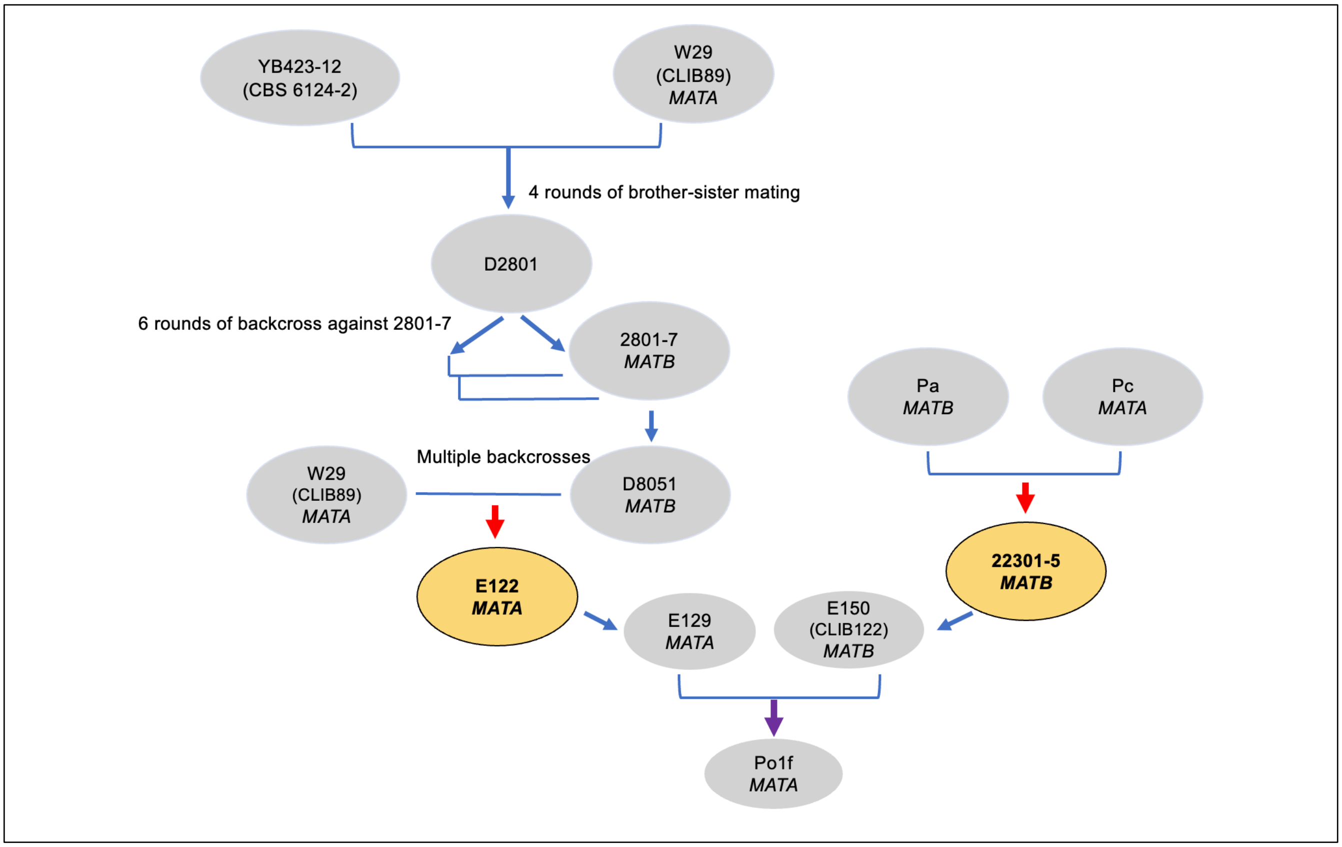
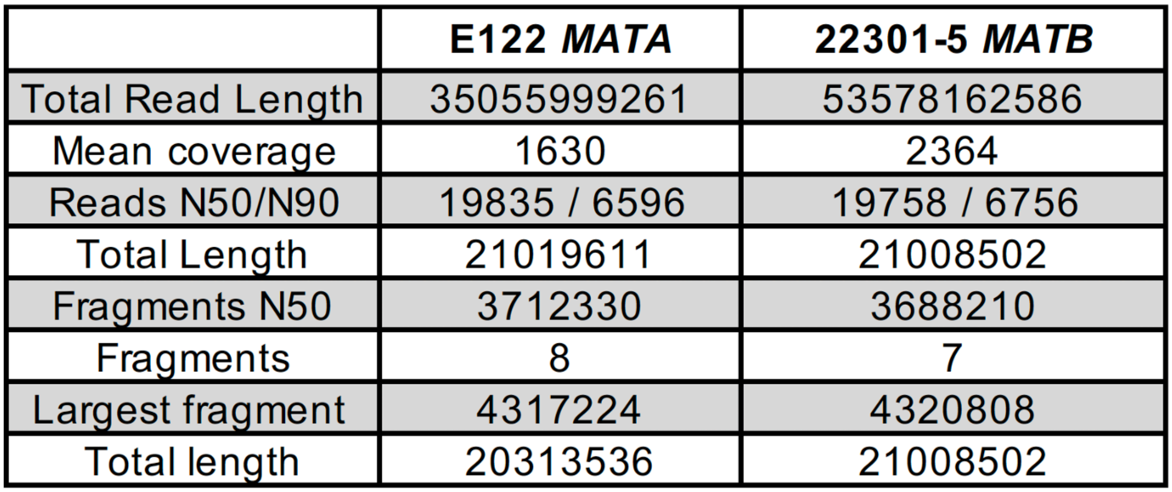


Figure 1: **Breeding and backcrossing strategy for strain development in *Yarrowia lipolytica*.** The flowchart illustrates the lineage and mating strategy employed to derive key strains of *Y. lipolytica*. YB423-12 lys1.13 and W29 (CLIB89 *MATA*) underwent four rounds of brother-sister mating to generate the intermediate strain D2801. D2801 was subjected to six rounds of backcrossing against 2801-7 *MATB*, resulting in the development of the strain D8051 *MATB*. Parallel strategies were employed using Pa *MATB* and Pc *MATA*, which were crossed to form 22301-5 *MATB*. E122 *MATA* was derived from multiple backcrosses and further developed into strains like E129 *MATA*, E150 (CLIB122 *MATB*), and Po1f *MATA*, widely used for research and industrial applications. Color-coded ovals indicate key final strains derived from these processes (e.g., E122 *MATA* and 22301-5 *MATB*). Blue arrows represent mating and backcrossing steps; red arrows indicate the lack of a *Ura3* marker, and purple arrows indicate the lack of *Xpr2* and *Axp* genes.



