Università degli Studi di Modena e Reggio Emilia

Department of Life Sciences

Ph.D. School in Agri-Food Sciences, Technologies and Biotechnologies
(XXXI Ph.D. Cycle)

Molecular Mechanisms of Cell Identity and Hybrid Sterility in an Allodiploid *Zygosaccharomyces* Yeast Unravelled by Synthetic Biology and *de novo* Genome Assembly from Miseq and Minion Platforms

Supervisor:
Dott.ssa Lisa Solieri

Co-Supervisor:
Dott. Stefano Cassanelli

Ph.D. Student:
Melissa Bizzarri

Coordinator of the Ph.D. School:
Prof. Alessandro Ulrici

Academic Year 2017-2018

Acknowledgements

Many thanks to everyone in the Unimore UMCC lab past and present: Paolo, Lisa, Stefano, Tikam, Veronica, Tommaso, Gabriele, Alexandra, Andrea, Marcello, Giulia, Francesco, Laura, Maria and Luciana and anyone else I may have forgotten. You made it a lot of fun and it was a pleasure to work with all of you.

Many thanks also to Prof. Hana Sychrovà and everyone in her lab in Prague: Micha, Vicent, Hanka, Barbora. I had a great time in Czech Republic and learned more in three months than I have done in a long time.

Also thanks to all my collaborators along the way, on the various projects that did work out, did not work out and will work out in the future: Leszek and Jan from Warsaw that helped me in sequencing and assembling two challenging genomes.

Special thanks to Lisa Solieri and Stefano Cassanelli for being great supervisors and mentors - critical, insightful and the kind of people that is easy to learn from. Thanks for all the opportunities to travel and meet people as well as to learn.

Finally, I want to dedicate this to my parents. They always believed that education was the way to a happier life and always backed me whatever I wanted to do. Many thanks for everything.

ABSTRACT

Many interspecies hybrid yeasts are exploited in food biotechnology, but most of them are asexual and can replicate only mitotically, hampering breeding programs. *Zygsaccharomyces* clade includes several natural interspecies hybrids; currently two of them have been recently sequenced, opening a new era in functional and comparative genomics of these non-conventional yeasts. Recent studies have proposed that interspecies hybrids belonging to the *Zygosaccharomyces bailii* species complex can regain a complete sexual cycle by damage of one allele at the mating-type (*MAT*) locus. This evolutionary route is referred to as whole-genome duplication (WGD) after interspecies hybridization, and was supposed to be the mechanism occurred in an ancestor of *Saccharomyces cerevisiae* to rescue fertility.

Zygosaccharomyces rouxii is an industrial osmo and halo-tolerant promising cell factory that occupies a phylogenetic peculiar position in the hemiascomycetes lineage since it diverged from *S. cerevisiae* before WGD. The aim of this thesis is to elucidate the mechanisms underlying sterility in *Z. rouxii* model hybrid ATCC 42981 by developing synthetic biology tools and third genome sequencing technologies (3GS). Firstly, we investigated the life cycle of the *Z. rouxii* haploid type-strain CBS 732^T, focusing on how mating-type switching acts as a plastic process that causes intra-strain genetic instability and phenotypic novelties in haploid cells derived from different homothallic CBS 732^T stock cultures. Secondly, we investigated the sex determination system of the *Z. rouxii* allodiploid and sterile ATCC 42981 strain and used it as a model to study how a chimeric mating-type gene repertoire contributes to hybrid reproductive isolation. We performed draft genome sequencing of ATCC 42981 combining Illumina MiSeq and Oxford Nanopore MinION platforms to decipher the genetic basis of the transcriptional network incompatibility underlying hybrid sterility and of the adaptation to high salt and sugar concentrations in non-conventional yeasts.

The initial assembly had 33 nuclear scaffolds with a cumulative size of 20.9 Mb; 10,254 genes were annotated by similarity to the closest *Z.rouxii* haploid relative CBS 732^T using Exonerate v2.2.0. Comparison with the type-strain showed that ATCC 42981 assembled genome had a 2.14 time larger size and almost twice genes number. The assembly completeness was supported by the content of near-universal single copy orthologs (BUSCO v3: 98.28%). Gene prediction and annotation were further refined combining *ab initio* and evidence based approaches. We annotated 10,821 protein-coding genes using the Yeast Genome Annotation Pipeline (YGAP) and 10,510 using Funannotate pipeline v1.3.4. Three haplotypes were dissected, one of them was identical to that of CBS 732^{T}, relaying on scaffold synteny mapping, as well as ortholog and homeomolog genes; suggesting a recursive hybridization model. The draft genome sequence assists us in resolving incongruences in mating-type like (*MTL*) loci organization between two different stock cultures of strain ATCC 42981, namely our infertile in-house stock culture and the fertile Japanese sister stock JCM22060. Two different hybrid *de novo* genome assemblies were combined with *in vitro* long PCR approach to demonstrate that ATCC 42981 exhibits a different assessment of *MTL* loci compared to JCM22060, retaining two transcriptionally active *MAT* cassettes. When *MAT*α locus was deleted, the ATCC 42981 *MAT***a**/- hemizygous mutant did not rescue the ability to either sporulate or mate, as incomplete *HML*α silencing masks the loss of heterozygosity at the *MAT* locus induced by gene deletion. Overall, our findings contribute to elucidate the mechanisms of hybrid sterility and the circuits that determine cell identity in *Z. rouxii* pre-WGD species.

Keywords: *Zygosaccharomyces rouxii*, mating-type, hybrid, sterility, MinION, genome sequencing, genome annotation, heterozygous genome assembly, orthologs.

RIASSUNTO

Molti lieviti ibridi interspecifici trovano applicazione nelle biotecnologie alimentari, tuttavia, molti sono incapaci di riprodursi per via sessuale, precludendo l'utilizzo di programmi di breeding. Il complesso *Zygosaccharomyces* comprende diverse specie naturali interspecifiche e, tra queste, due sono state da poco sequenziate, aprendo una nuova era nella genomica funzionale e comparativa dei lieviti non convenzionali. Studi recenti ipotizzano che ibridi interspecifici di *Zygosaccharomyces bailii* ripristinano un ciclo sessuale completo in seguito al breakage di un allele del locus codificante il mating-type (*MAT*). Questo meccansimo evolutivo, responsabile del ripristino della fertilità nell'ancestore di *Saccharomyces cerevisiae*, risiede nella whole-genome duplication (WGD) avvenuta dopo l'ibridizzazione interspecifica. *Zygosaccharomyces rouxii* è una cell factory promettente per l'elevata tolleranza ad elevate concentrazioni di soluti. Occupa una posizione filogenetica peculiare nell'evoluzione degli emiascomiceti perché è il primo a divergere da *S. cerevisiae* prima della WGD.

Scopo di questa tesi è elucidare i meccanismi molecolari causa della sterilità nel lievito ibrido modello *Z. rouxii* ATCC 42981, sviluppando strumenti di biologia sintetica e utilizzando le nuove tecniche di sequenziamento del genoma di terza generazione. Inizialmente, abbiamo studiato il ciclo vitale del ceppo tipo *Z. rouxii* CBS 732^T, focalizzandoci su come lo switching del mating-type agisca da meccanismo plastico che determina instabilità genetica intra individuo e novità fenotipiche in cellule aploidi derivanti da differenti colture stock del ceppo omotallico CBS 732^T. Poi, abbiamo caratterizzato il sistema determinante il sesso nell'allodiploide e sterile ATCC 42981, utilizzandolo come modello per studiare come un assetto chimerico di geni *MAT* possa contribuire all'isolamento riproduttivo negli ibridi. Il sequenziamento del genoma di ATCC 42981, integrando due piattaforme di sequenziamento diverse, Illumina MiSeq e Oxford Nanopore MinION, ci ha aiutato a decifrare le basi genetiche della sterilità negli ibridi e dell'adattamento a stress ambientali nei lieviti non convenzionali. L'assemblaggio iniziale è risultato in 33 scaffolds con una dimensione totale del genoma di 20.9 Mb; 10,254 geni sono stati annotati con Exonerate v2.2.0 per similarità al ceppo aploide più vicino CBS 732^T. Dal confronto con quest'ultimo è emerso che il genoma di ATCC 42981 è 2.14 volte più grande di quello del CBS 732^T e contiene quasi il doppio dei geni. La predizione e annotazione dei geni è stata ulteriormente migliorata combinando metodi *ab initio* e basati sull'evidenza. 10,821 geni codificanti proteine sono stati annotati con Yeast Genome Annotation Pipeline (YGAP) e 10,521 usando la pipeline di Funannotate v1.3.4. Sulla base della sintenia degli scaffolds, dei geni ortologhi e omeomologhi (BUSCO v3: 98.28%), tre aplotipi sono stati identificati, uno di questi quasi identico al CBS 732^T, suggerendo un modello di ibridizzazione ripetuta. Il sequenziamento del genoma ha aiutato anche a risolvere l'organizzazione strutturale incongruente dei loci *MAT* tra due differenti colture stock di ATCC 42981: la nostra sterile e quella giapponese fertile chimata JCM22060. L'assetto dei loci *MAT* è divergnte e, a differenza di JCM22060, ATCC 42981 ne trascive due. Abbiamo dimostrato che la delezione del locus *MAT*α in ATCC 42981 non è sufficiente a ripristinare la sporificazione o il mating, dato che l'incompleto silenziamento di *HML*α maschera la perdita di eterozigosità nel locus *MAT* a seguito della delezione. Nel complesso, i nostri risultati contribuiscono a chiarire i meccanismi responsabili della sterilità negli ibridi e i circuiti che determinano l'identità cellulare nella specie pre-WGD *Z. rouxii*.

Parole chiave: *Zygosaccharomyces rouxii*, mating-type, ibrido, sterilità, MinION, sequenziamento, annotazione, assemblaggio genomi eterozigoti, ortologhi.

THESIS OVERVIEW

This PhD thesis seeks to decipher the molecular and genetic basis of hybrid sterility and dysregulation of cell type identity in an allodiploid non-conventional yeast belonging to the *Zygosaccharomyces* clade, through the combination of synthetic biology, genetic engineering, genome sequencing and comparative genomics. This work is divided into nine chapters, which are briefly introduced here.

Chapter 1 provides an overview of the evolution and diversity of sexual reproduction systems among the hemiascomycetous yeasts. The mating-type genetic organization and the evolution of mating-type switching mechanism in the Saccharomycetaceae family are briefly described, focusing on the pre-WGD species. Finally, the genomic and phenotypic variability within the *Z. rouxii* complex is presented using the allodiploid and sterile ATCC 42981 strain as a promising model to investigate how a hybrid mating-type system affects phenotypic traits, such as meiosis, self- and out-cross fertility, as well as salt stress response.

Chapter 2 introduces the main 2nd and 3rd generation DNA sequencing technologies used in this work and bioinformatics tools exploited for data analysis. In the last decade, the advent of next generation sequencing technologies and, more recently, of the long-read sequencing technologies have marked the start of a new era in biological and genomics research. This is connected with the revitalization of bioinformatics, especially in the areas of data analysis and interpretation.

Chapter 3 provides an overview of the synthetic biology tools and bio-bricks currently available for yeast genetic manipulation, focusing on the challenges in genetic and metabolic engineering of non-conventional yeasts. An outline of the *Z. rouxii* derived mutants currently available is presented.

Chapter 4 describes the *MAT*α to *MAT***a** switching, which led to mixed cell populations in two independent derived cultures of the *Z. rouxii* haploid type-strain CBS 732^T. We have analysed three different aspects of this rearrangement: 1) the 'copy and paste mechanism', which replaced Yα with Y**a** segment using the *HMR***a** locus as donor cassette; 2) the transcriptional regulation of

HO gene which encodes for the endonuclease required for mating-type switching; 3) morphology and self/outcross fertility behaviour of both switched populations. Our results showed that in haploid cells derived from *Z. rouxii* CBS 732^{T} type-strain, mating-type switching is a plastic process that generates a new *MAT***a**2 copy variant and leads to intra-strain genetic and phenotypic variability.

Chapter 5 investigates the molecular basis of sterility in *Zygosaccharomyces* allodiploids, considering ATCC 42981 as a model. Since *MAT* active loci are critical to determine fertility and cell type identity, we characterized its sex determining system addressing important biological questions, such as which are the mechanisms involved in pre-zygotic barriers between species; how hybrid transcriptional complexes co-operate in allodiploids and how the hybrid progenitor of WGD species restores fertility and perform genome doubling. Finally, we provided strong evidences that a chimeric **a**1-α2 heterodimer rewires the diploid-specific regulatory circuit toward the haploid state and significantly drives to hybrid sterility.

Chapter 6 is focused on the effort to expand the biobricks available for the genetic manipulation of Z*. rouxii* by developing two centromeric and two episomal vectors that harbour the antibiotic resistance selection markers *KanMX*^{R} and *NAT*^{R}, respectively. In addition, pGRCRE, another newly constructed plasmid is described, which contains *NAT*^{R} as selectable marker and can be used for *cre*-recombinase-mediated marker recycling in prototrophic deletion mutants generated with *loxP-kanMX-loxP* deletion cassette. ATCC 42981 G418-resistant mutants, generated by replacing the *MAT*α expression locus with the *loxP-kanMX-loxP* cassette, was used to validate pGRCRE as system to rescue the *kanMX-loxP* module in *Z. rouxii* prototrophic strains.

Chapter 7 deals with the announcement of the draft genome sequences of the halotolerant and allodiploid strains *Z. rouxii* ATCC 42981 and *Z. sapae* ABT301^{T}. In particular, it describes the evaluation of different genomic DNA extraction methods to identify the most efficient one to obtain long fragments with 3rd generation sequencing platform MinION (ONT). A hybrid assembly approach that combines Illumina short reads with Nanopore long reads is used to resolve the frequent repetitive and problematic regions of these highly heterozygous strains. Details about the assembly pipeline implemented to assembly these hybrid genomes are also given.

Chapter 8 describes how the newly released draft genome sequence of the allodiploid strain ATCC 42981 serves as a fundamental support to resolve incongruences in mating-type like (*MTL*) loci organization between two different stock cultures of strain ATCC 42981, namely our infertile in-house stock culture and the fertile Japanese sister stock JCM22060. Differently from JCM22060, ATCC 42981 retains two transcriptionally active *MAT* cassettes. We showed that when *MAT*α locus was deleted, the *MAT***a**/- hemizygous mutant did not rescue the ability to either sporulate or mate, as incomplete *HML*α silencing masks the loss of heterozygosity at the *MAT* locus induced by gene deletion.

Chapter 9 deals with the application of two different strategies for ATCC 42981 gene prediction and gene annotation: YGAP exploits conserved syntheny and aa homology across pre- and post-WGD yeast species, while Funannotate combines *ab initio* (*self*- and *trained*-algorithms) and evidence-based approaches. GO functional annotation was achieved by cluster analysis of putative gene products (OrthoVenn). ATCC 42981 genome complements were distinguished in a *Z. rouxii*-like T subgenome and a divergent P subgenome through syntheny mapping with *Z. rouxii* reference type-strain CBS 732^T (best reciprocal blast with SynChro). Homeomologs relationships between T and P subgenome scaffolds were explored using cluster analysis and YGAP pipeline. ATCC 42981 putative proteins were mapped on KEGG biological pathways with a particular focus on genes involved in meiosis and in flavour compound production. Comparative genomics with CBS 732^T showed that most of the genes involved in both pathways were duplicated in accordance with the hybrid allodiploid nature of ATCC 42981. These findings suggest that ATCC 42981 pre zygotic barriers are related to transcriptional and/or post translational incompatibilities between T and P subgenomes, rather than a mere gene loss.

The **Appendix** compiles some studies to which I have contributed. These collaborative projects not only provided me with interesting data to analyse, but also validated some of the findings I present in this work.

Finally, **Supplementary Materials**, organized in different sections corresponding to each chapter, can be found at the end of the Thesis.

CONTENTS

PART I INTRODUCTION

CHAPTER 1: EVOLUTION OF HEMIASCOMYCETOUS YEASTS AND SEXUAL REPRODUCTION SYSTEMS

Cassanelli S., Bizzarri M., and Solieri L. (2016). Recent Advances in Understanding Yeast Genetics of Sex Determination. Fungal Genomics & Biology, 6(1). doi:10.4172/2165-8056.1000e122.

In this chapter, I provide general information on the hemiascomycetous yeasts, in particular, the model organism *S. cerevisiae*. I also summarize what is known about the non-conventional yeast *Z. rouxii*. In addition, I review the literature on budding yeast genetics of sex determination and provide some background on the role of mating-type system and, therefore, the *MAT* loci in determining hybrid sterility.

1.1 BUDDING YEASTS IN THE 'COMPARATIVE GENOMICS ARENA'

The rapidly growing amount of sequenced eukaryotic genomes offers the unique opportunity to explore mechanisms governing genome evolution not only by comparing representatives of distant evolutionary phyla, but also species of the same phylogenetic branch where individual evolutionary events can be more readily identified.

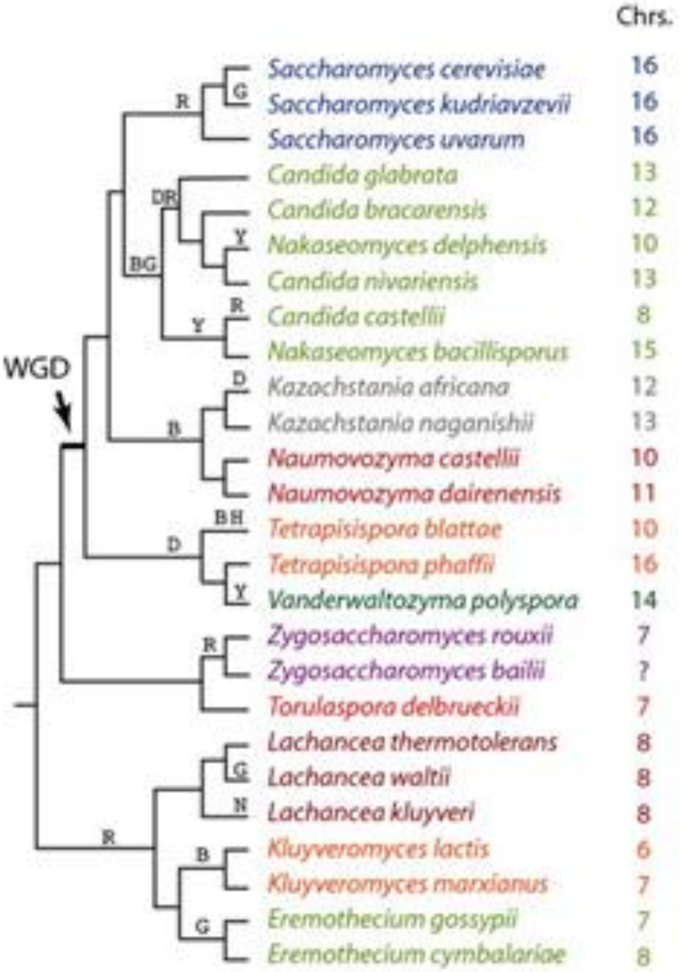
Yeasts of the subphylum Saccharomycotina, the so-called budding yeasts, offer the largest number of sequenced species for a single clade, and represent the cutting edge of evolutionary genomics of eukaryotes (Dujon, 2010 and references herein). Furthermore, they are of great importance to humans not only as pathogens, but also for numerous essential ecosystem services. Some straightforward advantages of these unicellular organisms are that they are easy and fast to cultivate, in large populations and in inexpensive media, with short generation time. Yeasts contain compact and small genomes (9-20 Mb with 4700-6500 genes) with few introns and have a reduced number of transposable elements compared with higher eukaryotes (Liti and Louis, 2005). In addition, the Saccharomycotina yeasts are more genetically divergent from one another than the entire phylum Chordata, which includes vertebrates (Dujon, 2006), but they still retain extensive synteny across these deep phylogenetic distances (Byrne and Wolfe, 2006;

Scannel et al., 2007). All these aspects make budding yeasts highly interesting and accessible models for exploring eukaryotic genome evolution by comparative genomics.

Full genome sequences are available for less than a 10th of the over 1,500 species of budding yeasts described so far (Dujon and Louis, 2017). From presently available genome sequences, the Saccharomycotina sub-phylum splits four major sub-groups into of yeasts: (i) the Saccharomycetaceae, by far the most extensively studied family; (ii) the "CTG clade" a diversified subgroup made of yeast species using an alternative genetic code; (iii) the"methylotroph clade" exemplified by a few recently sequenced yeasts bearing novel signatures; and (iv) several species belonging to distinct and probably distant lineages altogether regarded as "basal" to the Saccharomycotina subphylum but actually very heterogeneous.

The majority of available sequenced genomes are from the family Saccharomycetaceae and fall into one of two clusters (Souciet et al., 2009). The *Candida* cluster consists primarily of *Candida* species but also includes yeasts such as *Debaryomyces hansenii*, while the second cluster is comprised primarily of species from the *Saccharomyces* and *Kluyveromyces* genera. The major division within the Saccharomycetaceae family is between those yeasts whose common ancestor underwent a whole-genome duplication (WGD) and those that diverged prior to this event. We term these post-WGD and pre-WGD yeasts respectively. The WGD event (Wolfe and Shield, 1997) took place approximately 100 million years ago, separating *Saccharomyces* lineage and other post-WGD genera such as *Kazachstania*, *Naumovozyma*, *Nakaseomyces*, *Tetrapisispora* and *Vanderwaltozyma*, from pre-WGD lineages (Dujon et al., 2010) (**Fig. 1.1**). The ancestral organism that underwent WGD contained about 5000 genes. The WGD increased this number transiently to about 10 000 genes, but most of the extra copies of genes were not retained and instead they became 'lost'-that is, one of the two genes in each pair became deleted, usually without any other rearrangements in the local area of chromosome (Scannell et al., 2007). Post-WGD species now typically contain about 5500 genes, which includes 500 pairs of genes (ohnologs) that were formed by the WGD; the other 4500 loci were not retained in duplicate and became single-copy again.

Figure 1.1. Phylogenetic relationships of main species in the subphylum Saccharomycotina (adapted from Peter and Schacherer, 2015). Numbers of chromosomes are indicated on the right side. The topology of the cladogram is from Kurtzman 2003; Hedtke et al., 2006; Gordon et al., 2011; Gabaldon et al., 2013). Different colors indicate different genera. Letters above branches indicate inferred points of loss of the DAL (D), GAL (G), BNA (B), HIS (H), RNAi (R; loss of all Argonaute genes or all Dicer genes), NHEJ (N) and dynein (Y) pathways. 'Chrs', number of chromosomes. Different colors indicate different genera. Note that *Tetrapisispora* does not appear to be monophyletic (Gordon et al., 2011).


Chrs.
WGD
R
G
Saccharomyces cerevisiae 16
Saccharomyces kudriavzevii 16
Saccharomyces uvarum 16
DR
BG
Candida glabrata 13
Candida bracarensis 12
Y
Nakaseomyces delphensis 10
Candida nivariensis 13
Y
R
Candida castellii 8
Nakaseomyces bacillisporus 15
D
B
Kazachstania africana 12
Kazachstania naganishii 13
Naumovozyma castellii 10
Naumovozyma dairenensis 11
BH
D
Tetrapisispora blattae 10
Tetrapisispora phaffii 16
Y
Vanderwaltozyma polyspora 14
R
Zygosaccharomyces rouxii 7
Zygosaccharomyces bailii ?
Torulaspora delbrueckii 7
Lachancea thermotolerans 8
G
Lachancea waltii 8
N
Lachancea kluyveri 8
R
B
Kluyveromyces lactis 6
Kluyveromyces marxianus 7
G
Eremothecium gossypii 7
Eremothecium cymbalariae 8

1.1.1 Mechanisms of yeast genome evolution and role of interspecies hybridization

Conventionally, point mutations and chromosomal rearrangements have been described as main forces contributing to the evolution of gene repertoire in budding yeasts. Point mutations results in gene inactivation by pseudogenization of coding sequences, as well as in gene gain by *de novo* gene creation from non-coding sequences. Chromosomal rearrangements fall into two categories: balanced rearrangements, such as chromosome fusion/fission, inversion, and translocations, modify gene order and orientation and can potentially create novel gene combinations or induce gene disruptions at the breakpoints; unbalanced chromosomal rearrangements, including deletions and duplications of chromosomal regions, lead to reduction and expansion of the gene repertoire, respectively.

Comparative genomics demonstrated that polyploidization, a state resulting from doubling of a genome within a species (autopolyploidy) or the merging between different species (allopolyploidy), also contributes to yeast genome evolution and ultimately drives phenotypic adaptation and speciation. Recently, phylogenomics analysis has shown that the WGD event occurred in yeast ancestor more than 100 million years ago was an allopolyploidization or interspecies hybridization between two divergent parental lineages (Marcet-Houben and Gabaldón, 2015). One of these parental lineages was most closely related to a clade containing *Zygosaccharomyces* and *Torulaspora* (ZT), whereas the other was closer to a clade containing *Kluyveromyces*, *Lachancea*, and *Eremothecium* (KLE). The ZT and KLE clades are the two major groups of non-WGD species in family Saccharomycetaceae (Marcet-Houben and Gabaldón, 2015).

Interspecific hybridization naturally occurs when two different but closely related sympatric species mate producing viable hybrids. Natural hybrids have been frequently found in food environments (Masneuf et al., 1998). The most commonly used lager brewer's yeast, *Saccharomyces carlsbergensis*, is a triploid with a diploid *Saccharomyces eubayanus* and haploid *Saccharomyces cerevisiae* genome content (Walther et al., 2014). Natural hybrids between *Saccharomyces sensu stricto* species have been also described from wine and cider fermentation at low temperature (Morales and Dujon, 2012 and references herein). Other lineages of budding yeasts contain natural hybrids other than *Saccharomyces sensu stricto*. The osmotolerant yeast *Millerozyma* (*Pichia*) *sorbitophila* has been recently proven to be an allodiploid between two highly divergent species of *Millerozyma* (Kurtzman and Suzuki, 2010).

When an organism doubles its genome, it establishes a reproductive barrier between itself and its ancestor, providing raw material for the divergence of gene functions between homologs (pairs of paralogous genes produced by the WGD) (Wolfe et al., 2015). Biological responses to this doubling are globally defined as genome and transcriptome shocks. Genome shock describes genome changes that occur in response to polyploidization, such as gene loss, chromosome *mis*-pairing, and rearrangements between sub-genomes. Chromosome rearrangements and gene losses progressively erode the genetic pools of initial parental subgenomes, leading to the unbalanced distribution of gene sets into the offspring. This phenomenon is also known as loss of heterozygosity (LOH) (Dujon and Louis, 2017). Transcriptome shock includes rearrangements of gene expression networks following the mixing of two dissimilar genomes, each with their own set of transcription factors, *cis*-regulatory elements, and their own chromatin profiles.

Stable hybrid lines resulting from genome and transcriptome shocks often display some benefits, such as heterosis, increased robustness, improved by-product production, and enhanced stress resistance, compared with the parental species. Thus, interspecific hybridization via spore-to spore mating or rare mating has been used in laboratory to construct artificial strains with novel phenotypes and adaptive properties (Giudici et al., 2005; Steensels et al., 2014). On the other hand, genome and transcriptome stabilization after the merger of two parental subgenomes may cause negative epistasis, leading to hybrid depression, sterility (inability to produce viable spores) or infertility (unability to sporulate). Chromosomal rearrangements or allele-allele incompatibilities can produce these epistatic interactions and contribute to hybrid sterility/infertility. Overall, these considerations open the question about how yeasts develop several different and versatile life cycles, diploid/haploid state sensors, cell identity and, ultimately, the choice among distinct fates, such as mitotic propagation, cell-to-cell mating and meiosis.

1.2 SEX CHROMOSOMES EVOLUTION

Reproduction by means of sex is a common and topic theme in the transmission of genetic information between generations in multicellular and unicellular eukaryotes. The gametes can be either similar (isogamy) or dissimilar (anisogamy) in size and form, and it is the last one that can give rise to the existence of males and females as separate sexes. Sex can be determined by the environment or be genetically controlled by one or more loci located either on an autosome or on sex chromosomes. The latter represent a fascinating example of evolutionary convergence since they have evolved independently several times in different lineages, but in spite of this, they show many common features. While in plants and animals, sexual identity is governed by entire sex chromosomes, in fungi it is determined by single positions on chromosome called mating type (*MAT*) loci.

Sexual reproduction between two different genomes gives many advantages; the most important one is to create new and favorable adaptive allelic combinations. Life cycle and the possibility to undergo meiosis and mating are two main driving forces that affect genetic variation within a yeast population, such as the heterozygosity level. For unicellular yeasts, breeding systems represent an important means, enhancing the chances to survive to drastic environmental changes. For instance, when nutrient sources are depleted, diploid cells are able to generate spores that can survive long starvation periods or other non-favorable conditions, while haploid cells must find a partner with opposite mating-type to produce spores resistant to hostile environments. This suggests that sex is an important means, which increases the rate of adaptation to harsh environments in yeasts (Goddard et al., 2005). Furthermore, recombination is a fundamental mechanism that prevents the accumulation of deleterious mutations in sexual population. Fitness difference occurs between the progeny and depends upon the mode of reproduction and mating system. Yeasts show great variation regarding their preference for a diploid or haploid life style, leading to consider the possible advantages and disadvantages conferred by either life style. Diploid life style is favored, for example, for the protective effect of diploidy on the deleterious consequences of haploid mutations or the ability to preserve allelic variations (Knop, 2006). On the other hand, for haploids there might be quicker and more efficient systems to maintain beneficial alleles or to lose disadvantageous ones in populations.

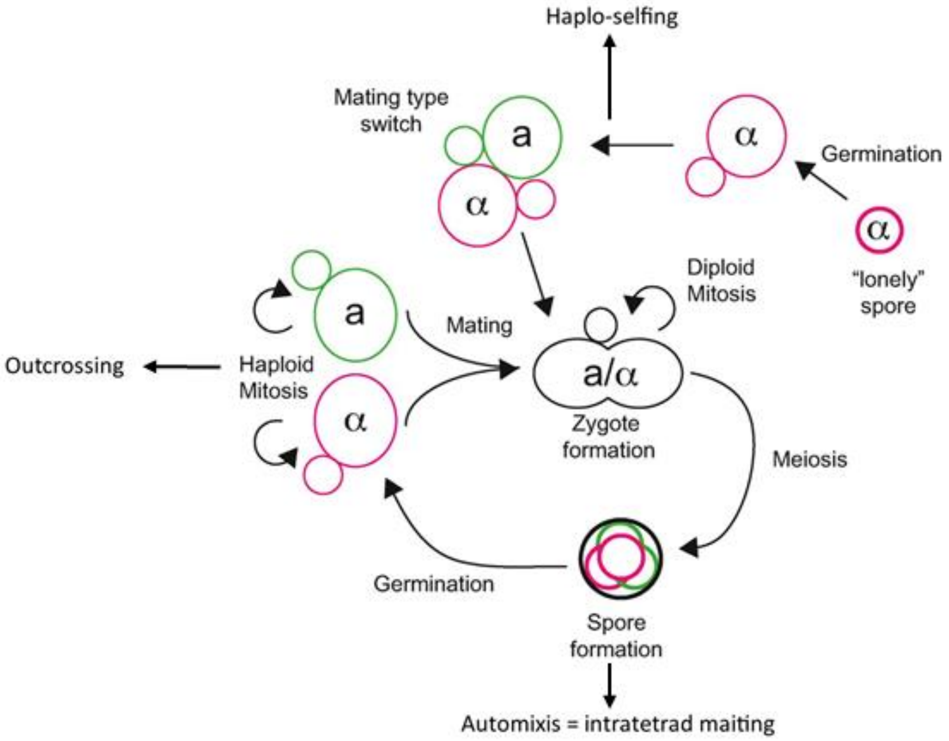
1.2.1 Yeast life cycle and the logic of the cell type-specification circuit

S. cerevisiae is a budding yeast, *i.e*. it reproduces by forming a smaller cell with new cell components, rather than enlarging and then dividing (fission yeasts) and undergoes a regular mitotic cycle (G1, S, G2 and M phases). Yeasts reproduce both in haploid and diploid forms, which allow for both sexual crossing and clonal division (budding). Sexual reproduction in yeast is initiated by the recognition of a mating partner and cell fusion, followed by nuclear fusion (karyogamy) to form diploid cells that clonally reproduce or, alternatively, undergo meiosis and produce haploid progeny. Typically, mating in yeasts starts by the fusion of haploid cells that are similar in size and shape (isogamy), which produces zygotes (Knop, 2006). Fusion involving diploid cells is also possible, resulting in polyploids (Albertin et al., 2009). Virtually all yeast species in the hemiascomycete lineage exist in three cell types: haploids of two isogamous mating-types, **a** and α, and the product of their mating, the meiosis competent **a**/α diploid (Herskowitz 1988; Madhani, 2000). The two types of haploid are often called mating-types because they describe mating behaviour: mating occurs only between **a** cells and α cells, because these cells express **a**-specific and α-specific genes respectively. These sets of genes include cell-type specific mating pheromones (**a**-factor and α-factor respectively), transporters for the export of the relevant mating-factors and receptors for mating factors of the opposite mating type (Johnson, 1995). Thus, **a** cells secrete a-factor which is detected by an **a**-factor receptor expressed on the surface of α cells and *vice versa*. Once mating factors are detected, the cell cycle is arrested, shmoos (mating projections) are produced by the mating cells and cytogamy is initiated. The later stages of this process are shared between the two haploid cell types and are regulated by a set of haploid-specific genes that are expressed in both **a** and α cells but repressed in diploids.

The haploid spores are produced when a *MAT***a**/*MAT*α diploid stops the cell cycle in G1 phase and undergoes meiosis if nutrients deficiency occurs or under particular environmental conditions. The resulting haploid spores are contained in the same envelope called ascus and are characterized by a mating-type, either **a** or α (Knopp, 2006).

In the model yeast *S. cerevisiae* sexual reproduction can involve a single partner (homothallism) or two compatible partners (heterothallism) via three events: i) mating of unrelated haploids derived from diploid unrelated cells (amphimixis or outcrossing); ii) mating between spores from the same tetrad (automixis); iii) and mother daughter mating upon mating-type switching (haplo-selfing) (Billiard et al., 2012; Hanson and Wolfe, 2017) (**Fig. 1.2**). Mating-type switching is the process by which a haploid **a** cell can become a haploid α cell, by changing its genotype at the *MAT* locus from *MAT***a** to *MAT*α, or *vice versa*. In *S. cerevisiae* the life cycle alternates between haploid and diploid states in order to prevent aneuploidy. In heterothallic haploid yeasts, outcrossing represents the only possible strategy to establish a provisional diploid state because syngamy can only occur between haploid cells carrying different alleles at the mating-type locus. These yeasts have a haploid life style where mating occurs before spore formation (Knop, 2006).

Figure 1.2. Schematic life cycle of *S. cerevisiae*. (Adapted from Hanson and Wolfe, 2017).

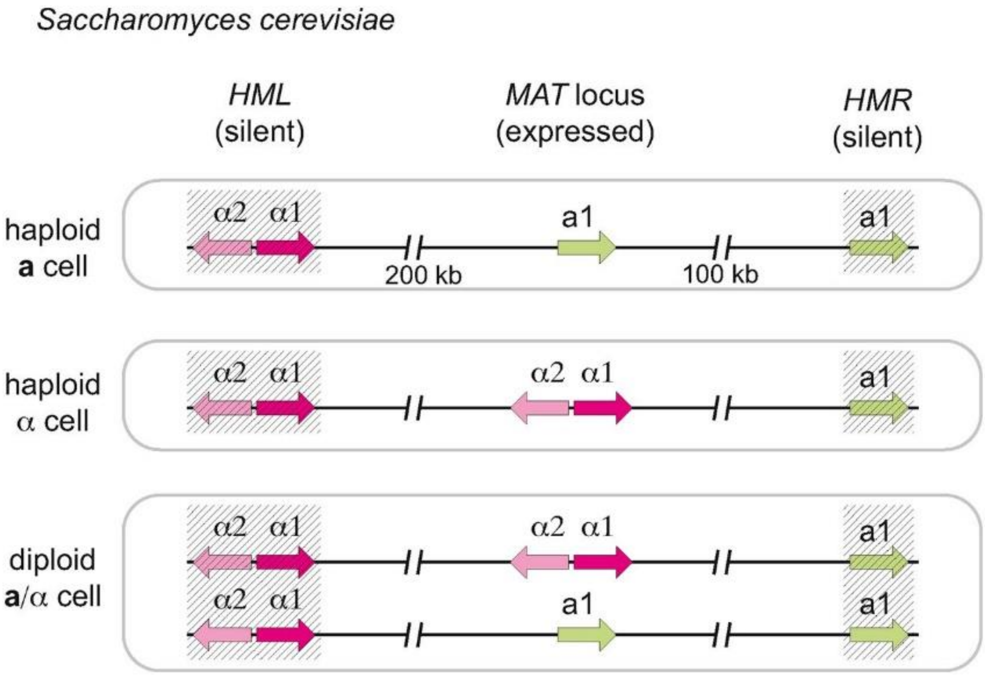


Heterothallic diploid yeasts possess automixis and amphimixis as alternatives to sexual reproduction. In these yeasts, the spore formation can occur with/without mating and their lifestyle is referred to as haploid-diploid inbreeding style. Haplo-selfing is a syngamy event, which can occur between genetically identical haploid cells (clones), and it is a reproductive mode possible only for homothallic yeasts. Homothallic yeasts possess three alternatives to sexual reproduction, such as automixis, anphimixis and haplo-selfing. Therefore homothallic yeasts lack of discrimination at syngamy, *i.e.* each haploid is compatible with all other haploids in the population, whereas heterothallic yeasts may restrict the number of potentially compatible partners for any gamete. There are several mechanisms that may confer homothallism in fungi: most often, each haploid carries two active mating-type alleles (Coppin et al., 1997), whereas in other species syngamy can simply occur between haploid cells carrying and expressing the same single mating-type allele (Alby et al., 2009; Fraser et al., 2005; Lin et al., 2005; Metzenberg and Glass, 1990). Only unicellular fungi perform haplo-selfing by switching the mating-type haploid individuals. Switching has arisen in two yeast lineages represented by *S. cerevisiae* (sub-phylum Saccharomycotina) and *Schizosaccharomyces* (sub-phylum *Taphinomycotina*) (Haber, 2012 and references herein). *S. cerevisiae* ability to "switch" the mating-type allele is related to the presence of the endonuclease HO (HOmothallism), which catalyzes the first step of *MAT* conversion process (Haber, 1998). Consequently, haplo-selfing results from syngamy between cells genetically identical, except for the mating-type locus.

1.2.2 The mating-type loci in *Saccharomyces cerevisiae* control cell type fate

In *S. cerevisiae*, the two idiomorphs *MAT*α and *MAT***a** harboured by the *MAT* locus are located in the middle of the right arm of the chromosome III, nearly 100 Kb from both the centromere and the telomere. The *MAT* locus is divided into five regions (W, X, Y, Z1, and Z2). Mating-type-specific Y**a** (for *MAT***a**) and Yα (for *MAT*α) sequences are about 650 and 750 bp, respectively, differentiate the two alleles by providing many promoters and Open Reading Frames (ORFs) for proteins that regulate many aspects of the cell sexual behaviour. The Y**a** and Yα sequences are surrounded by 700 bp W and X regions, a 230 bp Z1 region and a 90 bp Z2 region. Furthermore, *S. cerevisiae* carries two additional copies of mating-type genes, *HML*α (Hidden *MAT* Left) and *HMR***a** (Hidden *MAT* Right), at distant locations on the same chromosome as the *MAT* locus (**Fig. 1.3**).

Figure 1.3. Gene organization at the *MAT*, *HML*, and *HMR* loci on *S. cerevisiae* chromosome III. (Hanson and Wolfe, 2017). Shading indicates genes whose transcription is repressed.

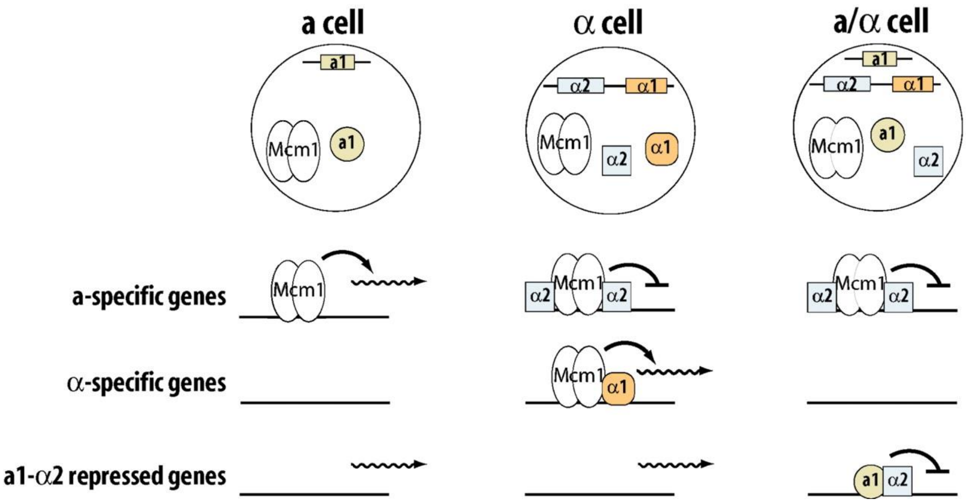


These genes are, however, not expressed, as both *HML* and *HMR* are surrounded by silencer sequences (designated E and I) that recruit the specialized histone deacetylase, Sir2, and other silencing factors (Sir1, Sir3 and Sir4) to create a heterochromatic domain that is not transcribed. *HML*α and *HMR***a** serve as donors during the recombination process called mating-type switching. In *S. cerevisiae* haploid *α* cells *MAT*α genes encode two proteins, α1 and α2. The *MAT*α1 gene codes for the HMG-domain transcription activator α1 (previously referred to as a "α-domain" protein but now recognized as a divergent HMG domain; Martin et al., 2010). This activator interacts with the constitutively expressed protein Mcm1 in order to recruit the transcription factor Ste12 necessary for the activation of α-specific genes (αsgs), such as those encoding α factor mating pheromone and the receptor Ste3 for detecting **a**-factor. *MAT*α2, instead, encodes the helix-turn-helix homeodomain protein α2 that acts with Mcm1 to form a repressor, which inhibits **a**-specific genes (**a**sgs), comprising those encoding the **a**-factor mating pheromone and the receptor Ste2 for α-factor. Chromatin Immuno-Precipitation (CHIP) and microarray-based transcriptional profiling have identified six genes to be targets of this repressor, all of them being associated with the **a**-phenotype.

The **a**-like mating behaviour, on the contrary, appears in cells where either *MAT*α1/*MAT*α2 promoter or the entire *MAT*α locus is deleted. Differently from a non-mating *MAT*α/*MAT***a** cell, in fact, a *MAT***a**/*MAT*Δ cell shows an α-mating behaviour thanks to the unrepressed **a**sgs. This is explained by the fact that, instead of requiring an **a**-specific activator, **a**sgs (*MAT***a**1 in particular) are activated by Mcm1 and Ste12, which are constitutively expressed in all cell types, and need α2/Mcm1 dimer to be repressed. Therefore, in *S. cerevisiae*, the **a**-cell type is the default type, and yeast cells lacking a *MAT*α locus will mate with haploid α cells.

The third cell type, the diploid **a**/α cell, does not mate, because **a**sgs , αsgs and haploid specific genes (hsgs) are turned off. In *S. cerevisiae* each of the three cell types is specified by a unique combination of transcriptional regulators. This transcriptional circuit has served as an important model for understanding basic features of the combinatorial control of transcription and the specification of cell type. **a**/α diploid cells have *MAT*α1 and *MAT*α2 genes at the *MAT* locus on one chromosome, and *MAT***a**1 on the other, which results in formation of the **a**1-α2 heterodimer of the two homeodomain proteins. This heterodimer, which has been showed to be active on 19 total genes, represses the transcription of hsgs and suppresses αsgs through repression of *MAT*α1. Transcription of **a**sgs in diploids is repressed by α2-Mcm1 as in haploid α cells (**Fig. 1.4**).

Figure 1.4. Cell-type regulation in *S. cerevisiae* (Galgoczy et al., 2004). The basic regulatory scheme summarizes the target genes directly regulated by the *S. cerevisiae* mating-type locus.



The **a**sgs and αsgs regulated by the *MAT* locus primarily include genes for pheromones, their receptors, and signalling proteins required for recognition of cells with the opposite mating-type (Johnson 1995; Galgoczy et al., 2004). By using genome-wide chromatin immunoprecipitation, transcriptional profiling, and phylogenetic comparisons, Galgoczy and colleagues investigated the complete cell-type-specification circuit for *S. cerevisiae*. The main genes belonging to this circuit and differentially expressed in the three cell-types are listed in **Table 1.1**.

Table 1.1. Summary of the circuit controlled by the *MAT* locus: the complete set of genes regulated by the mating-type transcriptional regulators.