	E122 *MATA*	**22301-5 *MATB***
Total Read Length	35055999261	53578162586
Mean coverage	1630	2364
Reads N50/N90	19835 / 6596	19758 / 6756
Total Length	21019611	21008502
Fragments N50	3712330	3688210
Fragments	8	7
Largest fragment	4317224	4320808
Total length	20313536	21008502

Table 1:**Comparative genome assembly statistics for *Yarrowia lipolytica* strains E122 (*MATA*) and 22301-5 (*MATB*).**

Total Read Length: Total bases sequenced for each strain; Mean Coverage: Average sequencing coverage (sequencing depth); Reads N50/N90: Median (N50) and 90th percentile (N90) read lengths; Total Length: Total assembled genome length in base pairs; Fragments N50: Fragments: Number of assembled fragments; Largest Fragment: Largest assembled fragment in base pairs; Total Length (Alternate): Total length in base pairs across all fragments. These statistics provide a detailed comparison of genome assembly quality, highlighting the structural integrity of the two *Yarrowia lipolytica* strains.

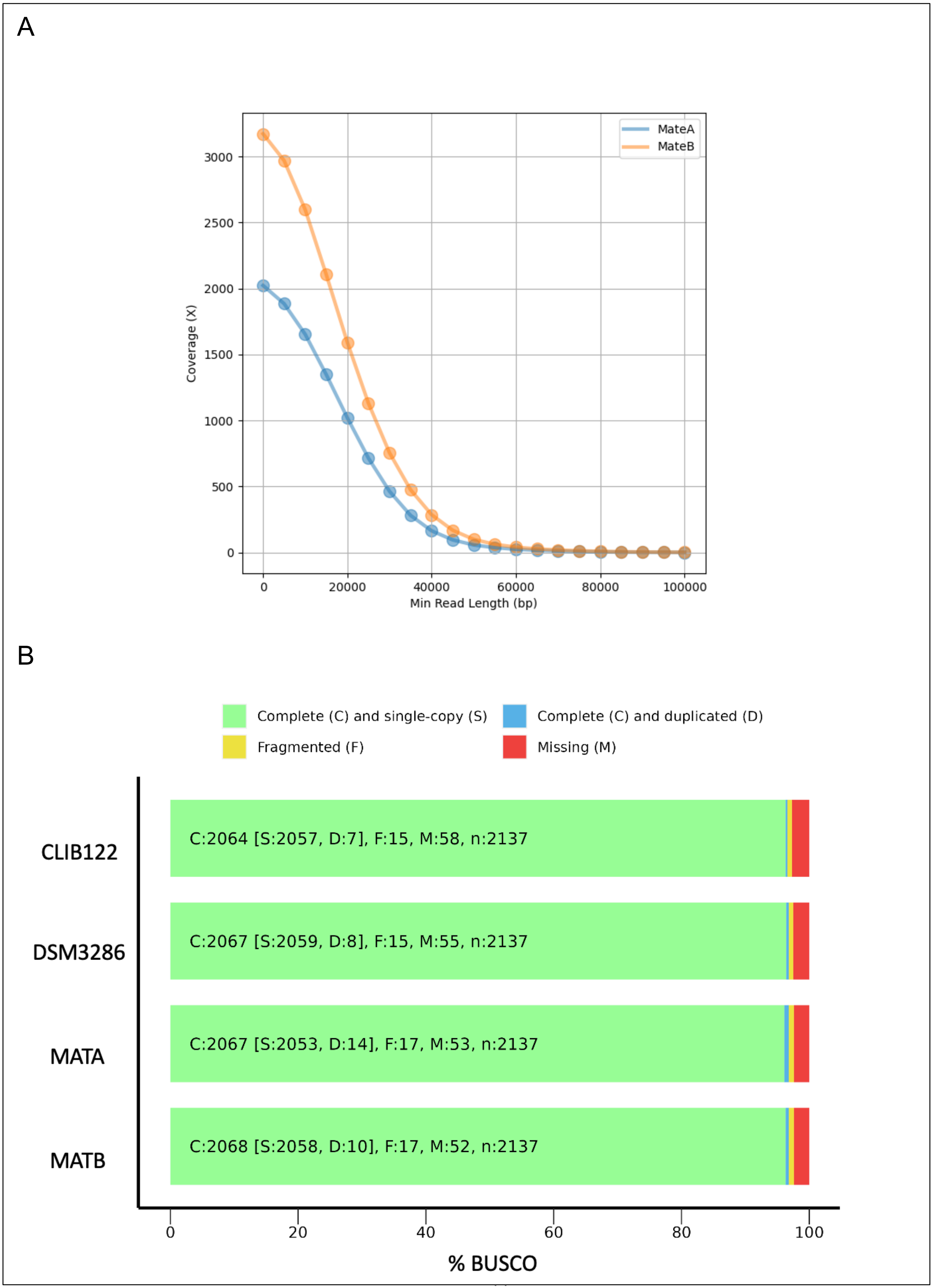

A
MateA
MateB
Coverage (X)
3000
2500
2000
1500
1000
500
0
0
20000
40000
60000
80000
100000
Min Read Length (bp)
B
🟩 Complete (C) and single-copy (S)
🟦 Complete (C) and duplicated (D)
🟨 Fragmented (F)
🟥 Missing (M)
CLIB122
C:2064 [S:2057, D:7], F:15, M:58, n:2137
DSM3286
C:2067 [S:2059, D:8], F:15, M:55, n:2137
MATA
C:2067 [S:2053, D:14], F:17, M:53, n:2137
MATB
C:2068 [S:2058, D:10], F:17, M:52, n:2137
0
20
40
60
80
100
% BUSCO

Figure 2. Comparison of genome assembly metrics between *Yarrowia lipolytica* strains E122 (*MATA*) and 22301-5 (*MATB*). A. Coverage versus minimal read length distribution: Coverage (X) is plotted against minimal read lengths for both strains. The blue line represents E122 *MATA*, while the orange line represents 22301-5 *MATB*. The higher coverage for 22301-5 *MATB* indicates a more profound sequencing effort compared to E122 *MATA*. The gradual decline in coverage with increasing read length reflects the expected distribution of read sizes. **B.** BUSCO analysis for genome completeness across four *Yarrowia lipolytica* strains (CLIB122, DSM 3286, E122 *MATA*, and 22301-5 *MATB*). C: Number of complete BUSCO genes, divided into S: Single-copy genes and D: Duplicated genes. F: Number of fragmented BUSCO genes. M: Number of missing BUSCO genes. N: Total number of BUSCO groups analyzed (2137). Each bar represents the percentage distribution of these categories for a strain. Most BUSCO groups are complete and single-copy (light green), reflecting high genome assembly quality. Slight variations in duplicated (blue), fragmented (yellow), and missing (red) categories highlight subtle differences in genome assemblies among the strains.

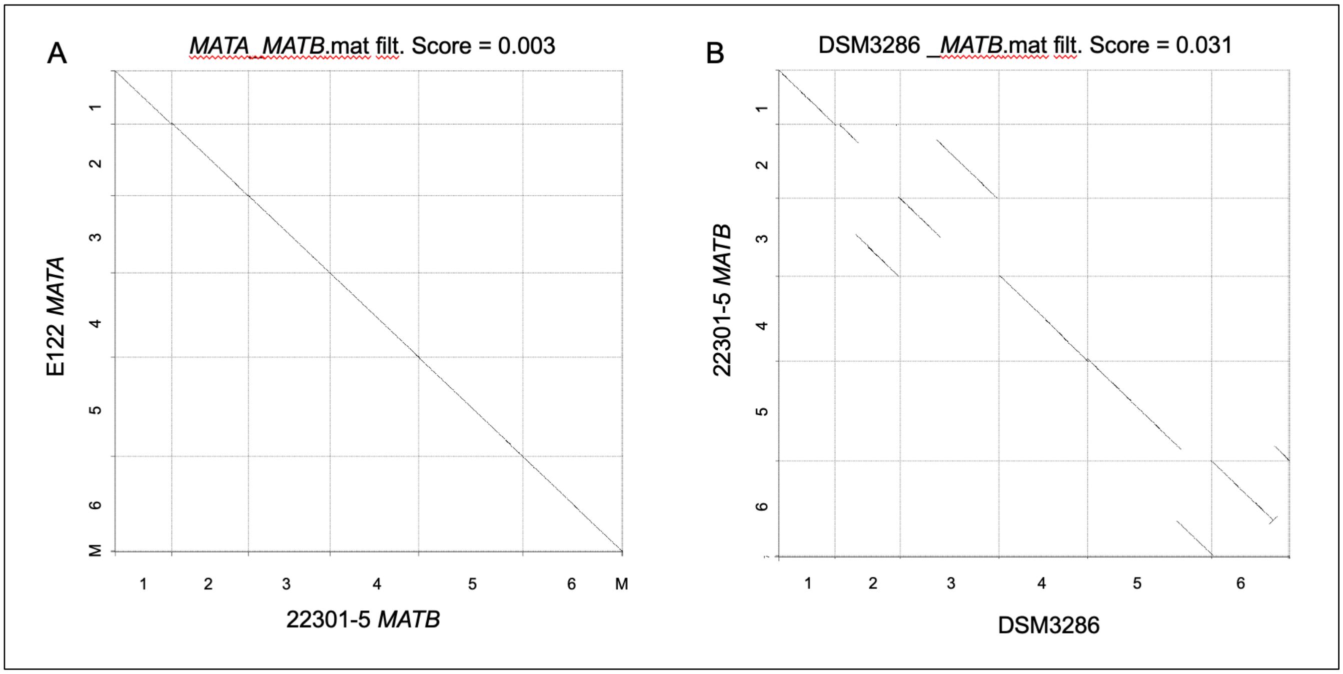
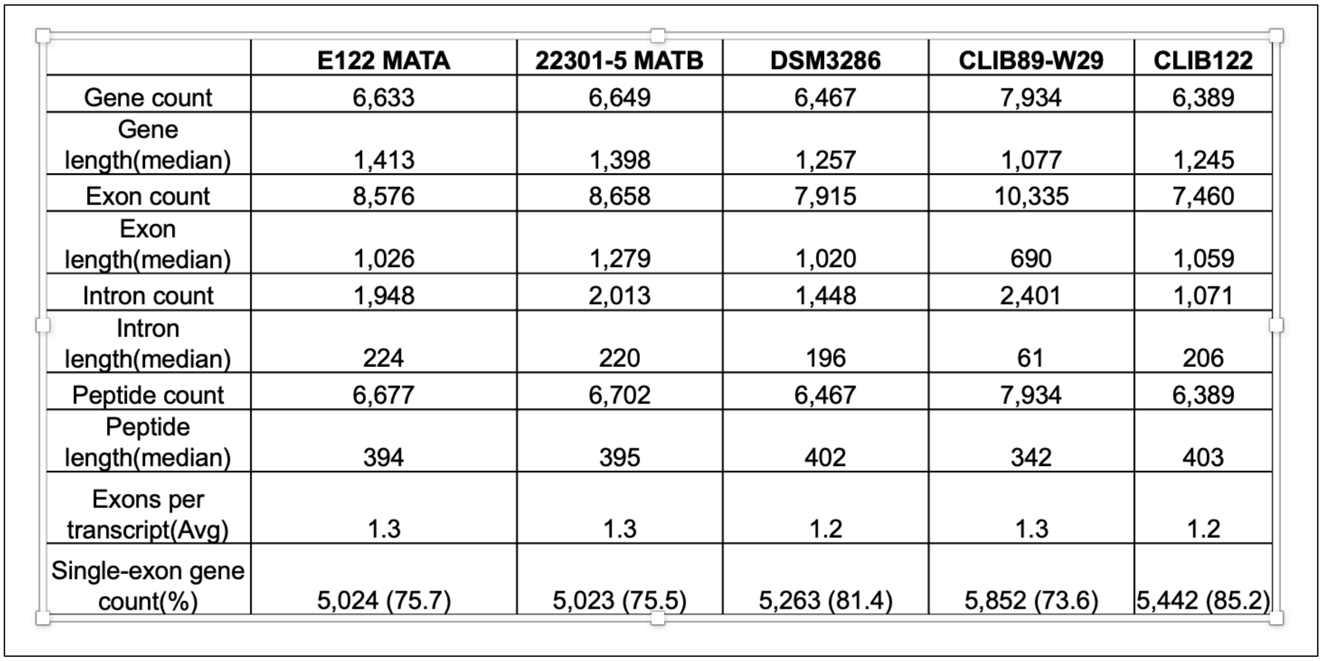