Gene category	Gene	Function	Biological process
a-Specific Genes (α2-Mcm1 repressed genes)	*STE2*	α-factor receptor	mating
	STE6	**a**-factor transporter	mating
	*MFA1**	Structural genes for **a**-factor precursor	mating
	*MFA2**		
	BAR1	α factor protease	mating
	AGA2	Adhesion subunit of **a**-agglutinin of **a**-cells	mating
α-Specific Genes (α1-Mcm1 repressed genes)	*STE3*	**a** factor receptor	mating
	MFα1	α-factor mating pherormone	mating

	MFα2		
	SAG1	**a**-cell agglutinin	mating
Subset of Haploid-Specific Genes (**a**1-α2 repressed)	*HO*	Mating cassette recombination endonuclease	mating-type switching
	MATα1	Mating-type transcriptional activator	mating-type regulation
	MATα2		
	STE12	Transcriptional factor generally activated by pherormones	mating
	*GPA1***	Pherormone signalling G-protein α	mating
	*STE4***	Pherormone G-protein β	mating
	*STE18***	Pherormone G-protein γ	mating
	*FAR1***	Pherormone-induced cell cycle arrest	mating
	MFα1	α-factor mating pherormone	mating
	CCW12	Cell wall mannoprotein involved in agglutination	mating
	*RME1***	Transcriptional repressor of meiosis	meiosis
	HOG1	Osmotic response MAPK	Osmotic sensing
	*FUS3***	Pherormone signalling MAPK	mating
	FUS1	Membrane protein required for cell fusion	mating
	*AXL1***	Pherormone-induced morphogenesis and fusion	mating
	SST2	GTPase-activating protein for Gpa1	mating
	STE5	Pheromone signaling MAPK scaffold	mating
	AMN1	Mitotic exit regulation; haploid-specific clumping	?
	TEC1	Transcription factor targeting filamentation genes and Ty1 expression	chronological cell aging
	*NEJ1***	Non-homologous end-joining DNA repair	DNA repair
	*RDH*54	Recombinational repair	DNA repair

*Both genes must be inactivated to cause a mating defect; **haploid-specific genes repressed by **a**1-α2 heterodimer.

In 2004, Galgoczy and collegues identified six **a**-specific genes and five α-specific genes that are reported in **Table 1**. Each of these genes is a direct target of the mating-type-encoded regulators α2 and α1, respectively, and are tightly shut off in the inappropriate cell types. Moreover, they demonstrated that 19 genes directly regulated by **a**1-α2 heterodimer are characterized by a range of repression values, with some genes being merely turned down in the **a**/α cell type.

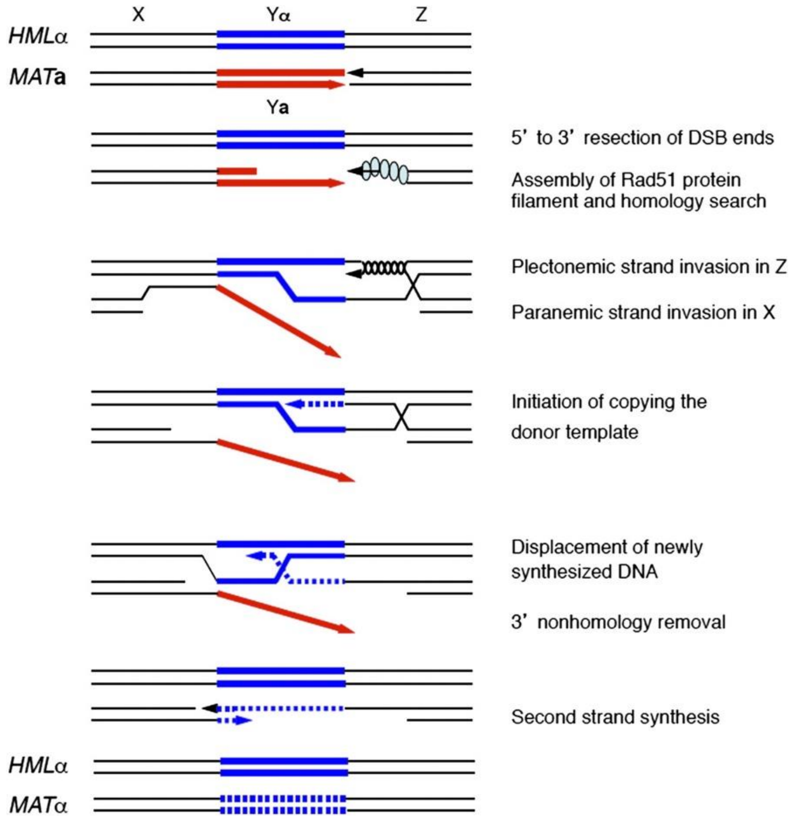
As expected, *MAT* locus also controls the possibility for the cell to enter meiosis. Under the appropriate nutritional conditions, a diploid cell ceases cell division (vegetative growth) and enters the meiotic cell cycle to produce the four haploid meiotic products. Entry of the diploid into meiosis requires that the **a**1-α2 repressor inactivates the expression of *RME1* gene (Repressor of Meiosis), allowing the cell to undergo meiotic division and spores production. *RME1* is a zinc-finger transcription factor involved in inhibiting meiosis by repressing the expression of *IME1* (Initiator of Meiosis). Furthermore, it has a positive effect on pseudohyphae formation, invasive growth and adhesivity, by regulating the transcription of *FLO11*, a cell-wall-expressed glycoprotein important for cell adhesion (Van Dyk et al., 2003; Magwene et al., 2011). Other important genes controlled by mating-type are those involved in non-homologous end joining (NHEJ). In eukaryotic cells double strand breaks (DSBs) can be repaired either by Homologous Recombination (HR, whether provided with another copy of the gene) or by NHEJ. Although both molecular processes are efficient in haploid cells, HR, being involved largely (90% of times) in cycle arrest in G1 makes NHEJ to become predominant. More in particular, these cells are unable to undergo homologous recombination because kinase Cdk1 is inactive. However, in diploid **a**/α cells *NHEJ1* is turned off by repressor α1-**a**2, which acts upon *NHEJ1* gene in addition to the partial repression of *LIF1* gene, another *NHEJ* component. It has been hypothesized that avoiding error-prone NHEJ is functional to the meiotic process, where up to 100 double strand breaks can occur, leading to a high risk of mutation.

1.2.3 Ho-catalysed mating-type switching

As previously described above, *S. cerevisiae* has two additional copies of mating-type genes, *HML*α and *HMR***a**, at distant locations on the same chromosome. Their main role is to serve as donors during the mating-type switching, a process that allows a *MAT***a** cell to switch to *MAT*α or *vice versa* (secondary homothallism), leading to the mother-daughter mating (haplo-selfing).
Mating-type switching in *S. cerevisiae* is often called gene conversion, but it is more accurately described as a synthesis-dependent strand annealing (SDSA) process because of the non-homology of the Y regions between the outgoing and incoming alleles (Ira et al., 2006). During mating-type switching, the gene content at the *MAT* locus of a haploid cell is unidirectionally replaced by copying a reserve version of the *MAT* genes of the opposite allele at the *HMR*/*HML* loci (Haber, 2012; Lee and Haber, 2015) (**Fig. 1.5**). DNA repair at *MAT* locus is guided by the Z and X regions that are almost identical between *MAT, HML*, and *HMR*. In *S. cerevisiae* the Z and X regions occur in three copies in parallel orientation: the Z region contains the 3’ end of the *MAT*α1 gene, the X region contains the 3’ end of *MAT*α2 gene, and the 5’ end of the neighbouring

chromosomal gene *BUD5*. In *S. cerevisiae* gene conversion begins only when the HO endonuclease, encoded by the *HO* gene on chromosome IV, cleaves the Y-Z junction in the *MAT* locus; the Y region is degraded and the Z and X sequences direct the switching process (that takes about 1 hour) in which both strands of DNA at *MAT*-Y are newly synthesized using *HML* or *HMR* silent cassettes as a template for repair. *HO* gene is expressed only in cells that have budded once, which means that only mother cells are able to switch mating-type. An *HO*-induced DSB is resected by 5′ to 3′ exonucleases or helicase endonucleases to produce a 3′-ended ssDNA tail, on which assembles a Rad51 protein filament, which is required along with other members of the Rad52 epistasis group for recombinagenic repair of damage-induced DSBs in budding yeast. In the *MAT*-Z region, strand invasion can form an interwound (plectonemic) joint molecule that can assemble DNA replication factors to copy the Yα sequences. Unlike normal replication, the newly copied strand is thought to dissociate from the template and, when sufficiently extended, anneal with the second end, still blocked from forming a plectonemic structure by the long non-homologous single-stranded Y**a** sequences.

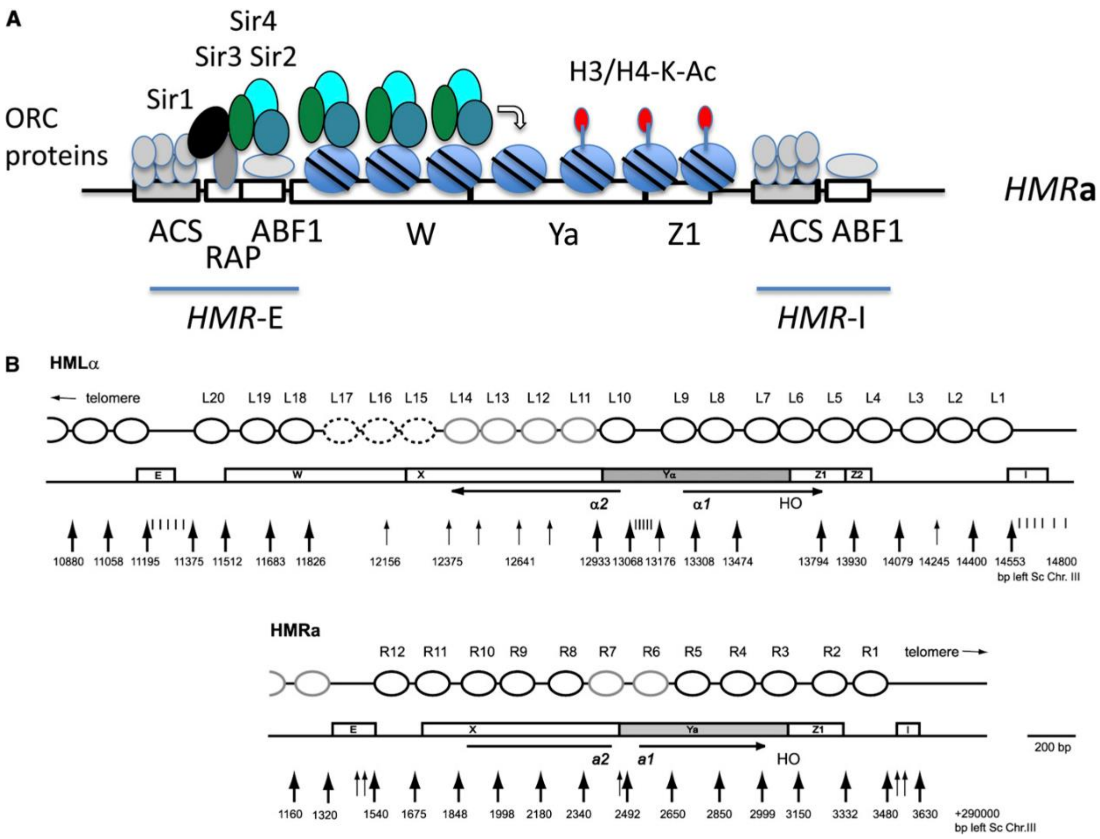
Figure 1.5. Mechanism of *MAT* switching. Key steps in the switching of *MAT***a** to *MAT*α by a synthesis-dependent strand-annealing (SDSA) mechanism are outlined (reviewed by Pâques and Haber, 1999).


X
Yα
Z
HMLα
MATa
Ya
5' to 3' resection of DSB ends
Assembly of Rad51 protein filament and homology search
Plectonemic strand invasion in Z
Paranemic strand invasion in X
Initiation of copying the donor template
Displacement of newly synthesized DNA
3' nonhomology removal
Second strand synthesis
HMLα
MATα

These sequences are clipped off once strand annealing occurs, by the Rad1-Rad10 flap endonuclease, so that the new 3′ end can be used to primer extend and copy the second strand of the Yα sequences. Consequently, all newly synthesized DNA is found at the *MAT* locus, while the donor is unaltered. A small fraction of DSB repair events apparently proceed by a different repair mechanism involving the formation of a double Holliday junction (see Pâques and Haber, 1999 for details). Switching is a slow process, taking ∼70 min, and is >1000 times more error-prone than normal DNA replication (Hicks et al., 2010, 2011). Even though the newly synthesized *MAT* DNA will generally be replaced the next time the cell switches mating-type, the high error rate nevertheless imposes an evolutionary cost because sometimes the errors will render the *MAT* genes non-functional. Fine-tuned regulatory networks and a Sir complex-dependent epigenetic control assure that Ho cuts only its recognition site, placed at the junction between the Y and Z segments of the *MAT* locus. Based on this mechanism, the mating-type switching requires a total of 6 steps (**Fig. 1.6**): (1) *HML* and *HMR* are "silent cassettes" that store α-specific and **a**-specific sequence information, respectively, but are transcriptionally inactive due to chromatin modification; (2) an endonuclease (*HO*) that creates a double-strand break (DSB) at the *MAT* locus, which then is repaired using *HML*α or *HMR***a** as a donor; (3) a mechanism (Sir1 and Sir2, 3, 4 proteins) for repressing transcription and *HO* cleavage at the silent loci; (4) two triplicated sequences (the Z and X regions) that guide repair of the DSB; (5) a donor-bias mechanism (the recombination enhancer, RE) to ensure that switching happens in the correct direction, (6) and a cell lineage-tracking mechanism (Ash1 mRNA localization) to ensure that switching occurs only in particular cells, such as the mother but not the daughter cell. The *MAT* heterozygosity is fundamental in the switching of the mating-type genes. In fact, haploid **a**- and α-cells express endonuclease gene *HO*, while diploid heterozygous cells do not. Those strains, which are able to undergo this kind of process are known as homothallic, while the opposite, being heterothallic strains, lack this possibility. Within a single cell division *S. cerevisiae* has the capacity to shift rapidly from expressing α-specific genes to expressing the **a**-specific program; this has been proved as freshly switched *MAT***a** cells respond to α-factor, with arrest in G1 of cell cycle. This means that these cells already express **a**-specific genetic program, with Ste3 pheromone receptor (for **a**-factor) expression turned off and that they do not produce α-factor anymore. It also means that **a**-factor and Ste2 transmembrane receptor (for α-factor) expression has to be already turned on. The process of switching involves a programmed and site-specific DSB at the *MAT* locus by *HO* endonuclease and the replacement of either Y**a** or Yα with the sequence contained in two silenced accessory loci (*i.e*. *HML* and *HMR*), by means of homologous recombination. Both *HMR* and *HML* feature cis-acting silencer sequences known as *HML-E*, *HML-I*, *HMR-E* and *HMR-I*, which interact with trans-acting factors (Sir proteins, histones, Rap1 protein, and various chromatin modifiers). Short regions (3 Kb) of heterochromatin are formed by the interaction of these many factors, and the two above said loci end up in a highly ordered nucleosomic structure, which is thus also inaccessible to nucleases as *HO* (**Fig. 1.6**).

The process is directional, in which the sequences at *MAT* are replaced by copying new sequences from either *HMLα* or *HMR***a**, while the two donor loci remain unchanged by the transaction. There are tightly mechanisms that ensure that the *HO* gene is expressed only in haploid mother cells and only at the G1 stage of the cell cycle (Nashmith, 1987). *MAT* switching represents one of the most fascinating process in eukaryotic cell biology and provides a powerful model to study the determination of cell lineage. Only half of the cells in a colony are able to switch mating-type in any one cell division (**Fig. 1.2**). A germinating haploid spore grows, produces a bud, and divides without changing its mating-type. Then, in the next cell division cycle, the older mother cell and its next (second) daughter change mating-type while the first daughter buds and divides without any change (Haber and George, 1979). In addition, a cell lineage pattern ensures that only half of the cells in a population switch at any one time, to guarantee that there will be cells of both mating-types in close proximity.

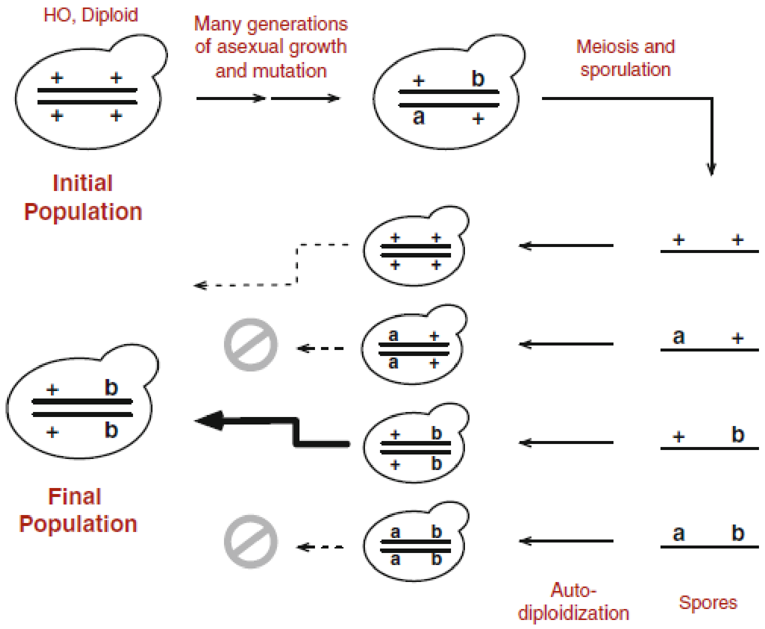
Figure 1.6 Silencing of *HMR* and *HML* cassettes (adapted from Haber and Wolfe, 2005). (**A**) Establishment of silencing at *HMR*-E. The process of silencing is illustrated. Proteins bound to the three elements of the *HMR*-E silencer recruit Sir1 that in turn recruits the Sir2-Sir3-Sir4 complex. The NAD+-dependent HDAC Sir2 deacetylates lysines on histones on the N-terminal tails of H3 and H4, which allows the Sir3-Sir4 to bind and stabilize the position of the nucleosome. Sir2 can then deactylate the next nucleosome and silencing spreads further. Here the spread of silencing is shown progressing in one direction and from one of the two silencing elements. Actually, silencing spreads from both *HMR*-E and *HMR*-I and in a limited fashion to the flanking regions. (**B**) Highly positioned nucleosomes in *HML* and *HMR* are represented, as determined by Weiss and Simpson 1998, and Ravindra et al., 1999.



1.2.4 From Mortimer's "genome renewal" to the "lonely spore" scenario

Mating-type switching is a highly regulated and complex process, so it must confer a benefit to yeast or it would not have been maintained by natural selection. The "Genome Renewal" hypothesis was firstly postulated by Robert Mortimer (Mortimer, 1994) and later revisited by Magwene (2014) explains patterns of genetic variation observed in the wild population of *S. cerevisiae* and the evolutionary role of mating-type switching (**Fig. 1.7**). Mortimer and colleagues found out that most yeast strains isolated from vineyards, despite being homothallic and thus being able to undergo haplo-selfing or autodiplodization, featured one or more heterozygous loci. *S. cerevisiae* would propagate vegetatively as diploid, accumulating mutations and thus raising its level of heteroygosity over generations (Mortimer et al., 1994; Mortimer, 2000). More recently, genome sequencing of many yeast strains confirmed Mortimer's results, showing high degrees of heterozygosity in industrial yeasts.

Figure 1.7. A schematic illustration of Mortimer's Genome Renewal Hypothesis (adapted from Magwene, 2014).



The vegetative asexual proliferation is a much more prevalent way of reproduction, with on average only one meiotic cycle for every 1000 mitotic divisions (Ruderfer et al., 2006; Tsai et al., 2008; Zörgö et al., 2012). During these asexual reproductive cycles, spontaneous mutations, such as point mutations, InDels, transposon insertions, and recombination events, can occur.

Rare sexual cycles involving meiosis followed by mating-type switching and haplo-selfing would have facilitated the loss of deleterious alleles and fixed beneficial ones in a homozygous diploid, thus leading to "Genome Renewal". Magwene argued that high levels of heterozygosity in *S. cerevisiae*, coupled with selfing during rare sexual cycles, can facilitate rapid adaptation to novel environments (Magwene, 2014). Indeed, for highly heterozygous homothallic strains, the adaptive evolutionary landscape has a high degree of "accessibility" because large regions of genotypic and phenotypic space can be sampled and tested in offspring by local selection, getting the most favourable allele combinations to be fixed. This process occurs even more rapidly when a population is founded by clonal reproduction of a single or a few related individuals. Another benefit gained by switching is under the name of "lonely spore" scenario (Herskowitz, 1988; Gordon et al., 2011). Gordon and colleagues proposed that the goal of switching is to maximize the ability of a young haploid colony to make new spore if nutrient level falls. Switching mechanism enables isolated spores to germinate in very poor environments, replicate for a few cell cycles as permitted by the environment, and then to resporulate. In contrast, species that cannot switch would need to be more cautious about germinating. Over time, species that can switch are predicted to have a growth advantage over species that cannot switch, because they can risk germinating in poorer environments and so germinate earlier in improved environments. Yeasts are dispersed to new habitats when they are eaten and excreted by insects (Reuter et al., 2007). Although ascospores (sets of four haploid spores formed by meiosis of a diploid) are structures that assist yeast cells to survive passage through the insect digestive tract, digestion by the insect may remove the ascus wall and causes some spores to become isolated (Stefanini et al., 2012). If an isolated spore germinates, it has no way of making new spores unless it finds a mating partner of the opposite mating-type. Switching provides a partner, allowing cells to become diploid and able to make new spores.

1.2.5 Evolution of the mating-type system in the Saccharomycetaceae family

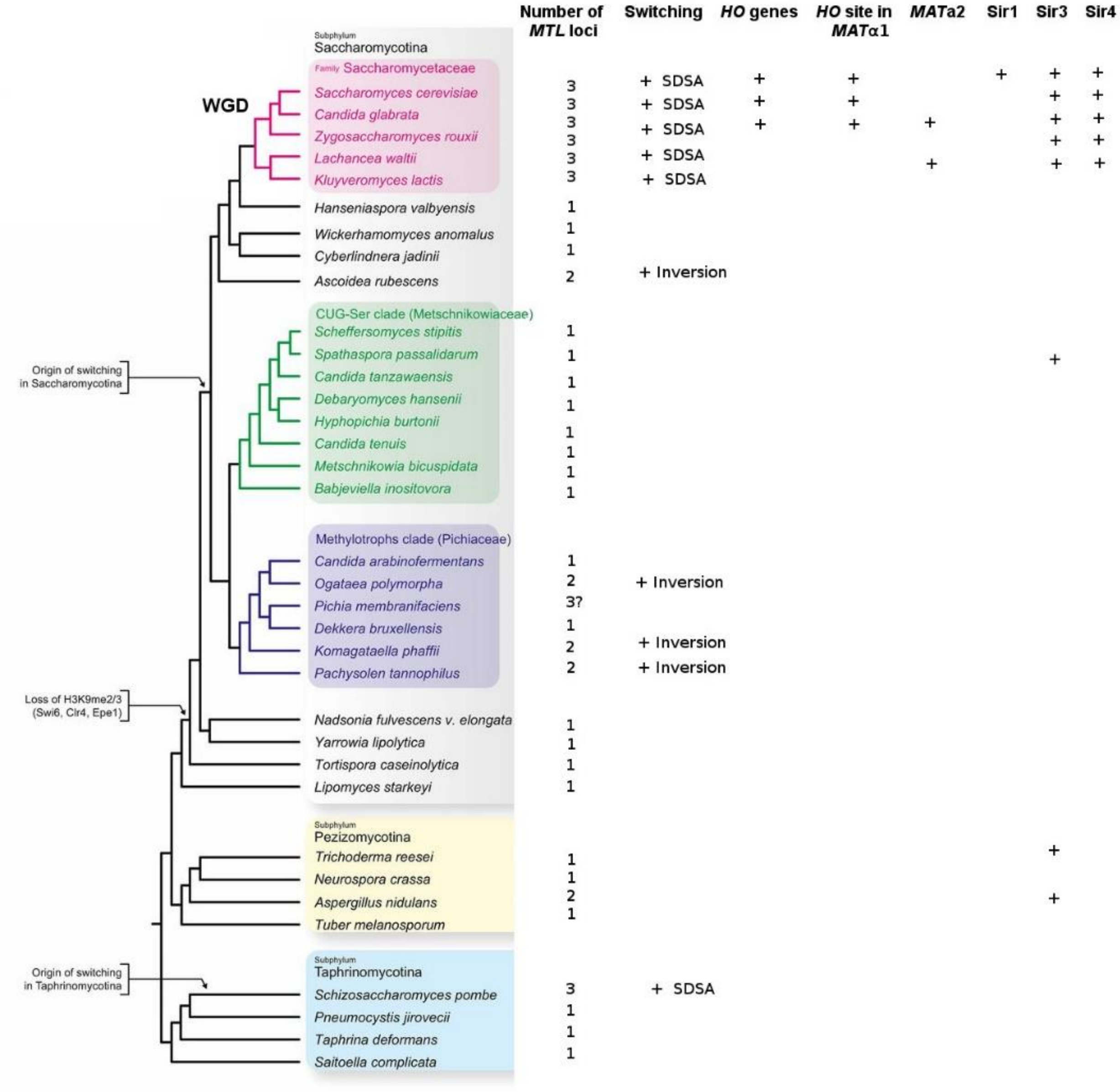
Polyploidy, a state in which the chromosome complement has undergone a sudden increase by genome doubling, is a major force in evolution and has been extensively exploited for improving food yeasts. WGDs are rare evolutionary events with profound consequences. They double an organism’s genetic content, creating a reproductive barrier between it and its ancestors and providing raw material for the divergence of gene functions between homologs (pairs of paralogous genes produced by the WGD) (Wolfe et al., 2015). They must overcome a suite of biological responses to this merger, known as genome and transcriptome shocks. Genome shock describes genome changes that occur in response to polyploidization, such as gene loss, chromosome mis-pairing, and rearrangements between the sub-genomes. Transcriptome shock includes rearrangements of gene expression networks following the mixing of two dissimilar genomes, each with their own set of transcription factors, cis-regulatory elements, and their own chromatin profiles.

Phylogenomic studies revealed that in subphylum Saccharomycotina mating system has evolved from an obligate heterothallic system (as seen in *Yarrowia lipolytica*), to heterothallism with low-switching frequency (as seen in *Kluyveromyces lactis*) and finally to a HOcatalyzed homothallic switching (as seen in *S. cerevisiae*), via a three-step event (Butler et al., 2004).

The first step was the origin of the *HML* and *HMR* cassettes (three-cassette system), which occurred in Saccharomycetaceae family after it had diverged from the GTG clade (including

Debaryomycetaceae and the *C. albicans* clade) and the methylotrophic yeasts *Hansenula polymorpha* and *Pichia pastoris* (**Fig. 1.8**).

Figure 1.8. Phylogenetic tree of phylum Ascomycota showing major clades, *MAT*-locus organization, and known or inferred mating-type switching mechanisms (adapted from Hanson and Wolfe, 2017). Based on Riley et al. (2016), with placement of *Amanita rubescens* as in Shen et al. (2016), mating-type switching does not occur in species with only one *MAT*-like locus or in *Aspergillus nidulans*, which is a primary homothallic species. Abbreviations; WGD, whole genome duplication; SDSA, synthesis-dependent strand annealing (or gene conversion).


Number of MTL loci
Switching
HO genes
HO site in MATα1
MATa2
Sir1
Sir3
Sir4
Subphylum Saccharomycotina
Family Saccharomycetaceae
WGD
Saccharomyces cerevisiae 3 + SDSA + + + + +
Candida glabrata 3 + SDSA + + + +
Zygosaccharomyces rouxii 3 + SDSA + + + + +
Lachancea waltii 3 + SDSA + +
Kluyveromyces lactis 3 + SDSA + + +
Hanseniaspora valbyensis 1
Wickerhamomyces anomalus 1
Cyberlindnera jadinii 1
Ascoidea rubescens 2 + Inversion
CUG-Ser clade (Metschnikowiaceae)
Scheffersomyces stipitis 1
Spathaspora passalidarum 1 +
Candida tanzawaensis 1
Debaryomyces hansenii 1
Hyphopichia burtonii 1
Candida tenuis 1
Metschnikowia bicuspidata 1
Babjeviella inositovora 1
Origin of switching in Saccharomycotina
Methylotrophs clade (Pichiaceae)
Candida arabinofermentans 1
Ogataea polymorpha 2 + Inversion
Pichia membranifaciens 3?
Dekkera bruxellensis 1
Komagataella phaffii 2 + Inversion
Pachysolen tannophilus 2 + Inversion
Loss of H3K9me2/3 (Swi6, Clr4, Epe1)
Nadsonia fulvescens v. elongata 1
Yarrowia lipolytica 1
Tortispora caseinolytica 1
Lipomyces starkeyi 1
Subphylum Pezizomycotina
Trichoderma reesei 1 +
Neurospora crassa 1
Aspergillus nidulans 2 +
Tuber melanosporum 1
Origin of switching in Taphrinomycotina
Subphylum Taphrinomycotina
Schizosaccharomyces pombe 3 + SDSA
Pneumocystis jirovecii 1
Taphrina deformans 1
Saitoella complicata 1

Recently, Hanson and colleagues (2014) included these last yeasts in the comparative genomics analysis of *MAT* loci the methylotrophic yeasts, which belong to the Saccharomycotina subphylum, but diverge from the Saccharomycetaceae family and the members of the clade GTG. These last yeasts are homothallic, but lack of *HO* and possess two *MAT* loci, which are switched by reversible inversion of a chromosomal section with *MAT***a** genes at one end and a *MAT*α gene at the other end (Hanson et al., 2014). This inversion (or flip-flop)-based recombination mechanism moves genes between expressed and non-expressed sites. Hanson and co-workers proposed that the three-cassette system evolved from a two-cassette flip-flop model.

Species having silent cassettes, for example *Lachancea waltii* (Di Rienzi et al., 2011), are probably able to switch mating-types using the homologous recombination machinery, however, in some clades, a second evolutionary step increased the rate and/or precision of switching by directing a DSB to the *MAT* locus in cells that are about to switch. This second step occurred independently in different species having silent cassettes, and consists in the acquisition of a specialized machinery for increasing the rate and/or accuracy of switching by directing a DSB to the *MAT* locus. In *Saccharomyces* and their closest relatives *Zygosaccharomyces* species (Solieri et al., 2013a), the DSB is catalysed by the *HO* endonuclease. Rajaei and colleagues (2014) demonstrated that in *K. lactis* the DBS is made by the trasposase Kat1, which catalyzes the excision of a mobile genetic element from the *MAT*α idiomorph during the switching from *MAT*α to *MAT***a**. This mobile element contains a gene, α3 which is only present in *Kluyveromyces* (Rajaei et al., 2014). In methylotrophic yeasts *H. polymorpha* and *P. pastoris*, which have a two-cassette system, but lack of homologs of *HO* endonuclease, the inversion of *MAT* locus can be induced when two strains with the same mating-type are crossed, but it is not clear whether the mechanism of inducible inversion is mediated by a specific recombinase or by a general recombination machinery (Hanson et al., 2014).

The third event in the evolution of *S. cerevisiae* mating-type switching mechanism is the loss of an additional HMG domain gene (*MAT***a**2), which codes for an HMG DNA-binding protein. The *S. cerevisiae MAT* locus idiomorphs code for only three proteins: the homeodomain proteins α1 and α2 and the ‘‘**a**-domain’’ protein 1. Homologs of these three genes are found in nine hemiascomycete species. The additional gene *MAT***a**2 is present in the *MAT***a** idiomorphs of several species, including *C. albicans*, *S. kluyveri* and *Z. rouxii* (**Fig. 1.8**). In this vein, Tsong and colleagues (2006) proposed an evolutionary model of transition from positive to negative regulation of **a**sgs. In *C. albicans* persists the ancestral network where **a**sgs are activated in cells by the **a**2-Mcm1 heterodimer, whereas in *K. lactis* an additional **a**sgs repressor Mcm1-α2 appeared in α cells. *S. cerevisiae* lineage recently lost the *MAT***a**2 gene, acquiring a α2-repressing mode of **a**sgs by α2- Mcm1-α2 complex in α cells.

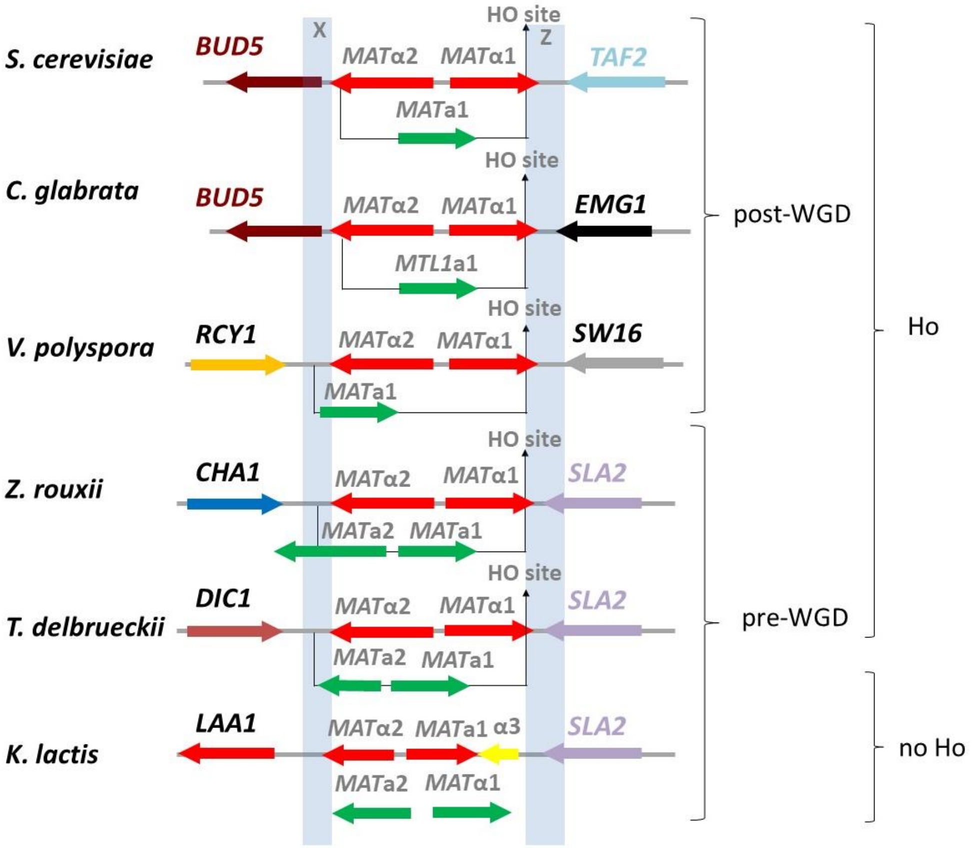
In all species that have silent cassettes, DNA repair at *MAT* locus is guided by two regions (the Z and X regions) that are almost identical among *MAT*, *HML*, and *HMR* loci. Butler and colleagues (2004) observed that in *S. cerevisiae*, *C. glabrata*, and *K. delphensis* the cleavage site for *HO* endonuclease is located within *MAT*α1 gene, because the 3' end of this gene is located in the Z region (**Fig. 1.9**). Similarly, in *Saccharomyces castellii* the Y/Z junction occurs within *MAT*α1 gene and the genome sequence includes *HO* gene. In more distant species, such as *S. kluyveri*, *K. lactis*, *Y. lipolytica* and *C. albicans MAT*α1 gene is located completely within the unique Yα region of *MAT*α1 and the sequences at the Y/Z junctions are heterogeneous and do not resemble the *HO* site. Considering the phylogenetic relationship among the species, we can infer that the Y/Z boundary has appeared inside *MAT*α1 gene in recent times, because of the gain of the *HO* cleavage site. Comparative genome approaches pointed out that the *MAT* and *HML* loci are linked *in cis* on the same chromosome, probably due to the conservation of the Recombination Enhancer (RE) site among species (Faber et al., 2005; Gordon et al., 2011). The RE, which has so far only been found in *S. cerevisiae*, is located between *HML* and *MAT* loci and it seems to increase the frequency of productive switching by unfairly influencing the choice of donor, operating by binding the α2 protein.

Another interesting topic related to the evolution of mating-type system is how the heterochromatin-based gene silencing rapidly evolved in *S. cerevisiae* and its close relatives among the budding yeasts. This system is based on four Sir proteins, among which Sir1 is perhaps the most enigmatic. In particular, in *S. cerevisiae*, Sir1 is a histone deacetilase essential to mediate the *HM* loci silencing together with the SIR complex; in fact the failure to recruit Sir1 is thought to account for the instability of subtelomeric silencing relative to *HM* loci (Chien et al., 1993). Abnormal expression of cryptic *HMR*/*HML* loci has been described in *Vanderwaltozyma polyspora*, the *Z. rouxii* closest relative, which branched after the WGD (Roberts and Van der Walt, 1959). Consequently, *V. polyspora* haploid cells behave as **a**/α diploid and appear mating-incompetent for many generations only to subsequently restore silencing. Significantly, this inability in *V. polyspora* has been correlated to the lack of Sir1 histone deacetilase. Like *V. polyspora*, *C. glabrata* is another species close to *Z. rouxii*, which lacks of a *SIR1* ortholog (Gábaldon et al., 2013). The defection of a complete silencing system lead to the expression of *MAT***a** gene in *C. glabrata MAT*α cells (Muller et al., 2008) and makes *HML* more prone to *HO* cleavage at the Y/Z junctions (Boisnard et al., 2015).

In 2016, Ellahi and Rine examined the evolution of Sir-based silencing, focusing on Sir1, using the budding yeasts *S. cerevisiae* and the pre-WGD species *Torulaspora delbrueckii*, which is a budding yeast evolutionarily well positioned to explore some of the most enigmatic questions concerning the origins of Sir-based silencing, and especially the role of Sir1. Chromatin immunoprecipitation followed by deep sequencing (ChIP-Seq) analysis of Sir proteins revealed that in *T. delbrueckii HML* and *HMR* loci have a different topography of chromatin than was observed in *S. cerevisiae*., where Sir1, enriched at the silencers of *HML*α and *HMR***a**, was absent from telomeres and did not repress subtelomeric genes. In *T. delbrueckii* they found the most ancestral form of *S. cerevisiae SIR1*, a paralog named *KOS3* (*K*in *o*f *S*ir1 3), that is located ~1 Kb away from the copy of *HMR* cassette. In contrast to *S. cerevisiae SIR1*'s partially dispensable role in silencing, *KOS3* plays a key role in *HML*/*HMR* silencing (Ellah and Rine, 2016). Another pre-WGD species, *Z. rouxii*, possesses the archetypal member of the *SIR1* family, *KOS3* (Gallagher et al., 2009).

In 2011, Gordon and colleagues, by comparing the *MAT* loci organization in 16 species belonging to the family Saccharomycetaceae, inferred the "ancestral" gene order inferred to have existed just before WGD occurred (**Fig. 1.9**). In the ancestral genome nomenclature, *HML* and *MAT* are on chromosome 1, with *HML*α1 and *HML*α2 being the first two genes on this chromosome and the *MAT* locus located about 120 genes farther along. The genes ancestrally flanking the 5' and 3' end of *MAT* are *DIC1* and *SLA2*, an arrangement that seems to be quite old and stable because it is conserved in several pre-WGD species. Ancestral chromosome 1 was duplicated during WGD, giving rise to two daughter chromosomes. We call one of them the "*MAT* chromosome" because it lost its copies of these loci. After WGD, both chromosomes underwent further rearrangements and large deletions, beginning at the *MAT* locus and extending in the Z direction. There are many evidences that switching errors, which accumulate along evolutionary lineages, have a profound effect on the structure and organization of the *MAT*-containing chromosome in post-WGD species. Over evolutionary time, chromosomal genes located immediately beside *MAT* have continually been deleted, truncated or transposed to other places in the genome in a process that has gradually brought silent cassettes *HMR* and *HML* into proximity with *MAT* locus, resulting in a higher risk for removal. During this process, the triplicated sequence regions, called Z and X, have continually replaced to allow *HML* and *HMR* to be used as templates for DNA repair at *MAT* locus during mating-type switching. A general tendency to delete DNA beside the *MAT* locus exists in pre-WGD species as well as post-WGD species; however, the effects of the deletion process are much more drastic in post-WGD species. The deletion and transposition events may have been caused by evolutionary accidents during mating-type switching, combined with natural selection to keep *MAT* and *HML* on the same chromosome. As mating-type switching is an accident-prone process that removes genes and erodes the flanking chromosomal DNA it is likely to impact on the biology of the species in which it occurs. Furthermore, error-prone DNA synthesis occurs during switching. Therefore, unlike recombination, switching does not create or maintain any genetic diversity and it occurs both in species that grow primarily as diploids and in the others that grow primarily as haploids. The most important benefit brought by switching mechanism may be that it provides a way for an isolated germinating spore to reform spores if growth conditions are too poor or to test environment of uncertain quality.

Figure 1.9. Schematic organization of the *MAT* locus in six species (adapted from Butler, 2004). For each species, the main horizontal line shows the organization of the α idiomorph, and the **a** idiomorph is represented by the offset box below it. The recognition site of Ho endonuclease in *MAT*α1 is marked when present. α idiomorph are coloured in red; **a** idiomorph are indicated in green.



1.3 FOCUS ON THE *ZYGOSACCHAROMYCES ROUXII* COMPLEX

The *Z. rouxii* complex includes industrially important halotolerant and osmotolerant yeasts that participate in both the elaboration and spoilage of foodstuff. These yeasts are able to grow in such high salt and/or sugar concentrations, which hamper the growth of many other species, included *S. cerevisiae*. They have been identified as potential spoilage agents for different food products, such as fruit juices, soft drinks, high-sugar syrups, acetic preserves, wine, and cider (Statford, 2006; Dakal et al., 2014). The *Z. rouxii* complex strains exhibit a near continuity in salt tolerance (Solieri et al., 2014a), as well as remarkable genome size and karyotype variability (Solieri et al., 2008), and unusual rDNA heterogeneity (Dakal et al., 2016; Solieri et al., 2013a).

Z. rouxii is also very attractive for genome evolution study since this clade diverged from the *Saccharomyces* lineage after gaining the *HO* gene, but before WGD event occurring in the ancestor of *Saccharomyces* clade (Wolfe and Shields, 1997; De Montigny et al., 2000; Dakal et al., 2014). Thus, it represents the pre-WGD species most closely related to the model yeast *S. cerevisiae* and it has been also indicated as near to the parental lineage ZT involved in allopolyploidization leading to the *Saccharomyces* lineage (Marcet-Houben and Gabaldón, 2015; Wolfe et al., 2015).

1.3.1 Genotypic and molecular diversity within the *Z. rouxii* complex

Although the physiological and biochemical processes of *Z. rouxii* have been studied extensively (Van Zyl et al., 1990; Kurtzman and Fell, 1998; Jansen et al., 2003; Martorell et al., 2007), relatively little is known about the diversity in this branch at genetic and molecular levels, mainly due to the lack of tools and bio-bricks for its genetic manipulation and engineering. In 2013, Solieri and colleagues investigated the *Z. rouxii* complex, encompassing strains that in other works have been studied separately and comparing them in a comprehensive way (**Table 1.2**). The strains considered in this thesis were described as a complex of haploid and diploid heterogeneous species. Three major groups were delineated: i) the haploid and mating-competent *Z. rouxii* (type-strain CBS 732^T); ii) the diploid species *Zygsaccharomyces sapae* isolated from Traditional Balsamic Vinegar (TBV), for which ABT301^T is the type-strain (which rarely undergo meiosis); iii) a subgroup of allodiploid and sterile strains with uncertain taxonomical position, including someone retrieved from salt environment, such as ATCC 42981, CBS 4837 and CBS 4838 strains (Solieri et al., 2008, 2013a). Gordon and Wolfe (2008) found that

ATCC 42981 genome originated from a recent allopolyploidization event between two phylogenetically divergent sub-genomes. They named them T and P sub-genomes and hypothesized that they derived from one lineage that is over 99% identical to CBS 732^T and another one not yet identified (presumably *Z. pseudorouxii* nom. inval. NCYC 3042), respectively. Any traces of gene losses were found in ATCC 42981 genome, suggesting that the allopolyploidization event was so recent that its genome has not had enough time to decay. However, ATCC 42981 karyotype cannot be simply considered as an additive result between the putative parental counterparts, suggesting that some structural rearrangements have occurred in the ATCC 42981 genome compared to *Z. rouxii* and *Z. pseudorouxii* (Pribylova et al., 2007; Gordon and Wolfe, 2008). Recently, whole genome sequencing of another *Z. rouxii* strain, CBS 4837, suggested that also this osmotolerant/halotolerant yeast possesses an allopolyploid genome (Sato et al., 2017). Unlike haploid *Z. rouxii* CBS 732^T strain, *Z. sapae* and aneuploid/allodiploid strains display an unusual rDNA heterogeneity with regard to the internal transcribed spacers (ITS), rRNA regions and/or the D1/D2 domains of large subunit rDNA (LSU; James et al., 2005; Solieri et al., 2013a). They showed that the majority of strains are unusually heterogeneous in their ribosomal DNA, but the pattern of this heterogeneity varies significantly between *Z. sapae* and the mosaic lineage. The co-occurrence of rRNA gene variants in the genome of a single individual suggests relaxation of concerted evolution in a recombination-driven process that is responsible for homogenizing rRNA gene repeats (Birky, 1996).