Figure 3. Assembly alignment comparison: **A.** Dot plot of the final assemblies for E122 *MATA* and 22301-5 *MATB* shows synteny between the two genomes. The diagonal alignment highlights the high degree of sequence conservation, with a calculated score of 0.003, indicating minor structural or sequence variations between the two genomes. This data highlights the quality of sequencing and genome assembly while comparing structural differences between these two strains of *Y. lipolytica*. **B.** Dot plot of the sequences for 22301-5 *MATB* and DSM 2386. The diagonal alignment highlights the genome rearrangements between the two strains, thereby reflecting a lower calculated score of 0.031.



	E122 MATA	**22301-5 MATB**	**DSM3286**	**CLIB89-W29**	**CLIB122**
Gene count	6,633	6,649	6,467	7,934	6,389
Gene length(median)	1,413	1,398	1,257	1,077	1,245
Exon count	8,576	8,658	7,915	10,335	7,460
Exon length(median)	1,026	1,279	1,020	690	1,059
Intron count	1,948	2,013	1,448	2,401	1,071
Intron length(median)	224	220	196	61	206
Peptide count	6,677	6,702	6,467	7,934	6,389
Peptide length(median)	394	395	402	342	403
Exons per transcript(Avg)	1.3	1.3	1.2	1.3	1.2
Single-exon gene count(%)	5,024 (75.7)	5,023 (75.5)	5,263 (81.4)	5,852 (73.6)	5,442 (85.2)

Table 2: Comparative Gene and Transcriptomic Features of *Yarrowia lipolytica* Strains. This table presents the genomic and transcriptomic characteristics of four *Yarrowia lipolytica* strains (E122 *MATA*, 22301-5 *MATB*, DSM3286, and CLIB122), highlighting differences in gene structure and composition. The *Yarrowia lipolytica* strains exhibit notable variations in gene structure, with 22301-5 *MATB* having the highest gene count (7,769) and DSM3286 the lowest (6,439), while E122 *MATA* has the longest median gene length (1,473 bp). The high proportion of single-exon genes (78.2–85.2%) and low exons per transcript (~1.2–1.3) suggest a predominantly condensed genetic structure, with CLIB122 having the most compact gene architecture and E122 MATA showing greater structural complexity. This data provides insights into structural genomic variation and transcriptional complexity across these *Y. lipolytica* strains.

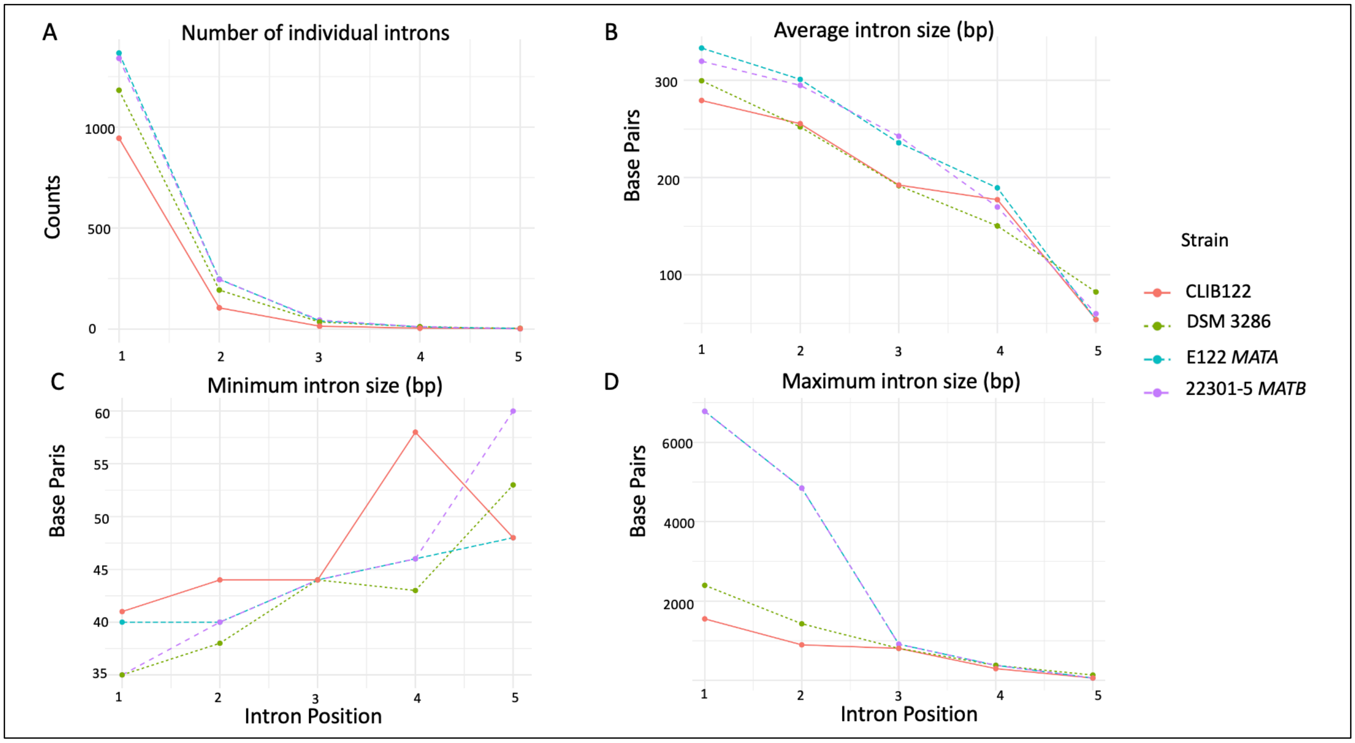


Figure 4. Comparative analysis of intron properties across *Yarrowia lipolytica* strains CLIB122, DSM 3286, E122 *MATA*, and 22301-5 *MATB*.

A. A line plot of the number of individual introns observed at each gene position. The first intron position is the most prevalent across all strains, with a sharp decline in frequency for subsequent positions. **B.** Line plot showing the average size of introns for each strain across different intron positions. 22301-5 *MATB* exhibits the highest average intron size for more 3' intron positions compared to other strains. **C.** Line plot of the minimum size of introns at each intron position within the genes, with E122 *MATA* and 22301-5 *MATB* showing an increase in size for more 3' intron positions compared to other strains. **D.** Line plot showing the maximum intron size for each intron position within the genes. E122 *MATA* and 22301-5 *MATB* shows the highest number of introns for the initial positions, followed by a steep decline.

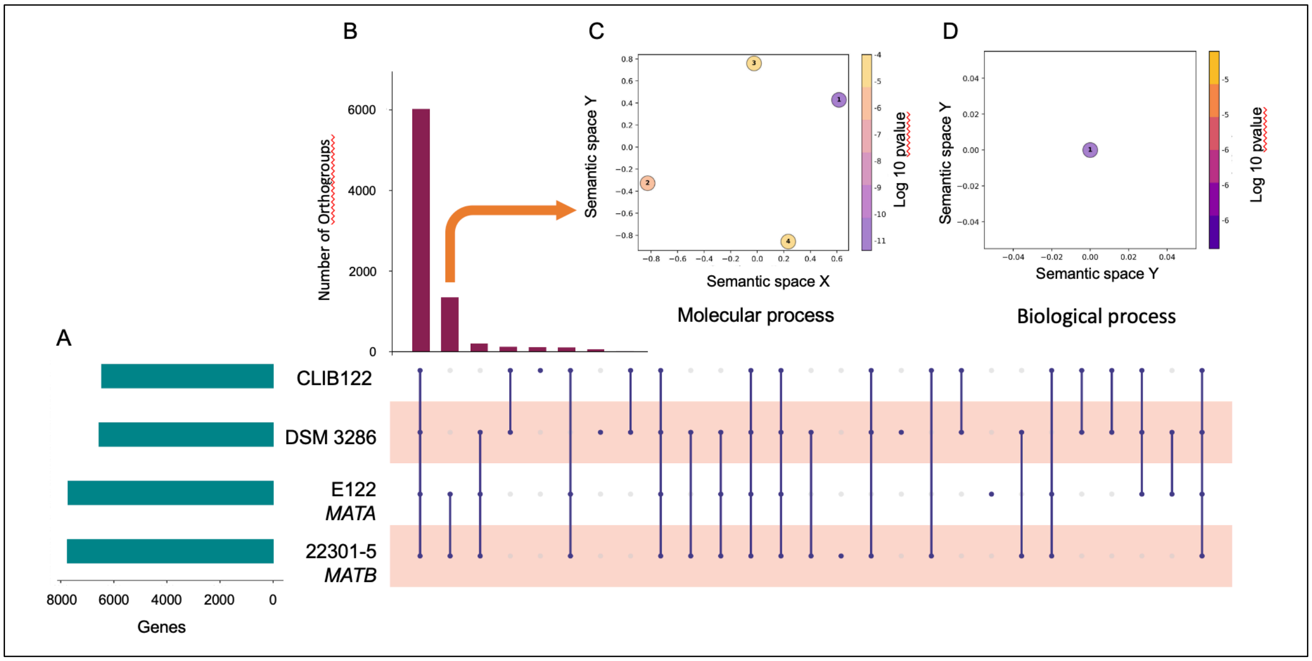
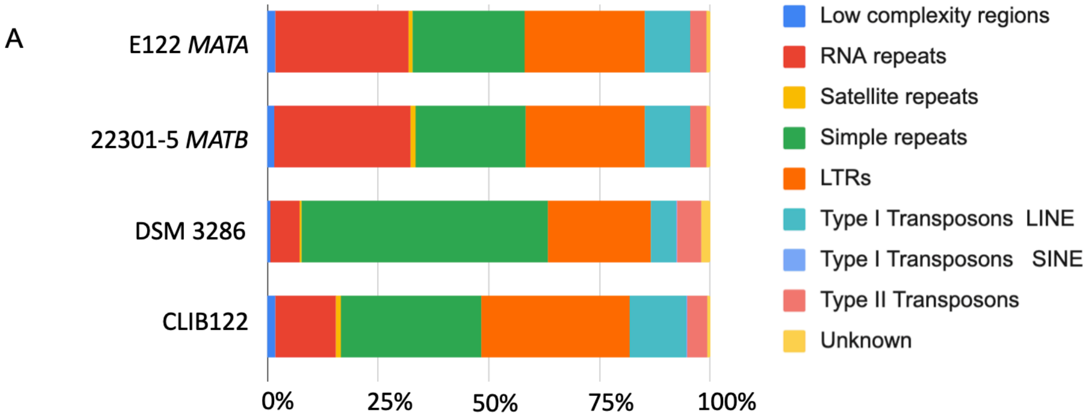


Figure 5. Comparative analysis of orthologous gene groups, functional enrichment, and gene distributions among four *Yarrowia* strains: CLIB122, DSM 3286, E122 *MATA*, and 22301-5 *MATB*. A. Left, the green bar graph shows the number of predicted genes using an unfiltered analysis for each of the four strains**. Right,** a comparison of gene orthology across the strains, with vertical lines connecting shared orthologous genes shared between the indicated strains. Unconnected points indicate unique contributions. **B.** Bar Chart showing the number of genes in each of the orthologous groups shown in A, right panel, that are shared among the strains. Most orthologs are shared across all strains (left most bar in panel B). The second bar in panel B represents orthologs shared between E122 *MATA* and 22301-5 *MATB* but not found in DSM 3286 or CLIB122 sequences. C. Insert shows a scatter plot showing functional enrichment analysis in the molecular process category identifies significant GO terms such as "methylmalonate-semialdehyde dehydrogenase activity" (1), "TBP-class protein binding" (2), and "oxidoreductase activity" (3). Points are colored by the significance (log10 p-value) **D.** Insert shows a scatter plot of Biological Process): Enrichment in the biological process category highlighted "DNA-templated transcription initiation" as a key process. This multi-dimensional analysis underscores the conserved and divergent functional pathways and genetic architecture of these *Yarrowia* strains