Table 1.2. Overview of the main molecular and genetic properties of strains belonging to *Z. rouxii* complex (adapted from Solieri et al., 2013a). Abbreviations: *Zr*, *Zygosaccharomyces rouxii*-like copy; *Zs*, *Zygosaccharomyces sapae*-like copy.

		Z. sapae		**Mosaic lineage**			
Properties	CBS 732^{T}	ABT301^{T}	ABT601	OUT7136	CBS 4838	CBS 4837	ATCC 42981
Genome size	9.8-12.7 †	28.1±1.3	39.0±0.3	19.57±0.47	22.5±0.20	21.7±0.33	21.9±0.20
Ploidy	Haploid	Diploid	Diploid	Aneuploid	Aneuploid	Aneuploid	Diploid
Chromosome no.	6	10	11	8	8	8	8
Markers							
ZSOD2	*ZrSOD2-22*	*ZrSOD2-22 ZrSOD22*	*ZrSOD2-22 ZrSOD22*	*ZrSOD22-ZrSOD2*	*ZrSOD22-ZrSOD2*	*ZrSOD22-ZrSOD2*	*ZrSOD22-ZrSOD2*
HIS3	Zr	2 (Zr+Zs)	2 (Zr+Zs)	2 (Zr+Zs)	2 (Zr+Zs)	2 (Zr+Zs)	2 (Zr+Zs)
ITS	Zr	2 (Zr+Zs)+1 †	2 (Zr+Zs)+1	Zr	2 (Zr+Zs)+1	2 (Zr+Zs)+1	2 (Zr+Zs)+1
LSU D1/D2	Zr	Zs	Zs	Zs	2 (Zr+Zs)+1	2 (Zr+Zs)+1	2 (Zr+Zs)+1
COX2	Zr	Zs	Zs	Zr	Zr	Zr	Zr

*Genome size in Mb.† Genome size of CBS 732^{T} was estimated at 9.8 Mb as result of the final assembly of genome project (Souciet et al., 2009) and of 12.7 Mb according to PFGE determination (Solieri et al., 2008). ± Additional recombinant copy.

A previous study showed that repeats of the 5S ribosomal genes of filamentous fungi species are dispersed among the genomes and escaped the concerted evolution model (Rooney and Ward, 2005). The results outlined a high variability in the rRNA composition among species and the possible explanations could be various: these strains may be either aneuploidy or diploid, with each chromosome of the pair of homologous chromosomes bearing one rRNA variant; there are different tandem repeat variants arranged in their rRNA gene arrays located on the same chromosome; there are divergent rRNA gene arrays dispersed among different chromosomes. Further analysis showed that *Z. rouxii* species are characterized by hypervariable karyotypes, different levels of ploidy and harbour mosaic genomes with two copies of many genes, suggesting that the genome has rearranged very fast upon the separation of single lineages. Changes in ploidy and karyotype have been demonstrated to introduce potentially significant physiological effects, providing high or low fitness benefits mainly in highly stressful environments (Mable and Otto, 2001; Zeyl et al., 2003; Anderson et al., 2004). Unlike haploid *Z. rouxii* CBS 732^T strain, *Z. sapae* and the allopolyploid group displayed copy number variations of some genes, such as the housekeeping markers *ZrSOD* and *HIS3*. Furthermore, *Z. sapae* possesses diploid genome bigger than those of other strains belonging to the *Z. rouxii* complex (Gordon and Wolfe, 2008; Solieri et al., 2013a). Sequence analysis of individual genes confirmed that the parental strains contributing to the mosaic genome closely resemble the *Z. rouxii* type-strain CBS 732^T and *Z. pseudorouxii* nom. inval. (James et al., 2005; Gordon and Wolfe, 2008). Overall, these findings led us to suppose an evolutionary model in which the presumptive diploid-like status originated prior to sorting of *Z. sapae* and the mosaic lineage into different groups. Under this hypothetical evolutionary scenario, outcrossing between two divergent haploid cells would have resulted in a diploid ancestor, which lost the ability to undergo meiosis and gave rise to stable diploid lineages (namely *Z. sapae* and the mosaic lineage), which reproduced clonally and evolved independently. A possible explanation for this unexpected diversity could be the resulting mode of propagation alternative to classical sexuality described in *Saccharomyces* species, which can account for this genome complexity making the *Z. rouxii* complex prone to genome mosaicism and reticulate evolution.

1.3.2 Phenotypic variability within the *Z. rouxii* complex

Modern yeast biotechnology places a large emphasis on exploring potential biotechnological applications of so-called non-conventional yeasts, such as *Pichia*, *Zygosaccharomyces* and

Kluyveromyces (Porro and Branduardi, 2009). Within the genus *Zygosaccharomyces*, *Z. rouxii* has a leading role in food industry as both fermentation-driving biocatalyser and food-spoiling agent. *Z. rouxii* strains have also been exploited as cell factories for producing enzymes characterized in the model organism *S. cerevisiae*, which shows moderate halo- and osmo-tolerance (Kashyap et al., 2002). In contrast, little efforts have been put in determining the molecular physiology in *Z. rouxii* (Watanabe et al., 1995, 2004; Iwaki et al., 1998, 1999; Pribylova et al., 2008).

The *Z. rouxii* complex shows variability with respect to not only gene copy number variation, karyotype variability and change in ploidy, but also to phenotypic and functional diversity in stress responses (Solieri et al., 2014a). By subjecting several strains of the *Z. rouxii* complex to a pattern of different environmental perturbations, encompassing high concentration of alkali metal cations, glycerol consumption and growth at 37°C, it was underlined inter-strain stress response variation within the *Z. rouxii* complex, partially supporting the previous segregation of strains in two different clusters, that is, the allodiploid/aneuploid group (ATCC 42981, CBS 4837 and CBS 4838 strains), the diploid *Z. sapae* and the haploid CBS 732^T strain. No differences were detected in response to osmotic pressure mediated by sugars, whereas *Z. sapae* resulted less responsive to salt adaptation than allodiploid strains. Probably because the allodiploid ATCC 42981, isolated from miso paste, has two copies of functional genes involved in the production of glycerol as a compatible solute to protect the cell against lysis and efflux of Na+ from cells in high concentrations of salt (James et al., 2005; Solieri et al., 2006; Solieri et al., 2007; Gordon and Wolfe, 2008). These redundant genes in an allopolyploid strain can contribute to survival under high-osmotic conditions, such as the process of brewing soy sauce. These results imply that the two groups differ in mechanisms counteracting high ionic strength and low aw and confirm that physiological response to sugar stress may not be as effective for salt stress (Lages et al., 1999). In a previous work, Solieri and colleagues (2013b) proposed that the inbreeding system could lead to aneuploidy and allodiploidy in *Zygosaccharomyces*. As genetic and phenotypic variability may be induced by stress (as reviewed by Berman and Hadany, 2012), they hypothesized that *Zygosaccharomyces* reproduction is a hypermutagenic process that contributes to stress adaptation by generating progenies with different genetic and phenotypic outcomes. The generation of divergent lineages could be a successful strategy, developed under stressful conditions, to increase the probability to achieve descendants improved in adaptation to hostile environments. Further investigation is required to see whether there is a link among genetic and phenotypic variation and inbreeding reproduction inside the *Z. rouxii* complex requires.

1.3.3 *MAT* loci: hypermutational hotspots in *Z. rouxii*

Z. rouxii complex diverged from the *Saccharomyces* lineage after gaining the *HO* gene, but before WGD event, therefore chromosomal rearrangements involving mating-type conversion may be detected in its genome. The availability of complete genome sequence for *Z. rouxii* type-strain CBS 732^T provides an ideal opportunity to analyze the consequences of *HO* gene acquisition.

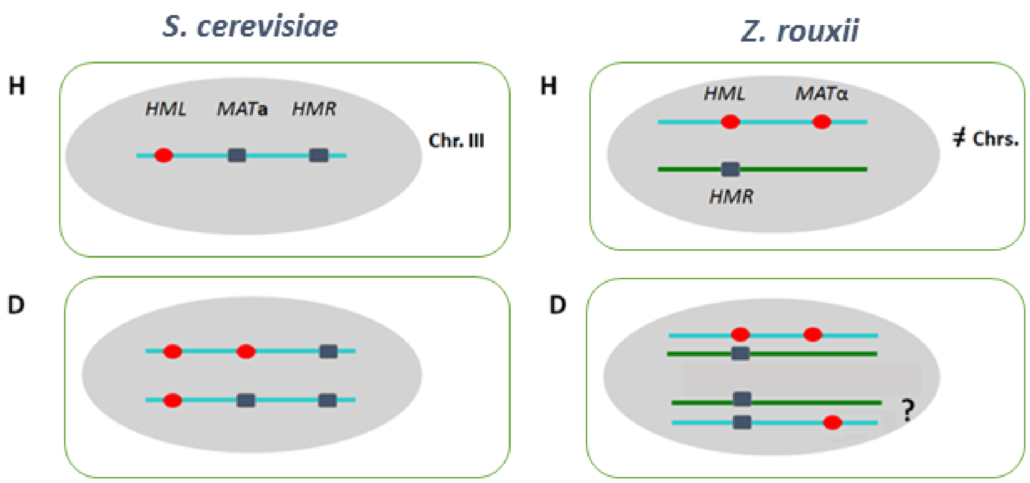
Solieri et al. (2014) and Watanabe et al. (2013) demonstrated that *Z. rouxii* inter-strain karyotype variability is related to sex-chromosome diversity caused by ectopic recombination at the *MAT*, *HML* and *HMR* loci during mating-type switching. This genotypic variability could favour phenotypic diversity and adaptation to hostile environments.

Differently from *S. cerevisiae*, in *Zygosaccharomyces* yeasts the *HMR* locus and the *MAT-HML* linkage are located on distinct chromosomes. This means that *MAT*-like (*MTL*) loci are susceptible to ectopic and inter-chromosomal recombination events between two non-homologous chromosomes in haploids and two pair of non-homologous chromosomes in diploids (**Fig. 1.10**). Furthermore, as being an error prone mechanism, switching contributes to strong variability in organization and structure of sex chromosomes at many levels. Haploid strains frequently exhibit abnormal genotypes with redundant number of *MAT*-like cassettes flanked by variable genes, resulting from *HO*-independent inter-chromosomal translocations between *MAT*-like loci.

Watanabe and colleagues (2013) demonstrated that reciprocal translocation at the *MTL* loci was responsible of genomic instability in CBS 732^T stock of the type-strain of the *Z. rouxii* species. Furthermore, translocation events make the *MTL* flanking regions variable in strains of the same species. Finally in the *Z. sapae* type-strain ABT301^T, with **a**ααα genotype, an unusual cassette configuration without a *HMR* silent cassette (Solieri et al., 2013), makes difficult the α-**a** switching and dys-regulates cell-type identity (Bizzarri et al., 2016).

Although the pre-WGD species are known to retain a *MAT*-locus organization similar to *Torulaspora delbrueckii* (*DIC1-MAT-SLA2*), *Z. rouxii* CBS 732^T strain shows a *MAT* organization divergent from that of other non-WGD species (*CHA1-MAT-SLA2*). This new arrangement is the consequence of a translocation, occurred in *Z. rouxii* type-strain, joined the X side of *MAT* to a telomeric region containing *CHA1* (Gordon et al., 2011). This *CHA1-MAT* linkage represents a peculiar structural feature of *MAT* locus in *Z. rouxii*, as it has never been detected in the genome of other pre-WGD hemiascomycetous yeasts, leading to infer that if these deletions, truncations and transpositions are related to the acquisition of *HO* endonuclease, a remnant of degraded *MAT* organization may be detected among the *Z. rouxii* population.

Figure 1.10. Comparison between the different three-cassette system organization found in *S. cerevisiae* and *Z. rouxii*. In *S. cerevisiae* diploids (D), the two blue bars indicate two non-homologous chromosomes; while in *Z. rouxii* diploids (D) the green and blue bars represent two pair of non-homologous chromosomes in diploids. Red dots are the α idiomorphs, while the dark boxes stand for the **a** idiomorphs. Abbreviations: H, haploid; D, diploid; Chr, chromosome.



More recently, Watanabe et al. (2013) analyzed the *MAT* locus organization in the *Z. rouxii* population of haploid strains; in particular, the authors considered a pool of haploid *Z. rouxii* strains as well as two different stocks of the *Z. rouxii* type-strain, namely CBS 732^T and the Japanese stock NBRC 1130^T. PCR analysis was carried using specific primer sets that annealed to *MAT*, *HML* and *HMR* flanking regions (outside of the X and Z regions) characterized in *MAT*, *HML* and *HMR* cassettes of CBS 732^T genome that has been sequenced. PCR results can be represented not only by flanking regions described in CBS 732^T genome (*CHA1-MAT-SLA2*), but also by other combinations, confirming that the flanking sequences of *MAT*, *HML* and *HMR* cassettes are highly variable among the *Z. rouxii* populations. These results imply that the *MAT*, *HML* and *HMR* loci represent translocation and mutation hotspots in *Z. rouxii* haploid strains, suggesting that chromosomal rearrangements could play key roles in phenotypic variation and genome evolution. Perhaps this reciprocal translocation has occurred in the early passage culture of CBS 732T in the past 80 years because the type-strain, deposited in Centraalbureau voor Schimmelcultures (CBS) collection and distributed to others, was originally isolated by Sacchetti in 1932. Furthermore, they found out that not all the PCR products confirmed CBS 732^T genome sequencing data and that in four haploid strains, namely NBRC 1733, NBRC 0686, NBRC 0740 and NBRC 1053 the terminal region of the chromosome containing *HMR* locus was replaced with the chromosomal region on the left of the *MAT* or *HML* loci (Watanabe et al., 2013). The loss of the region on the right of the *HMR* locus in these strains was confirmed by PCR analysis, which detected only α amplicons, suggesting that they lost **a** information and were forced to behave like heterothallic. Overall, these outcomes indicate that the acquisition of *HO* endonuclease increased the frequency of genotypic switching in *Z. rouxii*. However, the inter-chromosomal rearrangement accompanying ectopic recombination between *MAT* and *HMR* or between *HML* and *HMR* cassettes is independent of the HO endonuclease acquisition, at least under experimental conditions tested by Watanabe and colleagues. Intriguingly, they found that the difference in switching frequency between *Z. rouxii* wild type and Δ*HO* cells is negligible when compared with the difference detected between *S. cerevisiae HO* and ho strains (10^6 order of magnitude; Hicks et al., 1977). These data suggest that *HO* gene could play a different role in regulating mating-type switching between *S. cerevisiae* and *Z. rouxii*.

1.4 Dobzhansky-Muller incompatibility

Saccharomyces sensu stricto yeast species are generally distinct based on low viability of spores produced by hybridization. Whereas mating between members of the same *S. cerevisiae* strain produces spores with viabilities of close to 100% (Greig et al., 2002) and spores produced by mating between *S. cerevisiae* strains often show viabilites of ~80% (Greig et al., 2002), mating between *S. cerevisiae* and *Saccharomyces paradoxus* or other *Saccharomyces* species typically result in <1% viable of spores (Greig et al., 2002). The bases of the reproductive barriers among *Saccharomyces sensu stricto* yeasts have been investigated intensely over the last few years. A variety of different mechanisms can be responsible for hybrid infertility, such as chromosomal rearrangements, Dobzhansky-Muller incompatibilities between epistatically interacting genes and sequence divergence acted on by the mismatch repair system.

The Dobzhansky-Muller model (DM model) posits that after an ancestral lineage diverges to create two daughter lineages incompatible changes arise in alternative members of a pair of interacting loci. Thus, in one lineage, one of the genes diverges from the ancestral sequence and in the second lineage the other gene diverges from the ancestral sequence. In the case that the diverged versions of both genes are brought together in a hybrid, they will interact in such a way as to reduce fitness through a mechanism that is not specified. It is important to underline that the incompatibility can be either dominant or recessive. In the former case, the presence of the two diverged genes will reduce fitness irrespective of what other genes are present. In the latter case however, the existence of an incompatibility can be masked by the presence of an ancestral type sequence at both loci (*e.g.* in an F1 hybrid).

1.4.1 Allopolyploid sterility and the DM model

As result of increased allelic variation and modification of transcriptional networks, allopolyploids display traits of tremendous agro-economic and evolutionary value, such as heterosis (the increased levels of growth exhibited by allopolyploids relative to their parents) and better stress adaptation to changing or sub-optimal environments compared to the parents. However, allopolyploids cannot be used in breeding programs for novel strain development or in genetic analysis due to their sterility. Despite the evolutionary and biotechnological importance, the molecular bases of pre- and postzygotic reproductive isolation remain elusive in yeast allopolyploids. For example, *Saccharomyces* species are postzygotically isolated: interspecies F1 hybrids are viable but sterile, producing only about 1% viable gametes that are generally highly aneuploid. When chromosomes from different parents are sufficiently diverged, they cannot crossover during meiosis and so fail to segregate accurately. Classical chromosomal speciation cannot be the cause of hybrid sterility, because all *Saccharomyces sensu stricto* species have 16 chromosomes with a total of just four, two or zero rearrangements relative to S. cerevisiae that have no correlation with genetic distance (Fischer et al., 2001; Orr and Turelli, 2001).

In animals (especially *Drosophila*), genic incompatibility is thought to be the primary cause of hybrid sterility, as predicted by Dobzhansky (1937) and Muller (1942). In the DM model, the extent of the reproductive isolation relies on epistatic interactions between genes. Negative epistasis occurs where independently fixed mutations in allopatric populations could not properly function together when combined in hybrids. In particular, isolated populations fix different beneficial alleles at different loci, which results in reproductive isolation if the alleles are incompatible when in the same genetic background. In the most simple form, a two-allele two-locus model, genotypes AAbb and aaBB have normal fitness but hybrid genotypes such as aabb have reduced fertility, as the a and b alleles at the two loci are incompatible. The extent of incompatibility can vary with the ‘dominance’ of the effect, in that genotypes aabb, Aabb and AaBb may differ in epistatic interactions and hence incompatibility (Orr and Turelli, 2001).

Despite several efforts, a few types of DM genetic incompatibilities have been identified at inter-species level in yeast so far. Lee et al. (2008) demonstrated that incompatibility of nuclear and mitochondrial genomes causes hybrid sterility between species *S. cerevisiae* and *Saccharomyces bayanus*. Another candidate locus for DM incompatibility could be the *MAT* cassette regulating cell type identity. Greig et al. (2002) deleted the *S. cerevisiae* copy of the *MAT* locus in F1 hybrid diploids. Diploids with two active copies of the *MAT* locus do not have mating-type, but with only one copy of the *MAT* locus, a diploid behaves as a gamete that can divide, switch mating-type and auto-fertilize, producing an allotetraploid.

A proper regulation of the expression, activation, and interaction of *MAT* locus genes is also or essential for growth and differentiation in yeasts, since they are master regulators of cell-type identity. For instance, in *C. albicans* sexual reproduction governs the switching between morphological forms and it is correlated to the shift from opportunistic to pathogenic status (Heitman, 2006). Furthermore, reproductive gene isolation and incompatibilities among *MAT*-related genes are invoked as sources of reproductive isolation during fungi speciation, together with cito-nuclear incompatibilities (Solieri, 2010). Therefore, knowledge on functional mating system and the way to restore it when impaired in sterile hybrids will thus reduce the amount of efforts required for breeding process and assist the study of genetic determinants of important industrial traits and may provide proper disease management strategies and can elucidate the speciation and evolution of life history in ascomycetes.

CHAPTER 2: SEQUENCING AND SEQUENCE DATA ANALYSIS

In this Chapter, I present a review of the main improvements in DNA sequencing technologies. Particular attention is paid to the most recent 3rd generation sequencing technologies, since they are the only platforms that allow the resolution of mosaic structures typical of hybrid genomes, the phase reconstruction of parental complements and a reliable identification of gene losses or redundancies by limiting miss-assemblies due to collapsing of homeomologous and/or repetitive segments.

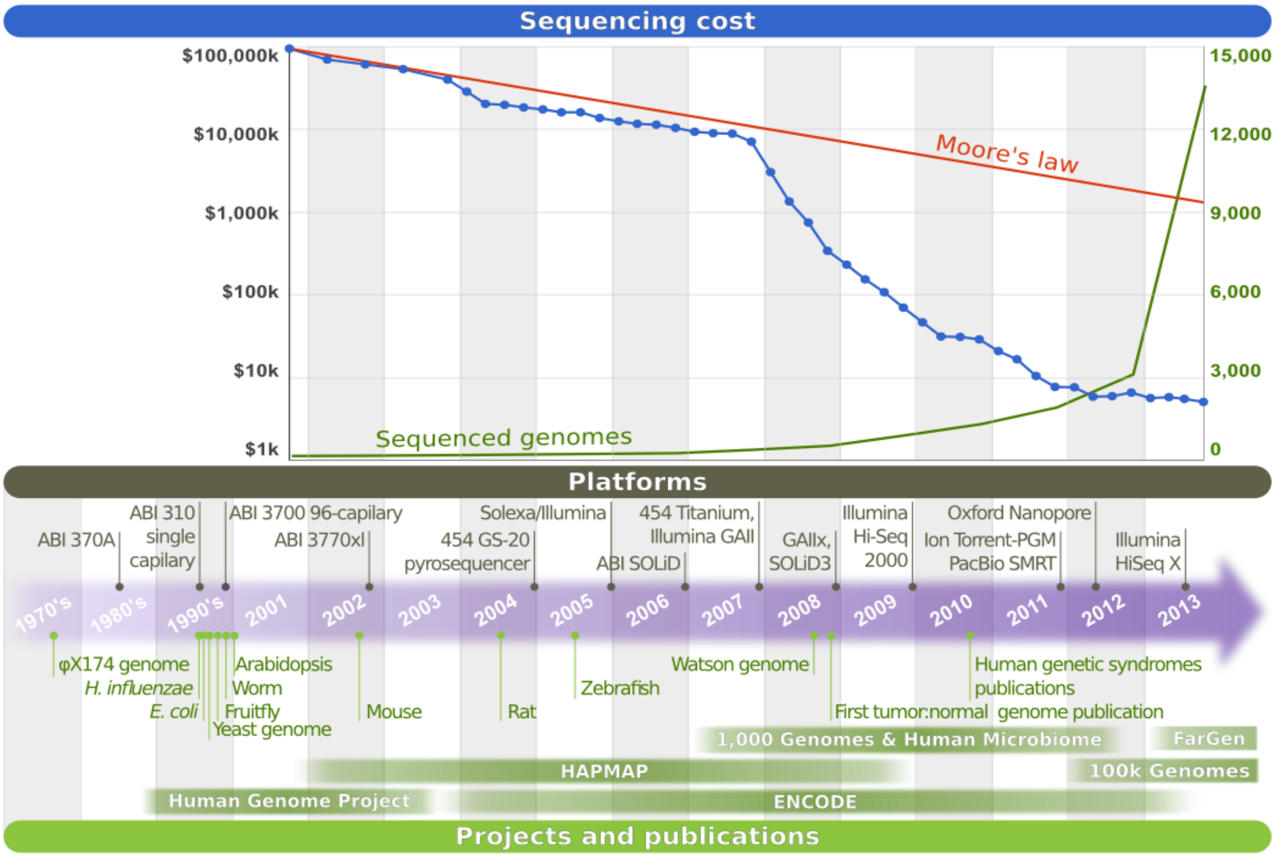
2.1 FIRST GENERATION SEQUENCERS

In 1977, Sanger developed the first rapid DNA sequencing method based on the selective incorporation of chain-terminating dideoxynucleotides by DNA (Sanger et al., 1977a). Soon after, the very first DNA genome, that of the bacteriophage ϕX174 was sequenced (Sanger et al., 1977b). In the 1980s' researchers realized that genome sequencing of larger, non-viral genomes would require automatization. The research community started to work together with the private sector on the development of automated sequencing machines. ABI 370A, the first automated sequencer, was released in 1986 (**Fig. 2.1**). From then on, the field of genomics started to evolve rapidly. Large genome sequencing trials of *Mycoplasma capricolum, Escherichia coli*, *Caenorhabditis elegans*, and *S. cerevisiae* started. In 1995, the first free-living organism, the pathogenic bacterium *Haemophilus influenzae* was sequenced (Fleischmann et al., 1995), followed by the first eukaryote, *S. cerevisiae*, one year later (Goffeau et al.,1996) (**Fig. 2.1**).

The Human Genome Project (HGP) was devised during the 1980s' and formally the 15-years project started in 1990 with a budget of $3 billion (De Lisi, 2008). In addition, Celera Corporation started in parallel, a privately funded project in 1998. The human genome draft, covering 83% of the genome, was published in 2001 by both, public and private initiative (Lander et al., 2001; Venter et al., 2001). The complete human genome was released in 2003, two years earlier than originally planned (International Human Genome Sequencing Consortium, 2004).

For the past 30 years, Sanger-sequencing technology has remained the most commonly used DNA sequencing technique until date. This method reached its zenith in the development of single tube chemistry with fluorescently marked termination bases, heat stable polymerases, and automated capillary electrophoresis; but then reached a plateau in terms of technical development.

Figure 2.1. Sequencing evolution. Changes in the sequencing cost of the human genome (blue) and in the number of sequenced genomes (green) over the past decade are given at the top. Timeline representing milestone sequencing platforms and timing of major sequencing projects are given in the middle and at the bottom, respectively.



2.2 NEXT GENERATION SEQUENCERS

In the first years after the HGP completion, the evolution of sequencing cost followed Moore's law, dropping by half every two years (**Fig 2.1**).

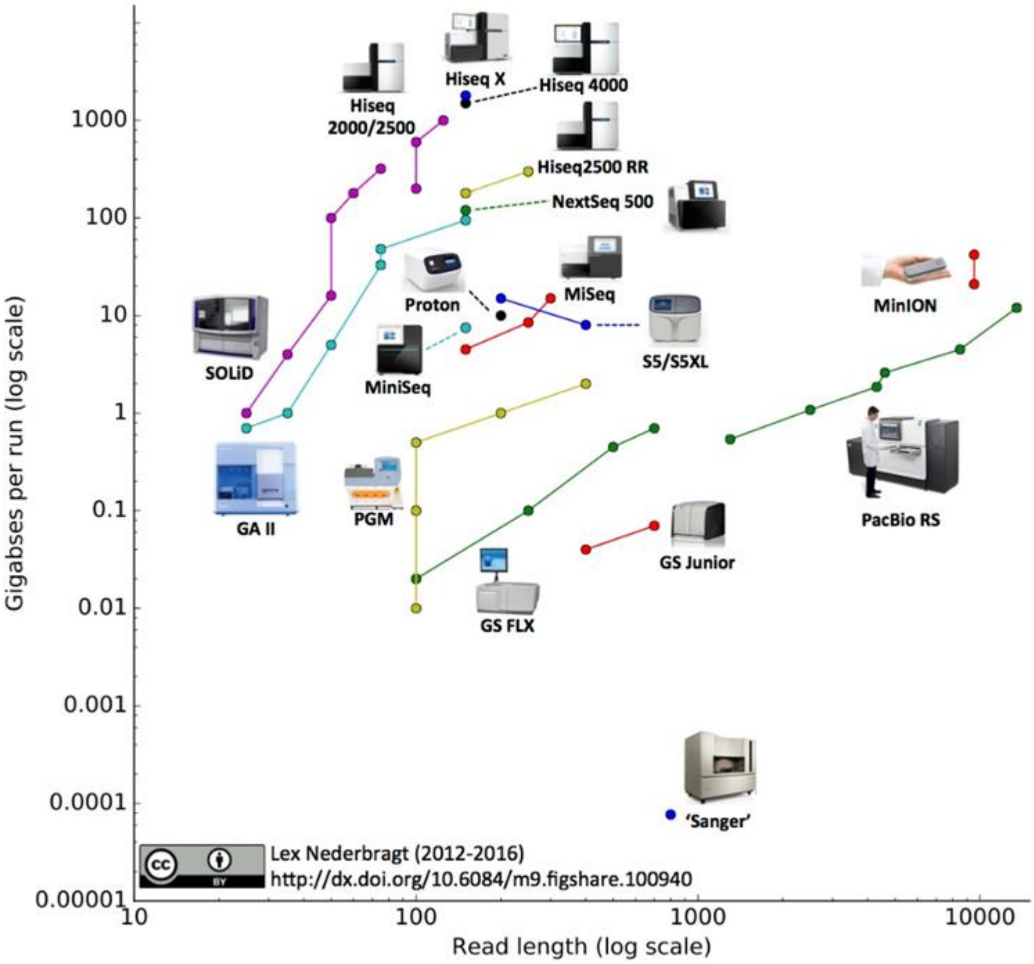
Nevertheless, the 3 billion bases long entire euchromatic human genome was sequenced in 13 years at a cost of approximately $3 billion. This enormous cost of sequencing per run and low throughput of the Sanger sequencing technique are limiting its use for large scale whole genome sequencing projects (Ronaghi, 2001).

With the ultimate goal of deciphering complete genes and entire genome, the requirement of high-throughput sequencing grew by an unpredicted extent. Novel approaches evolved to provide sequence data around a hundred times faster and cheaper than the dominant sequencing data provider Sanger. Second generation sequencers brought a drastic drop in sequencing prices. These approaches fall under the broad definition of "next generation sequencing" (NGS).

Second generation NGS technologies were born at the dawn of twenty-first century in the year 2000 with the foundation of 454 Life Sciences (originally 454 Corporation) by Jonathan Rothberg. At the same time, other sequencing platforms, such as Solexa (Illumina) and SOLid (ABI/Life Technologies), were also introduced into the market (Hert et al., 2008). Despite these platforms differ configurationally, they share many common features and are based on similar work flows for the production and analysis of sequencing libraries (Shendure and Ji, 2008). First, the sample nucleic acids have to be sheared in order to reach a size compatible with sequencing (typically <500 bp). Second, DNA adapters containing unique sequences are attached at both ends of the sheared DNA molecules. These adapters subsequently allow the DNA fragments to be singled out, either on beads or on a slide ("flowcell"), enabling them to be sequenced in parallel. A comparison between throughput metrics for the different platforms is showed in **Figure 2.2**.

Sanger sequencing is regarded as the foundation for the genomic research, but next generation sequencing techniques has dramatically improved the breadth and depth of our knowledge and understating of the genome function and dynamics, as more genomes are sequenced, analysed and compared (Mardis, 2007). NGS technologies rely heavily on automation and high-throughput technologies that are capable of processing millions of sequence reads in parallel fashion in very short time duration without significant loss of accuracy (Metzeker, 2010). This massively parallel throughput may require only few (one or two) instruments runs to accomplish sequencing experiment (Mardis, 2007, Morozova and Marra, 2008; Riesenfeld et al., 2004).

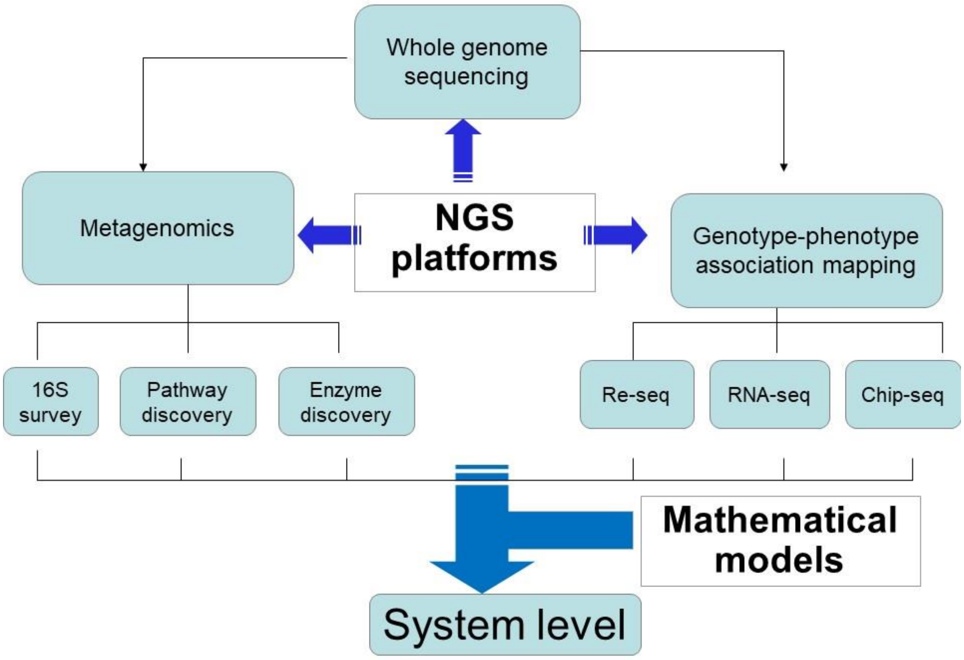
Figure 2.2. Summarises the developments in next generation sequencing. Throughput metrics for the different platforms since their first instrument version came out: raw bases versus read length.



DNA sequencing, besides decoding DNA sequence, enables the analysis of RNA that is reverse-transcribed to DNA or any other type of molecules or phenomenon that can be bound to DNA. Already in the early Sanger sequencing era, the complementary DNA (cDNA) libraries, started to be sequenced to assist genome annotation (Adams et al., 1991). But the burst of new DNA sequencing applications started with the second generation sequencing era: high-throughput, high sensitivity and high dynamic range of new sequencers allowed for the development of new applications.

The reduction in cost and time for generating DNA sequence data has resulted in a range of new successful applications, such as whole genome sequencing, resequencing, RNA-sequencing to delineate the cellular transcriptome, ChIP-sequencing to identify binding sites of DNA-associated proteins, as well as in the study of ancient DNA (Poinar et al., 2006). Moreover, the advent of NGS approaches has greatly increased the ability of researchers to profile food microbial metagenomics and to investigate the molecular mechanisms of interesting functionalities in food ecosystem. An overview of the main applications of NGS technologies in the field of food microbiology is shown in **Figure 2.3.**

Figure 2.3 Overview of the main applications of "next generation sequencing" (NGS) to address food microbiology questions.



Food microbiology deals with the study of micoorganisms that have both beneficial and deleterious effects on the quality and safety of food products. The fast and low-cost NGS approaches have revolutionized microbial taxonomy and classification, characterization of food pathogens and spoilage microorganisms and screening of potent starter cultures for food processing (Coenye et al., 2005). Application of NGS to microbial genomics in relation to food biotechnology is not just limited to predict the prevalence of microorganisms in food samples and to assign phylogeny, but also to provide in-depth molecular basis of how microorganisms respond to different food-associated conditions which, in turn, will offer tremendous opportunities to prevent and control undesirable growth and survival of microorganisms in food products. Unitedly, NGS has facilitated the development of new genome-assisted approaches for correlating genotype and phenotype to better understand microbial behaviour in food ecosystems.

2.2.1 What sets NGS apart from conventional sequencing technology?

Even if Sanger sequencing is considered the foundation of genomics, research in the field of microbial genomics, transcriptomics, and metagenomics was greatly limited by unavailability of an efficient technology for high-throughput screening and sequencing large sets of genome data. The limitations of Sanger sequencing technology drove the research for more scalable and lower-cost sequencing solutions.

Second generation sequencers as vector-independent methods for library synthesis offered several sequencing advantages compared to capillary sequencers (**Table 2.1**), mainly:

i) Only a small amount of input DNA is needed;

ii) Sample preparation is faster, less laborious and cheaper;

iii) It is free from cloning associated biases (Liu et al, 2009; Mardis, 2007);

iv) Sequencing biases are reduced due to vector replication in bacterial hosts (Farris and Olson 2007; Mardis 2007);

v) Limited risk of library sequencing underrepresentation owing to unintended expression; of toxic products from the cloned fragment in the bacterial vector (Kimelman et al., 2012);

vi) Sequencing occurs more rapidly since complementary strand synthesis and base detection are simultaneous processes.

On the other hand, PCR amplification is sensitive to GC content variation and may introduce DNA polymerase-related errors (Quail et al., 2012). Despite this, the complementary strand synthesis and base detection is a simultaneous process, and occurs in a nucleotide-by-nucleotide stepwise fashion. In contrast, in capillary sequencing DNA synthesis is performed first and the base detection from electrophoresis gel happens later. Finally, second generation sequencers allow the sequencing of both ends of DNA fragments. As a result, so-called pair-end or mate-pair reads with known distance between the pair can be generated. These libraries, beside improving the genome assembly by decreasing the high error rates that prevent assembly software to resolve large structural rearrangements and to disambiguate repeat regions, allow detection of chromosomal rearrangements. Paired ends are obtained from the ends of random and small DNA fragments and the resulting data allow the scaffolding of contigs in the absence of contiguous coverage of intervening sequences (Bentley, 2006). In mate-pair sequencing, random DNA fragments are circularized, thereby combining previously distant ends. Typically, mate-paired methods generates a longer insert size compared to paired-end (150-500 bp on average), with insert sizes measuring between 2 and 20 Kb.

Considering that amplification of templates is necessary in NGS, PCR amplification steps are often associated with PCR bias and the subsequent possibility of introducing base sequence errors or favouring certain sequences over others. These potential drawbacks can be avoided only if a single DNA molecule is used directly for sequencing without undergoing PCR amplification steps. In this direction, new competitive and revolutionary technologies recently appeared on the market. They are so-called 3rd generation sequencing technologies and hold great promise in terms of offering rapid and cost-effective sequencing of gene/genome from a single DNA molecule (Schadt et al., 2010). They includes sequencing platforms available commercially from Helicos Biosciences (HeliScope) (Braslavsky et al., 2003), PacBio (Ansorge, 2009) and the newest one ONT (**Table 2.1**).

Table 2.1. Sequencing platforms comparison. Characteristics of first (Sanger Sequencing), second (454, Illumina, SOLiD and HeliScope) and third (PacBio and MinION) generation sequencers are provided. Abbreviations: na, not applicable; stPCR, standard PCR; emPCR, emulsion PCR.

Platfoms	Chemistry	Starting DNA	Bases/template	Read Length	PCR Amplification	Read Accuracy	Reads/run	TH/run	Run time	Advantage	DisAdv	Applications
Sanger Sequencing	Asynchronous with base-specific terminator	0.5-1mg	700 (Margulies et al., 2005)		stPCR	99.40-99.90% (Margulies et al., 2005)	Few 1,000 bp	1Mb	2hrs	Length	PCR biases, low degree of parallelism, high cost of sequencing	Gene/genome sequencing
Roche/ 454	Sequencing-by-synthesis (Pyrosequencing)	1µg for shotgun library and 5µg for pair-end	~400	~400	emPCR	99.96% (Margulies et al., 2005)	1,000,000	0.4-0.6Gb	7-10hrs	High read length	Asynchronous synthesis, homopolymer run, base insertion and deletion errors, emulsion PCR is cumbersome and technically challenging PCR biases	*De novo* genome sequencing, RNA-seq, Resequencing/targeted re-sequencing
Illumina	Polymerase-based sequencing-by-synthesis	<1µg for single or pair-end	~75(35-100)	36	Bridge Amplification	99.50%	40,000,000	3-6Gb	3-4days	High read length	PCR biases	*De novo* genome sequencing, RNA-seq, Resequencing/ targeted re-sequencing, metagenomics, ChIP
SOLiD v4	Ligation-based sequencing	<2µg for shotgun library and 5-20µg for pair-end	35-50	35	emPCR	99.99%	85, 000,000	10-20Gb	7days	Two-base encoding-a error correcting scheme	Emulsion PCR is cumbersome and technically challenging PCR biases	Targeted re-sequencing, transcript counting, mutation detection, ChIP, RNASeq *etc*.
HeliScope	Polymerase (Asynchronous extension)	<2µg, single end only	35	<1000	Single Molecule	97-99.80%	1,000,000,000	28Gb	8days	Sequencing from single DNA molecule, no PCR, high raw sequencing accuracy, low raw sequencing error	Asynchronous synthesis, homopolymer run, high instrument cost, short read lengths	Resequencing, transcript counting, ChIP, RNA-Seq
PacBio RS II	Phospho-linked Fluorescent Nucleotides	~1.5µg (Ideally 2-3µg)	800-1,000	1000-1200	SMRT	99.99%	100,000,000	100Gb/Hr	8hrs	SMRT Dephaing not an issue, No PCR, long read	High instrument cost Low number of sequence read per run	*De novo* genome sequencing, RNA-seq, resequencing/targeted re-sequencing, metagenomics, ChIP
MinION (ONT)	Nanopore sequencing	From 10pg (ideally 4-5 µg)	na	Up to 1Mb (1,000,000 bp)	na	70-90%	100,000,000		1min-48hrs	Longest read length; portable; affordable	Low read accuracy; high error rate in base calling	*De novo* genome sequencing, RNA-seq,

2.2.2 Practical application of NGS to food microbiology

Role of NGS in testing food safety and authentication

One of the principal challenge for the food industry is to produce safe foods with the desired functionalities using minimal processing technologies. Whole genome characterization of undesirable microorganisms in foodstuff is the first step towards prevention of food spoilage.

Whole genome sequencing initiatives from around the world have resulted in the complete sequencing of a large number of food-borne pathogens, including most of the bacterial pathogens associated with significant public health threats. Therefore, NGS could have important applications in reducing the risks of food-borne diseases due to the huge improvements in the rate at which the whole genome of food microorganisms from different species and from strains belonging to the same species can be generated.

The ability of NGS to authenticate food products offers a means to monitor and identify products for consumer protection and regulatory compliance. There are numerous molecular techniques and physical technique available to identify adulterants in processed food such as single strand conformational polymorphism (fish species in fresh processed fish) (Hold et al., 2001), small sequence length polymorphism (long grain rice in basmati rice) (Bligh, 2000), restriction fragment length polymorphism (common wheat in durum wheat pasta) (Bryan et al., 1998), DNA methylation (neuronal tissues in muscle tissue containing processed meat) (Woolfe and Primrose 2004), ELISA (meat, fish, milk and juices) (Asensio et al., 2008), RAPD and RFLP (different potatoes varieties).

However, the disadvantage of NGS is that it is unsuitable for quantitative detection of fraudulent substitutions in some food samples such as processed meat where it is unable to differentiate between neuronal tissue and muscular tissue in processed meat samples because both tissues contains identical DNA having identical gene sequence (Woolfe and Primrose, 2004).

With the increasingly widespread use of NGS, it is reasonable that data generated by NGS and other "omics" techniques (*i.e.*, transcriptomic and metabolomics) will be integrated by mathematical algorithms into a system model at the species and "meta"-species levels, so that environmental and processing parameters will be predictive of species composition in food (**Fig. 2.3**).

NGS-assisted starter optimization

Selection and dominance of a starter culture on indigenous population in fermented food can speed up fermentation significantly and increase sensorial properties. While producing food products by fermentation, the quality of raw material, selection of bacterial combinations to be used as starter cultures and controlled fermentation conditions should be properly considered during fermentation process (Oguntoyinbo et al., 2011). The move toward processed foods, fermented foods and other booming endeavours such as discovery of new food and energy sources are driving the use of high throughput sequencing technologies. It is well-established that population structure and dynamics studies of microorganisms during fermentation is rather difficult to perform using traditional culture-dependent methods due to the problems associated with processing large numbers of samples (Humblot and Guyot, 2009). At the starting of the fermentation process, considerable microbial diversity may be expected in culture, whereas during course of time only few representative species remain in culture. Pyrosequencing has enabled to present an overall community structure and population dynamics of microorganisms in a fermented food (Pearl millet slurries) (Humblot and Guyot, 2009). Recently the outbreaks in the peanut butter and peanut paste products associated with *Salmonela* caused massive illness across United States. Traditional heat-killing steps have shown to be ineffective in the processing of lipid-rich matrices including peanut. In this regard upon inclusion of high-pressure processing steps in peanut products manufacturing, the population of *Salmonela* was dramatically reduced as confirmed by pyrosequencing.

Another field in which NGS-assisted starter optimization is maturing is the study of wine yeast. An annotated genome sequence for *S. cerevisiae* is available, which provides a framework for genome-scale metabolic network reconstruction (Borneman et al., 2007).

Studies on fermented foods using next-generation sequencing techniques are highly valued in the field of food biotechnology and had offered a better understanding of food ecosystems, in which fermented food processing was greatly influenced by diverse microbial communities (Roh et al., 2010).

Future prospects and conclusion

The rapid developments in next-generation sequencing technologies have allowed us to obtain very high-definition genome snapshots, and these will, undoubtedly, significantly increase our insights in transcriptional and post-transcriptional events in microorganisms. Application of next-generation sequencing technique in food biotechnology has dramatically changed the way we perform detection and subtyping of microorganisms, characterization food pathogens and spoilage microorganisms and screening potent starter cultures for food processing. High throughput sequencing can play a key role in whole genome-assisted optimizing of food starter cultures. It is reasonable that future challenges will be aimed at achieving connectivity between data generated by NGS and other “omics” techniques in the context of time and space. This integration will provide comprehensive genetic maps of important food traits, as well as predictive models of the contribution of individual microorganisms in the development of food quality and safety.