B

Strain	Low complexity regions	RNA repeats	Satellite repeats	Simple repeats	LTRs	Type I Transposons LINE	Type I Transposons SINE	Type II Transposons	Unknown	Total
E122 *MATA*	22	357	12	301	321	123	1	42	11	1190
22301-5 *MATB*	19	376	12	302	329	123	1	44	11	1217
DSM 3286	4	34	2	287	121	29	1	29	10	517
CLIB122	17	128	12	296	314	121	1	43	7	939

Figure 6. Comparative repeat composition in the genomes of *Yarrowia lipolytica* strains *MATA*, *MATB*, DSM 3286, and CLIB122. A. The stacked bar chart represents the proportional distribution of various repeat classes, including: Low complexity regions (blue); RNA repeats (red); Satellite repeats (yellow); Simple repeats (green); LTRs (orange); Type I Transposons (LINE) (teal); Type I Transposons (SINE) (light blue); Type II Transposons (pink); Unknown repeats (yellow). Each bar represents the total percentage of genomic content attributed to these repeat types in the corresponding strain, highlighting variations in transposable elements and repetitive sequences across strains. This analysis reveals genome-wide repeat diversity and relative abundance, including transposons and low-complexity regions. **B.** The counts of the various repeat elements described in A in the four strains.

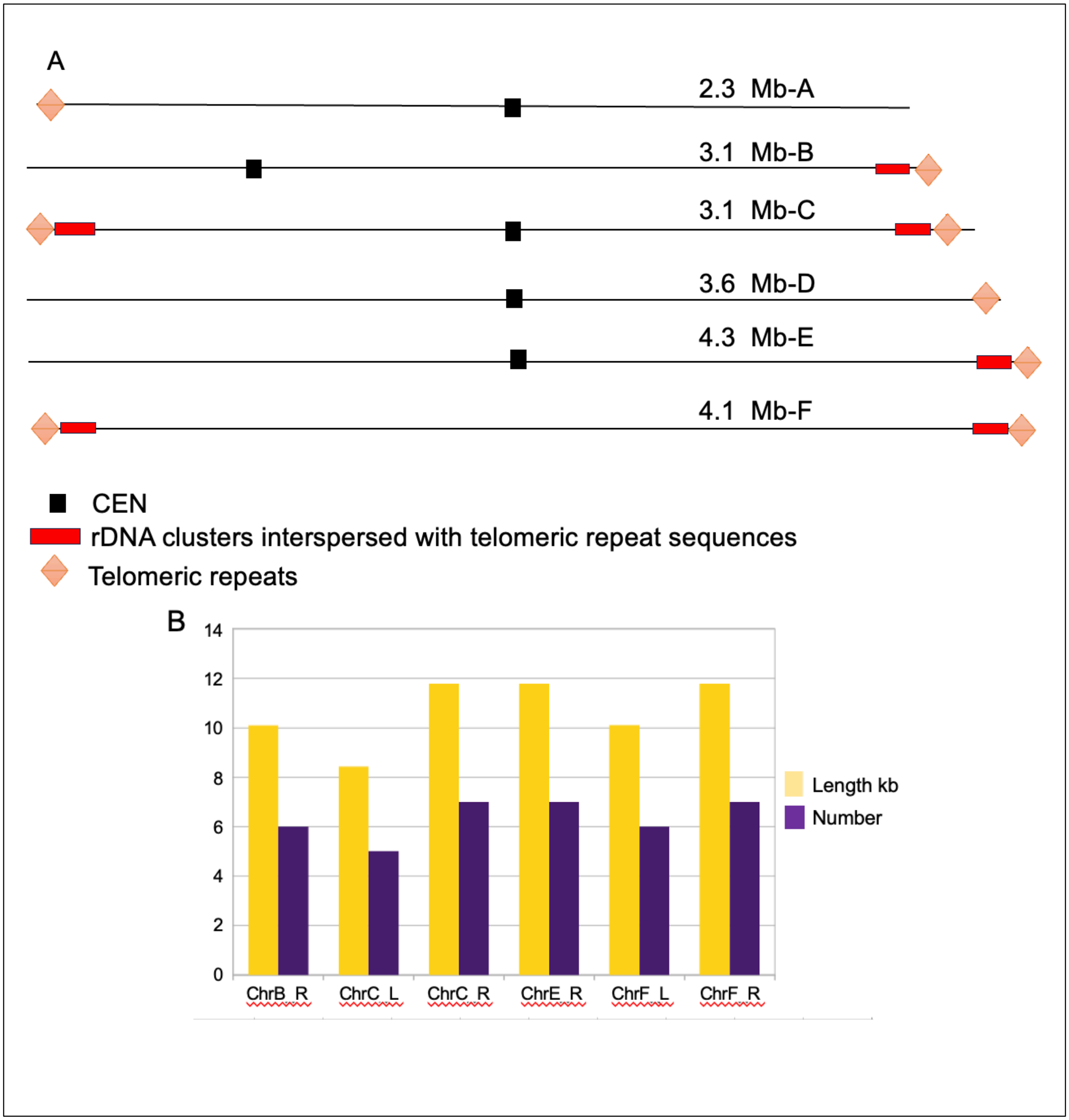


Figure 7. Analysis of rDNA Repeats.

A. Genome landscape of six chromosomes of *Y. lipolytica* mapped rDNA (red bar) and telomere (orange diamond) sequences. The known centromere sequences are shown with black bars. **B.** This bar chart compares the Length (in kilobases, yellow bars) and Count (purple bars) of annotated features for specific regions across chromosomes in the genome. The x-axis represents distinct chromosome regions, including ChrB-R (Right arm of Chromosome B), ChrC-L (Left arm of Chromosome C), ChrC-R (Right arm of Chromosome C), ChrE-R (Right arm of Chromosome E), ChrF-L (Left arm of Chromosome F), ChrF-R (Right arm of Chromosome F). The yellow bars indicate the cumulative length of the regions (in kilobases), while the purple bars indicate the total count of repeats identified within these regions.

REFERENCES

1. Gonçalves,F.A.G., Colen,G. and Takahashi,J.A. (2014) Yarrowia lipolytica and Its Multiple Applications in the Biotechnological Industry. *Sci. World J.*, **2014**, 476207.

2. Mamaev,D. and Zvyagilskaya,R. (2021) Yarrowia lipolytica : A multitalented yeast species of ecological significance. *Fems Yeast Res*, 10.1093/femsyr/foab008.

3. Bankar,A.V., Kumar,A.R. and Zinjarde,S.S. (2009) Environmental and industrial applications of Yarrowia lipolytica. *Appl. Microbiol. Biotechnol.*, **84**, 847.

4. Groenewald,M., Boekhout,T., Neuvéglise,C., Gaillardin,C., Dijck,P.W.M. van and Wyss,M. (2014) Yarrowia lipolytica: Safety assessment of an oleaginous yeast with a great industrial potential. *Crit. Rev. Microbiol.*, **40**, 187–206.

5. Barth,G. and Gaillardin,C. (1997) Physiology and genetics of the dimorphic fungus Yarrowia lipolytica. *Fems Microbiol Rev*, **19**, 219–237.

6. Magnan,C., Yu,J., Chang,I., Jahn,E., Kanomata,Y., Wu,J., Zeller,M., Oakes,M., Baldi,P. and Sandmeyer,S. (2016) Sequence Assembly of Yarrowia lipolytica Strain W29/CLIB89 Shows Transposable Element Diversity. *Plos One*, **11**, e0162363.

7. Fumasoni,M. and Murray,A.W. (2020) The evolutionary plasticity of chromosome metabolism allows adaptation to constitutive DNA replication stress. *Elife*, **9**, e51963.

8. Mekouar,M., Blanc-Lenfle,I., Ozanne,C., Silva,C.D., Cruaud,C., Wincker,P., Gaillardin,C. and Neuvéglise,C. (2010) Detection and analysis of alternative splicing in Yarrowia lipolytica reveal structural constraints facilitating nonsense-mediated decay of intron-retaining transcripts. *Genome Biol*, **11**, R65.

9. Hu,Y., Tareen,A., Sheu,Y.-J., Ireland,W.T., Speck,C., Li,H., Joshua-Tor,L., Kinney,J.B. and Stillman,B. (2020) Evolution of DNA replication origin specification and gene silencing mechanisms. *Nat Commun*, **11**, 5175.

10. Marahrens,Y. and Stillman,B. (1992) A yeast chromosomal origin of DNA replication defined by multiple functional elements. *Science (New York, NY)*, **255**, 817–823.

11. Bell,S.P. and Labib,K. (2016) Chromosome Duplication in Saccharomyces cerevisiae. *Genetics*, **203**, 1027–1067.

12. 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.

13. Rusche,L.N. and Hickman,M.A. (2020) Sex in Fungi. 10.1128/9781555815837.ch11.

14. Rusche,L.N., Kirchmaier,A.L. and Rine,J. (2003) The establishment, inheritance, and function of silenced chromatin in Saccharomyces cerevisiae. *Annual review of biochemistry*, **72**, 481–516.

15. Hickman,M.A., Froyd,C.A. and Rusche,L.N. (2011) Reinventing heterochromatin in budding yeasts: Sir2 and the origin recognition complex take center stage. *Eukaryotic cell*, **10**, 1183–1192.

16. Smith,J.S., Brachmann,C.B., Celic,I., Kenna,M.A., Muhammad,S., Starai,V.J., Avalos,J.L., Escalante-Semerena,J.C., Grubmeyer,C., Wolberger,C., *et al.* (2000) A phylogenetically conserved NAD+-dependent protein deacetylase activity in the Sir2 protein family. *Proc. Natl. Acad. Sci.*, **97**, 6658–6663.

17. Bell,S.P., Kobayashi,R. and Stillman,B. (1993) Yeast Origin Recognition Complex Functions in Transcription Silencing and DNA Replication. *Science*, **262**, 1844–1849.

18. Hickman,M.A. and Rusche,L.N. (2010) Transcriptional silencing functions of the yeast protein Orc1/Sir3 subfunctionalized after gene duplication. *Proceedings of the National Academy of Sciences*, **107**, 19384–19389.

19. Hou,Z., Bernstein,D.A., Fox,C.A. and Keck,J.L. (2005) Structural basis of the Sir1–origin recognition complex interaction in transcriptional silencing. *Proc. Natl. Acad. Sci.*, **102**, 8489–8494.

20. Grunstein,M. and Gasser,S.M. (2013) Epigenetics in Saccharomyces cerevisiae. *Cold Spring Harbor Perspectives in Biology*, **5**, a017491.

21. Maria,H. and Rusche,L.N. (2022) The DNA replication protein Orc1 from the yeast Torulaspora delbrueckii is required for heterochromatin formation but not as a silencer-binding protein. *Genetics*, **222**, iyac110.

22. Hu,Y. and Stillman,B. (2023) Origins of DNA replication in eukaryotes. *Mol Cell*, **83**, 352–372.

23. Hyrien,O., Guilbaud,G. and Krude,T. (2025) The double life of mammalian DNA replication origins. *Genes Dev.*, **39**, 304–324.

24. Devillers,H. and Neuvéglise,C. (2019) Genome Sequence of the Oleaginous Yeast Yarrowia lipolytica H222. *Microbiol. Resour. Announc.*, **8**, 10.1128/mra.01547-18.

25. Bigey,F., Pasteur,E., Połomska,X., Thomas,S., Coq,A.-M.C.-L., Devillers,H. and Neuvéglise,C. (2023) Insights into the Genomic and Phenotypic Landscape of the Oleaginous Yeast Yarrowia lipolytica. *J. Fungi*, **9**, 76.

26. Liu,L. and Alper,H.S. (2014) Draft Genome Sequence of the Oleaginous Yeast Yarrowia lipolytica PO1f, a Commonly Used Metabolic Engineering Host. *Genome Announc.*, **2**, e00652-14.

27. Luttermann,T., Rückert,C., Wibberg,D., Busche,T., Schwarzhans,J.-P., Friehs,K. and Kalinowski,J. (2021) Establishment of a near-contiguous genome sequence of the citric acid producing yeast Yarrowia lipolytica DSM 3286 with resolution of rDNA clusters and telomeres. *Nar Genom Bioinform*, **3**, lqab085-.

28. Devillers,H., Brunel,F., Połomska,X., Sarilar,V., Lazar,Z., Robak,M. and Neuvéglise,C. (2016) Draft Genome Sequence of Yarrowia lipolytica Strain A-101 Isolated from Polluted Soil in Poland. *Genome Announc.*, **4**, e01094-16.