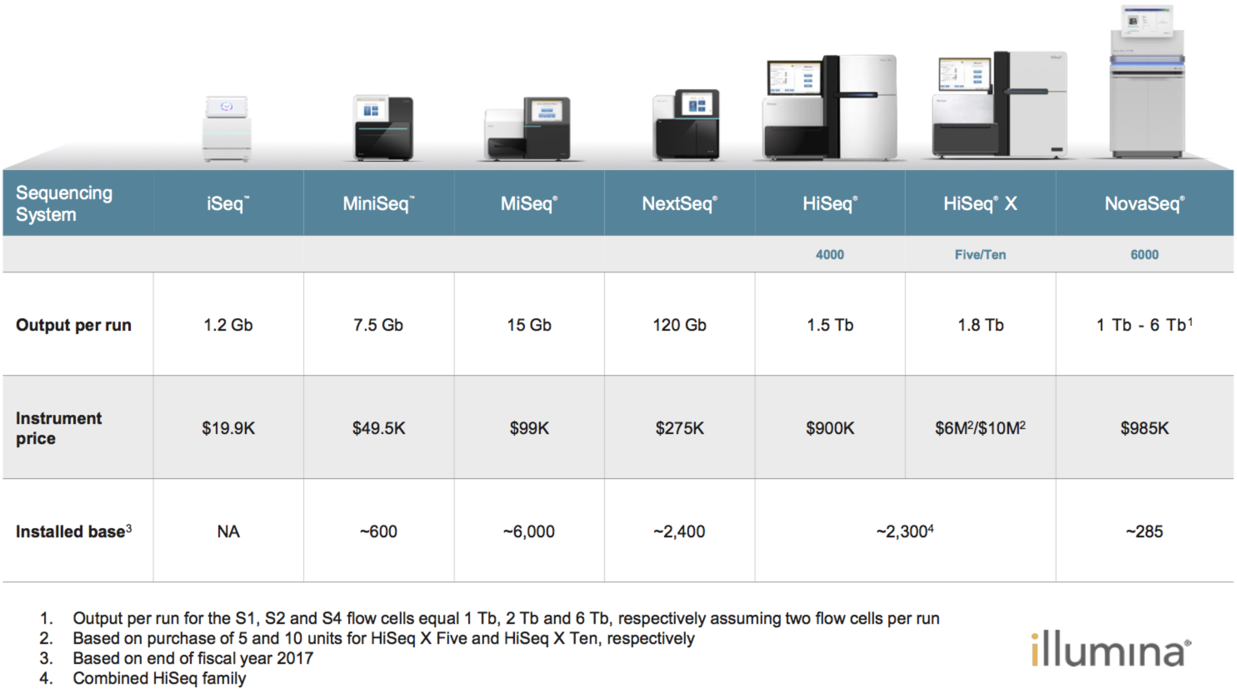
2.2.3 MiSeq Illumina platform

Next-generation sequencing (NGS) technology has revolutionized genomic and genetic research. The pace of change in this area is rapid with three major new sequencing platforms having been released in 2011: Ion Torrent’s PGM, Pacific Biosciences’ RS and the Illumina MiSeq. All the platforms have library preparation protocols that involve fragmenting genomic DNA and attaching specific adapter sequences. Illumina currently produces a suite of sequencers (MiSeq, NextSeq 500, and the HiSeq series) optimized for a variety of throughputs and turnaround times. The Illumina sequencing technology is based on what is called sequencing by synthesis (SBS) technology. Essentially, the genome is broken up into smaller pieces, which are then attached to a surface inside of the sequencer's flow cell. The MiSeq and HiSeqs are the most established platforms. The MiSeq is designed as a fast, personal benchtop sequencer, with run times as low as 4 hr and outputs intended for targeted sequencing and sequencing of small genomes. The HiSeq, on the other hand, is engineered for high-throughput applications, yielding current outputs of 1 Tb in 6 days. The Illumina MiSeq desktop sequencer allows high flexibility and easy access for smaller projects like targeted re-sequencing, 16S metagenomics, small genome sequencing, with a very short turn-around time. Key advantages of the MiSeq include long reads; highest flexibility and fast run times; low run cost/base; best-established technology and broad range of biological applications.

With read lengths of currently up to 2 × 300 bp, high throughput and low sequencing costs, Illumina's MiSeq is becoming one of the most utilized sequencing platforms worldwide. The platform is manageable and affordable even for smaller labs. This enables quick turnaround on a broad range of applications such as targeted gene sequencing, metagenomics, small genome sequencing and clinical molecular diagnostics.

In 2014, Illumina introduced the NextSeq, which is similar to the MiSeq and designed as a fast benchtop sequencer for individual labs. This system also employs a novel two-channel sequencing strategy. In this approach, cytosine is labeled red, thymine is labeled green, adenine is effectively yellow (labeled with a mixture of red and green), and guanine is unlabeled. In contrast to the four-channel strategy used in the MiSeq and HiSeq platforms, two-channel sequencing requires only two images for nucleotide detection, reducing data processing times and increasing throughput. Despite the reduced complexity, the overall error rates (<1%) are similar to the more established HiSeq machines (Reuter et al., 2015) (**Fig. 2.4**).

Figure 2.4. Overview of commercially available Illumina platforms.



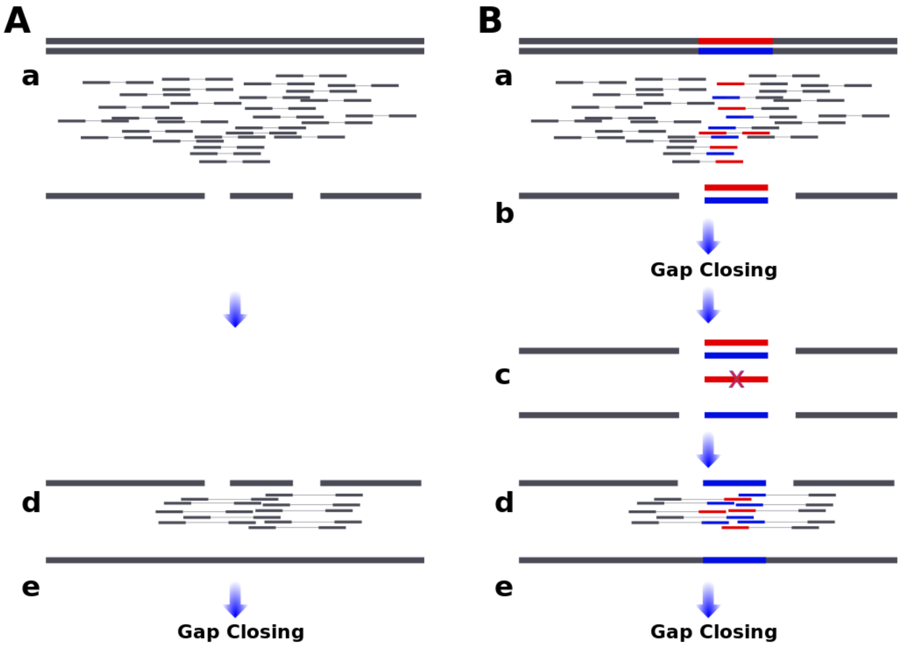
Sequencing System	iSeq™	MiniSeq™	MiSeq®	NextSeq®	HiSeq®	HiSeq® X	NovaSeq®
					4000	Five/Ten	6000
Output per run	1.2 Gb	7.5 Gb	15 Gb	120 Gb	1.5 Tb	1.8 Tb	1 Tb - 6 Tb[1]
Instrument price	$19.9K	$49.5K	$99K	$275K	$900K	$6M[2]/$10M[2]	$985K
Installed base[3]	NA	~600	~6,000	~2,400	~2,300[4]		~285

1. Output per run for the S1, S2 and S4 flow cells equal 1 Tb, 2 Tb and 6 Tb, respectively assuming two flow cells per run
2. Based on purchase of 5 and 10 units for HiSeq X Five and HiSeq X Ten, respectively
3. Based on end of fiscal year 2017
4. Combined HiSeq family

2.2.4 NGS data analysis

In the Sanger-sequencing era, the sequence generation was a bottleneck. Nowadays the limiting step moved from sequencing itself to the subsequent data analysis. This is related to both, high throughput and data characteristics, as short sequences are difficult to assemble or align unambiguously, and downstream analyses are needed before the data can be interpreted. Genome sequencing provides gigabytes of data, but the sequencing results need to be analysed *de novo* (*de novo* genome assembly) or compared to a reference genome (re-sequencing). In the first scenario, the reference set of chromosomes is generated. Genome assembly from NGS reads is challenging and multiple solutions have been developed. The most successful programs, like SOAPdenovo (Luo et al., 2012), Velvet (Zerbino et al., 2009), SPAdes (Bankevich et al., 2012) or ABySS (Simpson et al., 2009) are based on de Bruijn graphs. A directed graph representing genome fragments (contigs) is created connecting the reads based on exact sequence overlaps of a given length (K-mer). Subsequently, contigs are further joined into scaffolds using pair-end or mate-pairs libraries (scaffolding). Finally, the gaps may be closed using the same paired-end and mate pairs libraries (**Fig. 2.5**, **panel A**).

Figure 2.5. Genome assembly from short reads. Standard (**A**) and heterozygous (**B**) genome assembly pipelines are compared. Heterozygous regions in diploid chromosome are marked in red and blue. Heterozygous genome assembly pipeline consists of five steps. a) Standard de novo assembly is performed and b) optionally gaps are closed. Obtained assembly is larger than expected and fragmented because two alternative contigs are recovered from heterozygous region (blue and red), while single contig is recovered from homozygous regions (grey). Further scaffolding of such assembly is impossible, as homozygous contigs can be joined to any of heterozygous contigs (blue and red). c) To overcome this, redundant contigs from heterozygous regions are removed (here the red contig) and d) homogenised assembly is further scaffolded. e) Finally, gaps are closed.



The predominantly used assembling programs described above are not suitable for highly heterozygous diploid genomes since the increased complexity of the *de Bruijn* graph (DBG) structure makes assembly a substantial challenge. Different studies showed that in most of the DBG-based assemblers the scaffold NG50 values are dramatically reduced when the heterozygosity is >0.5%. DBG-based assembling programs are not able to overcome one of the primary obstacle of *de novo* hybrid genomes assembly, *i.e* the existence of heterozygosity between diploid chromosomes (You et al., 2013; Zheng et al*.,* 2013). For diploid samples, during the building up of the DBG, different k-mers derived from the heterozygous regions corresponding to each homologous chromosome are created and used in the graph structures. In correspondence of the borders between homozygous and heterozygous regions, junctions are created in the graph and subsequently bubble structures are generated. Most of the existing DBG-based assemblers try to solve this problem by simplifying these structures, failing in both efficiently eliminating errors and resolving repeats.

Therefore, new methods are required to address the increasing demand for sequencing of non-model microorganism, such as non-conventional yeasts as *Z. rouxii*. Moreover, the genome reconstruction process in allodiploid strains is even more challenging due to the peculiar heterozygosity and chimeric genetic organization. An additional, but not less important, underlying difficulty of allodiploid genome reconstruction is the lack of an appropriate diploid reference.

To overcome these difficulties alternative assembling strategies have been recently proposed. For example, Platanus is a *de novo* assembler released in 2014 and represents a novel and efficient approach for the assembly of Gb-sized highly heterozygous genomes since it can to reconstruct genomic sequences of highly heterozygous diploids from massively parallel shotgun sequencing data (Kajitani et al., 2014).

Platanus works assembling short reads into contigs by constructing DBG with automatically optimized k-mer sizes followed by the scaffolding of contigs based on paired-end information. During both the contig assembly step and the scaffolding step, the complicated graph structures that result from the heterozygosity are simplified and heterozygous regions containing structural variations, repeats, and/or low-coverage sites, are captured (Kajitani et al., 2014). During the contig assembly and the scaffolding steps, Platanus attempts to assemble each haplotype sequence separately into a single contig/scaffold, resulting in a mosaic genome where the heterozygosity is preserved at the price of contig fragmentation.

Other assemblers, even not specifically designed for hybrid genomes, such as MaSuRCA, may retain the heterozygosity level by compressing overlapping reads into super-reads of 3–13 kb using an Eulerian *de Bruijn* graph and then assembling these super-reads in contig using an overlap-layout-consensus–based algorithmn, avoiding, in this way, a sharp decrease in its scaffold NG50 (Zimin et al., 2013). However, in assembling real data from various organisms, Platanus often returned much larger scaffold NG50 values than those from MaSuRCA, possibly due to the presence of more complex variants in the actual data set. Moreover, MaSuRCA required a higher execution time and memory resources for assembly compared to Platanus owing to the super-reads creation step. With the advent of 3rd generation sequencing technologies, the reconstruction of hybrid genomes took advantage from combining short- and long-reads data (hybrid assembly) to improve the scaffolding of fragmented contigs, thought new assembly pipelines have to be developed.

2.3 3RD GENERATION SEQUENCING TECHNOLOGIES: NANOPORE AND SINGLE-MOLECULE REAL TIME SEQUENCING

The development of novel genome sequencing methods has been the major driving force behind the rapid advancements in genomics of the last decades. Many more genomes were sequenced thanks to the advent of second generation sequencing, which provided researchers with the required throughput and cost efficiency. Recent years saw the dawn of what can be considered a third generation, that allows reading of single DNA molecules in long consecutive stretches without need of amplification (de Lannoy et al., 2017). Currently, two are the dominating methods of this new generation: nanopore sequencing and single molecule real time (SMRT), championed by Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), respectively.

2.3.1 MinION platform and physical basis of DNA sequencing using nanopores

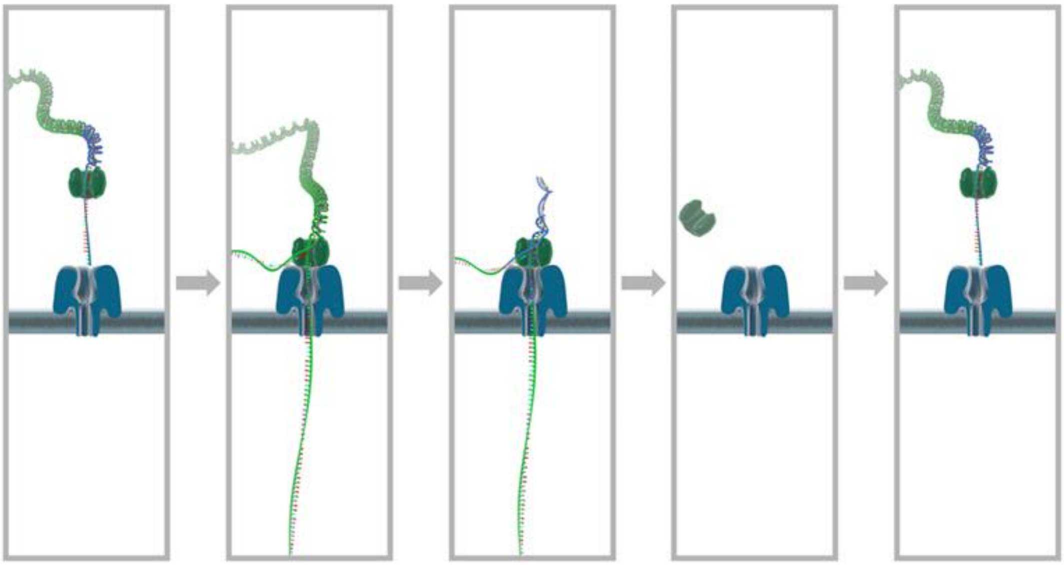
Nanopore sequencing is a promising new venue in biology research. Since the introduction of the first commercially available Nanopore sequencing device, ONT's MinION in 2014, the field of Nanopore sequencing has advanced rapidly; both new applications and improvements to existing ones are published on a regular basis. The advantages of the MinION over other sequencing devices are numerous: inexpensive, small, capable of producing long reads and free from the need of nucleotide labelling or amplification. Considering all these characteristics, it is conceivable that the MinION will make cost-effective, fast and portable *de novo* whole genome sequencing of even complex genomes possible in the future.

Currently, the most prominent obstacles for *de novo* sequencing using the MinION are the lower signal-to-noise ratio, stochasticity introduced by its biological components, and the resulting high error rate of the reads. Improving basecalling accuracy would improve assembly quality and, more important, allow more computationally efficient assembly.

Nanopore sequencing is a novel sequencing approach that exploits an electrical potential applied across an insulating membrane in which a single small pore is inserted.

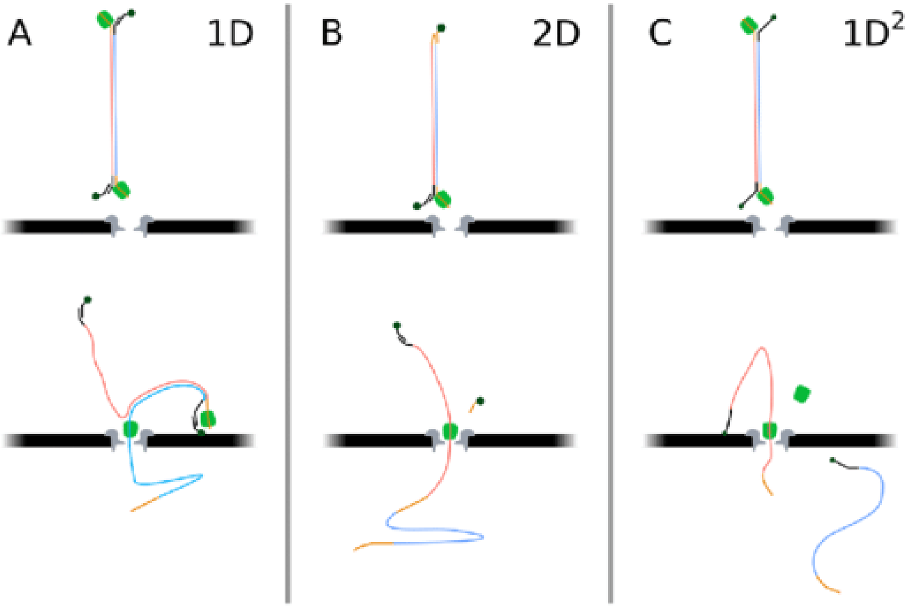
A DNA strand is pulled through the pore and the sequence is inferred from the characteristic way in which the passing base combinations influence the current (de Lannoy et al., 2017). More in detailed, a microscopic opening wide enough to allow single-stranded DNA to pass the nanopore is introduced in an insulating membrane between two compartments filled with saline solution and an electric potential is applied across it. DNA strands are then added to one compartment and allowed to diffuse toward the nanopore, where they are captured by the electric field and threaded through the pore. While a strand is passed through, the characteristic way in which the bases influence the electric current through the nanopore is measured. These measurements can then be decoded to retrieve the sequence of the DNA strand (**Fig. 2.6**). During MinION sequencing run, the potential over the membrane is kept stable, while the electrical current is sampled at a frequency in the kHz range. This signal serves as the basis for basecalling since it is characteristic for the bases moving towards through the pore.

Figure 2.6. Sequencing of a DNA strand using nanopores. From left to right, double-stranded DNA with attached motor protein attaches to a pore protein in an insulating membrane. The applied potential pulls one strand through the pore, while the motor protein unzips the DNA in a step-wise fashion. After the DNA has been unzipped completely and one strand has passed through, the complex detaches from the pore entrance and the pore is ready to receive another strand. Image from Oxford Nanopore Technologies Ltd.



There are three categories of DNA reading chemistries for the MinION (**Fig. 2.7**). The first one is called 1D chemistry: only the first strand of a dsDNA stretch (by definition, the template strand) is threaded through the pore and read. In the second method, referred to as 2D-sequencing, the 3' end of the template strand and the 5' end of its complement are covalently connected using a hairpin adapter, allowing the complement strand to be pulled in automatically after the template strand. This approach was replaced by the 1D^2-sequencing in May of 2017. This 2D successor chemistry also allows reading both strands, but rather than attaching the two, the complement strand is tethered to the membrane while the template strand is sequenced. This method, that combines reads of both strands, has been proved to significantly increase sequencing accuracy.

Figure 2.7. Schematization of the three categories of DNA sequencing methods for the MinION. (**A**) In the 1D chemistry, only the template strand (blue) is threaded by its motor protein (green) and read. The complement strand (red) is discarded at the cis side of the pore. The tethers (dark-green) allow for selection of properly ligated complexes during sample preparation and attach to the membrane to increase the availability of strands near pores during sequencing. (**B**) In the abandoned 2D chemistry, a hairpin covalently connected template and complement strand, thus allowing sequencing of the complement strand immediately after the template strand. (**C**) 1D^2 chemistry, the successor of 2D, is the currently implemented method to sequence both strands: the complement strand is tethered to the membrane while the template is sequenced. After the template strand is threaded through, the complement strand is drawn in and the tether is pulled loose. Figure based on Jain et al., 2015.



2.3.2 Nanopore data analysis

Once nanopore reads have been basecalled (*i.e* the current signal has been translated into the underlying DNA sequence), they may be *de novo* assembled using assembly tools that can make use of the long read length wile mitigating the error-prone nature of the reads. This is often followed by the last step of "polishing", in which a better consensus between the assembly and the raw reads is sought (de Lannoy et al., 2018). One of the technical advantages of ONT data is the read length, which offers great prospects for genome assembly. As reviewed by Henson et al. (2012), generally, assemblers are based on several different types of algorithms, such as greedy, overlap-layout-consensus (OLC), *de Bruijn* graph (DBG), and string graph.

The NGS platforms such as Illumina's HiSeq and MiSeq have played a dominant role in genomic research and applications and it is foreseeable that short read data will continue to be a very important part of data sources for years to come. With the emergence of third generation sequencing (3GS), some assemblers start to include long reads into the pipeline to improve assemblies primarily generated from short reads (Lu and Giordano, 2016). Some examples of hybrid assemblers are ALLPATHS-LG (Gnerrea et al., 2011), SPAdes (Bankevich et al., 2012) and DBG2OLC (Ye et al., 2012). The latter one employs an algorithm that starts with linear unambiguous regions of a DBG, and ends up with linear unambiguous regions in an overlap graph (used in the OLC framework). Due to this property, it was called DBG2OLC. Differently from previous approaches, this assembler uses the NGS assembly to lower the computational burden of aligning 3GS sequences rather than just polishing 3GS data. In this way, the sequencing gaps in NGS data may be covered by the data from 3GS and *vice versa*. The utilization of NGS data also lowers the required sequencing depth of 3GS, saving about half of the sequencing cost. In conclusion, we can say that DBG2OLC is an efficient hybrid assembler that simultaneously utilizes NGS and 3GS data to reduce the high error rate of long reads of large genomes, to detect and correct structural errors generated by chimeric long reads and finally to assemble and polish 3GS reads.

2.3.3 Fungal genomics' history until Nanopore sequencing

The last decade has witnessed a revolution in the genomics of the fungal kingdom that has been driven by the evolution of genome sequencing technology. The genomes of almost twenty different hemiascomycete yeasts have been sequenced (Wolfe, 2004) since the genome of *S. cerevisiae* was published a decade ago (Goffeau et al., 1996). This milestone revolutionized work in yeast and enabled the first global studies of eukaryotic gene function and expression. The number of available genomes and the remarkable conservation of gene order means that yeast comparative genomics can be used to address hypotheses about genome evolution that would be impossible in any other system. Over 40 complete fungal genomes have been publicly released, generating an explosion in fungal genomics that has greatly expanded our knowledge of the genetic and physiological diversity of these organisms, providing a great opportunity to study the biology and evolution of this medically, industrially, and environmentally important kingdom. In addition, fungi also serve as model organisms for all eukaryotes (Galagan et al., 2005). The advent of high throughput and cost-effective NGS sequencing and assembly technologies produced fungal genome sequences with unparalleled accuracy and long-range contiguity at ever-reduced cost. These methods represent an advance over the clone-by-clone approaches used to sequence the first eukaryotic genomes. Despite these advances, a number of challenges remain. Repetitive sequences, that are numerous in fungi and usually associated with telomeres, centromeres and rDNA arrays represent the single biggest difficulty in assembling whole genome shotgun (WGS) sequence data. A special case of repeated sequences is diploid genome, in which the extent of heterozygosity can vary dramatically across chromosomal regions. Highly heterozygous genomes can lead to more fragmented assemblies, or create doubt about the homology of the contigs. Consequently, allelic differences are difficult to distinguish from distinct paralogs. When possible, these complications can be avoided or minimized by sequencing a haploid reference of the species, but in many cases, such as with *C. ablicans* and additional eight *Candida* species (Jones et al., 2004; Braun et al., 2005), sequencing a diploid is inevitable. To resolve the assembly of repeats, reads need to be long enough to also include the unique sequences flanking the repeats. It can therefore be a good idea to use long-read technologies that go under the name of 3GS, if you know that you are working with a genome with a high content in repeats. In particular, Nanopore sequencing revolutionize fungal genomics, since the long reads generated allow to circumvent difficulties of assembly linked to repetitive and homopolymeric stretches that are abundant especially in highly heterozygous diploid genomes. In fact, one of the major advantages of the ONT is the possibility of sequencing very long DNA fragments, which span repetitive regions. Despite the availability and the constant improvement of these new techniques, progress in sequencing fungal genomes is currently quite slow. An overview of the complete or near-complete fungi genomes sequenced and assembled until now with the ONT MinION sequencing platform is presented in **Table 2.2**.

Table 2.2. List of available Saccharomycetes genomes and transcriptome sequenced with ONT MinION platform.

Species/strain	Data type	Scope	Accession n°	Sequencing center(s)
S. eubayanus CBS12357T	Transcriptome or Gene expression	Multiisolate	PRJNA481526	Delft University of Technology
S. cerevisiae IMX1812	Raw sequence reads	Monoisolate	PRJNA478763	Delft University of Technology
S. eubayanus CBS12357	Genome sequencing and assembly	Monoisolate	PRJNA450912	Delft University of Technology
S. cerevisiae S288c and SK1; *S. paradoxus* N44 and CBS432.	Genome sequencing and assembly	Monoisolate	PRJEB19900	University of South Carolina (SC)
C. vartiovaarae *(Torulopsis vartiovaarae)*	Genome sequencing and assembly	Monoisolate	PRJEB19912	ZF-Screens B.V.
S. cerevisiae S288C	SRA Experiments	Monoisolate	PRJEB14774	Glenville State College (GSC)
Dekkera bruxellensis	Genome sequencing and assembly	Monoisolate	PRJEB21262	University of San Diego (USD)
S. cerevisiae CEN.PK 1137D	Raw sequence reads	Monoisolate	PRJEB22218	University of Arkansas for Medical Sciences (UAMS)
S. cerevisiae CEN.PK 1137D	Genome sequencing and assembly	Monoisolate	PRJNA398797	University of Arkansas for Medical Sciences (UAMS)

2.4 OVERVIEW OF EUKARYOTIC GENOME ANNOTATION

Genome assembly is just a set of very long strings composed of As, Cs, Ts, Gs and Ns, that represent the sequence of chromosomes. Thus, genome assembly as such contains no biological knowledge. To gain such knowledge, potential coding and regulatory regions have to be detected in a process known as genome annotation. Prediction of eukaryotic genes is quite challenging from the sequence itself, as only a minor fraction of the genome sequence is coding, and genes have often complicated exon-intron structure. For that, gene prediction is usually assisted with information about known genes from closely related species. In addition, RNA-Seq reads are often used in order to improve detection of transcripts and exon-intron boundaries (Reid et al., 2014). Gene prediction in *Zygosaccharomyces* species is less challenging than in higher eukaryotes, as intergenic regions are rather short (76.4% of the haploid *Z. rouxii* type-strain CBS 732^T genome is coding) and there are very few protein coding genes with introns (0.38% in *Z. rouxii*). The final step is functional annotation. Gene functions are typically assigned from homologs from close relative species. Orthologs are recommended, as orthologous genes are more likely to share the same function than paralogs (Sonnhammer et al., 2014). In addition, functions can be assigned using sequence features such as structural and functional domains.

2.4.1 Step 1: gene prediction

The process of correctly determining the location and structure of the protein coding genes in a genome is called "gene prediction" and it is fairly well understood with many successful algorithms being developed over the past decades. In general, there are three main approaches to predict genes in a genome: intrinsic (or *ab-initio*), extrinsic and the combiners (Del Angel et al., 2018). The great advantage of *ab initio* gene predictors is that, in principle, they need no external evidence to identify a gene or to determine its intron-exon structure. However, these tools have some limitations since they use organism-specific genomic traits to distinguish genes from intergenic regions and to determine intron-exon structures (Yandell and Ence, 2012). If your genome is not very closed to an organism for which precompiled parameters are available, the gene predictor needs to be trained on the genome of interest.

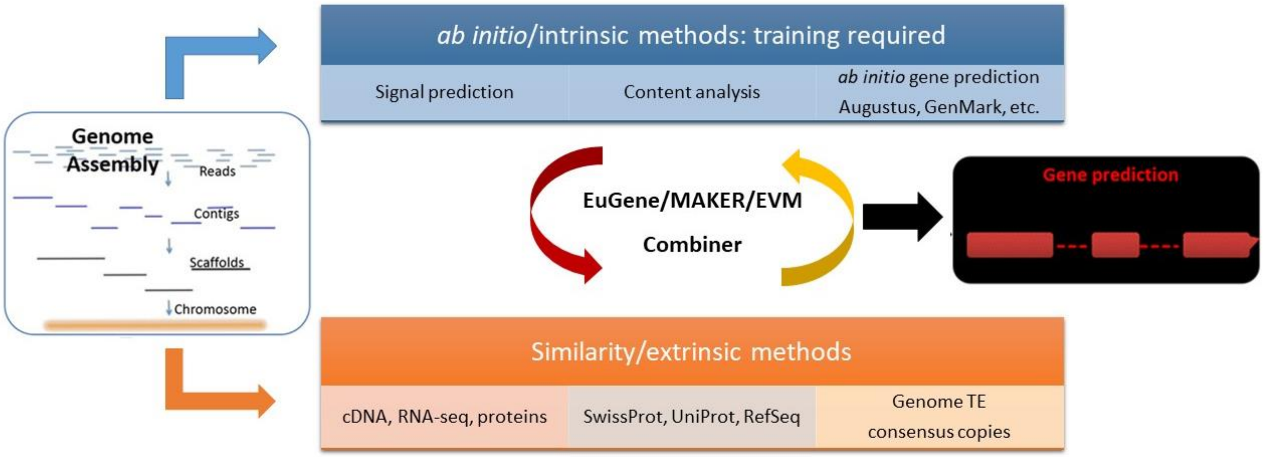
The intrinsic (or *ab-initio*) approach focuses only on information that can be extracted from the genomic sequence itself, such as coding potential and splice site prediction. This method is labor intensive as requires the building of statistical models and software training and optimization. First of all, it is crucial to have a good training set, *i.e.* a set of structurally well annotated genes used to build models and to train gene prediction software. Considering the peculiarity of each genome, these models and software must be specific to each genome and thus need to be rebuilt and retrained for each new species. This is, however, also the big advantage of this approach, as it is capable of predicting fast evolving and species specific genes (Del Angel et al., 2018).

The most used *ab initio* tools are Augustus, based on a Generalized Hidden Markov Model, a probabilistic model of a sequence and its gene structure (Stanke et al., 2006) and GeneMark-ES/ET, a self-training, but sometimes less-accurate, algorithm (Reid, 2018).

On the other hand, the extrinsic way is much more universally applicable, as it is based on polypeptide sequences that are already described and available in databases (*e.g*. NCBI non-redundant, protein, RefSeq, UniProt), which creates a wealth of information to be exploited in the gene prediction process. Transcript information and RNA-seq data, if available for the target organism or for close related species, play an even more reliable role in this approach. In this case, the most exploited tools are Blast-based algorithms and Exonerate, which is a generic sequence alignment tool, freely available (http://www.ebi.ac.uk/~guy/exonerate/) for pairwise sequence comparison. It allows you to align sequences using a many alignment models, either exhaustive dynamic programming or a variety of heuristics. These latter approaches run quickly, but their complexity makes them more difficult to implement.

Finally, *ab initio* and extrinsic tools can be complemented by using *combiners* that are probably the most popular and widely used gene prediction approach (**Fig. 2.8**). For example, Funannotate (https://github.com/nextgenusfs/funannotate) is a pipeline for gene annotation in which at the gene prediction core there is EVidence Modeler (EVM), a combiner that takes several different gene prediction inputs and outputs consensus gene models (Palmer, 2016; Haas et al., 2008). Other commonly used combiners are MAKER and EuGene that is an open integrative gene finder for eukaryotic and prokaryotic genomes (Cantarel et al., 2008; Foissac et al., 2008). However, they are not all working in the same way, some simply aim to pick the most appropriate model or build the consensus out of the provided input data, others have a more integrated approach in which the *ab initio* prediction can be modified by the given extrinsic data (Del Angel et al., 2018).

Figure 2.8. Simplified scheme of how it works a structural genome annotation using Combiners (adapted from De Angel et al., 2008). On the left, the diagram shows a typical assembly process that gives scaffolds or chromosomes ready to be annotated. These scaffolds are then annotated using two different methods. The first method is called *ab*-initio/intrinsic and requires a known set of training genes. Once the *ab initio* tool has been trained it can be used to predict other similarly structured genes. The second extrinsic approach relies on experimental evidence such as CDSs, protein sequences or RNA-seq to build gene models. Combiners (such as MAKER or EuGene) can then incorporate all of these results, eliminate incongruences, and present gene models best supported by all methods.



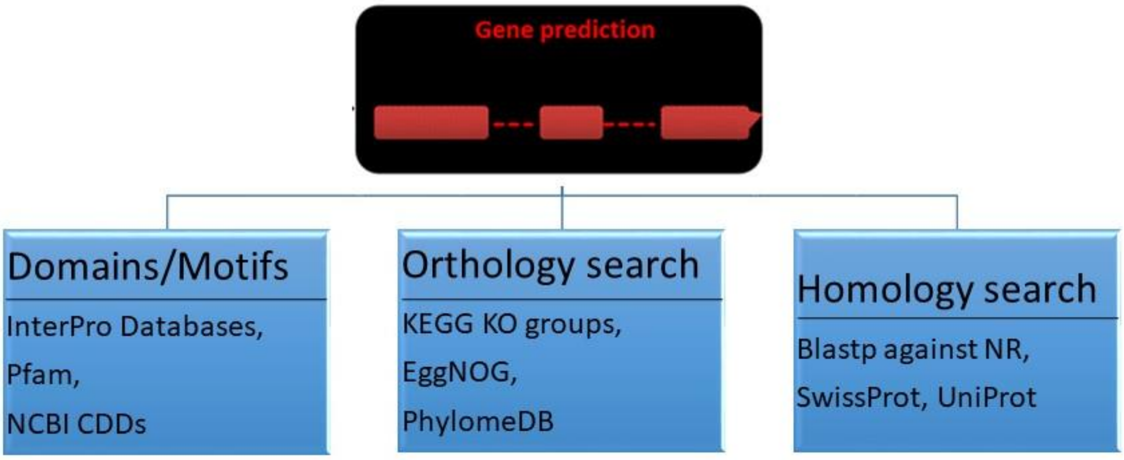
2.4.2 Step 2: functional annotation

The ultimate goal of the functional annotation process (**Fig. 2.9**) is to assign biologically relevant significance to the predicted proteins, and to the features, they derive from (*e.g.* gene, mRNA). This step is particularly relevant nowadays in the context of the NGS and 3GS era due to the capacity of sequencing, assembling, and annotating full genomes in short periods of time (Del Angel et al., 2018). Functional annotation can be done searching similarity between the sequence of interest and other sequences in different public repositories, for example running BLASTP against UniProt/SwissProt.

However, assigning results merely based on sequence similarity can be misleading, as two evolutionary independent sequences, which share some common domains could be considered homologs (Galperin and Koonin, 1998). Therefore, whenever possible, it is better to use orthologous sequences for annotation purposes rather than simply homologous sequences (Kirstensen et al., 2011). The Yeast Genome Annotation Pipeline (YGAP), an automated system specifically designed for new yeast genome sequences lacking transcriptome data, follows this approach (Proux-Wéra et al., 2012). YGAP does automatic *de novo* annotation, exploiting the existence of a large number of gene sequences ("pillars") conserved among fungal species and maintained in the Yeast Gene Order Browser (YGOB) database (as well as the syntenic arrangement of coding regions among a large number of fungi). The basic premises underlying YGAP's approach are that data from other species already tell us what genes we should expect to find in any particular genomic region and that we should also expect that orthologous genes are likely to have similar intron/exon structures.

Finally, functions can be assigned using sequence features such as structural and functional domains. Such strategy is used by InterProScan 5 (Jones et al., 2014) that runs several tools and annotate genes with multiple features, including protein and transmembrane domains, protein localisation, putative enzymatic functions, etc.

Figure 2.9. Functional annotation pipelines (adapted from Del Angel et al., 2018). The schema is showing a typical functional annotation process that implements three parallel routes for the definition of functions. The first refers to proteins domains and motifs, the second to orthology search and finally the third is applied to homology search. At the end, the output from the three different sources is put together for more reliable predictions.



CHAPTER 3: SYNTHETIC BIOLOGY FOR NON-CONVENTIONAL YEASTS

This section provides an introduction and overview of the main synthetic biology tools currently available for the genetic manipulation of non-conventional yeasts. In particular, I focus on the *Z. rouxii* clade that is very attractive for biotechnological purposes and for genome evolution studies, but tools and methodologies for its genetic engineering are still lacking compared to those for *S. cerevisiae*. In order to circumvent this issue and investigate the role of *MAT* loci in governing cell type identity of *Z. rouxii* prototrophic hybrid strains, we developed new biobricks, plasmids, transformation protocols and knock-out mutants that are extensively described in **Chapter 6**.

3.1 CHALLENGES IN GENETIC ENGINEERING OF NON-CONVENTIONAL YEASTS

Synthetic biology coupled with traditional molecular genetic techniques can enable the rapid prototyping and optimization of microbial cell factories for industrial applications. Compared to commonly used bacteria, yeasts show many industrially attractive traits for biotechnological purposes (Chen and Nielsen, 2013; Kim et al., 2012; Liu et al., 2013; Wildt and Gerngross, 2005).
A few of these attributes include:

- Growth to high cell densities on a wide variety of carbon sources;
- Robustness against several environmental stresses;
- Ability to perform a variety of post-translational modifications;
- Potential to compartmentalize reactions in organelles;
- High secretion capability;
- Lack of susceptibility to infectious agents, like bacteriophage.

Filamentous fungi also share many of these advantageous characteristics, but compared to yeasts they are often more difficult to transform with exogenous DNA (Kawai et al., 2010). Hyphal developmental growth prevents the simple bioreactor cultivation (Gibbs et al., 2000). Consequently, yeasts are widely used in both traditional and modern biotechnology for the production of foods, beverages, enzymes, fine chemicals and pharmaceutical reagents. *S. cerevisiae* and related species are particularly well-known because of their importance in making fermented beverages, but there is a wide diversity of yeasts, *i.e.* within the *Kluyveromyces*, *Zygosaccharomyces*, *Pichia*, *Debaryomyces* and *Yarrowia* genera, that have roles in biotechnology (**Fig. 3.1**). The particular subject of this Thesis is one of these yeasts, *Z. rouxii*, since, as extensively described in **Chapter 1**, this clade exhibits unique and advantageous phenotypes, which make it suitable to produce fermented food and food additives (**Table 3.1**).

The vast majority of yeast synthetic biology tools have been developed in the model yeast *S. cerevisiae* due to its well-annotated genome, genetic tractability, and overall its ease of use (Chen and Nielsen, 2013). Metabolic engineering of non-conventional yeasts is more challenging in comparison with *S. cerevisiae*, because less is known about their metabolism and genomics, and the availability of advanced genetic engineering tools is limited. As yeast applications in the biotechnology sector are constantly evolving, there is an increasing interest in applying modern molecular tools to understand and improve non-conventional yeasts. The present introduction will be focused on describing metabolic engineering tools and bio-bricks for non-conventional yeasts, with a special focus on our selected model organism *Z. rouxii*.

Figure 3.1. Overview of the main non-conventional yeasts, their properties and relevant applications (adapted from Wagner and Alper, 2016)**.**

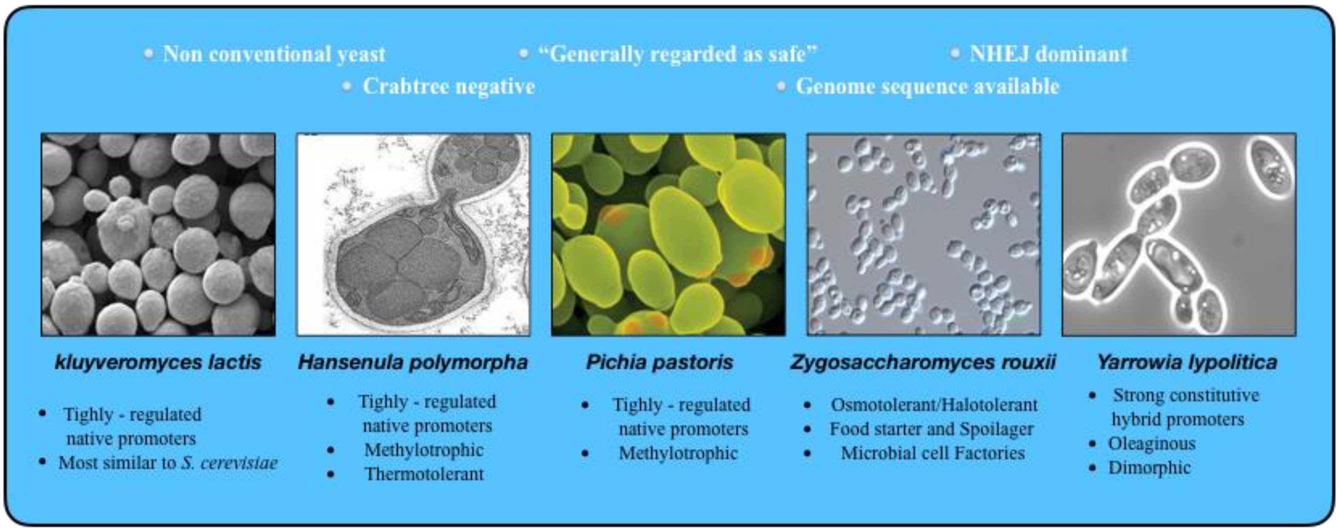

• Non conventional yeast
• "Generally regarded as safe"
• NHEJ dominant
• Crabtree negative
• Genome sequence available
kluyveromyces lactis
• Tighly - regulated native promoters
• Most similar to S. cerevisiae
Hansenula polymorpha
• Tighly - regulated native promoters
• Methylotrophic
• Thermotolerant
Pichia pastoris
• Tighly - regulated native promoters
• Methylotrophic
Zygosaccharomyces rouxii
• Osmotolerant/Halotolerant
• Food starter and Spoilager
• Microbial cell Factories
Yarrowia lypolitica
• Strong constitutive hybrid promoters
• Oleaginous
• Dimorphic

Table 3.1. Examples of industrially relevant products made using non-conventional yeasts.

Species	Product	Product type	Product use	References
H. polymorpha	Hepatitis B surface antigen (HBsAg)	Protein/peptide	Vaccine (hepatitis B virus)	Janovic et al., 1991
	Hirudin	Protein/peptide	Biopharmaceutical (anticoagulant)	Weydemann et al., 1995
	Phytase	Enzyme	Food/beverage additive, agricultural supplement	Mayer et al., 1999
	Hexose oxidase	Enzyme	Food/beverage additive (bread dough modifier)	Fraatz et al., 2013
	Triacylglycerol lipase	Enzyme	Food/beverage additive	Fraatz et al., 2013
K. lactis	Chymosin	Enzyme	Food/beverage additive (rennet component)	Van den Berg et al., 1990
	Lactase	Enzyme	Food/beverage additive (lactose depletion)	Van Ooyen and Albert, 2006
	Glycolic acid	Organic acid	Renewable chemical precursor	Bianchi et al., 2001
	Lactid acid	Organic acid	Food/beverage additive, renewable chemical precursor	Bianchi et al., 2001
P. pastoris	Ecallantide	Protein/peptide	Biopharmaceutical (hereditary angioedema)	Markland et al., 1996; Cicardi et al., 2010
	Insulin glargine	Protein/peptide	Biopharmaceutical (insulin analogue)	Kannan et al., 2009
	Insulin - like growth factor 1 (IGF-1)	Protein/peptide	Biopharmaceutical (growth hormone)	Brierley et al., 1998
	Phospholipase C	Enzyme	Plant oil processing	Löbs et al., 2017
Y. lipolytica	Citric acid	Organic acid	Food/beverage additive, renewable chemical precursor	Förster et al., 2007
	Erythritol	Sugar, alcohol	Food/beverage additive.	Mirończuk et al., 2014
	α-Ketoglutaric acid	Organic acid	Renewable chemical precursor	Yovkova et al., 2014
	Lipids (triacylglycerol)	Fatty acid	Renewable chemical precursor	Blazeck et al., 2014; Tai and Stephanopoulos, 2013
	Lycopene	Carotenoid	Food/beverage additive, supplement	Matthäus et al., 2013

	Omega-3 eicosapentaenoic acid	Fatty acid	Biopharmaceutical, supplement	Xu et al., 2013
Z. rouxii	4-hydroxyfuranone derivates	Flavour compounds	Soy sauce additive	Hauck et al., 2003
	HEMF	Flavour and aroma compounds	Soy sauce additive	Hauck et al., 2003
	ethyl acetate	Aroma compounds	Soy sauce additive	Hauck et al., 2003
	4-ethylguaiacol	Aroma compounds	Soy sauce additive	Cao et al., 2010

Metabolic engineering entails the ability to express a gene from an expression cassette and the possibility to knockout native genes. In the first case, host strategies utilize episomal plasmids and auxotrophic markers to provide selective pressure for the maintenance of heterologous DNA containing the expression cassette. Alternatively, heterologous gene expression may include the integration of expression cassette into the host's genome. Similarly, gene disruption can be performed by the replacement of targeted gene with a deletion cassette.