29. Madzak,C. (2021) Yarrowia lipolytica Strains and Their Biotechnological Applications: How Natural Biodiversity and Metabolic Engineering Could Contribute to Cell Factories Improvement. *J Fungi*, **7**, 548.

30. Gaillardin,G.B. and C. and Wolf (2019) Yarrowia lipolytica.

31. Mansour,S., Bailly,J., Landaud,S., Monnet,C., Sarthou,A.S., Cocaign-Bousquet,M., Leroy,S., Irlinger,F. and Bonnarme,P. (2009) Investigation of Associations of Yarrowia lipolytica, Staphylococcus xylosus, and Lactococcus lactis in Culture as a First Step in Microbial Interaction Analysis. *Appl. Environ. Microbiol.*, **75**, 6422–6430.

32. Kolmogorov,M., Yuan,J., Lin,Y. and Pevzner,P.A. (2019) Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.*, **37**, 540–546.

33. Martin and M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal [S.I.]*.

34. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

35. Danecek,P., Bonfield,J.K., Liddle,J., Marshall,J., Ohan,V., Pollard,M.O., Whitwham,A., Keane,T., McCarthy,S.A., Davies,R.M., *et al.* (2021) Twelve years of SAMtools and BCFtools. *GigaScience*, **10**, giab008.

36. Walker,B.J., Abeel,T., Shea,T., Priest,M., Abouelliel,A., Sakthikumar,S., Cuomo,C.A., Zeng,Q., Wortman,J., Young,S.K., *et al.* (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS ONE*, **9**, e112963.

37. Alonge,M., Soyk,S., Ramakrishnan,S., Wang,X., Goodwin,S., Sedlazeck,F.J., Lippman,Z.B. and Schatz,M.C. (2019) RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.*, **20**, 224.

38. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

39. Pertea,M., Pertea,G.M., Antonescu,C.M., Chang,T.-C., Mendell,J.T. and Salzberg,S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.

40. Kainth,A.S., Haddad,G.A., Hall,J.M. and Ruthenburg,A.J. (2023) Merging short and stranded long reads improves transcript assembly. *PLOS Comput. Biol.*, **19**, e1011576.

41. Pertea,G. and Pertea,M. (2020) GFF Utilities: GffRead and GffCompare. *F1000Research*, **9**, ISCB Comm J-304.

42. Campbell,M.S., Holt,C., Moore,B. and Yandell,M. (2014) Genome Annotation and Curation Using MAKER and MAKER-P. *Curr. Protoc. Bioinform.*, **48**, 4.11.1-4.11.39.

43. Huang,Y., Niu,B., Gao,Y., Fu,L. and Li,W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.

44. Trincado,J.L., Entizne,J.C., Hysenaj,G., Singh,B., Skalic,M., Elliott,D.J. and Eyras,E. (2018) SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.*, **19**, 40.

45. Solovyev,V., Kosarev,P., Seledsov,I. and Vorobyev,D. (2006) Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.*, **7**, S10.

46. Stanke,M., Diekhans,M., Baertsch,R. and Haussler,D. (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, **24**, 637–644.

47. Haas,B.J., Delcher,A.L., Mount,S.M., Wortman,J.R., Jr,R.K.S., Hannick,L.I., Maiti,R., Ronning,C.M., Rusch,D.B., Town,C.D., *et al.* (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.*, **31**, 5654–5666.

48. Jones,P., Binns,D., Chang,H.-Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G., *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.

49. Olson,A.J. and Ware,D. (2021) Ranked choice voting for representative transcripts with TRaCE. *Bioinformatics*, **38**, 261–264.

50. Stabenau,A., McVicker,G., Melsopp,C., Proctor,G., Clamp,M. and Birney,E. (2004) The Ensembl Core Software Libraries. *Genome Res.*, **14**, 929–933.

51. Vilella,A.J., Severin,J., Ureta-Vidal,A., Heng,L., Durbin,R. and Birney,E. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.

52. Simão,F.A., Waterhouse,R.M., Ioannidis,P., Kriventseva,E.V. and Zdobnov,E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

53. Pérez-Wohlfeil,E., Diaz-del-Pino,S. and Trelles,O. (2019) Ultra-fast genome comparison for large-scale genomic experiments. *Sci. Rep.*, **9**, 10274.

54. Rosas-Quijano,R., Gaillardin,C. and Ruiz-Herrera,J. (2008) Functional analysis of the MATB mating-type idiomorph of the dimorphic fungus Yarrowia lipolytica. *Current Microbiol*, **57**, 115–120.

55. Kurischko,C., Schilhabel,M.B., Kunze,I. and Franzl,E. (1999) The MAT A locus of the dimorphic yeast Yarrowia lipolytica consists of two divergently oriented genes. *Mol. Gen. Genet. MGG*, **262**, 180–188.

56. Dujon,B., Sherman,D., Fischer,G., Durrens,P., Casaregola,S., Lafontaine,I., Montigny,J. de, Marck,C., Neuvéglise,C., Talla,E., *et al.* (2004) Genome evolution in yeasts. *Nature*, **430**, 35–44.

57. McCarthy,C.G.P. and Fitzpatrick,D.A. (2019) Pangloss: A Tool for Pan-Genome Analysis of Microbial Eukaryotes. *Genes*, **10**, 521.

58. Lee,S.C., Ni,M., Li,W., Shertz,C. and Heitman,J. (2010) The Evolution of Sex: a Perspective from the Fungal Kingdom. *Microbiol. Mol. Biol. Rev.*, **74**, 298–340.

59. Haber,J.E. (2012) Mating-Type Genes and MAT Switching in Saccharomyces cerevisiae. *Genetics*, **191**, 33–64.

60. Hanson,S.J. and Wolfe,K.H. (2017) An Evolutionary Perspective on Yeast Mating-Type Switching. *Genetics*, **206**, 9–32.

61. Dujon,B.A. and Louis,E.J. (2017) Genome Diversity and Evolution in the Budding Yeasts (Saccharomycotina). *Genetics*, **206**, 717–750.

62. Kueng,S., Oppikofer,M. and Gasser,S.M. (2013) SIR proteins and the assembly of silent chromatin in budding yeast. *Annual review of genetics*, **47**, 275–306.

63. Blander,G. and Guarente,L. (2004) The SIR2 Family of Protein Deacetylases. *Biochemistry*, **73**, 417–435.