In non-conventional yeasts, these genetic manipulations are hampered by three main challenges. Firstly, synthetic biological parts or bio-bricks are low in number, such as promoters, terminators, and replication elements. Additionally, the generation of targeted integration cassettes and episomal vectors requires several PCR reactions, molecular cloning steps, and time, thus limiting the throughput of genetic manipulations. Another limitation is the low availability of selectable markers for the screening of transformants carrying the desired genetic modifications. Even in *S. cerevisiae*, where a relatively large number of selection markers are available, the construction of multiple successive genetic modifications remains a challenge, as the number of genetic manipulations typically equals the number of selection markers introduced in the host.

Transformation with a linear DNA fragment containing a selectable marker entails that the cassette is targeted to a specific site in the genome by homology to the site of interest. Native DNA repair pathway based on HR is responsible for this site specificity. *S. cerevisiae* has a highly efficient HR DNA repair system, which makes the control over the integration loci very easy, avoids disrupting essential genes, and allows for integration into a site with a consistent expression profile. Consequently, integration cassette via HR includes very short homology targeting regions and can also be used to knockout native genes. In non-conventional yeasts, the NHEJ DNA repair pathway is favoured over the HR and genome engineering by HR is inefficient. As a result, engineering of non-conventional yeasts is frequently accomplished by random integration of transformed cassette, which can lead to unwanted disruptions of open reading frames or other genomic elements. In addition, expression levels of heterologous cassettes have been shown to be highly dependent on the integration site, and so random integration can result in variable expression across transformants.

3.2 Survey of metabolic engineering bio-bricks

3.2.1 Promoters

Promoters are responsible for driving gene expression and are one of the first key synthetic parts required in a host system. Due to their critical role in expression cassette design, promoters are likely the most characterized and engineered genetic part in many yeast systems. In this vein, promoters that span high strength and a range of expression are necessary to provide coarse, quantitative and temporal expression control.

The field of yeast synthetic promoter engineering has been growing quickly, especially in the model yeast *S. cerevisiae*. Certainly, more advanced and engineered promoter elements can enable sophisticated control and/or coordination of a pathway, network, or circuit (Liang et al., 2012; Teo and Chang, 2014) in a minimal space. However, most of the field is dominated by the use of native yeast promoter scaffolds, resulting in relatively large promoter elements (Alper et al., 2005; Nevoigt et al., 2006). The promoters typically used in non-conventional yeasts are endogenous promoters, which have not been engineered to enhance or alter performance, or alternatively *S. cerevisiae* promoters. In *H. polymorpha* expression cassettes are commonly constructed using the strong inducible methanol oxidase promoter (PMOX) (Krasovska et al., 2007), but constitutive elements such as the glyceraldehyde-3-phosphate dehydrogenase promoter (PGAP) have also been utilized (Heo et al., 2003). In *K. lactis* heterologous protein production strategies have frequently relied on *S. cerevisiae* promoters such as PGAL1 or PPGK (Van Ooyen and Albert, 2006). The acceptable performance of these promoters in *K. lactis*, including glucose-repression and galactose-induction for P_{GAL1} (Horwitz, 2015), suggests a high level of promoter element transferability between these species. The most common *P. pastoris* promoter is derived from the strong, inducible alcohol oxidase 1 gene (PAOX1) (Tschopp, 1987). Several semi-rational promoter engineering efforts have been used to improve this element. For example, in silico transcription factors were used to identify seven cis-acting elements that contribute to the tight glucose repression and strong methanol induction of PAOX1 (Hamilton and Abremski, 1984).

3.2.2 Terminators

Transcriptional terminators display a mechanistic role in transcription, but also influence mRNA stability (Geisberg, 2014; Mischo et al., 2013). However, the impact of terminators on mRNA abundance and protein output is often underappreciated in comparison to promoters. Most commonly used *S. cerevisiae* expression vectors exploit a small set of previously identified, non-optimal native terminators such as CYC1t or ADH1t (Mumberg et al., 1995). *H. polymorpha* expression cassettes commonly include the endogenous methanol oxidase terminator (MOXt), amine oxidase terminator (AMOt) (Saraya, 2012) or heterologous *S. cerevisiae* terminators such as ADH1t (Cox et al., 2000). *K. lactis* vectors frequently use the endogenous beta-galactosidase terminator LAC4t, while *S. cerevisiae* terminators, such as GPD1t, ADH1t andADH2t, are also functional and commonly used. The endogenous AOX1 terminator is generally used in *P. pastoris* expression cassettes (Sreekrishna et al., 1993), nevertheless *S. cerevisiae* terminators such as CYC1t can also be used in *Y. lipolytica* (Blazeck et al., 2011).

3.2.3 Vector replication elements

Episomal (high-copy) and centromeric (low-copy) plasmids are replicating vectors highly used in metabolic engineering for protein expression.

Yeast episomal plasmids (YEps) are shuttle vectors as they can replicate in *E. coli* and in *S. cerevisiae*. They consist of the following parts: 1) elements from 2µ plasmid; 2) one yeast selectable marker; and 3) pBR322 backbone harbouring the bacterial ampicillin and tetracycline resistant genes *Amp*^{R} and *Tet*^{R}. Episomal plasmids provide a convenient test bed for genetic constructs by serving as a quick, intermediate proof of concept step between initial prototype construction and full chromosomal integration. Moreover, a replicative vector can provide additional workflow flexibility to an expression system. Transformation efficiencies for these plasmids tend to be much higher than for genomic integration, and plasmids often can afford a higher copy number depending on the selection marker used (Karimi et al., 2013; Orr-Weaver et al., 1981).

Centromeric vectors contain a species-specific centromere sequence (CEN) and autonomously replicating sequence (ARS), therefore they are often referred to as CEN/ARS vectors and are stably maintained at a lower copy number than episomal ones.

3.3 SELECTABLE MARKERS AND STRATEGIES FOR THEIR RECYCLING

Selection markers enable the selection of mutants carrying the desired genetic modifications and can be classified in two main categories: auxotrophic markers, which restore growth of specific mutants, and dominant markers, which confer completely new functions to their host (**Table 3.2**). Both types suffer from substantial drawbacks. The use of auxotrophic markers is restricted to auxotrophic strains, that is, strains carrying mutations in one gene leading to a strict requirement for a specific nutrient (Pronk, 2002). This constrain is augmented for industrial strains that are typically prototrophic and for which the aneuploidy or polyploidy state makes the construction of auxotrophic strains a laborious task (Pronk, 2002). The expression in a single strain of multiple dominant marker genes, under the control of strong promoters, may result in protein burden and other negative effects on host strain physiology (Nacken et al., 1996; Gopal et al., 1989). Additionally, for industrial strains dedicated to food applications such as the production of nutraceuticals, the lack of heterologous DNA is highly desired.

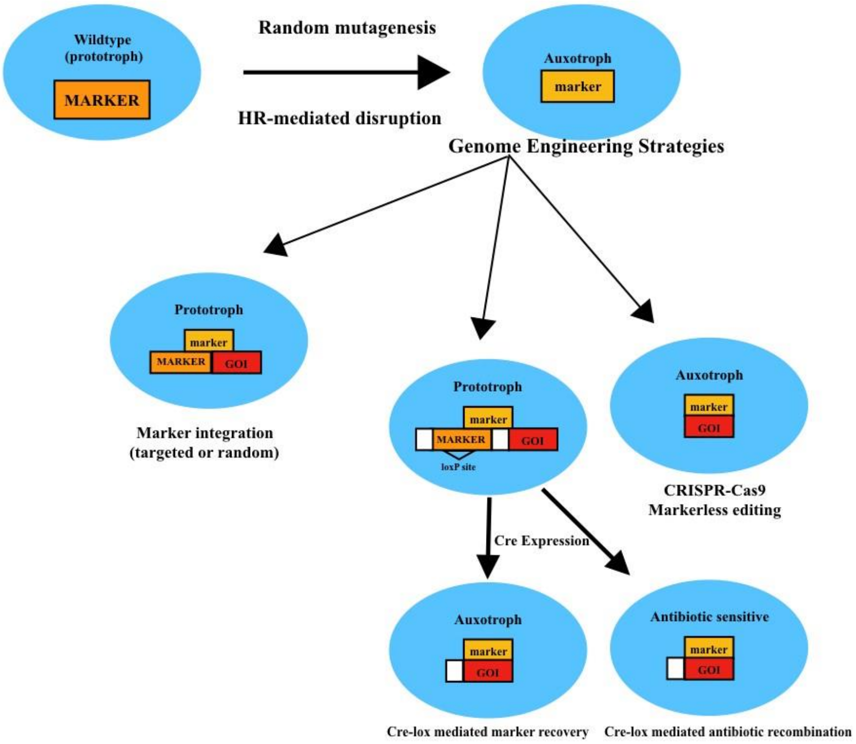
Table 3.2. Different selectable markers used in laboratory and industrial *S. cerevisiae* strains (adapted from Solis-Escalante, 2013). Abbreviation: 5-FOA, 5-fluorootic acid.

Marker gene	Mode of action	Recyclable/Method	References
Auxotrophic markers			
URA3	Repairs uracil deficiency	Yes/negative selection with 5-FOA	Alani et al., 1987; Längle-Rouault and Jacobs, 1995
KlURA3	Repairs uracil deficiency	Yes/negative selection with 5-FOA	Shuster et al., 1987
CaURA3	Repairs uracil deficiency	Yes/negative selection with 5-FOA	Losberger and Ernst, 1989
HIS3	Repairs histidine deficiency	No/-	Wach et al., 1997
HIS5	Repairs histidine deficiency	No/-	Wach et al., 1997
LEU2	Repairs leucine deficiency	No/-	Brachmann et al., 1998
KlLEU2	Repairs leucine deficiency	No/-	Zhang et al., 1992

LYS2	Repairs lysine deficiency	Yes/negative selection with alpha-aminoadipate	Chattoo et al., 1979
TRP1	Repairs tryptophan deficiency	No/-	Brachmann et al., 1998
ADE1	Repairs adenine deficiency	No/-	Nakayashiki et al., 2001
ADE2	Repairs adenine deficiency	No/-	Brachmann et al., 1998
MET15	Repairs methionine deficiency	Yes/negative selection with methyl-mercury	Singh and Sherman, 1974; Brachmann et al., 1998
Dominant markers			
KanMX	Resistance to G418	No/-	Wach et al., 1994
ble	Resistance to phleomycin	No/-	Gatignol et al., 1987
Sh ble	Resistance to zeocin	No/-	Drocourt et al., 1990
hph	Resistance to hygromycin	No/-	Gritz and Davies, 1983
Cat	Resistance to chloramphenicol	No/-	Hadfield et al., 1986
CUP1	Resistance to Cu^{2+}	No/-	Henderson et al., 1985

To avoid interference by selection markers, marker-free strains are preferred in both academia and industry. Two main strategies have been developed for gene disruption and subsequent marker-recycling (**Fig. 3.2**). One approach relies on the HR machinery of the host. As reported above, HR and NHEJ are the two processes for maintenance of genome integrity after DNA damage, in most eukaryotic cells, HR being the preferred repair mechanism in *S. cerevisiae* (Aylon and Kupiec, 2004; Jasin and Rothstein, 2013). By flanking a marker gene with repeated sequences and cultivating mutants in nonselective media, it was observed that mitotic recombination could remove the marker, albeit at a low frequency (10^{-4}-10^{-3}). Cells that underwent this process can be easily screened using negative selection or counter-selection (Alani et al., 1987). This approach requires the availability of a growth condition under which the presence of the selection marker is lethal and the presence of direct repeats flanking the marker to enable mitotic recombination and thereby marker excision. In non-conventional yeasts the prevalence of NHEJ as DNA damage repair system, makes this method ineffective. The second approach relies on the expression of heterologous or endogenous recombinase systems, among which the most commonly used system is Cre-*loxP*.

Figure 3.2. Schematic diagram of the generation and utilization of auxotrophic markers for yeast engineering. (Adapted from Löbs et al., 2017). Random mutagenesis of host DNA or homologous recombination of a cassette that inactivates an essential gene for nutrient synthesis can be used to produce stable auxotrophic strains. The presence of an auxotrophic trait allows more advanced genome editing and pathway engineering tools to be applied in the yeast species of interest. Shown here are 1) targeted and random integration using a selectable marker (bottom, left), 2) Cre-*loxP*-mediated marker recovery (bottom, middle), and 3) marker less editing by CRISPR-Cas9 (bottom, right).



3.3.1 URA3/URA5 5-fluoro-otic acid counterselection and the URA3/5 blaster cassettes

In *S. cerevisiae*, L-glutamine is converted into uridine monophosphate in sequential reactions catalyzed by five enzymes encoded by *URA2*, *URA4*, *URA1*, *URA5* and *URA3* genes. Mutations in ura2, ura4 and ura1 leave *S. cerevisiae* cells sensitive to 5-fluorootic acid (5-FOA), but lead to a requirement for uracil in the medium. However, ura3 and ura5 mutants are resistant to 5-FOA as well as being uracil auxotrophs (Boeke et al., 1984; Dave and Chattoo, 1997).

In *S. cerevisiae* a typical *URA3* blaster cassette includes a functional yeast *URA3* gene flanked by 1.1 Kb direct repeats of a bacterial sequence a *Salmonella* hisG sequence and portions of the target gene which can then be used for the disruption (Alani et al., 1987). Once introduced into the genome, the hisG direct repeats may undergo mitotic recombination to eliminate the *URA3* gene, leaving behind a single copy of the hisG repeat sequence at the site of the original integration in the target gene (Alani et al., 1987). The construct is inserted into a cloned target gene of interest and then introduce the resulting disruption into the yeast genome by HR. An appropriate DNA fragment containing the disruption plus flanking homology can be obtained by restriction enzyme digestion. After introducing such fragments into yeast by transformation, stable integrants can be isolated by selection for Ura+. The feature that makes this construct especially useful, is that recombination between the flanking direct repeats occurs at a high frequency (10^{-4}) in vegetatively grown cultures. After excision, only one copy of the repeat sequence remains behind. Thus, in the resulting strain, the Ura+ selection can be used again, either to disrupt a second gene in a similar fashion or for another purpose (Alani et al., 1987).

3.3.2 TRP1/5-fluoroanthranilic acid counterselection and TRP1 blaster cassettes

The *S. cerevisiae TRP1* gene encodes phosphoribo-sylanthranilate isomerase involved in tryptophan pathway (Toyn et al., 2000). 5-fluoroanthranilic acid (5-FAA) is an antimetabolite for the tryptophan pathway, and is toxic for prototrophic yeast due to its antimetabolic conversion to 5-fluorotryptophan. Auxotrophic trp1/trp5 mutants can survive in 5-FAA-containing medium (Toyn et al., 2000). Therefore, 5-FAA can be used for the selection of tryptophan auxotrophic strains, and for the counterselection of *TRP1* in applications that involve plasmid manipulations and *TRP1* blaster cassettes.

3.3.3 Cre-*loxP* recombinase system

The Cre recombinase has been described as “the universal reagent for genome tailoring” (Nagy, 2000). Cre recombinase (cyclization recombination) belongs to the integrase or tyrosine recombinases family and was firstly discovered in the bacteriophage P1 (Hamilton and Abremski, 1984). It is a 38-kDa monomeric protein in solution that binds cooperatively a 34-bp DNA target called *loxP* (locus of X-over of P1). The *loxP* site is composed of two 13 bp inverted repeats separated by an asymmetric 8 bp core sequence. The recombination is catalysed between the two *loxP* sites. Concerning the mechanistic features of recombination, two Cre monomers bind to each *loxP* site, resulting in a tetrameric enzyme that appears with each subunit to one of the specific sequences of the two recombination sites. These subunits work in pairs, not by directly cutting all four sequences, but they work alternately. In particular, a conserved, active tyrosine site (tyr^{324}) from one of the monomers on each *loxP* DNA molecule cleaves the DNA backbone, forming a covalent, 3' phosphotyrosine bond, leaving a free, 5' hydroxyl (OH) on one strand of each DNA double helix. The 5' OH's perform a nucleophillic attack on the phophotyrosines from the partner DNA substrates, yielding a Holliday junction intermediate. A second round of tyr324-catalyzed breakage, followed by strand joining reactions (nucleophillic attack of free OH's on phosphotyrosines) resolves the Holliday junction into recombinant products (Sauer, 1987). As result, recombination between two directly repeated sites on the same DNA molecule determines the excision of the DNA segment lying between the sites.

The first application of the Cre-*loxP* system in yeasts was described for *S. cerevisiae* and entailed lox sites flanking the *LEU2* gene (Sauer, 1987). Subsequently, Güldener and coworkers (1996) constructed the gene disruption cassette *loxP-kanMX*-l*oxP*, which combined the advantages of the heterologous *kan*^{R} marker with those of the Cre-loxP system. Since then, several deletion cassettes containing drug resistant dominant markers have been constructed and the Cre-*loxP* system has been extended and, when necessary, adapted to several other yeast strains, namely *K. lactis* (Steensma and Linde, 2001; Güldener, 2002), *Y. lipolytica* (Fickers, 2003), *C. albicans* (Dennison et al., 2005), *S. pombe* (Hentges, 2005; Iwaki and Takegawa, 2004), *H. polymorpha* (Krappmann et al., 2000) and *Kluyveromyces marxianus* (Orr-Weaver et al., 1981).

The Cre-*loxP* system displays an important limitation: each recombination catalysed by Cre leaves a scar composed of the recombinase recognition site. When used repeatedly, for instance in serial gene deletion experiments, the scars spread over the chromosomes promote recombination upon Cre induction, resulting in chromosomal translocations (Delneri et al., 2000). Although a few mutated recognition sequences have been engineered to prevent the occurrence of unwanted genomic rearrangements (Delneri et al., 2000), this instability limits the potential of external recombinase-based systems for extensive strain construction programs.

3.4 Tools for *Zygosaccharomyces rouxii* genetic manipulation

In **Chapter 1** a detailed phenotypic and genetic description of the *Zygosaccharomyces* yeasts, objective of the present Thesis, was given. We saw that among them, *Z. rouxii* is a very attractive protoploid yeast both for biotechnological application and for genome evolution studies.

Our knowledge of *Z. rouxii* cell properties at the molecular level lags far behind that of *S. cerevisiae*, mainly due to a lack of tools for *Z. rouxii* genetic manipulation. Therefore, most of the *Z. rouxii* genes and their products that have been studied so far were identified and/or characterized through the heterologous expression in *S. cerevisiae* mutants.

One of the possible explanation of this situation was that *Z. rouxii* cells are highly resistant to routine and quick transformation procedures set up for *S. cerevisiae* due to a different cell wall composition. This obstacle was circumvented by optimizing the transformation procedures that efficiently introduce pDNA into *Z. rouxii* by electroporation (Pribylova et al., 2007; Watanabe et al., 2013). Furthermore, different strains belonging to the *Z. rouxii* clade exhibited different transformation efficiencies, making necessary to set up specific electroporation conditions for each tested strain (Pribylova et al., 2007).

Another reason for the low accessibility of *Z. rouxii* to genetic manipulations was that *S. cerevisiae* centromeric plasmids were not stably maintained in *Z. rouxii* cells (Pribylova et al., 2007). Furthermore, the 2 µm replicon of *S. cerevisiae* does not function in *Z. rouxii* (Araki et al., 1985; Ushio et al., 1988), making *S. cerevisiae* episomal plasmids not suitable for these non-conventional yeasts. Multicopy vectors were developed owing to the isolation of pSR1, a cryptic double strand plasmid of *Z. rouxii* similar to the 2 µm plasmid of *S. cerevisiae* (Araki et al., 1985). The replicon of pSR1 was exploited to build the first *Z. rouxii* episomal vector, pSRT5 (Ushio et al., 1988), which was in an *E. coli-Z. rouxii* shuttle vector consisting in the following three main parts: 1) a DNA fragment of pBR322; 2) a fragment of *Z. rouxii* pSR1 plasmid, and 3) a fragment of the *S. cerevisiae* LEU2 gene or a DNA fragment bearing Tn601, a selective marker which confers G-418 resistance (Ushio, 1988). To construct pSRT5, a 1.2 Kb pSR1 fragment was selected as a replicon since it was previously shown to contain ARS and elements for stable partitioning into subsequent generations (Jearnpipatkul et al., 1987). *Z. rouxii* episomal vectors were further improved by adding multiple cloning site (MCS). The resulting pZEU, pZEA and pZEL vectors harbour pSR1 as a replicon and *ScURA3*, *ZrADE2* or *ZrLEU2* as marker genes (Pribylova et al., 2007). Centromeric plasmids were constructed after the release of low coverage *Z. rouxii* CBS 732^T genome (De Montigny et al., 2000). Preliminary evidences showed that *ScARS1* was replicable in *Z. rouxii* (Ushio et al., 1988), but the pU1 harbouring *S. cerevisiae CEN* was unstable (Pribylova et al., 2007). By exploring Génolevures project I database, Prybilova and co-workers (2007) recognized centromeric regions on *Z. rouxii* chromosomes II and IV and used them to construct the first two low-copy and stably propagated centromeric plasmids for *Z. rouxii*, namely pZCA and pZCC (both containing *ScURA3* marker) (Pribylova et al., 2007). These plasmids are *Z. rouxii-E. coli* shuttle (*ori*, *Amp*^{R}) containing one MCS and *ZrCEN* and *ScARS1* as replicon. To further enrich the set of *Z. rouxii*-specific plasmids, two additional centromeric plasmids with different markers were derived from pZCA, namely pZCAA (containing *ZrLEU2*) and pZCAL (containing *ZrADE2*).

Finally, only few auxotrophic mutants are available for *Z. rouxii*. The first auxotrophic strains were *leu2* and *ura3* mutants obtained from ATCC 42981 and CBS 732^{T}, respectively (Ushio et al., 1988; Pribylova and Sychrovà, 2003). Prybilova and Sychrovà (2003) extended the pool of auxotrophic mutants by using *loxP-kanMX-loxP* deletion cassette. Recently, Watanabe and colleagues (Watanabe et al., 2017) used both *loxP-KanMX-loxP* and *loxP-ZeoMX-loxP* deletion cassettes to construct auxotrophic mutants derived from the opposite mating **a** and α strains CBS 4837 (=NBRC1876) and CBS 4838 (=NBRC1877), respectively. The resulting mutants were used in trackable mating experiments, to produce tetraploids. A list of the main auxotrophic strains currently available is reported in **Table 3.3**.

Table 3.3 *Z. rouxii* derived mutants.

Mutant Name	Source Strain	Genotype	Reference
DL1	CBS 732^{T}	*ura3 leu2*Δ::*kanMX*	Pribylova et al., 2007
DL2	CBS 732^{T}	*ura3 leu2*Δ::*loxP*	
DA1	CBS 732^{T}	*ura3 ade2*Δ::*kanMX*	
DA2	CBS 732^{T}	*ura3 ade2*Δ::*loxP*	
DLA1	CBS 732^{T}	*ura3 leu2*Δ::*loxP ade2*Δ::*kanMX*	
DLA2	CBS 732^{T}	*ura3 leu2*Δ::*loxP ade2*Δ::*loxP*	
CBS 4837 *ura3*Δ	NBRC1876	*ura3*Δ::*loxP-KanMX-loxP*/ *ura3*Δ::*loxP-ZeoMX-loxP* (*MAT*α)	Watanabe et al., 2017
CBS 4838 *ura3*Δ	NBRC1877	*ura3*Δ::*loxP-KanMX-loxP*/ *ura3*Δ::*loxP-ZeoMX-loxP* (*MAT***a**)	
CBS 4838 *ade2*Δ	NBRC1877	*ade2*Δ::*loxP-KanMX-loxP*/ *ade2*Δ::*loxP-ZeoMX-loxP* (*MAT***a**)	

Objectives

Objectives

- Investigation of the cell cycle in the *Z. rouxii* haploid type-strain CBS 732^{T} with focusing on its sex-determination system, the mating-type switching process and the main genotypic and phenotypic consequences.

- Extending the characterization of sex determination system and *MAT* loci structural organization in the *Zygosaccharomyces* allodiploids, considering the halo-tolerant hybrid strain ATCC 42981 as a study model. In particular, we investigated the role of a chimeric mating-type system in determining hybrid sterility.

- Expanding the synthetic biology tools available for genetic engineering of *Z. rouxii* prototrophic strains and optimization of the available transformation methods.

- Genome sequencing of ATCC 42981 and another industrially relevant hybrid species *Z. sapae* ABT301^{T} by a hybrid sequencing approach that corrects highly erroneous Nanopore long reads with high quality Illumina short reads.

- Gene prediction and functional annotation of ATCC 42981 genome focusing on genes involved in halo-tolerance and meiosis commitment.

PARTII RESULTS

CHAPTER 4: MATING-TYPE SWITCHING IS A PLASTIC PROCESS LEADING TO INTRA-STRAIN GENETIC AND PHENOTYPIC VARIABILITY IN HAPLOID CELLS DERIVED FROM *ZYGOSACCHAROMYCES ROUXII* CBS 732^{T} TYPE-STRAIN

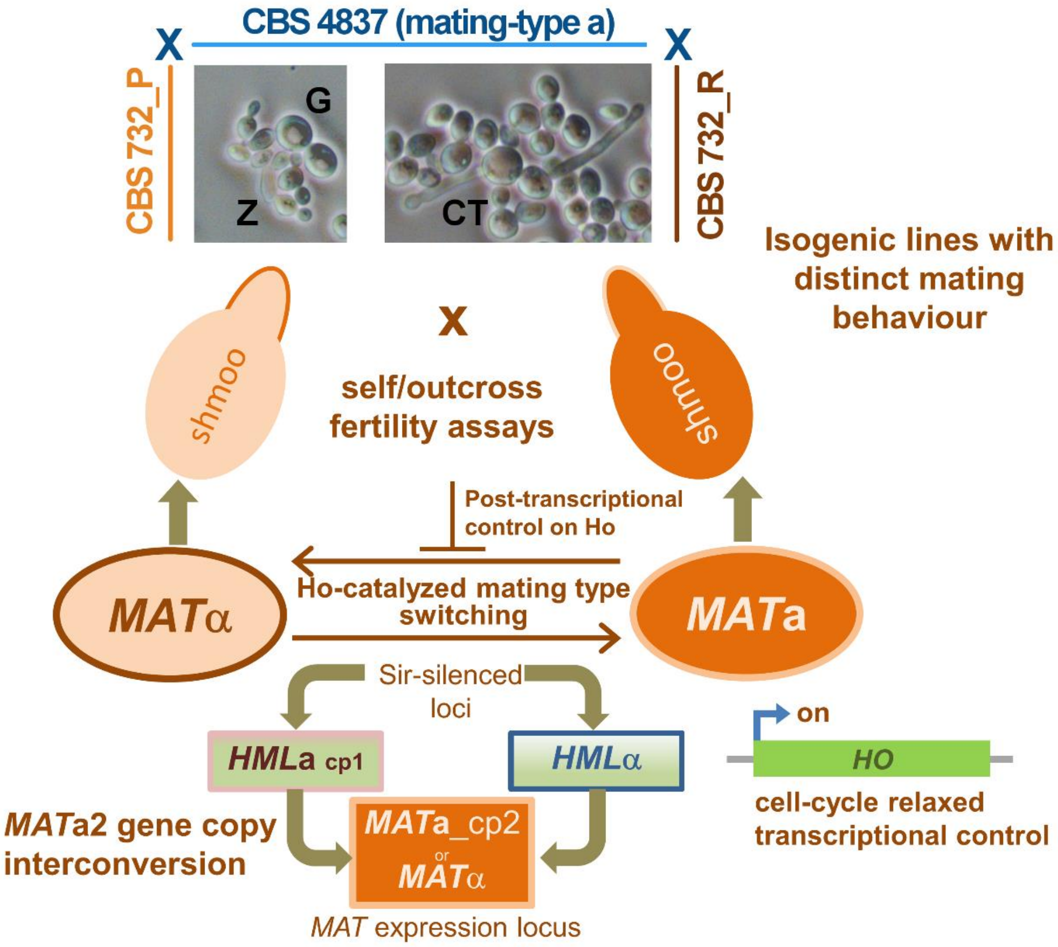
Bizzarri, M., Cassanelli, S., and Solieri, L. (2017). Mating-type switching in CBS 732^{T} derived subcultures unveils potential genetic and phenotypic novelties in haploid Zygosaccharomyces rouxii. FEMS Microbiology Letters, 365(2), fnx263.

Abstract

In haploid *Saccharomyces cerevisiae*, a complex recombination system regulates mating-type switching and requires one *MAT* expression locus, two donor cassettes (*HML* and *HMR*) and the *HO* endonuclease that catalyses gene conversion. *Zygosaccharomyces rouxii* is the most distant species from *S. cerevisiae* with a functional *HO*, but with a poorly understood mating-type switching. Here, we described that two subcultures of the type-strain CBS 732^{T} underwent the α to **a** genotype switching leading to mixed *MAT*α and *MAT***a** populations. Remarkably, during this event the donor cassette was copied into the *MAT* locus, except for its own 3' end, resulting in a new *MAT***a**2 gene copy different from the silenced *HMR***a**2. Moreover, CBS 732^{T} cells bypassed the cell-cycle control, which oversees *HO* transcription in *S. cerevisiae*, and expressed *HO* at the stationary phase. Despite *HO* dysregulation, mating-type switching seemed to occur rarely or belatedly during CBS 732^{T} colony formation in most of the tested conditions. When morphology and mating behaviour were analysed, two subcultures displayed distinct outcross fertility responses. Overall, our data support that mating-type switching causes genotype instability and phenotypic novelties in CBS 732^{T}, and open the question whether this mechanism is shared by other *Z. rouxii* haploid homothallic strains.

Keywords: mating-type switching; *HO* endonuclease; phenotypic novelty; *MAT* locus; genetic instability; *Zygosaccharomyces rouxi*

Graphical abstract


CBS 4837 (mating-type a)
X
X
CBS 732_P
CBS 732_R
G
Z
CT
Isogenic lines with distinct mating behaviour
X
shmoo
shmoo
self/outcross fertility assays
Post-transcriptional control on Ho
Ho-catalyzed mating type switching
MATα
MATa
Sir-silenced loci
HMLa cp1
HMLα
on
HO
cell-cycle relaxed transcriptional control
MATa2 gene copy interconversion
MATa_cp2 or MATα
MAT expression locus

Introduction

In *Saccharomyces cerevisiae*, the sexual differentiation circuit serves as a classic paradigm for the genetic control on cell type and ploidy in all eukaryotes. In this model, the mating-type (*MAT*) loci *MAT*α and *MAT***a** encode transcriptional factors responsible for α, **a** and α/**a** identity. The default mating-type is **a**: **a**-specific genes (**a**sgs) are repressed by α2 to obtain the α mating-type and the activation of α-specific genes (αsgs) controlled by α1 (Haber, 2012). Consequently, cells that express only the *MAT*α- or *MAT***a**-encoded DNA-binding proteins are haploid and mating-competent α-cells and **a**-cells, respectively. When strains with opposite mating-type mate (heterothallism), they generate a diploid α/**a** progeny incompetent for mating, because the **a**1/α2 heterodimer turns off the αsgs and **a**sgs. Additionally, diploid cells may arise from mating between cells of the same strain, due to mating-type switching (secondary homothallism), which allows *MAT***a** cells to change to *MAT*α, or *vice versa*. In *S. cerevisiae*, this process requires two identical, but silenced copies of *MAT* locus, *HML*α (Hidden *MAT* Left) and *HMR***a** (Hidden *MAT* Right), at distant locations on the same chromosome. These loci act as templates during intrachromosomal *MAT* gene interconversion.

Recombination is triggered by two blocks of conserved sequences flanking the *MAT*, *HMR* and *HML* loci (the X and Z regions) and starts with a double-strand break (DSB) at the *MAT* locus, catalysed by the homing endonuclease-derived protein HO (Hanson and Wolfe, 2017). Fine-tuned regulatory networks assure that only the G1 haploid mother cell switches after the daughter cell has budded (Chen and Gartenberg, 2015). Recently, comparative genomics revealed that sex determining mechanisms are unusually plastic in the hemiascomycetes and that 'evolutionary solutions' alternative to the standard model account for the high variability of lifestyle and mating-type switching found in this lineage. Like *S. cerevisiae*, species that underwent whole-genome duplication (post-WGD) contain both components of the switching machinery, *i.e*. an *HO* gene and three *MAT*-like cassettes, but with a different degree of specialisation, genetic and epigenetic control. *Candida glabrata*, the closest to *S. cerevisiae* among *Candida* species, is without the linkage between *HMR***a** and *MAT*/*HML*α loci and rarely switches genotype (Butler et al., 2004; Gabaldon et al., 2013). When forced to switching mating-type by the constitutive expression of heterologous *HO* genes, the cells died or underwent abnormal chromosomal rearrangements (Boisnard et al., 2015). In species that branched before the WGD (pre-WGD), the genetic circuits controlling cell identity and mating-type switching have been extensively rewired compared to *S. cerevisiae*. In pre-WGD species, the three *MAT*-like cassette system can be missing and switching can be *HO* independent (Butler et al., 2004; Fabre et al., 2005). In most pre-WGD species, *MAT***a** locus also contains the *MAT***a**2 gene coding for the *HMG* domain protein **a**2 (Butler et al., 2004; Gordon et al., 2011). In the haploid species *Kluyveromyces lactis*, **a**2 positively controls the **a**sgs (Tsong et al., 2003, 2006), while the transposase-like protein α3 supplants an unfunctional *HO* in catalysing the genotype switching (Barsoum, Rajaei and Astrom 2011; Rajaei et al., 2014). Overall, these findings support that the triplicated *MAT*-like cassettes originated at the base of Saccharomycetaceae family and initially underwent 'passive' mating-type switching at low frequencies by gene conversion; then *HO* gene was gained by 'domestication' of self-transposing genes (Butler et al., 2004; Koufopanou and Burt

2005; Rusche and Rine 2010). Under this evolutionary scenario, the haploid yeast *Zygosaccharomyces rouxii* represents one of the first pre-WGD species with the canonical three-cassette system and, differently from *K. lactis*, a putatively functional *HO* endonuclease (Butler et al., 2004; Solieri et al., 2014a). Like *C. glabrata*, *Z. rouxii* does not display the linkage between *HMR* and *MAT*/*HML* and the α to **a** idiomorph switching occurs via interchromosomal recombination. Additionally, haploid strains frequently exhibit redundant number of *MAT*-like cassettes flanked by variable genes, resulting from ectopic recombination between *MAT*-like loci (Watanabe, Uehara and Mogi 2013). A study about the interchromosomal rearrangements in *Z. rouxii* species suggested that CBS 732^T stock of *Z. rouxii* type-strain underwent reciprocal translocations between the *MAT* and *HMR* loci, changing the left side of *MAT* cassette from the typical pre-WGD *DIC1-MAT-SLA2* arrangement to *CHA1-MAT-SLA2* (Gordon et al., 2011). However, there are no data on the frequency of this organisation in other *Z. rouxii* strains. Among the *Z. rouxii* type-strain stocks, only CBS 732^T has genetic tools and genome sequencing project available (Prybilova et al., 2007; Souciet et al., 2009). Here, we found that two different derived subcultures of *Z. rouxii* CBS 732^T switched mating-type, and tentatively reconstructed how this event led to a new *MAT***a**2 gene. We investigated *HO* expression profile and analysed how mating-type switching affects the morphological and mating behaviour of these derived switched cultures.

Material and Methods

Yeast strains, media and culture conditions

We obtained CBS 732^T from CBS culture collection and stored it at -80◦C in Unimore Microbial Culture Collection (UMCC, Reggio Emilia, Italy; www.umcc.unimore.it) until *MAT* genotyping experiments. This CBS 732^T stock was termed CBS 732_R in this work. Another CBS 732^T culture (termed CBS 732_P) was from the stock stored in Sychrovà's lab in Prague (Czech Republic). Strains used in this work were detailed in **Table S1** (**see supplementary materials**). They were routinely cultured at 28◦C in YPD medium with or without 1.5% (w/v) agar and maintained at 4° C for the duration of experiments. Self-cross and outcross fertility assays were performed as previously described (Bizzarri *et al.,* 2016), with a few modifications. Briefly, CBS 732^T cells were grown alone or in mixed cultures either with CBS 4837 (**a**) or CBS 4838 (α) mating testers in liquid test media for 3 days at 27° C. Cells were collected from each tube, plated onto the corresponding medium supplemented with 2% (w/v) agar and incubated for 14 days at 27° C. Cells grown on YPD were used as control. After 4, 7 and 14 days, photomicrographs were made using a phase-contrast Nikon Eclipse 80i microscope with Nikon Digital Sight DS-5M digital camera. Frequencies of hyperelongated (H), zygote (Z), spore (S), giant (G), conjugative tube (CT) and shmoo (Sm) morphologies were measured after 7 days of incubation on MEA medium as follows. Cells were resuspended in physiological water at the final concentration of 10^7 CFU m/l (corresponding to OD_{600nm} = 0.35 - 0.40). After proper dilution, cells were examined under a light microscope as reported above. Percentage values are expressed as means of at least two replicates and represent the fractions of the total cells counted that had formed H, Z, S, G, CT or Sm structures, respectively. All media are detailed in **Table S2**.

DNA manipulation, standard PCR reactions and sequencing

Genomic DNA (gDNA) was extracted according to Sambrook, Fritsch and Maniatis (1989), and gDNA quantity and quality were evaluated spectrophotometrically using a NanoDrop ND-1000 device (Thermo Scientific, Waltham, MA, USA). PCR amplifications were carried out in 25 µl reaction volume containing 200 ng of template gDNA, using a T100 Thermalcycler (Bio-Rad). DreamTaq DNA polymerase (Thermo Scientific) was used according to the manufacturer's instructions, with the exception of >2 Kb long amplicons, which required Phusion Hot Start II Polymerase (Thermo Scientific). All primers were listed in **Table S3**. PCR products were sequenced on both strands (MWG, Heidelgerg, Germany). Sequences were assembled and edited using DNAStar Software (DNASTAR, Inc. Madison, WI, USA). Blastn and Blastp queries were performed at the NCBI server (Altschul et al., 1997). Nucleotide and amino acid sequences were aligned in Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/). Protein alignments were visualised with Jalview Version 2.10.1 (Waterhouse et al., 2009).

Colony PCR

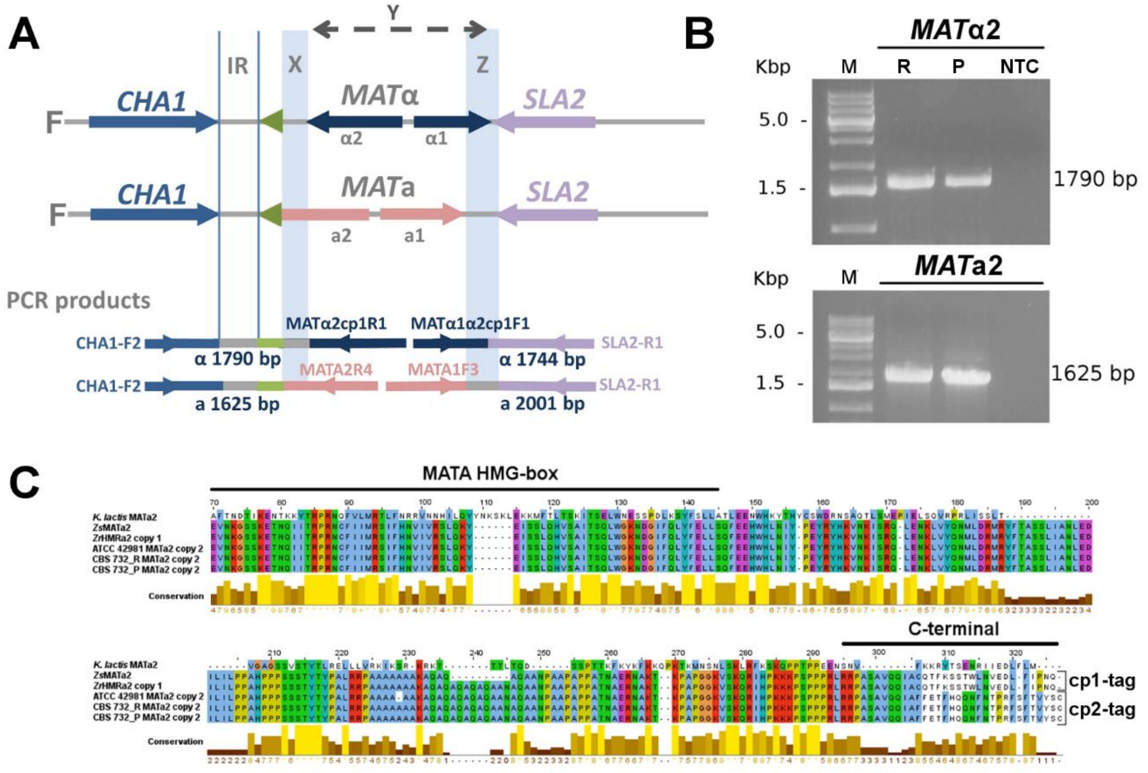
Mating-type was determined on individual colonies (36 colonies for CBS 732_R and 50 for CBS 732_P, respectively) using specific primers, which can discriminate the mating-type at the *MAT* locus by amplifying its upstream (**Table S3**). At least three colonies of each identified mating-type were then restreaked on YPD agar plate, allowing cells to grow into individual subcolonies, and at least 16 subclones per genotype were PCR-tested. DNA was extracted from 48-h-old cells using the LiOAc-SDS method (Looke, Kristjuhan and Kristjuhan 2011). One microlitre of LiOAc-SDS extracted DNA was PCR-amplified as reported above. Cycling conditions consisted of an initial denaturation at 94◦C for 2 min followed by 35 cycles of 1 min at 94°C, 1 min at 55°C and 90 s at 72°C, with a final extension step at 72°C for 10 min. RNA extraction, cDNA synthesis and RT-PCR RNA was extracted from exponentially and stationary growing cells, cultured in standard and salt-supplemented media (**Table S2**) (Solieri et al., 2016). RNAs were reverse transcribed using 0.5 µM oligo (dT) and RevertAid H Minus Reverse Transcriptase (Thermo Scientific) according to the manufacturer's instructions. cDNAs (25 ng) were amplified using DreamTaq polymerase with primers listed in **Table S3.**

Results and Discussion

Mating-type determination

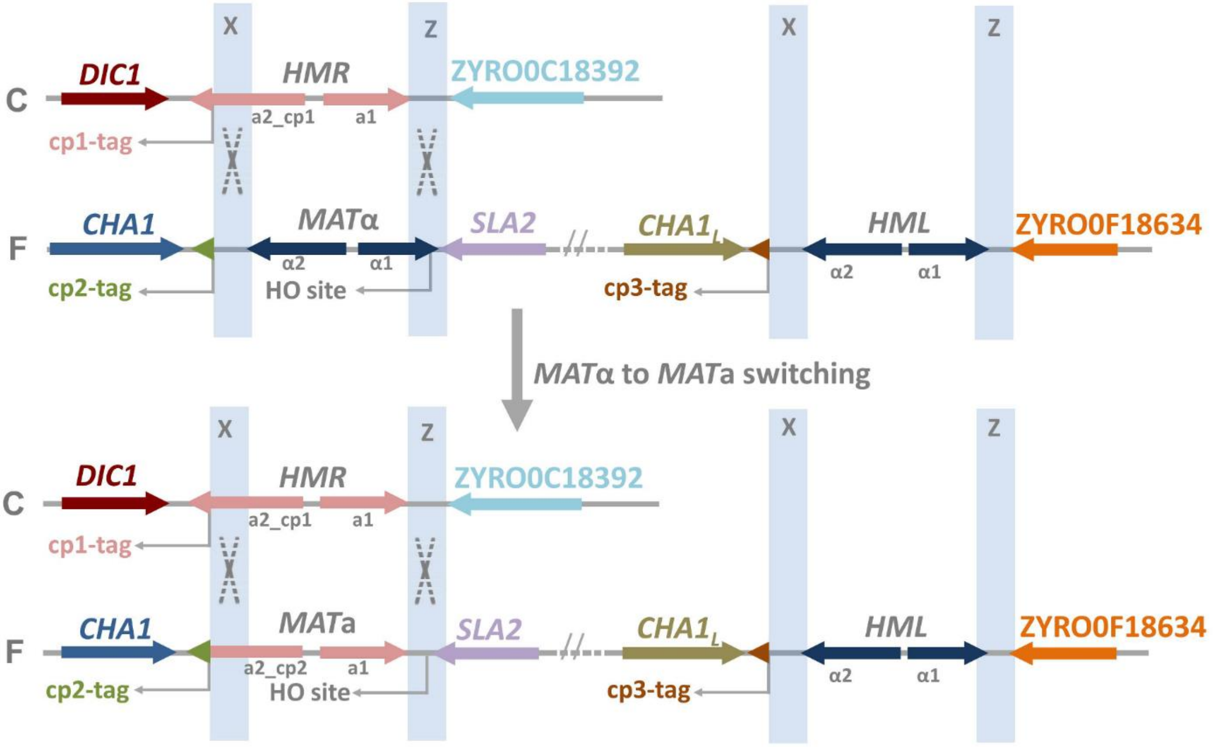
Z*ygosaccharomyces rouxii* genome sequence reports that strain CBS 732^T has a *MATα* expression locus on chromosome F flanked by *CHA1* at the left side and *SLA2* gene at the right side (Souciet *et al.,* 2009). To evaluate the *Z. rouxii MAT* genotype, we set up a PCR amplification using one reverse primer specific for Y*α* or Y**a** together with a forward primer on the common flanking gene *CHA1* (**Fig. 1A**). During the protocol setting, we used the in-house stock culture of *Z. rouxii* CBS 732^T (termed CBS 732 R) as *MATα* positive and *MAT***a** negative control. Unexpectedly, CBS 732_R yielded two strong PCR products with both Y*α*- and Y**a**-specific reverse primers (**Fig. 1B**). Our result was validated by repeating the *MAT* screening on another derived culture of strain CBS 732^T (termed CBS 732_P), which was stored in a different laboratory (**Fig. 1B**). We also checked the Y/Z junction (**Fig. 1A**) and confirmed the integrity of HO recognition site and that *SLA*2 gene flanked both Y**a** and Y*α* sequences at the 3' end (data not shown).

Figure 1. *MAT* genotype screening of CBS 732^T stocks and sequence comparison between MATa2 proteins. (**A**) Structure of *MAT*α and the *MAT***a** loci in CBS 732^T genomes and the idiomorph-specific primer pair combinations. (**B**) Resulting PCR amplification of idiomorph-specific products. (**C**) Alignment of MATa2 proteins from *K. lactis* MAT**a**2 (GenBank: XP452180), *Z. sapae* (*Zs*MATa2, GenBank: CDM87352), CBS 732T HMR**a**2 copy 1 (*Zr*HMR**a** copy 1, GenBank: XP002496430), ATCC 42981 MAT**a**2 copy 2 (GenBank: AMB17487), CBS 732_R MAT**a**2 and CBS 732_P MAT**a**2 copies 2 (this work). The horizontal black bars indicate the conserved MATA-HMG box and the divergent C-terminal specific tag (cp1- and cp2- tags). The amino acid identities were coloured according to the Clustal colour scheme, and Jalview (Waterhouse et al., 2009) provided the conservation index at each alignment position. Abbreviations: M, molecular weight marker; R, CBS 732_R stock; P, CBS 732_P stock; NTC, no-template control.



Watanabe, Uehara and Mogi (2013) reported that in the Japanese stock of the *Z. rouxii* type-strain, NBRC 1130^T, *DIC1* and *SLA2* genes flank *MATα* locus at the left and right side, respectively, while *CHA1* lies at the left side of the *HMR***a** cassette. Our findings could be due to a mixed culture with a subpopulation harbouring the *CHA1-HMR***a** arrangement, like NBRC 1130^T. Thus, PCR analysis was performed using specific primers that annealed to *MAT*-flanking regions *CHA1* and *SLA2,* and consequently the Y regions were screened by a seminested approach. Our results excluded the *MAT*-like loci arrangement found in NBRC 1130^T and supported the existence of two *CHA1-MAT***a***-SLA2* and *CHA1-MATα-SLA2* loci in CBS 732_R and CBS 732_P. RT-PCR with primers for the *MAT***a**1, *MAT***a**2, *MATα*1 and *MATα*2 targets confirmed that these genes are actively transcribed in both cultures (**Fig. S1 supplementary materials**). Sequencing of PCR products showed that both stocks had the *MATα* locus identical to that reported in CBS 732^T genome project, while the *MAT***a** locus differed from CBS 732^T *HMR***a** at the 5' end (data not shown). *Zygosaccharomyces rouxii* retains the *MAT***a**2 gene in this *MAT* portion (Butler *et al.,* 2004). MAT**a**2 proteins from both CBS 732^T derived cultures (termed copy 2) were 94.08% identical to that encoded by *HMR***a** (termed copy 1), while they were 100% identical (except for one single amino acid insertion) to **a**2 copy 2 protein characterised in *Z. rouxii* allodiploid strain ATCC 42981 (Bizzarri et al., 2016). Both **a**2 copies conserved the MATA-HMG box required for DNA binding, while differed each other for the C-terminal ends, therefore termed as copy 1 and copy 2 tags (cp1- and cp2-tags) (**Fig. 1C**). These tags are encoded by the 3'end of *MAT***a**2 gene at the left of the X region, whereas the MATA-HMG domain is encoded by the sequence across the Y**a** and X regions. The *MAT* locus structure was examined to tentatively infer the mechanism that generated a new **a**2 protein variant in switched *Z. rouxii* cells (**Fig. 2**). The *MAT***a**2 portion encoding the cp2-tag lies at the left of the X region. In *Saccharomyces cerevisiae*, HO cuts the Y-Z junction to induce a DSB and the Yα sequence to the left of the DSB is replaced by strand invasion and primer extension using a Y**a** donor sequence as template (Haber 2012). If *Z. rouxii* retains a similar mechanism, the 3' end of the invading strand primes for the new DNA synthesis, which copies the donor Y**a** sequences until the X region. The newly synthetised DNA joins to the X region at the *MAT* locus, so that the upstream region completes the *MAT***a**2 ORF with the cp-tag. In *Z. rouxii*, ectopic recombination changes the 5'-*MAT* flanking sequence and results into different *MAT***a**2 cp-tagging. For example, in NBRC 1130^T the α to **a**-idiomorph switching generated the opposite cp2- to cp1-tag conversion because *CHA1-HMR***a** cassette acts as donor instead of *DIC1-HMR***a** (Watanabe, Uehara and Mogi 2013). Intriguingly, the allodiploid ATCC 42981 harbours a third *MAT***a**2 variant (Bizzarri et al., 2016), which diverges from *MAT***a**2 copies 1 and 2 for the portion between the left of the X region and ZYRO0F18634 gene (cp3-tag) (**Fig. 2**). A crucial feature of this switching model is that *MAT***a**2 gene variants did not arise from interspecific hybridisation, but from ectopic recombination in *MAT***a** strains and ectopic recombination followed by mating-type switching in *MAT*α. In both cases, distinct **a**2 isoforms might have important and yet unexplored functional consequences on a-type determination, as *Z. rouxii*, like other pre-WGD species, conserves the ancestral mechanism of **a**sg activation mediated by **a**2 (Baker et al., 2012).

Figure 2. Inferred *MAT* locus organisation resulting from the *MAT*α to *MAT*a switching. The figure represents the organisation of *MAT*-like cassettes on chromosomes C and F: at the top according to CBS 732^T genome sequencing project (Souciet et al., 2009), at the bottom after *MAT*α to *MAT***a** switching. Blue shading denotes the X and Z regions. Ho endonuclease target site is also displayed inside the Z region. *MAT***a**2 copy specific-3' end tags are indicated as pink, green and brown tips (cp1-, cp2- and cp3- tags). Abbreviation: cp1, copy 1; cp2, copy 2; cp3, copy 3.



Mating-type colony screening

Colony PCR was used to confirm that both cultures contain switched haploid cells and to investigate the presence of heterozygous *MAT*α/**a** subpopulations. Cells from CBS 732_R and CBS 732_P were streaked onto rich medium and the resulting colonies were directly checked by PCR with *MAT***a** and *MAT*α-specific primers, respectively. The percentage of 'pure' switched colonies and 'mixed' colonies was then calculated (**Table S4**). We obtained a difference in *MAT* genotype distribution between CBS 732_R and CBS 732_P: 39% *MAT***a**, 58.3% *MAT*α/*MAT***a** and 2.8% *MAT*α colonies for CBS 732_R; 62% *MAT*α, 36% *MAT*α/*MAT***a**, and 2% *MAT***a** for CBS 732_P. In CBS 732_P, the *MAT*α idiomorph prevalence was consistent with the *Z. rouxii* genome project (Souciet et al., 2009). In CBS 732_R, we hypothesised that a single *MAT***a** colony was isolated from a starting *MAT*α and *MAT***a** mixed population by streaking on solid media and then clonally propagated, leading to a founder effect. Overall, these results indicate that in both CBS 732^T cultures there is a mix of *MAT*α and *MAT***a** haploid cells resulting from mating-type switching. This finding is further supported by the transcription of *MAT*α1 gene that is required by haploids to activate αsg program, while it is switched off by *MAT*α/**a** diploids (**Fig. S1**). To better understand the nature of mixed *MAT*α/**a** colonies, we performed *MAT* genotyping on subclones derived from another round of streaking. From mixed *MAT*α/**a** colonies, we obtained both pure and mixed subclones, the latter with a strong PCR product for the prevalent mating-type (α for CBS 732_P and **a** for CBS 732_R) and a weaker signal for the opposite one (**Table 1**). Although we cannot completely rule out the possibility of rare *MAT*α/**a** diploid cells, difference in PCR intensity supports that the streaking progressively allows to disaggregate cell clumps with opposite mating-type, or, alternatively, that the switching is rare or belated during the colony formation. This partially disagrees with that found in *S. cerevisiae*, where half of the haploid cells in a colony are able to switch mating-type in any one cell division (Haber 2012). We also determined *MAT* genotype in subclones derived from pure *MAT*α and *MAT***a** colonies. We did not observe backward switching between mating-types in all CBS 732_P subclones and in CBS 732_R subclones from pure *MAT*α colonies (**Table 1**). These results could be partially biased by limited sampling size. CBS 732_R *MAT***a** colonies mainly retained *MAT***a** genotype (73.5%), but rarely underwent switching, resulting in a few number of *MAT*α/*MAT***a** (20.4%) or pure *MAT*α (6.1%) subclones. These data confirm that mating-type interconversion is rare in these clones, at least in our conditions.

Table 1. Subcolonies *MAT* genotyping of pure switched and mixed colonies.

Strain stock	N° subclones	Colony genotype	Sub-clone *MAT* genotyping (%)		
			***MAT*a/α**	***MAT*a**	***MAT*α**
CBS 732_R	20	*MAT*a/α	73.0	9.1	18.2
	49	*MAT*a	82.0	0	18.0
	22	*MAT*α	0	0	100
CBS 732_P	17	*MAT*a/α	63.6	36.4	0
	16	*MAT*a	0	100	0
	16	*MAT*α	0	0	100

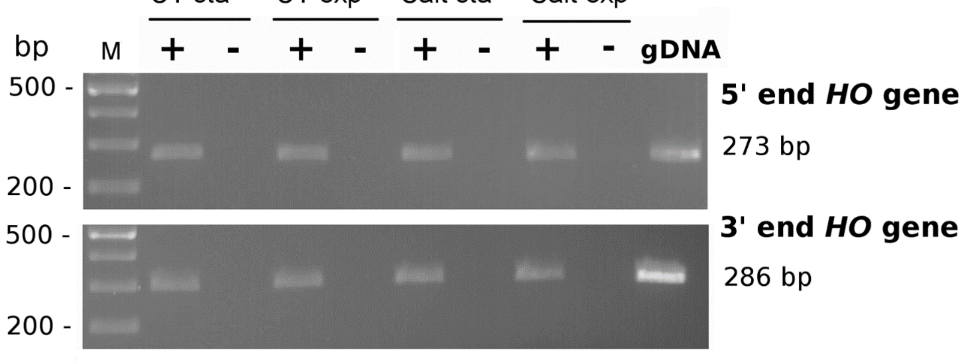
For each strain stock, at least two colonies of the indicated genotype were streaked out onto a fresh YPD plate. From each of them, a number of subclones were then PCR-tested with *MAT***a** and *MAT*α–specific primer pairs listed in Table S3. The *MAT***a**/α genotype was assigned to subclones with the *MAT***a** and *MAT*α-specific PCR signal ratio below 5-fold.

HO gene expression

Zygosaccharomyces rouxii HO conserves the eight intein motifs (termed A to H) lying at their C- and N- terminals, which form the relic of the protein-splicing domain in *S. cerevisie* HO (Solieri et al., 2014b). Considering the role of *S. cerevisiae* HO in initiating gene conversion at the *MAT* locus, we tested its expression in both CBS 732^{T} cultures. In addition to standard conditions, cells were grown under salt stress, since hyperosmotic stimuli are reported to induce switching, mating and sporulation in *Z. rouxii* (Mori and Onishi 1967; Mori 1973). Non-quantitative RT-PCR with *HO*-specific primers showed that exponentially and stationary growing cells actively transcribed HO gene in all the tested conditions (**Fig. 3**). Even if CBS 732^{T} subcultures have *HO* transcripts, colony genotyping suggests that they rarely undergo mating-type switching. These data could indicate that the *HO* expression is not crucial in triggering *MAT* switching or, alternatively, that HO is under post-transcriptional controls. To support the first hypothesis, there is the evidence that *Z. rouxii* Δ*HO* cells slightly decrease switching frequency compared to wild types (Watanabe, Uehara and Mogi 2013). By contrast, *S. cerevisiae HO* and *ho* strains exhibit difference in switching frequency of 106 order of magnitude (Hicks, Strathern and Herskowitz 1977). These lines of evidence suggest that *Z. rouxii* could switch mating-type by a mechanism partially independent from *HO* expression. In *S.*

cerevisiae, *HO* transcription is tightly regulated and its expression is detectable only in a short period of the late G1, after the mother cell has committed itself to another mitotic cell cycle (Nasmyth 1993), whereas it is repressed in diploid *MAT*α/*MAT***a** cells (Jensen et al., 1983). Moreover, *S. cerevisiae* switches mating-type both on minimal and rich medium, suggesting that *HO* expression is unaffected by stress conditions (Strathern and Herskowtiz 1979). Consequently, we expected that in an unsynchronised log-phase haploid population, half of the homothallic cells would actively express *HO* gene and switch mating-type at any one cell division, while the unbudded cells at the stationary phase would switch off HO (Strathern et al., 1982). Our results are consistent with the haploid status of CBS 732^{T} (Solieri et al., 2008) and partially disagree with the *HO* transcriptional profile described in *S. cerevisiae*. In both species, *HO* transcription is independent from the environmental conditions, but in *Z. rouxii HO* gene also escapes the growth-phase control, suggesting that this regulatory module has merged later in the evolution of hemiascomycetes or it was lost in *Z. rouxii*. In any case, functional complementation assays of *Z. rouxii HO* in *S. cerevisiae* Δ*HO* knockout mutants and GFP-tagged *HO* experiments could definitively elucidate whether HO plays a functional role in *Z. rouxii* mating-type switching.

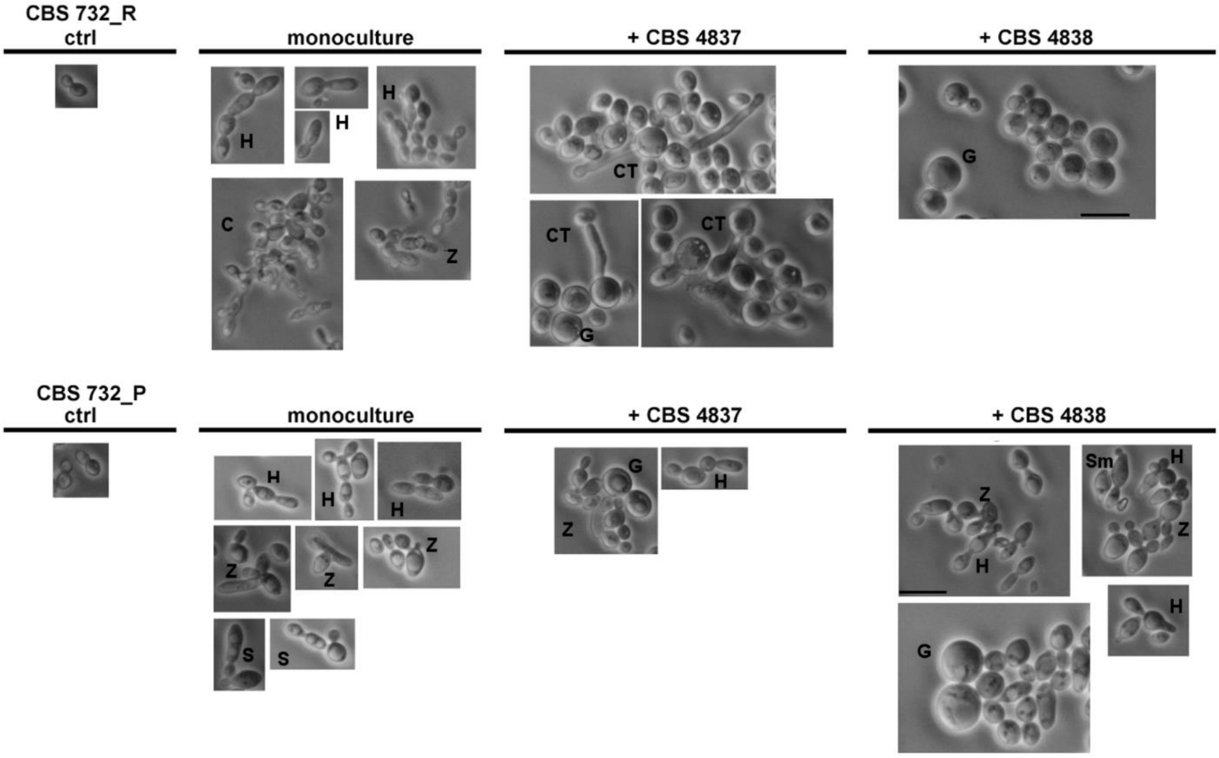
Figure 3. Expression pattern of CBS 732^{T} *HO* gene. The figure depicts amplified cDNAs generated with 5' end and 3' end of *HO* gene-specific primers from CBS 732^{T} salt-stressed and unstressed cells harvested at exponential and stationary growth phases. ± indicate reverse transcription positive and negative controls. gDNA amplification was used as positive PCR control. Abbreviations: M, molecular weight marker; CT, standard medium; salt, medium supplemented with 2.0 M NaCl; exp, exponential phase; sta, stationary phase.



Morphology and fertility assay

We checked the morphology and the mating ability of CBS 732_R and CBS 732_P grown as monoculture or in mixing with CBS4837 (**a**) or CBS 4838 (α) mating testers. **Figure 4** depicted the main morphological changes observed in fertility assays, while **Table S5** reported the percentage fractions (%) of cells exhibiting morphologies deviating from canonical ovoid budding cells. After 1 week of incubation, cells from single cultures were larger and hyperelongated on all test media compared to the control round-oval cells (**Fig. 4**). These round or moderately elongated cells often displayed hyperelongated buds. Budding pattern was difficult to determine by visual inspection owing to the clumping manifested by these strains, mainly in salt medium. However, CBS 732_R and CBS 732_P seemed to shift from axial to distal budding in test media relative to control condition, resulting in a filamentous-like growth. These morphological changes resembled those seen during pseudohyphal growth (Cullen and Sprague, 2012) or, alternatively, during chemotropic response to lowpheromone gradient (Erdman and Snyder 2001; Paliwal et al., 2007). Nonetheless, we did not find any rough and wrinkled colony correlated to pseudohyphae and invasive growth phenotype (Palkova and Vachova, 2006). Furthermore, we frequently observed that bud emergence in mother cells initiated in advance compared to daughter cells (**Fig. 4**). This budding pattern is distinct from that observed in cells forming pseudohyphal chains, where mother and daughter cells bud synchronously (Erdman and Snyder 2001). Therefore, hyperelongated morphology could be a signal of chemotropism due to a mixture of *MAT*α and *MAT***a** cells. In *Z. rouxii*, zygote resulting from mating produces new buds at either one or both ends, which remain connected to the zygotic mother (Mori 1973). Therefore, differently from *S. cerevisiae*, in *Z. rouxii* mating events were difficult to discern from filamentous-like growth. Nevertheless, a few zygotes emerging from the conjugation bridge were observed, mainly when CBS 732_R and CBS 732_P monocultures were grown on salt medium (**Fig. 4**). This is consistent with the previous observation that salt induces mating (Mori and Onishi 1967). However, we did not observe any conjugative asciwith dumbbell shape, as generally described after successful mating in *Zygosaccharomyces* spp. CBS 732_R and CBS 732_P behaved differently when mixed with CBS 4837 or CBS 4838 testers (**Fig. 4**). CBS 732_P cells exhibit hyperelongated morphology in response to CBS 4837 or CBS 4838. We observed both elongated and round shape cells in which the buds had an elongated or peanut shape. Shmoos were observed in co-culture with CBS 4838 and zygotes resulting from shmoo fusion were visible in co-cultures with CBS 4837 or CBS 4838. In both assays, giant round shape cells were also present. Mori (1973) described large cell clones as a result of mating followed by karyogamy between *MAT*α and *MAT***a** haploid cells or *MAT*α/*MAT*α and *MAT***a**/*MAT***a** diploid homozygous cells. Similarly, these giant cells could arise from budded zygotes after mating and karyogamy. Unlike CBS 732_P, CBS 732_R did not exhibit hyperelongated morphology in mixture with CBS 4837 or CBS 4838 (**Fig. 4**). Instead, cells were mainly round ovoid in all media tested and some giant round cells were also detected. In the presence of CBS 4837, starting from4 days, we rarely observed abnormal long projections (termed as conjugation tubes, CT) emanating from a round-like cell body (**Fig. 4**). These structures were much longer than the *S. cerevisiae* shmoo, and resembled those formed by *Candida albicans MTL***a**/*MTL***a** in mixture with *MTL*α/*MTL*α opaque-phase cells or in response to α signals (Miller and Johnson 2002; Daniels et al., 2003; Lockhart et al., 2003). Furthermore, many of these tubes seemed to form buds at the tube apices, presumably because no tubes of the opposite mating-type were encountered and/or pheromone was rapidly degraded by secreted aspartyl proteinases. After 14 days, the buds at the tube apices remarkably grew and giant cells were more abundant than after 7 days (17.01% vs 9.60%). Conjugation tubes were rarely produced by CBS 4837 (**a**) in monoculture (data not shown) and seemed to be inhibited by CBS 732_P. Why no conjugation tubes were observed when CBS 4837wasmixed with CBS 732_P has not been understood yet. These abnormal structures deserve further investigations, especially given that CBS 4837 genome has been recently described as allodiploid (Sato et al., 2017). This could mean that strains conventionally described as mating testers, such as CBS 4837 and CBS 4838 (Mori and Onishi, 1967), could have uncertain ploidy identity due to unpredictable interactions between two partially introgressed parental genomes (Bizzarri et al., 2016).

Figure 4. Fertility assay. The panels show selected differential phase-contrast microscopic images of CBS 732_R and CBS 732_P monocultures and mating mixtures. Abbreviations: H, hyperelongated cells forming filamentous-like chains and single cells with elongated buds; C, clumped cells in 6%NaCl-MEA medium (see **Table S2**); Z, conjugation bridge with emerging zygote; Sm, shmoo projections (single cells with mating projection); G, giant cells; S, putative conjugated asci with two spores; CT, abnormal conjugation tubes. Budded cells grown in YPD for 24h were used as control (ctrl) and put at the left. Scale bars represent 11 µm.



Concluding remarks

In conclusion, this study provides a model for mating-type switching in haploid homothallic yeast, which merged from the common ancestor of the Saccharomycotina subphylum, after the gain of *HO* endonuclease. We obtained isogenic, pure *MAT*α and *MAT***a** cultures suitable for breeding programs and pheromone based studies on cell-to-cell communication. In our model, novelty was generated by new alternative *MAT***a**2 copies, which share MATA-HMG-box but differ in cp-tags. Markedly, *Z. rouxii* regulates the **a**sgs by a hybrid regulatory circuit controlled by both **a**2 (activation) and α2 (repression) (Baker et al., 2012). One interesting questions is whether different cp-tags affect the affinity of **a**2 transcriptional factor towards gene target promoters. Similarly, the cell-cycle relaxation on *HO* transcription suggests a difference between *Z. rouxii* and *S. cerevisiae* on the role of *HO* in mating-type switching, which deserves further studies. Despite being preliminary, morphological heterogeneity between the CBS 732_R and CBSR 732_P suggests that mating-type switching could have consequences on phenotype beyond the *MAT* interconversion. *Zygosaccharomyces rouxii* homothallic haploid cells seem to elude the assumption that subcultures originating from the same strain should be considered genotypically and phenotypically identical. Indeed, recombination at the *MAT* locus could produce intrastrain genotype variation and lead to genetic instability and phenotypic novelties inside the progeny cells upon which selection can act.

Acknowledgements

Prof. Hana Sychrovà is highly acknowledged for providing the CBS 732_P derived culture. We thank Prof. Paolo Giudici for helpful discussion and comments on the manuscript.

Funding

This work was partially supported by AEB spa (Brescia, Italy).

Author Contributions

Conceived and designed the experiments: LS SC. Performed the experiments: LS MB SC. Analyzed the data: LS MB SC. Contributed reagents/materials/analysis tools: SC. Wrote the paper: LS MB.

CHAPTER 5: SEX DETERMINATION SYSTEM IN THE ALLODIPLOID AND STERILE *ZYGOSACCHAROMYCES ROUXII* ATCC 42981 YEAST

Bizzarri, M., Giudici, P., Cassanelli, S., and Solieri, L. (2016). Chimeric sex-determining chromosomal regions and dysregulation of cell-type identity in a sterile Zygosaccharomyces allodiploid yeast. PloS One. 11(4), e0152558. doi:10.1371/journal.pone.0152558.

Abstract

Allodiploidization is a fundamental yet evolutionarily poorly characterized event, which impacts genome evolution and heredity, controlling organismal development and polyploid cell-types. In this study, we investigated the sex determination system in the allodiploid and sterile ATCC 42981 yeast, a member of the *Zygosaccharomyces rouxii* species complex, and used it to study how a chimeric mating-type gene repertoire contributes to hybrid reproductive isolation. We found that ATCC 42981 has 7 *MAT*-like (*MTL*) loci, 3 of which encode α-idiomorph and 4 encode **a**-idiomorph. Two phylogenetically divergent *MAT* expression loci were identified on different chromosomes, accounting for a hybrid **a**/α genotype. Furthermore, extra **a**-idimorph-encoding loci (termed *MTL***a** copies 1 to 3) were recognized, which shared the same *MAT***a**1 ORFs but diverged for *MAT***a**2 genes. Each *MAT* expression locus was linked to a *HML* silent cassette, while the corresponding *HMR* loci were located on another chromosome. Two putative parental sex chromosome pairs contributed to this unusual genomic architecture: one came from an as-yet-undescribed taxon, which has the NCYC 3042 strain as a unique representative, while the other did not match any *MAT-HML* and *HMR* organizations previously described in *Z. rouxii* species. This chimeric rearrangement produces two copies of the *HO* gene, which encode for putatively functional endonucleases essential for mating-type switching. Although both **a** and α coding sequences, which are required to obtain a functional cell-type **a**1-α2 regulator, were present in the allodiploid ATCC 42981 genome, the transcriptional circuit, which regulates entry into meiosis in response to meiosis-inducing salt stress, appeared to be turned off. Furthermore, haploid and α-specific genes, such as *MAT*α1 and *HO*, were observed to be actively transcribed and up-regulated under hypersaline stress. Overall, these evidences

demonstrate that ATCC 42981 is unable to repress haploid α-specific genes and to activate meiosis in response to stress. We argue that sequence divergence within the chimeric **a**1-α2 heterodimer could be involved in the generation of negative epistasis, contributing to the allodiploid sterility and the dysregulation of cell identity.

Introduction

Ploidy variation has played a major role in the evolution of many extant eukaryotic lineages (Van de Peer, 2009). In yeasts, numerous studies have demonstrated how variations in the ploidy state took place frequently during evolutionary history (Kellis et al., 2003; Scannell et al., 2007). Allodiploid offspring often have strong selective disadvantages due to their sterility (Johnson, 2010). However, in some instances, the increased genome size and complexity of allodiploids may enhance heterosis and/or adaptive flexibility (Riesberg et al., 2003), particularly at the edges of the ancestral species' range, where they are more likely to encounter stress (Alipaz et al., 2005; Otto and Gerstein, 2008; Nolte and Tautz, 2010). Life history models note that there is an intricate interplay between ploidy variation and alterations of mating, meiosis and sporulation patterns (Zorgo et al., 2013). Consequently, the ploidy state affects the genetic composition at sex-determining loci, giving rise to the ploidy-dependent initiation of dedicated transcriptional programs (Heitman et al., 2009).

In the well-studied model organism *Saccharomyces cerevisiae*, the meiosis of diploid cells is triggered by environmental cues, such as nutrient depletion, and gives rise to four haploid meiotic spores with two distinct mating-types (*MAT***a** and *MAT*α). Haploid spores eventually re-establish diploid *MAT***a**/*MAT*α lines by one of three processes: (1) by mating with their own mitotic daughter cells after switching their mating-type in a process catalysed by the endonuclease *HO* (HOmothallism); (2) by mating with another spore created by the same meiotic event (intratetrad mating); or, more rarely, (3) by mating with an unrelated individual (outcrossing) (Knop, 2006). The **a**1-α2 protein heterodimer is one of the master regulators of cell identity and sexual development (as reviewed in Haber, 2012 and Granek et al., 2011, respectively). The **a**1 and α2 proteins are homeodomain transcriptional factors encoded by the *MAT***a**1 and *MAT*α2 genes at the active *MAT***a** and *MAT*α loci, respectively. In *MAT***a**/*MAT*α diploid cells, α2 interacts with **a**1 to bind DNA as a heterodimer and transcriptionally repress mating genes, preventing polyploidy and aneuploidy.

Furthermore, the **a**1-α2 heterodimer inhibits the expression of haploid-specific genes (h-sgs) and activates the developmental switch from mitosis to meiosis under appropriate stress stimuli (Goutte and Johnson, 1988; Herskowitz, 1988). Two h-sg targets of negative regulation by the **a**1-α2 heterodimer are *MAT*α1 (encoding the transcription factor that positively regulates the α-specific genes) and *HO* genes (encoding an endonuclease, which cleaves a specific DNA sequence at the *MAT* locus during the first step of mating-type switching) (Mathias et al., 2004). Furthermore, **a**1-α2 heterodimer controls the sexual development by inhibiting the *RME1* (*Repressor of Meiosis* 1) gene. In haploid cells, the RME1 transcriptional factor prevents the expression of the *IME1* (*Inducer of Meiosis* 1) gene (Covitz et al., 1991; Mitchell and Herskowitz, 1986) and positively regulates the transcription of adhesion-specific genes, such as the *FLO11* gene, in response to nutrient depletion (Van Dyk et al., 2003). In diploid cells, the **a**1-α2 heterodimer binds to the *RME1* gene promoter, repressing its transcription and thereby relieving *IME1* gene repression. The **a**1-α2 repressor complex directly controls entry into meiosis through the regulation of the *IME4* gene, which is required for the full expression of *IME1* (Shah and Clancy, 1992). The **a**1-α2 heterodimer negatively regulates the antisense (AS) long non-coding RNA (lncRNA) (Hongay et al., 2006; Gelfand et al., 2011). This lncRNA (also termed as *Regulator of Meiosis* 2 or *RME2*) is transcribed in haploid cells and blocks the expression of the *IME4* gene. In diploid cells, the **a**1-α2 complex represses *RME2* expression, allowing sense (S)-*IME4* to be transcribed under starvation conditions.

Yeasts of the non-whole genome duplication (non-WGD) *Zygosaccharomyces rouxii* species complex are adapted to grow in food with high solute concentrations and frequently experience variations in ploidy, resulting in different modes of reproduction (Solieri et al., 2013a). This complex includes the haploid heterothallic or homothallic (self-fertile) *Z. rouxii* species, which undergoes mating and subsequent meiosis under salt stimuli (Solieri et al., 2013; Mori, 1973); the diploid *Zygosaccharomyces sapae* species, which reproduces mainly by clonality and ascospores are rarely observed (Solieri et al., 2013b); and a group of anueploid/allodiploid strains of unclear taxonomical position (termed mosaic lineage) (Solieri et al., 2013a; Solieri et al., 2013b). Within the latter, the allodiploid ATCC 42981 strain has been extensively studied for its ability to withstand high concentrations of alkali metal cations and for its capability to produce glycerol under salt stress (Solieri et al., 2014a; Pribylova et al., 2007). From a molecular point of view, the ATCC 42981 genome displays rDNA heterogeneity (Solieri et al., 2008; Gordon and Wolfe, 2008), additional chromosomes compared to *Z. rouxii* (Pribylova et al., 2007), and diploid DNA (Solieri et al., 2008). Gordon and Wolfe showed that the ATCC 42981 genome contains two partially divergent complements, namely the T- and P- subgenomes, which could arise from a recent allodiploidization event between *Z. rouxii* carrying the T- subgenome and another not-yet-recognized species harbouring the P- subgenome (so far represented by the unique strain NCYC 3042 and referred to as *Zygosaccharomyces pseudorouxii* nom. inval. by James et al., 2005). Despite having a DNA diploid content, the ATCC 42981 strain has not been observed to undergo meiosis under different standard and stress growth conditions (Solieri et al., 2013b).

Because gene information retained at the *MAT* expression loci is essential to ensure appropriate haploid/diploid cell-type identity and functional cell development, several efforts have recently been made to characterize *MAT* loci in haploid *Z. rouxii* (Watanabe et al.,2013) and diploid *Z. sapae* (Solieri et al., 2014b), but not in the allodiploid ATCC 42981 strain. Like *S. cerevisiae*, haploid *Z. rouxii* wild strains possess a three-cassette system consisting of *MAT***a** or *MAT*α expression loci and two silent cassettes of both idiomorphs, *HMR* and *HML,* respectively, which act as donor sequences during the mating-type switching presumably catalysed by HO endonuclease (Butler et al., 2004; Souciet et al., 2009). *S. cerevisiae* maintains both *HML* and *HMR* cassettes at different locations on the same chromosome harbouring the *MAT* expression locus, whereas *Z. rouxii* only conserves the *MAT*/*HML* linkage on chromosome C and the *HMR* locus on chromosome F (Souciet et al., 2009). This structural organization implies that, differently from *S. cerevisiae*, *Z. rouxii* exploits the ectopic recombination between non-homologous chromosomes to switch the mating-type. Congruently, Watanabe et al., (2013) found that sex chromosomes represent hyper-mutational hotspots, with flanking regions of *MAT*, *HML* and *HMR* cassettes remaining highly variable among *Z. rouxii* haploid strains. Compared to *Z. rouxii*, the *Z. sapae* diploid species has one more copy of the *HO* gene and displays an unusual **a**ααα genotype, with a redundant number of divergent *MAT*α loci but without any *HMR* silent locus. This complex architecture of sex-determining chromosomal regions prevents mating-type switching due to the lack of the *HMR* cassette and hampers sexual development due to the imbalance in divergent mating-type genes.

Here, we determined the structure and functions of the three-locus sex-determining system in the allodiploid ATCC 42981 strain. The transcriptional profiling of these genes under meiotic-inducing salt stress revealed how the hybrid genetic configuration at the *MAT* loci contributes to allodiploid sterility. To the best of our knowledge, our report presents the first evidence that a chimeric sex-determination system accounts for the incomplete silencing of the h-sg program and contributes to the prevention of switching from mitosis to meiosis in allodiploid yeasts.

Materials and Methods

Strains, culture conditions and mating test

The *Zygosaccharomyces* strains used in this work are listed in **Table 1**. Strains were routinely cultured at 28°C in YPD (1% w/v yeast extract, 1% w/v peptone, 2% w/v dextrose) medium with or without 15% (w/v) agar and stored at 4°C. For long-term preservation, strains were stored at -80°C in YPD medium containing 25% glycerol (v/v) as a cryopreservative. To increase the probability of observing conjugated or un-conjugated spore-containing ascii, sporulation was tested by inoculating the early stationary phase culture of ATCC 42981 on five different media [YPD, YPDA, malt extract agar (MEA; Difco), MEA supplemented with 6% (w/v) NaCl (6%NaCl-MEA), and YNB5%GNaCl (1% w/v yeast extract, 5% w/v dextrose, 6.7 g/l yeast nitrogen base, 2.0 M NaCl)] for 3 weeks. To study sexual compatibility, 2-to-4-day-old cultures of the ATCC 42981 strain were incubated alone or in a mixture with mating-tester strains *Z. rouxii* CBS 4837 (mating-type **a**) or CBS 4838 (mating-type α) in both MEA and 6%NaCl-MEA media at 27°C for 3 weeks. Samples were examined microscopically every week using phase-contrast optics to detect conjugation. The YNB5%G [1% w/v yeast extract, 5% w/v dextrose, 6.7 g/l yeast nitrogen base (Difco)] and the YNB5%GNaCl (1% w/v yeast extract, 5% w/v dextrose, 6.7 g/l yeast nitrogen base, 2.0 M NaCl) media were used for RNA extraction from cells grown under unstressed and salt-stressed conditions, respectively.

Table 1. Details of strains used in the present study. Ploidy data were obtained from Solieri et al., 2013a; 2008, while the genotype of strain ABT301^{T} at the active *MAT* loci was obtained from Solieri et al., 2014b. Abbreviations: nd, not determined; TBV, Traditional Balsamic Vinegar.

Strains	Other collections	Source	Current taxonomical positions	Mating-type/thallism	Spore	Ploidy ratio
CBS 732^{T}	NCYC 568, NRRL Y-229	Grape must	*Z. rouxii*	*MAT*α*/*homothallic	-	1.3
CBS 4837	NYC 1682, NRRL Y2547	Miso	mosaic lineage	*MAT***a**/heterothallic	+	1.96
CBS 4838	NRRL Y2584	Miso	mosaic lineage	*MAT*α/heterothallic	+	1.90
ATCC	-	Miso	mosaic lineage	nd	nd	2.1
ABT301^{T}	CBS 12607,	TBV	*Z. sapae*	**a**ααα	+	2.0
NCYC 3042	CBS 9951	Soft drink	*Z. pseudorouxii*	nd		nd

PCR conditions and sequencing

Genomic DNA (gDNA) extraction was performed by a phenol-based method from stationary grown cells after mechanical lysis according to Hoffman and Winston [35]. The quantity of DNA was evaluated spectrophotometrically using a NanoDrop ND-1000 device (Thermo Scientific). All PCR reactions were performed on a T100 Thermalcycler (Bio-Rad) in a 25-µl reaction volume containing 200 ng of template gDNA. For PCR amplification < 2 Kb rTAQ DNA polymerase (Takara, Japan) and for PCR amplification ≥ 2 Kb Phusion Hot Start II Polymerase (Thermo Scientific, Waltham, MA) along with buffer HF (5x) or LA Taq DNA polymerase (Takara, Japan) with GC buffer I (2x), were used according to the manufacturer's instructions. All primers used in this study are listed in **Table S1** (see **supplememtary materials**); primers were designed with Primer 3 software (http://primer3.sourceforge.net/) and provided by either MWG (Heidelgerg, Germany) or BMR Genomics (Padova, Italy). The PCR products were resolved on 1.2% (w/v) agarose gels stained with ethidium bromide and their size was estimated by comparison with 100 bp or 1 Kb DNA Ladder Plus (Fermentas, USA) as molecular size markers. PCR products were purified using the DNA Clean & Concentrator^{TM}-5 (DCC^{TM}-5) Kit (Zymo Research) according to the manufacturer's instructions. Moreover, when required, DNA fragments were purified from 1% agarose gels using the Gene JET Gel Extraction Kit (Thermo Scientific, Waltham, MA). Finally, all PCR products were sequenced on both strands through a DNA Sanger sequencing process performed by either MWG (Heidelgerg, Germany) or BMR Genomics (Padova, Italy).

Genomic walking procedure

The overall strategy for determining mating-type-like (*MTL*) loci is depicted in **Fig. S1** (**panel A**). Briefly, three *MAT*α copy-specific primer pairs were designed on the *ZsMTL*α copies 1, 2 and 3 cassettes previously cloned in *Z. sapae* and used to clone the corresponding *MAT*α1 (partial) and *MAT*α2 (full-length) coding regions in the ATCC 42981 genome (**Fig. S1 and Table S1**). Complete *MAT*α1 loci were obtained in a second round of PCR walking using reverse primers (A, B, C and A_D) built on the 3'-end regions potentially flanking the *MAT* and *HML*/*HMR* cassettes (Watanabe et al., 2013; Solieri et al., 2014b) (**Fig. S1**). Similarly, three primer pairs were used to amplify putative *MTL***a** loci in the ATCC 42981 strain (**Fig. S1, panel** B). Two primer pairs were designed to specifically amplify the *ZsMAT***a**1 and *ZsMAT***a**2 genes, respectively, whereas a third primer pair was built to completely include *MAT***a**1 ORF and partially *MAT***a**2 ORFs, respectively. In order to further extend the *MAT***a**2 coding DNA sequence, a second round of PCR walking was performed by combining a *MAT***a**2-specific reverse primer and all the available forward primers (1, 2 and 3) built on the 5’-end regions flanking the *MAT* and *HML*/*HMR* cassettes (Watanabe et al., 2013) (**Fig. S1**). The DNA regions flanking *MTL***a** and *MTL*α loci were characterized through the semi-nested and direct PCR approaches as previously reported (Solieri et al., 2014b) (**Fig. S2**). The overview of the strategy employed for ATCC 42981 *HO* gene characterization is summarized in **Figure S3**, and a detailed primer list is given in **Table S1**.

Bioinformatics analyses

Sequences were assembled and edited using DNAStar Software (DNASTAR, Inc. Madison, Wisconsin USA). Multiple nucleotide and amino acid sequence alignments were performed using Clustal W2 (Larkin et al., 2007). Searches for nucleotide and protein sequence homologs were carried out in the GenBank database with Blastn and Blastp algorithms, respectively [37]. Phylogenetic analysis was conducted on aa sequences using MEGA6 (Tamura et al., 2013). The phylogenetic relationship was inferred using the neighbour-joining (NJ) method. Support percentages for the nodes of NJ-trees were computed using bootstrapping analysis with 1,000 replications and were shown next to the branches when ≥ 60%. For domain identification, Pfam-searches (Finn et al., 2014) were run on http://pfam.sanger.ac.uk/. Structure predictions were obtained with Jpred3 (Cole et al., 2008) and validated according to Martin et al., 2010. Sequences were submitted to the EMBL/GenBank databases under the accession numbers from KT598024 to KT598027 and from KT694298 to KT6942302.

PFGE-Southern blotting assays

Chromosomal DNA preparation in plug, pulsed-field gel electrophoresis (PFGE), and Southern blotting assays were performed as previously reported (Solieri et al., 2008; Solieri et al., 2014b). Primers engaged in probe synthesis for Southern blot analysis are listed in **Table S1**.

Subgenome assignment of mating-type and *HO* gene copies

PCR subgenome genotyping was performed using gDNA extracted from Z. *pseudorouxii* NCYC 3042 as a T- subgenome template. PCR amplification of all *MAT*α1, *MAT*α2, *MAT***a**1, *MAT***a**2, and *HO* gene copies present in ATCC 42981 genome was carried out with the primer pairs listed in **Table S1**. To avoid false negative results, we tested two alternative primer pairs for each target gene. ATCC 42981 gDNA was also included as a positive control. *Z. rouxii* CBS 732^{T} gDNA was not included in PCR reactions as one of the two putative parental genetic complements because its genome project was available.

RNA extraction and RT-PCR

Zygosaccharomyces cells were pre-cultured in YPD medium for 24 h at 28°C under shaking conditions (150 rpm), washed in physiological solution (9 g/l NaCl), and used to inoculate two sets of baffled Erlenmeyer flasks (E-flasks) containing 70 ml of YNB5%G (standard growth condition) and YNB5%GNaCl (hyperosmotic growth condition) media, respectively (initial OD_{600nm} 0.02-0.04). Inoculated E-flasks were incubated at 28°C under shaking conditions (150 rpm) and yeast growth was spectrophotometrically monitored at 600 nm two times a day. Cells at the stationary phase (no change in OD measurement detected in at least three consecutive readings) were harvested and frozen at -80°C. RNA extractions were carried out using the ZR Fungal/Bacterial RNA MicroPrep^{TM} Kit (ZymoReasearch, Irvine, California) according to the manufacturer's instructions. Purified RNAs were submitted to additional DNase digestion in solution and cleaned up with the RNA Clean & Concentrator^{TM}-5 Kit (ZymoReasearch, Irvine, California).

Total RNAs were reverse transcribed using 0.5 µM oligo (dT) and RevertAid H Minus Reverse Transcriptase (Thermo Scientific, Waltham, USA) according to the manufacturer's instructions. In the case of *IME4* transcript analyses, instead of oligo (dT) the forward and reverse primers ZrIME4F1 and ZrIME4R1 were used for the cDNA synthesis of AS-*IME4* lncRNA and S-*IME4* mRNA, respectively (**Table S1**). Then a common pair of internal primers (ZrIME4_F2/ZrIME4_R2; **Table S1**) was used to amplify the same cDNA fragment from each template, if any. Experiments were carried out with three biological replicates and non-reverse transcribed (NRT) controls were performed for each biological replicate. cDNA template (25 ng) was PCR-amplified with primers listed in **Table S1** and successful amplification was checked by electrophoresis in 2% agarose gel in 0.5X TBE Buffer.

Quantitative PCR (qPCR) assays

For relative expression level analysis, qPCR reactions were performed using 1 ng/µl of cDNA, 0.3 µM of each primer, and the Maxima^{TM} SYBR Green/ROX qPCR Master Mix (Fermentas, USA) according to the manufacturer's instructions. Primers were designed to selectively amplify *MAT***a**1 exons as well as different copies of *MAT*α1, *MAT*α2 and *HO* genes (**Table S1**). PCR efficiency was *in silico* predicted for each primer set using the open source tool Primer Efficiency (http://srvgen.upct.es/index.html; Mallona et al., 2011). Preliminary, the presence and the estimated size of amplicons were checked by RT-PCR on cDNA from stationary-phase cells (three biological replicates) under standard conditions, whereas unspecific amplifications and primer-dimer formation were checked by melting curve analysis after RT-qPCR assay. All qPCR reactions were run in the Applied Biosystems 7300 Real-Time PCR instrument (Applied Biosystem, Foster City, CA, USA). The *Z. rouxii* housekeeping gene *ACT1* (Zr*ACT1*; XM002497273) was used as a reference gene, according to Leandro et al., 2013. The relative expression of different gene transcripts was calculated by the ΔΔCt method and converted to the relative expression ratio (2−ΔΔCt) for statistical analysis (Livak and Schmittgen, 2001). For this purpose, a dilution series of standard points from pooled control cDNAs was exploited (concentration range 10-0.02 ng/µl) in technical triplicates. Amplification profiles, baselines and thresholds were analysed with 7300 SDS 1.4. PCR-Miner (Zhao and Fernald, 2005; Ruijter et al., 2009) was applied as an alternative method when the low gene expression level did not provide enough dynamic range to build a reliable standard curve. In this case, reaction efficiency and the fractional cycle number at threshold (Ct) were estimated by relying on the kinetics of individual PCR reactions containing 5 ng of cDNA template from three biological replicates (salt-treated and controls), including six technical replicates each.

Results

Morphological characterization and mating test

We first assessed the mating behaviour of the ATCC 42981 strain in pure and mixed cultures with the *Z. rouxii* mating partners CBS 4837 (mating-type **a**) and CBS 4838 (mating-type α), respectively. Examination under the microscope did not show any evidence of mating reaction of ATCC 42981 cells neither with *Z. rouxii* CBS 4837 or CBS 4838 tester strains, even after 3 weeks of incubation both on MEA and 6% NaCl-MEA media (data not shown). Even if rare mating events cannot be ruled out, this result suggests that the ATCC 42981 strain did not respond to *Z. rouxii* pheromone signalling or that there was an imbalanced or deficient organization in *MTL* loci. Based on our previous observations (Solieri et al., 2013b), ATCC 42981 cells grown on MEA medium displayed an adhesive phenotype with clamps on mother and daughter cells that remained attached to each other, but they were not able to form ascii in pure culture (data not shown). In contrast, *Z. sapae* ABT301^T rarely forms ascii at the same conditions (Solieri et al., 2013b). As salt has been reported to induce meiosis in *Zygosaccharomyces* yeasts (Mori, 1973), we tested the capability of the ATCC 42981 strain to form ascospores in YNB5%G NaCl and 6%NaCl-MEA media. No conjugated ascii were observed over time in any media tested. Furthermore, an increase in the adhesive phenotype was detected in cells grown under salt stress (data not shown).

Isolation and characterization of *MTL*α loci in ATCC 42981

Three *MTL*α loci, termed *ZsMTL*α copies 1 to 3, have been previously described in *Z. sapae* as regions containing phylogenetically distinct α1 and α2 genes, respectively (Solieri et al., 2014b). To establish whether strain ATCC 42981 also displays a similar genetic configuration, we exploited a PCR walking strategy. This assay relies on *MAT*α copy-specific primers designed on the corresponding *ZsMTL*α cassettes (Solieri et al., 2014b). PCR reactions were positive using *MTL*α copies 1 and 2-specific primers, whereas negative results were scored for any *MTL*α copy 3-specific primer set tested. We sequenced two distinct *MTL*α loci (hereafter referred to as *MTL*α copies 1 and 2) in ATCC 42981. Both loci consisted of two *MAT* genes encoding α1 and α2 proteins placed in opposite directions and separated by an intervening region of 343 bp. The *MAT*α1 genes from *MTL*α loci 1 and 2 (termed *MAT*α1 copies 1 and 2, respectively) were divergent from each other for 17 nt substitutions, whereas the *MAT*α2 from *MTL*α loci 1 and 2 (termed *MAT*α2 copies 1 and 2, respectively) displayed 16 nt substitutions. The deduced proteins MATα1 and MATα2 copies 1 were 87.00% and 78.67% identical to the deduced protein MATα1 and MATα2 copies 2, respectively. BLASTp search against the NCBI database showed that proteins MATα1 and MATα2 copies 1 are more similar to *Z. rouxii* orthologs (termed ZrMATα1 and ZrMATα2) than the paralogous proteins MATα1 and MATα2 copies 2 in the ATCC 42981 genome.

The NJ-based phylogenetic analysis was carried on MATα1 sequences from representative WGD and non-WGD species. As expected, ATCC 42981 MATα1 copy 2 protein did not group to *Z. rouxii* MATα1, but was instead clustered separately (bootstrapping value of 99%) together with ZsMATα1 copy 2 (**Fig. 1**). The alignment of ATCC 42981 MATα1 copies with *Z. rouxii* and *S. cerevisiae* MATα1 proteins revealed a region of high similarity inside the MATα-HMG domain (Martin et al., 2010) with three predicted conserved α helices (data not shown).

Phylogeny inferred from the MATα2 aa sequences of WGD and non-WGD species showed a tree topology congruent with the species relationships established using MATα1 sequences. The ATCC 42981 genome harbours two MATα2 variants that are related but phylogenetically distinct because of the high level of amino acid divergence (**Fig. 2**). In particular, ATCC 42981 MATα2 copy 1 clustered with *Z. rouxii* MATα2 (bootstrapping 100%), whereas MATα2 copy 2 was strictly related to *Z. sapae* MATα2 copy 2 (bootstrapping 100%). Both copies contained a conserved HD1 class homeodomain, consisting of a three-helix globular domain which contacts both major groove bases and the DNA backbone (Wolberger et al., 1991; Kues and Casselton, 1992) (data not shown). However, portions of the protein outside the homeodomain, which mediate interactions with accessory proteins, had a different degree of conservation.

Figure 1. Phylogenetic analysis of MATα1 proteins. The neighbour-joining (NJ) tree shows the phylogenetic relationships between the allodiploid strain ATCC 42981 and other hemiascomycetes inferred from MATα1 proteins. Numbers on branches indicate bootstrap support percentages (1,000 pseudoreplicates) higher than 60% from NJ. The red branch indicates the *Z. rouxii* yeast complex, which includes *Z. rouxii* MATα1, ATCC 42981 MATα1 copies 1 and 2, *Z. sapae* MATα1 copies 1, 2, and 3 sequences. The dark dot indicates WGD species, whereas the dark triangle indicates non-WGD species with the *HO* gene.

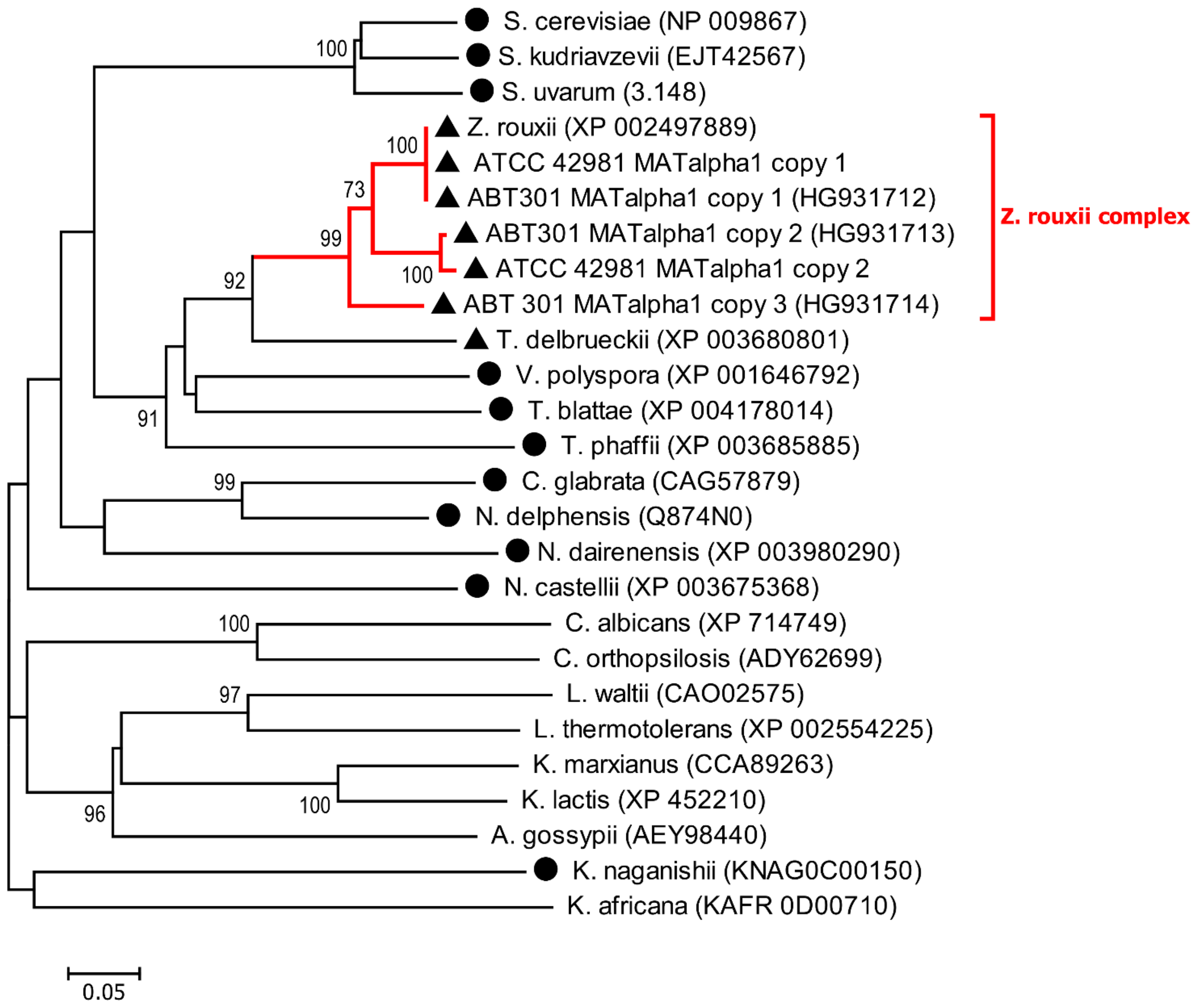
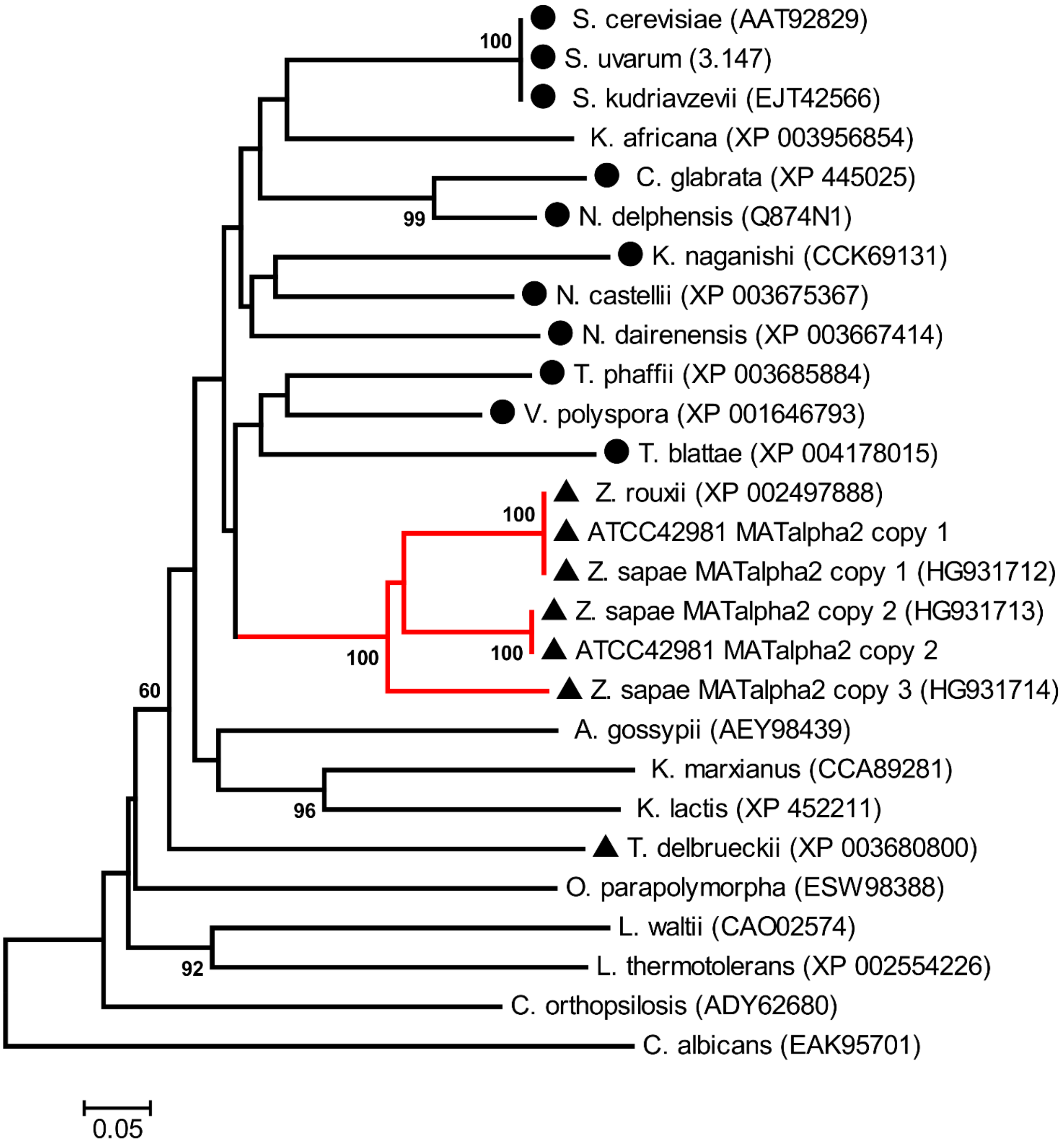

● S. cerevisiae (NP 009867)
● S. kudriavzevii (EJT42567)
● S. uvarum (3.148)
▲ Z. rouxii (XP 002497889)
▲ ATCC 42981 MATalpha1 copy 1
▲ ABT301 MATalpha1 copy 1 (HG931712)
▲ ABT301 MATalpha1 copy 2 (HG931713)
▲ ATCC 42981 MATalpha1 copy 2
▲ ABT 301 MATalpha1 copy 3 (HG931714)
▲ T. delbrueckii (XP 003680801)
● V. polyspora (XP 001646792)
● T. blattae (XP 004178014)
● T. phaffii (XP 003685885)
● C. glabrata (CAG57879)
● N. delphensis (Q874N0)
● N. dairenensis (XP 003980290)
● N. castellii (XP 003675368)
C. albicans (XP 714749)
C. orthopsilosis (ADY62699)
L. waltii (CAO02575)
L. thermotolerans (XP 002554225)
K. marxianus (CCA89263)
K. lactis (XP 452210)
A. gossypii (AEY98440)
● K. naganishii (KNAG0C00150)
K. africana (KAFR 0D00710)
Z. rouxii complex
100
100
73
99
100
92
91
99
100
97
100
96
0.05

Figure 2. Phylogenetic analysis of MATα2 proteins. The neighbour-joining (NJ) tree shows the phylogenetic relationships between the allodiploid strain ATCC 42981 and other hemiascomycetes inferred from MATα2 proteins. Numbers on branches indicate bootstrap support percentages (1,000 pseudoreplicates) higher than 60% from NJ. The red branch indicates *Z. rouxii* complex, which includes *Z. rouxii* MATα2, ATCC 42981 MATα2 copies 1 and 2, and *Z. sapae* MATα2 copies 1 to 3 sequences. The dark dot indicates WGD species, whereas the dark triangle indicates non-WGD species with the *HO* gene.


100
● S. cerevisiae (AAT92829)
● S. uvarum (3.147)
● S. kudriavzevii (EJT42566)
K. africana (XP 003956854)
● C. glabrata (XP 445025)
99
● N. delphensis (Q874N1)
● K. naganishi (CCK69131)
● N. castellii (XP 003675367)
● N. dairenensis (XP 003667414)
● T. phaffii (XP 003685884)
● V. polyspora (XP 001646793)
● T. blattae (XP 004178015)
▲ Z. rouxii (XP 002497888)
100
▲ ATCC42981 MATalpha2 copy 1
▲ Z. sapae MATalpha2 copy 1 (HG931712)
▲ Z. sapae MATalpha2 copy 2 (HG931713)
100
100
▲ ATCC42981 MATalpha2 copy 2
▲ Z. sapae MATalpha2 copy 3 (HG931714)
60
A. gossypii (AEY98439)
K. marxianus (CCA89281)
96
K. lactis (XP 452211)
▲ T. delbrueckii (XP 003680800)
O. parapolymorpha (ESW98388)
L. waltii (CAO02574)
92
L. thermotolerans (XP 002554226)
C. orthopsilosis (ADY62680)
C. albicans (EAK95701)
0.05

Isolation and characterization of *MTL*a loci in ATCC 42981

In *Z. sapae* the **a**-idiomorph encoding *MTL* locus harbours a *MAT***a**1-coding ORF (*ZsMAT***a**1) identical to the *Z. rouxii* orthologue (*ZrMAT***a**1) and a *MAT***a**2-coding ORF (*ZsMAT***a**2), which showed a 26-bp deletion compared to *ZrMAT***a**2, resulting in a 9-amino acid shorter protein (Solieri et al., 2014b). To identify *MTL***a** loci in the ATCC 42981 genome, a PCR strategy was used based on primers designed on the *MAT***a**1 and *MAT***a**2 gene sequences in *Z. rouxii* and *Z. sapae*. We found three *MTL***a** loci, each harbouring *MAT***a**1 and *MAT***a**2 genes in opposite directions separated by a 279-bp long intergenic sequence, which differed for a single SNP (T/G) compared to the corresponding region in *Z. sapae*. Sequence alignments and BLAST-type search revealed that all ATCC 42981 *MTL***a** loci displayed identical *MAT***a**1 genes (100% identity to *ZrMAT***a**1 and *ZsMAT***a**1), whereas differed from each other in the *MAT***a**2-coding ORFs. One *MAT***a**2-coding gene (referred to as the *MAT***a**2 copy 1) was 100% identical to *Z. rouxii* counterpart. Another *MAT***a**2 gene (termed the *MAT***a**2 copy 2) encoded an **a**2 protein, which diverged from ZrMAT**a**2 mainly at the C-terminal end (93.73% identity). A third *MAT***a**2 coding sequence, namely *MAT***a**2 copy 3, encoded a protein 94.41% and 94.24% identical to ZrMAT**a**2 and ZsMAT**a**2, respectively.

The NJ-tree confirmed the high degree of evolutionary conservation of MAT**a**1 proteins within the *Z. rouxii* complex (bootstrapping 100%) (data not shown). Furthermore, a search with the program Pfam revealed that ATCC 42981 MAT**a**1 proteins conserve the HD2 class homeodomain (Kues and Casselton, 1992; Anderson et al., 2000) (data not shown).

Phylogeny inferred from MAT**a**2 amino acid sequences showed that ATCC 42981 MAT**a**2 variants are phylogenetically distinct (**Fig. 3**). In particular, MAT**a**2 copy 1 clustered with *Z. rouxii* and *Z. sapae* MAT**a**2 (bootstrapping 100%), whereas MAT**a**2 copies 2 and 3 clustered together and were distinct from both ZrMAT**a**2 and ZsMAT**a**2. MAT**a**2 proteins from ATCC 42981 conserved a single MATA-HMG box, a class I member of the HMG-box superfamily of DNA-binding proteins (**Fig. 4**), coding a sequence spanning across the Y**a** and X regions.

Figure 3. Phylogenetic analysis of MATa2 proteins. The neighbour-joining (NJ) tree shows the phylogenetic relationships between the allodiploid strain ATCC 42981 and other hemiascomycetes inferred from MAT**a**2 proteins. Numbers on branches indicate bootstrap support percentages (1,000 pseudoreplicates) higher than 60% from NJ. The red branch indicates the *Z. rouxii* complex, which includes *Z. rouxii* MAT**a**2, *Z. sapae* MAT**a**2 and ATCC 42981 MAT**a**2 copies 1 to 3 sequences. The dark dot indicates WGD species, whereas the dark triangle indicates non-WGD species with the *HO* gene.

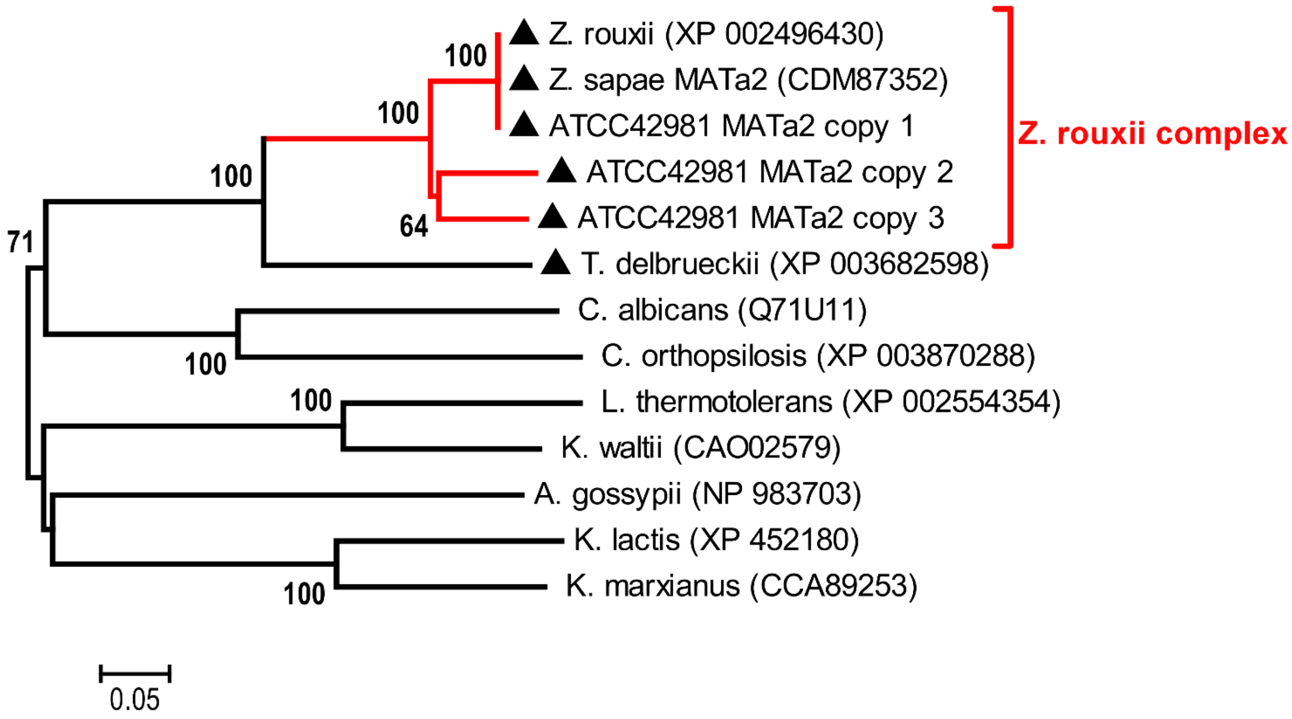
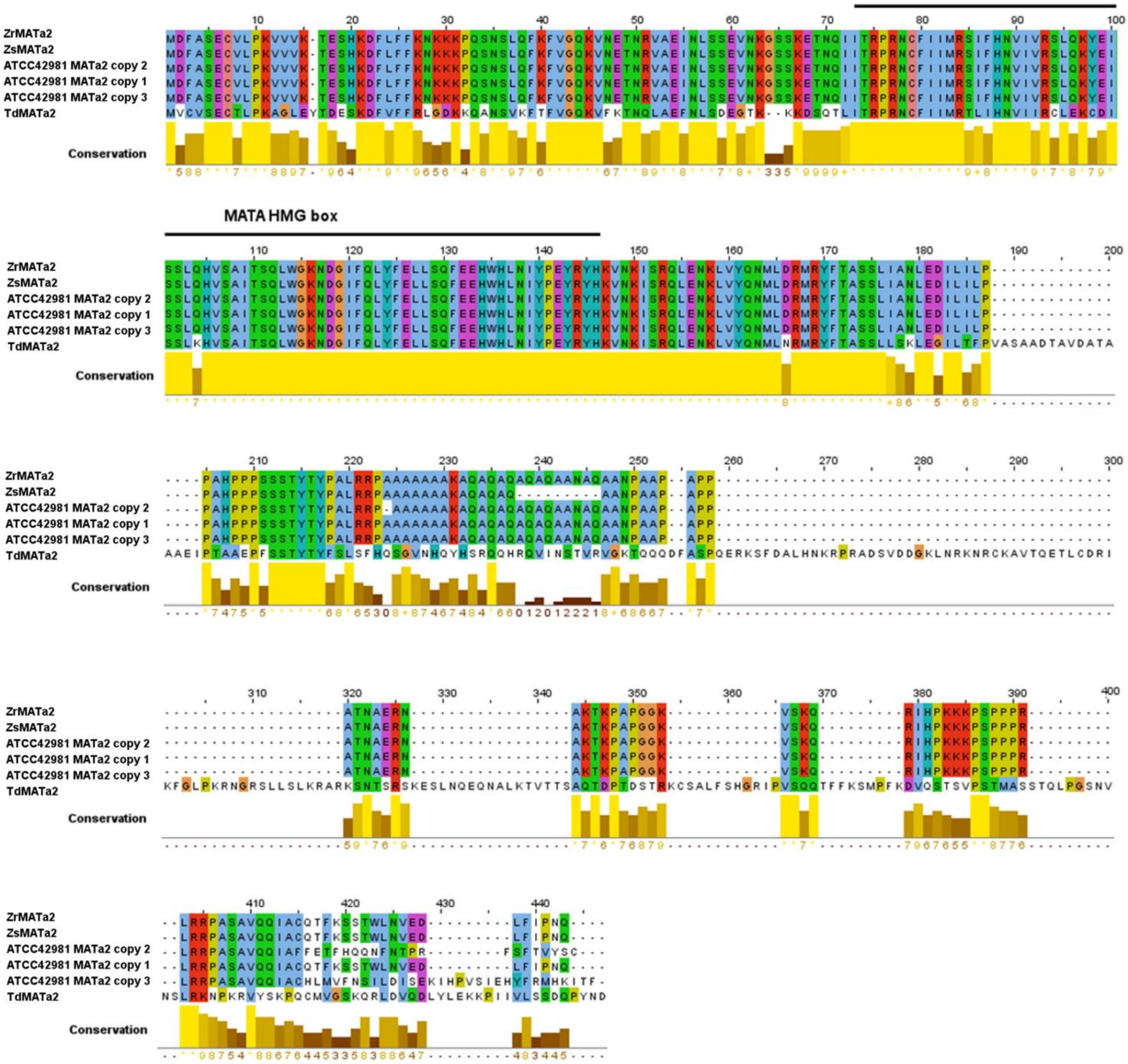


Figure 4. Sequence comparison of MATa2 proteins. Alignment of MAT**a**2 from *Z. rouxii* (ZrMAT**a**2, GenBank: XP002496430), *Z. sapae* (ZsMAT**a**2, GenBank: CDM87352), ATCC 42981 MAT**a** copies 1 to 3 and *Torulaspora delbrueckii* (TdMAT**a**2, GenBank: XP003682598). The MATA HMG domain, which binds the minor groove of DNA, is noted (horizontal black bar). In both alignments, the amino acid identities were coloured according the Clustal X colour scheme and the conservation index at each alignment position were provided by Jalview (Waterhouse et al., 2009).



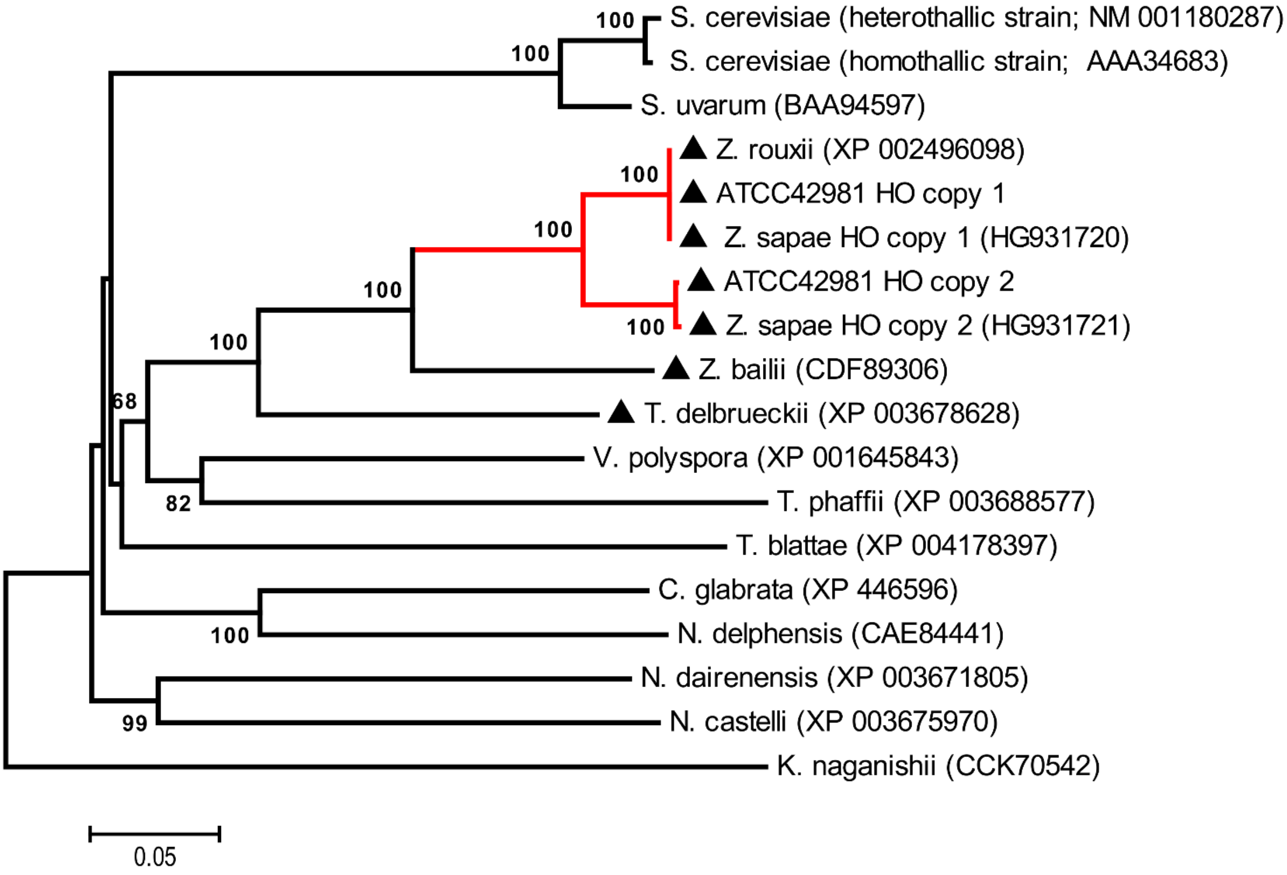
Characterization of *HO* genes

Using a PCR walking approach, we identified two 1797-bp full-length ORFs in the ATCC 42981 genome, termed *HO* copies 1 and 2, which diverged for 177 transitional and 49 transversional mismatches. The predicted proteins HO copies 1 and 2 shared 100% and 92.47% identity with *Z. rouxii* HO, while they were 100% and 99.83% similar to *Z. sapae* HO copies 1 and 2, respectively.

NJ-based phylogeny inferred from amino acid HO sequences confirmed that HO copy 2 from ATCC 42981 branched with *Z. sapae* HO copy 2 with a high level of support (bootstrapping 100%), while ATCC 42981 HO copy 1, ZsHO copy 1 and ZrHO clustered together into a separate clade (**Fig. 5**). These results showed that ATCC 42981 HO amino acid sequences are more divergent from each other than from the putative orthologs in *Z. sapae*. Amino acid alignment of both ATCC 42981 HOs with *S. cerevisiae* PI-*Sce*I (GenBank: CAA98762), *S. cerevisiae* HO and *Z. sapae* HO copies 1 and 2 showed the highest homology in conserved motifs characteristic of intein-encoded LAGLIDADG endonucleases (Belfort and Roberts, 1997; Stoddard, 2005; Hafez et al., 2012) (data not shown).

PFGE-Southern blotting showed that both *HO* copies of ATCC 42981 strain are on chromosome F (1.6 Mbp) (**Fig. S4**). This chromosomal assignment resembles that of CBS 732^T, but differs from that previously found in *Z. sapae* ABT301^T (Solieri et al., 2014b). We assumed that ATCC 42981 *HO* copies could be placed on the same chromosome or, alternatively, on two PFGE-co-migrating homeologous chromosomes.

Figure 5. Phylogenetic analysis of HO endonucleases. Neighbor-joining (NJ) tree shows evolutionary relationships between ATCC 42981 strain and other hemiascomycetes as inferred from HO proteins. Numbers on branches indicate bootstrap support (1,000 pseudoreplicates) from NJ. The red branch indicates *Z. rouxii* complex, which includes ZrHO, ZsHOs and ATCC 42981 HO sequences, whereas the dark triangle designates pre-WGD species.


100
S. cerevisiae (heterothallic strain; NM 001180287)
100
S. cerevisiae (homothallic strain; AAA34683)
S. uvarum (BAA94597)
▲ Z. rouxii (XP 002496098)
100
▲ ATCC42981 HO copy 1
100
▲ Z. sapae HO copy 1 (HG931720)
100
▲ ATCC42981 HO copy 2
100
▲ Z. sapae HO copy 2 (HG931721)
100
▲ Z. bailii (CDF89306)
68
▲ T. delbrueckii (XP 003678628)
V. polyspora (XP 001645843)
82
T. phaffii (XP 003688577)
T. blattae (XP 004178397)
C. glabrata (XP 446596)
100
N. delphensis (CAE84441)
N. dairenensis (XP 003671805)
99
N. castelli (XP 003675970)
K. naganishii (CCK70542)
0.05

Subgenome assignment of mating-type and *HO* genes

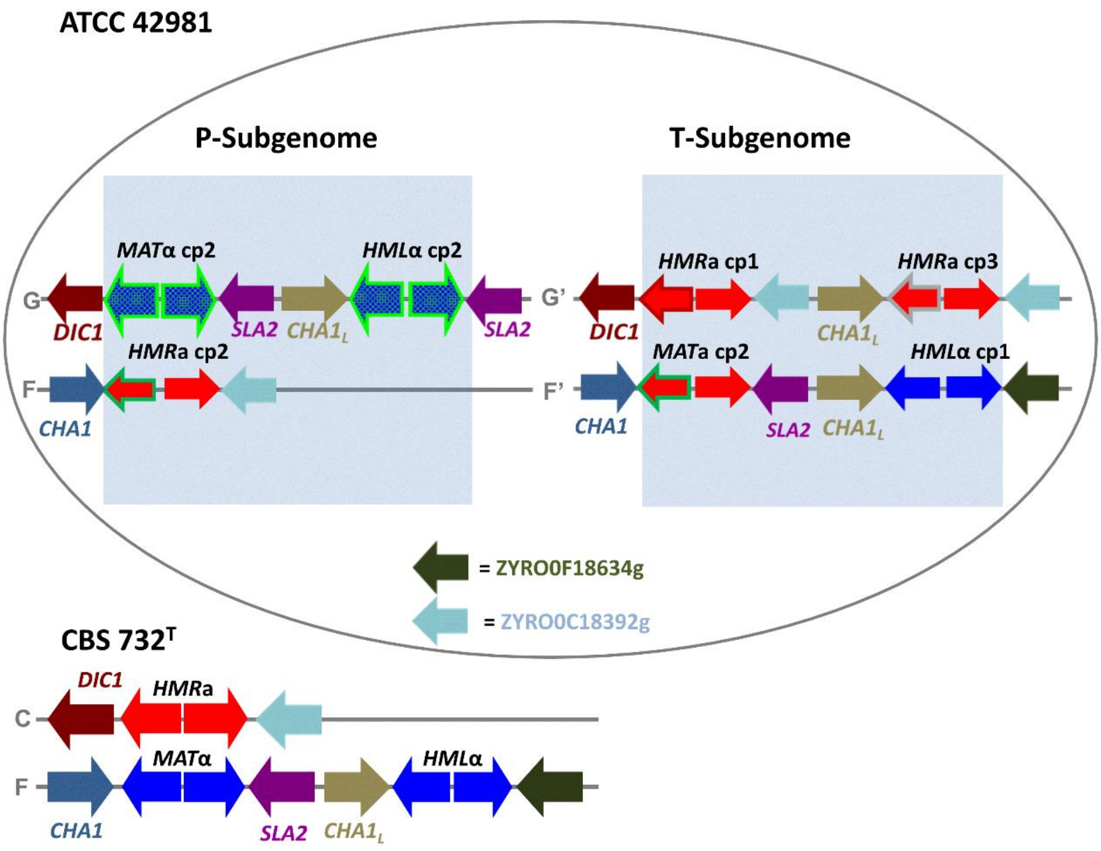
In order to infer the putative parental contributions to the ATCC 42981 *MTL*α, *MAT***a** and *HO* gene sets, we screened strain NCYC 3042 for all variants of *MAT***a**, *MAT*α, and *HO* genes found in ATCC 42981. According to PCR profiles, successful amplification was obtained for copy 2 sets of *MAT*α1, *MAT*α2 and *HO* genes in the NCYC 3042 strain, while negative results were obtained for *MAT*α and *HO* copy 1 genes (**Table S2**). PCR results were consistent with the high degree of similarity between the copy 1 set of *MAT*α and *HO* genes and *Z. rouxii* orthologs in the CBS 732^{T} genome project (**Figs 1, 2 and 5**) and suggested that *MAT*α and *HO* copies 2 originated from a *Z. pseudorouxii* parental genome. Interestingly, *MAT***a**2 copy-specific screening showed that the NCYC 3042 strain does not possess any *MAT***a**2 copies found in ATCC 42981 genome (**Table S2**).

Reconstruction of the three-cassette system

To assign the *MTL***a** and *MTL*α loci as *MAT* expression loci, *HMR* or *HML* silent cassettes (**Fig. 6**), we exploited direct PCR on *MTL* flanking regions and long-range PCRs that spanned left and right flanking regions of *MTL* loci (**Fig. S2**). Both approaches demonstrated that ZYRO0F18524g (termed *CHA1*$_L$ due to its similarity to *CHA1* gene) and ZYRO0F18634g ORFs are located at the 5' and 3' ends of the *MTL*α copy 1 locus, respectively. Because *HML*α silent cassettes are commonly downstream the *CHA1*$_L$ gene (Watanabe et al., 2013), we inferred the existence of a *HML*α silent cassette containing *MAT*α1 and *MAT*α2 copies 1 genes (referred to as *HML*α copy1) (**Fig. 6**). Long-range PCR products obtained with the primer pairs 2/A and 3/A were positively screened through *MTL*α copy 2-specific primers (**Fig. S2**). Sequencing showed that there are two *MTL*α copy 2 loci flanked by the *SLA2* gene at the 3' end and by either the *DIC1* gene or *CHA1*$_L$ at the 5' end, respectively. It has been reported that the *MAT*α expression loci retain the particular gene order *DIC1-MAT*α*-SLA2* in the majority of non-WGD species (Gordon et al., 2011). A similar synteny was found for *MTL*α copy 2, suggesting that ATCC 42981 possesses an actively transcribed *MAT*α copy 2 locus (**Fig 6**). Furthermore, *CHA1*$_L$ upstream to the *MTL*α copy 2 supports the presence of an additional *HML*α copy 2 cassette. Finally, we detected the following syntenic orders for *MTL***a** loci: *DIC1-MTL***a** copy 1-ZYRO0C18392g, *CHA1-MTL***a** copy 2-*SLA2*, *CHA1-MTL***a** copy 2-ZYRO0C18392g and *CHA1*$_L$*-MTL***a** copy 3-ZYRO0C18392g (**Fig. 6**). The gene organization of the *CHA1-MTL***a** copy 2-*SLA2* was syntenic with that of the *MAT*α expression locus in *Z. rouxii* CBS 732^T, whereas the gene orders *CHA1-MTL***a** copy 2-ZYRO0C18392g and *DIC1-MTL***a** copy 1-ZYRO0C18392g resembled those found in *HMR***a** silent loci of haploid *Z. rouxii* strains (Watanabe et al., 2013). In particular, *CHA1-MTL***a** copy2-ZYRO0C18392g resembled the *HMR***a** cassette found in a- and α-idiomorph mixed culture of strain NBRC 1130. Although the synteny *CHA1*$_L$*-MTL***a** copy 3-ZYRO0C18392g has been never found in *Z. rouxii* strains, *CHA1*$_L$ and ZYRO0C18392g ORFs generally surround silent donor cassettes. Overall, these evidences support the notion that ATCC 42981 possesses one *MAT***a** copy 2 expression locus and three *HMR***a** cassettes, referred to as *HMR***a** copies 1 to 3 (**Fig. 6**).

The chromosomal arrangement of *MAT*, *HMR* and *HML* cassettes was established by Southern blot analysis on PFGE-separated chromosome bands. Although PFGE-Southern blotting failed to clearly resolve the highest molecular weight chromosomes E, F and G spanning from 1.6 to 2.2 Mbp (**Fig. S5**), hybridization of PFGE-Southern blot with a *MAT*α1-specific probe (suitable to recognize both *MAT*α1 copies 1 and 2) resulted in a double band spanning from chromosome F to G for *Z. rouxii* CBS 732^T and strain ATCC 42981 (**Fig. S5**). However, differently from CBS 732^T, in strain ATCC 42981, the *MAT*α1-specific probe bound less chromosome F than chromosome G, leading to two bands with different signal intensities. This result suggests that chromosome G harbours more copies of the *MTL*α cassette than chromosome F. When the same analysis was carried out with a *MAT***a**1-specific probe, we obtained a double band spanning from chromosome F and G of ATCC 42981 karyotype (**Fig. S5**). These studies suggest that *MTL***a** loci reside on at least two of the chromosomes spanning from 1.6 and 2.2 Kb. According to PFGE-Southern blotting results, the ATCC 42981 strain arranges two expressed *MAT* loci (*MAT***a** and *MAT*α copies 2), three *HMR***a** and two *HML*α silent cassettes on two homeologous chromosome pairs (G/G' and F/F'). Taking into consideration that linked *MAT* and *HML* loci are located on a different chromosome compared to *HMR* (Watanabe et al., 2013; Gordon et al., 2011), we inferred the scenario depicted in **Fig. 6**. Chromosome G putatively contains the expression locus *MAT*α copy 2, which is linked to the *HML*α copy 2 cassette, while chromosome F hosts the *HMR***a** copy 2 silent cassette. The expression locus *MAT***a** copy 2 is putatively located on chromosome F', linked with *HML*α cassette copy 1. Chromosome G' provides the two remaining silent *HMR***a** copies 2 and 3 cassettes (**Fig. 6**).

Figure 6. Inferred genomic organization around *MAT*-like loci in the ATCC 42981 allodiploid genome. Two sex homologous/homeologous chromosome pairs are depicted, namely F/F' and G/G'. Chromosome G bears the *MAT*α copy 2 expression locus, which is linked to the putative silent cassettes *HML* copy 2, whereas chromosome F' bears the *MAT***a** copy 2 expression locus, which is linked to the putative silent cassette *HML* copy 1. Chromosomes F and G' harbour three **a**-idiomorph *HMR* loci, which differ for *MAT***a**2 genes. The *HMR* copy 2 locus is on chromosome F, while the *HMR* copies 1 and 3 loci are on chromosome G'. Blue arrows represent *Z. rouxii*-like (blue bordered) and *Z. sapae*-like (light green bordered) α-idiomorph loci. Red arrows indicate **a**-idiomorph loci with different *MAT***a**2 genes: copy 1 (dark red surrounded), copy 2 (dark green), copy 3 (grey), respectively. Chromosomal organization of the three-cassette system in *Z. rouxii* haploid strain CBS 732^T was reported for comparative purposes according to Souciet et al., 2009. *CHA1$_L$* indicates the ZYRO0F18524g locus, while dark green and light blue arrows indicate the ZYRO0F18634g and ZYRO0C18392g loci, respectively at the right side of mating-type loci.

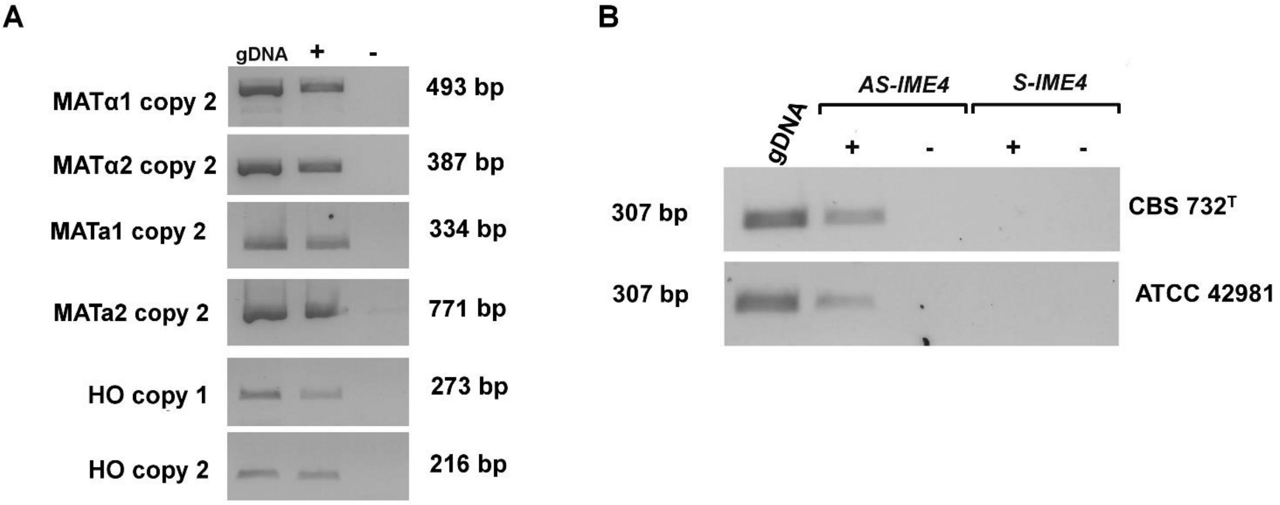


Gene expression analysis

Non-quantitative RT-PCR confirmed that the expression loci *MAT*α and *MAT***a** copies 2 are actively transcribed in ATCC 42981 cells grown in standard conditions (**Fig. 7**). Congruently to the cassette reconstruction, the *MAT*α1 and *MAT*α2 genes from *HML*α copy 1 were silent.

According to the regulatory circuit reported in *S. cerevisiae* diploid cells, we expected h-sg sets to be silenced by a functional **a**1-α2 heterodimer. In contrast, gene expression profile analysis revealed that the *MAT*α1 and *HO* genes were actively transcribed in the ATCC 42981 allodiploid strain. Similarly in haploid CBS 732^{T} unstressed cells *HO* gene was actively transcribed (data not shown). To verify whether other members of the h-sg set were expressed, we tested the presence of S-*IME4* and AS-*IME4* transcripts. RT-PCR analysis showed that salt-stressed ATCC 42981 diploid cells only transcribed AS-*IME4*, while S-*IME4* was not detected (**Fig. 7**). This pattern is similar to that displayed by haploid cells of the reference strain CBS 732^{T}, which does not show any PCR products for S-*IME4* specific RT-PCR (data not shown). The identity of AS-*IME4* PCR products was confirmed by direct sequencing. We concluded that ATCC 42981 is a diploid strain that only transcribes the AS-*IME4* transcript.

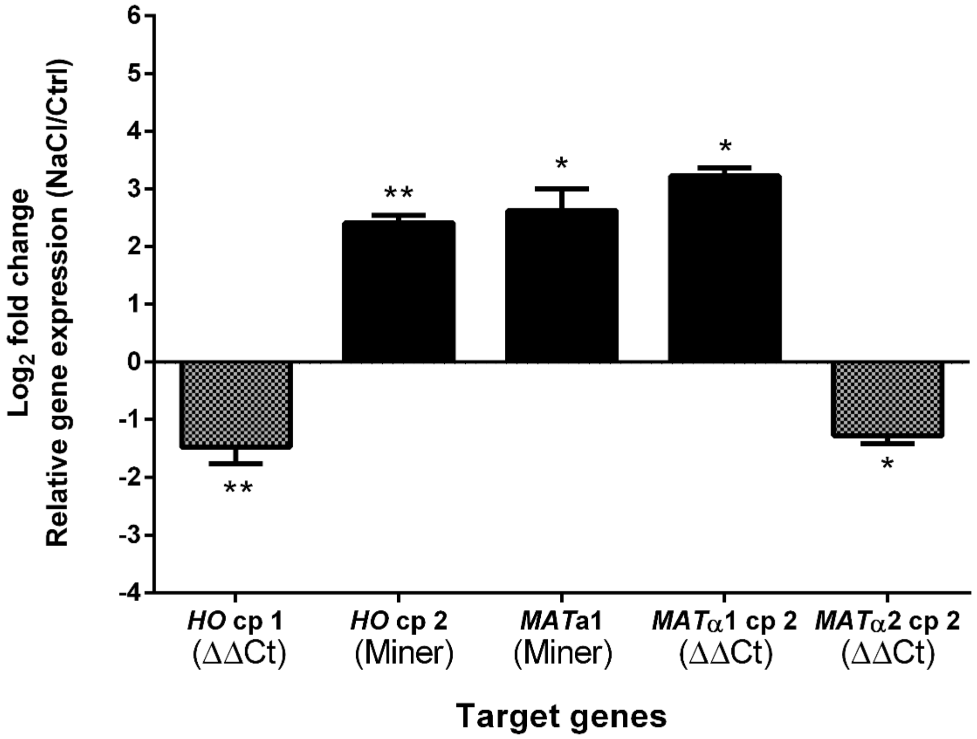
Figure 7. Expression pattern of mating-type, *HO* and *IME*4 genes. Panel A reports positive amplified cDNAs generated with *MAT* and *HO* copy variant-specific primers from ATCC 42981 unstressed cells in stationary growth phase. Panel B depicts *IME4*-specific PCR products generated from CBS 732^T and ATCC 42981 cDNA *IME4* antisense (AS-*IME4* lncRNA) and sense transcripts (S-*IME4* mRNA), respectively. +/- RT indicates addition of reverse transcriptase to the cDNA synthesis reaction. For each RT-PCR reaction gDNA was used as positive control. Abbreviations: AS, anti-sense long non-coding RNA; S, sense mRNA.



Transcriptional response to hyperosmotic stress

After hyperosmotic stress, ATCC 42981 cells exhibit a different expression profile for both mating-type and *HO* genes compared to controls. Salt stress induced higher transcript levels of *MAT***a**1 and *MAT*α1 copy 2 (6.2- and 9.3-fold, respectively). On the contrary, *MAT*α2 copy 2 expression was lower than that in controls (2.4-fold). *HO* copy 1 was slightly down-regulated (2.8-fold), whereas *HO* copy 2 was up-regulated (more than 5.3-fold) compared to controls (**Fig. 8**).

Figure 8. Differential expression by quantitative real-time PCR of mating-type and *HO* genes in ATCC 42981 cells after hyperosmotic stress. Expression of target genes was normalized on the reference *ZrACT1* (GenBank: XM002497273). Fold change was measured by ΔΔCt or PCR Miner methods and reported as the mean (± SEM) of three biological replicates. * indicates significant difference from controls as measured by independent Student's *t*-tests (**P*<0.05, ***P*<0.01).



Discussion

Previous studies demonstrated that outcrossing is a very rare event (approximately once every 50,000/110,000 generations) in yeasts as *S. cerevisiae* (Ruderfer et al., 2006), *Saccharomyces paradoxus* (Tsai et al., 2008) and *Lachancea kluyveri* (Friedrich et al., 2015). Nevertheless, it represents a potential source of phenotypic variability (novelty) and has been extensively exploited for the genetic improvement of microbial cell factories (Steensels et al., 2014). Although many inter-mating allopatric yeasts give rise to viable hybrids, post-zygotic isolating barriers (Hunter et al., 1996) prevent gene flow between species and contribute to the process of speciation (Coyne and Orr, 2004). Complete sets of orthologous genes are expected in yeast hybrids immediately after the merging of two parental genomes. These patterns can undergo extensive homogenization processes over evolutionary time through intragenic recombination, gene conversion and differential gene loss, shaping the offspring's genome architectures (Dunn et al., 2013; Wolf, 2015). Gordon and Wolfe (2008) did not find any traces of gene losses in the ATCC 42981 strain, arguing that the allopolyploidization was so recent that its genome has not had enough time to decay. These authors hypothesized that one parental subgenome resembles *Z. rouxii* CBS 732^T and the other *Z. pseudororuxii* (nom. inval.) NCYC 3042. However, ATCC 42981 karyotype cannot be simply considered as an additive result between the putative parental counterparts, suggesting that some structural rearrangements have occurred in the ATCC 42981 genome compared to *Z. rouxii* and *Z. pseudorouxii* (Pribylova et al., 2007; Gordon and Wolfe, 2008) . Similarly, the *Z. rouxii* CBS 732^T strain contains a *SOD2-22* gene variant (Kinclova et al., 2001), while *Z. pseudorouxii* NCYC 3042 has *SOD2* (James et al., 2005). The detection of *SOD2* and *SOD22* genes, but not the *SOD2-22* variant in the ATCC 42981 genome, further confirms that the subgenome complements are similar but not identical to those found in *Z. rouxii* CBS 732^T and in *Z. pseudorouxii* NCYC 3042.

In an allodiploid karyotype, two homeologous sex chromosomes are expected, bearing **a**- and α-idiomorph expression loci, respectively. Previous works demonstrated that in haploid *Z. rouxii* and diploid *Z. sapae*, *HMR* cassettes are located on a different chromosome compared to the *MAT-HML* linkage, accounting for the presence of two sex chromosomes (Watanabe et al., 2013; Solieri et al., 2014b). Therefore, in a recent allodiploid *Zygosaccharomyces* genome, we expected to identify two sex chromosome pairs. Accordingly, the ATCC 42981 strain displays an **a**/α genotype provided by the contribution of two sex chromosome pairs. The first sex chromosome pair consists of chromosomes G (*MAT*α copy 2-*HML*α copy 2) and F (*HMR***a** copy 2). This subgenome complement only partially resembles that found in *Z. pseudorouxii* NCYC 3042 strain (James et al., 2005) which might have contributed to this complement with *MTL*α copy 2. If *Z. rouxii* contributed to the second chromosome set, we would expect to detect a *Z. rouxii MAT***a**2 copy 1 gene at the expression locus. In contrast, chromosome F' (carrying *MAT***a** copy 2-*HML*α copy 1) and G' (carrying *HMR***a** copy 1-*HMR***a** copy 3) probably derived from a rearranged *Z. rouxii* genome complement [29], consistently with the mating-type loci being the hotspot of ectopic recombination in *Z. rouxii* strains (Watanabe et al., 2013). Furthermore, an extra *HMR***a** copy 3 was detected harbouring another partially divergent *MAT***a**2 ORF, which it might have derived from another non-*Z. rouxii* parental strain or resulted from an ectopic recombination event in the allodiploid ancestor. All these findings are not consistent with a simple additive set of mating-type loci recently assembled in the allodiploid genome and indicate that ATCC 42981 T- subgenome differs from its putative *Z. rouxii* CBS 732^T

counterpart. Watanabe et al., (2013) found that the copy 2 variant of *MAT***a**2 gene is at the *HMR* silent cassette of strain NBRC1130 which undergone mating-type switching, whereas copy 1 variant occurs in *MAT* expression locus. This genome assortment is specular to that found in ATCC 42981 T-subgenome and hints that different *MAT***a**2 copies co-exist within *Z. rouxii* species, harboured by *MTL* loci flanked by different genes at 5' end. We speculated that these copy variants could be relicts of past mating-type switching events. Interestingly, *Z. rouxii* possesses a hybrid regulation system targeting **a**-specific genes (**a**-sgs), which consists of **a**2-mediated activation and α2-mediated repression (Baker et al., 2012). In haploid cells, MAT**a**2 proteins with diverging C-terminal portions could be functionally equivalent in mediating the activation of the **a**-sg program.

Beyond divergent mating-type loci, the ATCC 42981 strain has also been reported to possess two divergent *HO* genes. Orthologous genes coming from each of the parental species should appear as paralogs in standard analyses because they are homologous genes encoded by the same genome (Wolfe, 2001). However, the degree of divergence identified between *HO* gene copies 1 and 2 excludes their origin as paralogs from a recent gene duplication event and hints that *HO* copy 2 originated from the *Z. pseudorouxii* parental genome. The pattern of neutral mutations detected in the endonuclease functional domains indicates that both *HO* genes have been exposed to the same selective pressure.

Gene regulatory networks, such as cell-type specification circuit, have been shown to evolve significantly over time (Tuch et al., 2008). When the divergent components of these circuits are forced to interact in a hybrid background, positive or antagonistic epistatic interactions may take place (Dettman et al., 2007). Among the mechanisms cited to explain the observed loss of hybrid fertility, the Bateson-Dobzhansky-Muller (BDM) model proposes that hybrid sterility results from the lack of interaction or the malfunctioning of interacting alleles derived from divergent genomes (Scannell et al., 2007; Brideau et al., 2006; Lee et al., 2008; Bayes and Malik, 2009). However, in some cases, inter-specific protein assemblies have been reported to generate novelties in protein-protein networks untested by selection in hybrid species (Orr and Turelli, 2001; Piatkowska et al., 2013). The reconstruction of the sex chromosome architecture in the ATCC 42981 genome indicates that the transcriptional regulators encoded at the *MAT* expression loci are phylogenetically divergent. In diploid cells, **a**1-α2 heterodimer acts as a master regulator of the shift from mitosis to meiotic growth under appropriate meiosis-inducing stimuli, as well as a repressor of the h-sgs under standard growth conditions. H-sgs contain the *MAT*α1 gene, which the **a**1-α2 heterodimer should repress by binding the bidirectional promoter located between *MAT*α1 and *MAT*α2 ORFs (Haber, 2012). In allodiploid ATCC 42981, the lack of MATα1 silencing could be due to: the lack of **a**1 and α2 proteins; the failure in the heterodimer formations; the inability of putative chimeric **a**1 and α2 subunits to positively interact with *cis*-regulatory promoter sequences. Similarly to ATCC 42981, *S. cerevisiae* diploids showing mutations in either **a**1 or α2 transcriptional factors fail to turn off *MAT*α1 and other h-sgs (Siciliano and Tatchell, 1984; Harashima et al., 1989). Additionally, Strathern et al., (1988) reported that a diploid mutant with a truncated *MAT*α2 defective allele displays a weak α phenotype and is unable to sporulate. Overall, these evidences suggest that in ATCC 42981 the cell-type specification circuit is ineffective in repressing the *MAT*α1 gene possibly due to, among other factors, a no functional chimeric **a**1-α2 heterodimer.

In diploids, the entry into meiosis requires the inhibition of lncRNA AS-*IME4* by a functional **a**1-α2 heterodimer Hongay et al., 2006). Conversely, our results show that the ATCC 42981 strain produces anti-sense transcripts for the *IME4* gene, leading to clonality as its unique mode of reproduction. The haploid-like transcriptional pattern displayed by stressed ATCC 42981 cells could also account for its increase of adhesive phenotype. In haploids grown under stress cues, a complex network of signalling modules and transcriptional factors induce clamp formation and pseudohyphal growth in order to enhance mating efficiency and chance of survival (Goossens et al., 2015). The *FLO11* gene has been reported to be a key determinant of the adhesion phenotype under the positive control of the RME1 transcriptional factor. In diploids, the *RME1* gene is silenced by a functional **a**1-α2 heterodimer, leading to the entry into meiosis in response to environmental cues. Because we observed more clamps in salt-stressed compared to unstressed ATCC 42981 cells, we argue that the chimeric **a**1-α2 heterodimer is also partially ineffective in repressing *RME1* transcription.

In *S. cerevisiae,* a redundant regulatory network accounts for the three-level regulation of mating-type interconversion catalysed by the *HO* endonuclease, namely cell-type control (*HO* gene is expressed in **a** or α haploid cells); mother-daughter control (*HO* is transcribed in the mother but not in the daughter cells); and cell-cycle control (*HO* is expressed during the late G1 phase of the cell cycle after the point of commitment to the next cell cycle) (Stemberg et al., 1987). Three transcriptional repressors are involved: the **a**1-α2 heterodimer, *SIN3* and *SIN6*, respectively. These three types of negative constraints must be relieved in order for the *HO* gene to be transcribed. Allodiploid ATCC 42981 cells express *HO* mRNAs, probably due to the lack of these types of controls or to partial incompatibilities among transcriptional factors. However, the presence of *HO* transcripts does not imply that ATCC 42981 undergoes mating-type interconversion. Indeed, the transcriptional analysis of *MAT* expression loci showed that only the expected *MAT***a** copy 2 and

*MAT*α copy 2 transcripts are detected in salt-stressed cells at the stationary phase. The failure of mating-type switching induced by salt stimuli could be due to either a *HO* post-transcriptional control or to the lack of a fully functional network controlling the DSB-initiated gene conversion. However, other effectors could be responsible for this event, as in haploid *Z. rouxii* strains *HO* deletion determines only a slight decrease in mating-type switching frequency (Watanabe et al., 2013). The *Z. rouxii* species complex emerged after the ancestor of hemiascomycetous yeasts had diverged from other families, such as *Kluyveromyces*. *K. lactis* has a non-functional copy of the *HO* gene (Fabre et al., 2005) and performs mating-type switching by an alternative transposase-mediated mechanism (Rajaei et al., 2014). As it is the first non-WGD clade with a functional *HO* gene, the *Z. rouxii* complex is likely to retain remnants of both mechanisms.

In haploid *Z. rouxii*, mating and the subsequent zygote formation occur immediately before sporulation mainly under salt stress (Mori, 1973; Mori and Onishi, 1967). Therefore, we investigated how the ATCC 42981 cells modulate the transcription of genes coding the **a**1 and α2 subunits, as well as that of their downstream h-sg targets *MAT*α1 and *HO* in response to long-term hypersaline stress. In diploid cells, the silencing of h-sg *MAT*α1 and *HO* and the positive regulation of the *MAT*α2 gene are controlled by a working **a**1-α2 heterodimer. The inability of the ATCC 42981 chimeric **a**1-α2 heterodimer to bind to the h-sg promoter regions may account for the observed up-regulation of *MAT*α1 and *HO* genes. Functional defects of the chimeric **a**1-α2 heterodimer could be due to gene incompatibility between two divergent subunits and/or to the transcriptional imbalance of their encoding genes. ATCC 42981 cells could attempt to overcome these functional deficiencies by over-expressing the components of the **a**1-α2 transcriptional factor. Congruently, we observed an up-regulation of the **a**1 transcript. A similar up-regulation of the other heterodimer subunit α2 should be expected. In contrast, *MAT*α2 gene expression appears to be down-regulated, even at a small level, suggesting an imbalance in the co-regulation of **a**1 and α2 subunits.

Concluding remarks

In conclusion, we demonstrated that allodiploid ATCC 42981 cells display a *MAT***a**/*MAT*α genotype with a chimeric sex-determination system originating from the co-existence of two different parental genome complements. The protein-protein interaction incompatibility between divergent **a**1 and α2 subunits could switch-off the meiosis commitment genes, contributing to ATCC 42981 allodiploid sterility. The presence of a chimeric **a**1-α2 heterodimer promotes an unusual haploid-like transcriptional profile in cells recovered at the stationary phase and after exposure to meiosis-inducing stimuli. To the best of our knowledge, this is the first cue that the BDM interaction between the divergent **a**1 and α2 subunits may act as a bottleneck, preventing genetic exchanges among *Zygosaccharomyces* species. Recently, a novel scenario has been proposed for yeast evolution, where two ancient non-WGD ancestral species have given rise to an allodiploid cell that has doubled its genome in order to restore fertility (with a possible interval of many mitotic generations between these two events) (Marcet-Houben and Gabaldon, 2015). Interestingly, one of the possible parents exhibits phylogenetic affinities with the non-WGD *Z. rouxii* clade. However, there are several critical open questions that still need to be answered, such as: how the genome duplication event took place and how the mechanism of restoring fertility operated (Wolfe, 2001). Allodiploid ATCC 42981 and other strains belonging to the *Zygosaccharomyces* mosaic lineage (Solieri et al., 2013a) could serve as promising models to shed light on the transcriptional network incompatibility underlying hybrid sterility at an incipient stage of speciation, and more in general, on yeast genome evolution.

Author Contributions

Conceived and designed the experiments: LS SC. Performed the experiments: LS MB SC. Analysed the data: LS MB SC. Contributed reagents/materials/analysis tools: SC PG. Wrote the paper: LS SC.

CHAPTER 6: DEVELOPMENT OF PLASMIDS HARBOURING ANTIBIOTIC RESISTANCE SELECTION MARKERS AND CRE RECOMBINASE FOR GENETIC ENGINEERING OF NON-CONVENTIONAL *ZYGOSACCHAROMYCES ROUXII* YEASTS

Bizzarri M., Dušková M., Sychrovà H., Cassanelli S., and Solieri L. (In preparation). Development of plasmids harbouring antibiotic resistance selection markers and Cre recombinase for genetic engineering of non-conventional Zygosaccharomyces rouxii yeasts.

Abstract

The so-called non-conventional yeasts are becoming increasingly attractive in food and industrial biotechnology, since, compared to the model species *Saccharomyces cerevisiae*, exhibit several advantages, which make them suitable for the generation of various products other than ethanol. Among them, we focused on *Zygosaccharomyces rouxii* that is known to be halotolerant, osmotolerant, petite negative and poorly Crabtree-positive. Overall, these phenotypic traits and its high fermentative vigour make this species very appealing for application in industrial purposes. Nevertheless, *Z. rouxii* exploitation in industrial and food bioprocesses has been hampered by the low availability of synthetic biology tools. Moreover, *Z. rouxii* suffers of some genetic intractability, which make difficult to genetically manipulate this yeast through the conventional transformation procedures. Centromeric and episomal *Z. rouxii* plasmids were successfully constructed, but they rely on a limited set of prototrophic markers and can be used only in specific auxotrophic strains mainly derived from the *Z. rouxii* haploid type-strain CBS 732^T. The majority of industrially promising *Z. rouxii* yeasts are prototrophic and allodiploid/aneuploid strains, such as ATCC 42981. In order to expand the biobricks available to genetically manipulate prototrophic *Z. rouxii* allodiplod strains, we newly developed two centromeric and two episomal vectors harboring *KanMX*^{R} and *NAT*^{R} as dominant selectable markers, respectively. We also constructed the first plasmid pGRCRE that allows the efficient Cre-recombinase-mediated markers recycle during multiple gene deletions. As proof of concept, ATCC 42981 G418-resistant mutants were constructed by replacing the *MAT*α

expression locus with the *loxP-kanMX-loxP* cassette and were subsequently used to validate pGRCRE as system to rescue the *kanMX-loxP* module in *Z. rouxii* prototrophic strains.

Keywords: *Zygosaccharomyces rouxii*, hybrid, mutants, Cre/*loxP*, synthetic biology, prototrophic yeast, electroporation, homologous recombination.

Introduction

Yeasts alternative to the model species *S. cerevisiae*, the so called non-conventional yeasts, are being increasingly attractive for traditional and industrial biotechnology. In food processing quality and healthy functionalities of food and beverages are improved by controlling sequential or simultaneous co-fermentation of *S. cerevisiae* and non-conventional yeasts, such as *Candida zemplinina*, *Torulaspora delbrueckii*, and *Zygosaccharomyces* spp. (Comitini et al., 2017).

In industrial biotechnology, non-conventional yeasts, such as *Hansenula polymorpha*, *Kluyveromyces lactis*, *Pichia pastoris*, and *Yarrowia lipolytica*, exhibit advantages, when compared with *S. cerevisiae*, which make them suitable for the generation of various products other than ethanol (Wagner and Alper, 2016).

The non-conventional yeast *Zygosaccharomyces rouxii* is known to be halotolerant, osmotolerant (Dakal et al., 2014), petite negative and poorly Crabtree-positive (Merico et al., 2007). The latter trait means that this yeast produces less ethanol in aerobic conditions than the Crabtree-positive *S. cerevisiae*, and it is more suitable than *S. cerevisiae* to produce by-products.

During soy sauce fermentation, *Z. rouxii* produces important aroma compounds, such as fusel alcohols (produced from the corresponding branched-chain amino acids via the Ehrlich pathway), 4-hydroxy-2(or 5)-ethyl-5(or 2)-methyl-3(2H)-furanone (HEMF) (Sasaki et al., 1991, 1996; Ohata et al., 2007), 4-ethylguaiacol (4-EG) and 4-ethylphenol (4-EP).

During balsamic vinegar production, *Z rouxii* converts sugars present in cooked grape must to ethanol, providing the substrate for acetic acid production by acetic acid bacteria (Solieri et al., 2012). Whole-cell processes using *Z. rouxii* strains were proposed for L-malic acid (Taing and Taing, 2007) and extra-cellular L-glutaminase production (Kashyap et al., 2002), as well as for the reduction of 3,4-methylenedioxyphenyl acetone to 3,4-methylenedioxyphenyl-(S)-isopropanol leading to 5H-2,3-benzodiazepine (Vincenzi et al., 1997). Carbonyl reductases were isolated from *Z. rouxii* for the asymmetric reduction of selected ketone substrates of commercial importance (Costello et al., 2000). Furthermore, the exceptional tolerance of *Z. rouxii* to several environmental stresses renders this species very appealing for application in industrial purposes.

Despite the increasing interest toward this species, the exploitation of *Z. rouxii* in industrial and food bioprocesses has been hampered by the low availability of synthetic biology tools. Compared with *S. cerevisiae*, *Z. rouxii* suffers from some genetic intractability, which make difficult to genetically manipulate this yeast. *Z. rouxii* is recalcitrant to conventional transformation procedures and species and/or strain-specific protocols had to be optimized to efficiently introduce pDNA into *Z. rouxii* by electroporation (Pribylova and Sychrovà, 2003; Pribylova et al., 2007a; Watanabe et al. 2010). Compared with *S. cerevisiae*, homologous recombination (HR)-mediated gene targeting is less efficient in *Z. rouxii* and can be achieved by including homology arms of at least 80 base pairs on the 5'end of primers against 40 base pairs required in *S. cerevisiae*. Prybilova et al. (2007b) firstly proved that PCR-generated *loxP-kanMX-loxP* cassette can be successfully used for gene deletion in *Z. rouxii*. Recently, Watanabe and colleagues (2017) exploited an alternative antibiotic selection marker to generate hybrid *Z. rouxii* deletion mutants, by PCR amplifying Zeocin-resistant gene from plasmid pUZ6, surrounded by two *loxP* sequences.

Another drawback in *Z. rouxii* genetic manipulation is that neither *S. cerevisiae* centromeric plasmids nor the 2 µm replicon of *S. cerevisiae* well worked in *Z. rouxii*, making necessary to build *ad hoc* shuttle plasmids. Centromeric and episomal *Z. rouxii*/*Escherichia coli* plasmids were successfully constructed, but they rely on a limited set of prototrophic markers (Sc*URA3*, Zr*LEU2*, Zr*ADE2*), and can be used only in specific auxotrophic strains mainly derived from the haploid strain *Z. rouxii* CBS 732^T (Pribylova et al., 2007b). By contrast, the majority of industrially promising *Z. rouxii* yeasts are prototrophic and allodiploid/aneuploid strains (Dakal et al., 2014). These difficulties can be circumvented using dominant drug resistance markers, such as the *E. coli* aminoglucoside 3' phosphotransferase *kan*^{R} that confers resistance to the aminoglycoside antibiotic gentamicin/G418 (Wach et al., 1994) or the *E. coli* streptothricin acetyl transferase sat1 which confer resistance to nouseothricin (ClonNAT or *NAT*^{R}) (Heim et al., 1989). However, the low number of available dominant selectable markers for *Z. rouxii* requires the usage of markers that can be subsequently excised and reused, mainly considering that genetic crossing in allodiploid strains is inefficient to generate strains carrying multiple gene deletions. A successful attempt to virtually recycle any desired marker was the development of the bacteriophage-derived *loxP*-Cre recombinase system (Hoess and Abremski, 1985; Sauer, 1987; Güldener et al., 1996, 2002). This system is considered "the universal reagent for genome tailoring" (Nagy, 2000), but requires the expression of a plasmid-borne recombinase and thereby necessitates an additional selection marker or extensive screening. The episomal plasmid pZCRE contains cre recombinase encoding gene under the *S. cerevisiae* inducible *GAL1* promoter and was successfully used for *kanMX-loxP* module pop-out in auxotrophic strains (Pribylova and Sychrovà, 2007). However, no cre recombinase expression vector harboring dominant selectable markers are available for marker recycling in *Z. rouxii* prototrophic strains.

In an effort to expand the genetic toolbox available in *Z. rouxii*, we generated two centromeric and two episomal vectors harboring *KanMX*^{R} and *NAT*^{R} as dominant selectable markers, respectively. We also constructed the plasmid pGRCRE, which contains *NAT*^{R} as selectable marker and can be used for *cre*-recombinase-mediated marker recycling in prototrophic deletion mutants generated with *loxP-kanMX-loxP* deletion cassette. As proof of concept, ATCC 42981 G418-resistant mutants were constructed by replacing the *MAT*α expression locus with the *loxP-kanMX-loxP* cassette and was used to validate pGRCRE as system to rescue the *kanMX-loxP* module in *Z. rouxii* prototrophic strains.

Materials and Methods

Strains, media and growth conditions

Yeast strains used in this work are listed in **Table 1.** Yeast cells were routinely propagated at 28°C in YPD medium with or without 1.5% (w/v) agar and maintained at 4°C on YPDA slants for the duration of experiments. Stock cultures were stored at -80°C with glycerol at final concentration of 25% (v/v) for long-term preservation. For antibiotic sensitivity tests, early stationary phase pre-cultures (grown on YPD liquid medium at 27°C) were harvested by centrifugation, washed in 1 ml sterile MilliQ water and adjusted to the same initial OD_{600nm} = 1.0 (corresponding to ~ 2 x 10^7 CFU/ml). Drops (5 µL) of serially 10-fold dilution were spotted aseptically on YPD was supplemented with increasing antibiotic concentration (400, 300, 250 and 200 µg/ml of G418 and 50, 25, 15, 10 and 5 µg/ml ClonNAT, respectively. Plates were incubated at 27°C for 48-72h and images were captured using an Epson Expression 10,000 XL scanner operating in transmitted light mode.

Escherichia coli Trans1-T1 (TransGen Biotech Beijing, China) was used as host strain for plasmid and routinely grown in standard Luria-Bertani (LB) medium supplemented with 100 mg/ml ampicillin (Sigma Aldrich, Milan, Italy). All media compositions are detailed in **Table 2**.

Table 1. Yeast strains used in this work. Abbreviations: na, not applicable.

Species	Strains	Features-Genotype	Source/References
S. cerevisiae	**BW31a**	Haploid strain *MAT***a**, *W303-1A ena1-4Δ::HIS3 nha1Δ::LEU2*	na; Kinclova-Zimmermannova et al*.,* 2006
	Y13925	Haploid strain derived from BY4742*; MATα; ura3Δ0; leu2Δ0; his3Δ1; lys2Δ0; YDL227c::kanMX4*	na; EUROSCARF
Z. rouxii	**ATCC 42981**	Allodiploid *MAT***a***/MATα*	Miso; Solieri et al., 2008
	***MATα*Δ ATCC 42981 clone_6**	Derived from ATCC 42981; *MATα*^{P}*::kanMX4*	na; this work
	***MATα*Δ ATCC 42981 clone_ 65**	Derived from ATCC 42981; *MATα*^{P}*::kanMX4*	na; this work
	***MATα*Δ ATCC 42981 clone_74**	Derived from ATCC 42981; *MATα*^{P}*::kanMX4*	na; this work
	***MATα*Δ ATCC 42981 clone_ 177**	Derived from ATCC 42981; *MATα*^{P}*::kanMX4*	na; this work

Table 2. Composition of the media used in this study. All reagents were purchased from Oxoid (Milan, Italy), with the exception of NaCl (Sigma Aldrich, Milan, Italy). Abbreviations: LiAC/SS carrier DNA/PEG, lithium acetate/single-stranded carrier DNA/PEG; h, hours; d, days.

Usage	Code	Composition	Growth conditions
Yeast routine growth; **RNA extraction**	YPD	1% w/v yeast extract, 2% w/v peptone, 2% w/v dextrose	30°C; 24-48h
Antibiotic sensitivity tests	YPD + ClonNAT	1% w/v yeast extract, 1% w/v peptone, 2% w/v dextrose, ClonNAT range: 50-5 µg/ml	27°C; 48-72h
	YPD + G418	1% w/v yeast extract, 1% w/v peptone, 2% w/v dextrose, G418 range: 400-200 µg/ml	27°C; 48-72h
LiAC/SS carrier DNA/PEG transformation method	2X YPD	2% w/v yeast extract, 4% w/v peptone, 4% w/v dextrose; optional 80 mg/L adenine hemisulfate for growth of adenine auxotrophic mutant BW31a	30°C; 24h
***S. cerevisiae* electroporation protocol**	YPDA + Sorbitol	1% w/v yeast extract, 2% w/v peptone, 2% w/v dextrose, 1,5% w/v agar, 1 M sorbitol; when required 5µg/ml or 7.5 µg/ml ClonNAT was added	26°C; 2/3 days
***Z. rouxii* electroporation protocol**	YPDA + NaCl	1% w/v yeast extract, 2% w/v peptone, 2% w/v dextrose, 300 mM NaCl, 1,5% w/v agar; when required 5µg/ml ClonNAT	26°C; 2/3 days
***E. coli* routine growth**	LB	1% w/v tryptone , yeast extract 5% w/v, 1% w/v NaCl, Adjust the pH to 7.0 with 5.0 N NaOH	37°C; 24h

DNA manipulations, standard and colony PCR reactions

Genomic DNA (gDNA) from yeast cells was isolated according to Hoffman and Winston (1987). Plasmid DNA (pDNA) from yeast and *E. coli* cells were obtained with Zymoprep Yeast plasmid Miniprep kit (Zymo Research, Orange, CA, USA) and GeneJET Plasmid Miniprep Kit (Thermo Scientific, Waltham, MA), respectively. DNA quantity and quality were evaluated electrophoretically and spectrophotometrically using a NanoDrop ND-1000 device (Thermo Scientific, Waltham, MA, USA). Zymoclean™ Gel DNA Recovery and DNA Clean & Concentrator™-5 Kits (Zymo Research, Orange, CA, USA) were used for the isolation of DNA fragments from agarose gels and for PCR amplicons purification, respectively. For plasmid isolation and purification of PCR products from gel or PCR mixtures we used Zymo Research kits (Zymo Research, Orange, CA, USA). Backbone PCR amplicons were sequenced by external sequencer service (BMR Genomic Padova, ITA). Restriction reactions were performed according to manufacturer's instructions (Thermo Fisher Scientific, Walthan, MA, USA). For colony PCR, DNA was extracted from 48h-old cells using the lithium acetate-SDS method (Lõoke et al., 2011). DreamTaq polymerase (Thermo Scientific, Waltham, MA) was used according to the manufacturer's instructions in 20 µl reaction volume containing 1 µl of LiOAc-SDS extracted DNA as template. PCR amplifications were carried out in a T100 Thermalcycler (Bio-Rad). *Escherichia coli* Trans1-T1 cells were transformed according to manufacturer's instructions. Standard primers and PCR conditions were detailed in **Table S1**.

Optimization of yeast transformation protocols

For *S. cerevisiae* four DNA transformation methods were tested with YEp352-SAT1 or pCg2XpH-N plasmids (**Table 3**). LiAC/SS carrier DNA/PEG method was according to Gietz and Schiestl (2007). Electroporation was carried out as follows. Briefly, an aliquot of overnight yeast culture was inoculated in 100 ml YPD at the final 0.005 OD_{600nm} of with and grow overnight to 1.3-1.5 OD_{600nm} (corresponding to $1x10^8$ cells/ml). Cells were harvested by centrifugation at 4,000*g* for 5 min at 4°C, washed twice in 40 ml ice-cold sterile water, and finally resuspended by vortexing in 20 ml ice-cold 1 M sorbitol. After recovery by centrifugation, cells were resuspended in 500 µl of ice-cold 1 M sorbitol and keep on ice. 100 µl of the cell suspension were transferred to a 1.5 ml microcentrifuge tube and 0.1 ug of pDNA was added before gently mixing with pipette several times. The mixture was transferred into a pre-chilled 2 mm gap width cuvette (code 5520; Thermo Scientific Waltham, MA) and submitted to electroporation in an electroporator 2510 device (Eppendorf, Hamburg, Germany) with setting voltage pulse to 1,500 V (constant pulse duration 5 ms). Immediately after electroporation, 500 µl of YPD + Sorbitol medium were added to cell suspension and incubated at room temperature for 15 min. Next, the mixture was transferred into a 1.5 ml microcentrifuge and supplemented with further 500 µl of RT YPD + Sorbitol. After incubation at 30°C for 3-4 h, cells were properly ten-fold diluted and plated into YPD + Sorbitol medium supplemented with appropriate antibiotic concentration and incubated at 30°C for 3-4 days.

Transformation efficiency was calculated according to **Formula 1**, where E_t is transformation efficiency, N_c is colony number and DF is dilution factor.

Formula 1.

$$E_t = \frac{(N_C \; x \; DF)}{\mu g \; pDNA}$$

Table 3. Protocols tested for *S. cerevisiae* cells transformation. All protocols were tested with *S. cerevisiae* haploid strain Y13925 (EUROSCARF).

Protocol n°	Inoculum medium	Transformation protocol	Plasmids	Selectable marker	Recovery step (Time h, T°C)	Plating medium
1	YPD	Electroporation	YEp352-SAT1	ClonNAT^R	4, 30	YPDA + 1M Sorbitol + 5 µg/ml ClonNAT
2	YPD	Electroporation	YEp352-SAT1	ClonNAT^R	3, 30	YPDA + 1M Sorbitol + 7.5 µg/ml ClonNAT
3	YPD	LiAC/SS carrier DNA/PEG method	YEp352-SAT1	ClonNAT^R	3, 30	YPDA + 5 µg/ml ClonNAT
4	2X YPD	LiAC/SS carrier DNA/PEG method	YEp352-SAT1 or pCg2XpH-N	ClonNAT^R	3, 30	YPDA + 5 µg/ml ClonNAT

For *Z. rouxii* transformation, electroporation protocol was modified from Pribylova and Sychrovà (2003) using pCg2XpH-N as tester plasmid, as detailed in **Table 4**. In all transformation experiments, plasmids were at the final concentration of 0.1 µg/µl (Llopis-Torregrosa et al., 2016; see **Table 3**). Electroporation was carried out using standard electroporation cuvettes (0.2 cm) (code 5520; Thermo Scientific, Waltham, MA) and Electroporator 2510 (Eppendorf, Hamburg, Germany) delivering square wave pulses with setting voltage pulse to 2,250 V (constant pulse duration 5 ms). All the transformation experiments were performed using 10^8 cells and verified both by diagnostic PCRs and phenotypic assays on YPDA supplemented with antibiotic. Primers for diagnostic PCRs are listed in **Table S1**. In case of transformation with pCg2XpH-N, fluorescence signal of yeast cells expressing pHluorin was observed under a Nikon Eclipse 80i Olympus AX 70 (Olympus Corporation) microscope using a U-MWB cube with a 450–480 nm excitation filter and 515 nm barrier filter or under Nomarski contrast.

Table 4. Protocol for *Zygosaccharomyces* transformation by electroporation. Protocol was modified from Pribylova and Sychrovà (2003). Main modifications are in bold.

Step	Description
1	Grow cells in 80 ml of YPD + NaCl till the exponentially phase at OD_{600nm} ~ 0.7-0.8
2	Wash cells with 25 ml ddH_2O (3,500*g*, 5 min; 4°C)
3	Resuspend cells in 16 ml of TE buffer supplemented with **20mM LiAC** and **25 mM dithiothreitol**. Incubate at 30°C for 30 min with shaking
4	Chill the cuvettes and 1M sorbitol on ice for next steps
5	**After cell recovery at 3,500*g* for 5 min (4°C), wash cells with 20 ml ice-cold 1M sorbitol**
6	**Repeat washing with 5 ml ice-cold 1M sorbitol**
7	Repeat washing with 800 µl ice-cold 1M sorbitol
8	Transfer 100 µl of cell suspension into a chilled electroporation cuvette and add 0.1 µg pDNA
9	Place the cuvette in the electroporation chamber and apply an electric pulse of 2250 V/cm, 5 ms
10	Immediately after the pulse, add 100 µl of ice-cold 1M sorbitol and incubate at room temperature for 10 min
11	Transfer cell suspension in 5 ml of YPD medium and incubate at 30°C for at least 2 h
12	After cell recovery at 3,500*g* for 5 min (4°C), resuspend cells in 200 µl ddH_2O
12	Plate transformed cells on YPD supplemented with appropriate antibiotic concentration

Plasmids construction

Plasmids were constructed by HR (Oldenburg et al., 1997) in *S. cerevisiae*, and named according to **Table 5**. Briefly, to replace *ScURA3* marker in pZEU and pZCA with the ClonNAT-resistance, *sat1* gene was PCR-amplified from YEp352-SAT1 (Krauke and Sychrovà, 2011) with primers pZEU/pZCA-URA3-

NAT-F and either pZEU-URA3-NAT-R or pZCA-URA3-NAT-R, respectively. The resulting *SAT1* cassette contained *C. albicans ACT1* exon 1 and intron in frame fused with *sat1* gene (Garraway et al., 1994; Reuß et al., 2004). To replace *ScURA3* marker with the G418 resistance in pZEU and pZCA, the *loxP–kanMX–loxP* module was PCR-amplified from pUG6 (Güldener et al., 1996) using the primer pZEU/pZCA-URA3-kanMX-F either with pZEU-URA3-kanMX-R or pZCA-URA3-kanMX-R.

To generate pGRCRE, the fragment containing the *cre* gene behind the *S. cerevisiae GAL1* promoter (pScGAL1) (abbreviated as pScGAL1+*cre*) was PCR-amplified from plasmid pZCRE (Pribylova et al., 2007a) with primers CRE-pGRB-SacI-XbaI-F1 and CRE-pGRB-XhoI-XbaI-R1. Before cloning, two copies of *ScPGK1* promoter-pHluorin module were excised from pCg2XpH-N (Llopis-Torregrosa et al., 2016) with *SacI* and *XhoI* (Thermo Scientific Waltham, MA). The linearized vector was gel-purified and cloned with the PCR insert pScGAL1+*cre* in *S. cerevisiae* BW31a, creating pGRCRE. To avoid mismatch, all PCR reactions used in cloning procedure were carried out either with Phusion Flash High-Fidelity PCR Master Mix or Phusion high Fidelity DNA polymerase (Thermo Scientific Waltham, MA), according to the manufacturer’s instructions. Primers used for plasmid construction are listed in **Table 6**. In all experiments, PCR insert/pDNA molar ratio was 1:3. Plasmids were validated by restriction analysis and diagnostic PCRs according to **Table S1**. Plasmids pZEU, pZCA, pUG6, pZCRE, pCg2XpH-N and YEp352-SAT1 were a gift from Hana Sychrovà.

Table 5. Plasmids used in this work. All plasmids are *Z. rouxii*/*E. coli* shuttle vectors (*ori*, Amp^R). Abbreviations: Zr, *Zygosaccharomyces rouxii*; Sc, *Saccharomyces cerevisiae.*

Plasmid name (GenBank number)	Description	History	References
pUG6 (AF298793)	*loxP*-flanked marker gene deletion cassette (*loxP*-pAg*TEF1-kanMX*-tAg*TEF1-loxP*); G418 resistance (G418^R) selectable marker	Plasmid pUG6 carrying *loxP–kanMX–loxP* module was constructed from plasmid pFA6-kanMX4 by integrating two 34 bp *loxP* sequences as direct repeats left and right of the *kanMX* module	Güldener et al*.,* 1996
YEp352-SAT1	2µ *ori*; nourseothricin resistance gene *sat1* (ClonNAT^R); Amp^R	YEp352 plasmid backbone; *ScURA3* marker was replaced with nourseothricin resistance gene *sat1* (ClonNAT^R)	Krauke and Sychrovà, 2011
pCg2XpH-N	*C. glabrata* centromere and autonomously replicating sequence (*ARS*)-based plasmid, containing two *S. cerevisiae PGK1* promoters (pScPGK1), two pHluorin genes, and nourseothricin resistance (ClonNAT^R) as selectable marker	Derived from pGRB2.2-pHluorin++, which contains two copies of the sequence encoding ratiometric pHluorin each under pScPGK1 promoter, and the *ScURA3* marker. In pCg2XpH-N, *ScURA3* marker gene was replaced with a nourseothricin resistance gene (*sat1*) from YEp352-SAT1	Reuß et al., 2004; Krauke and Sychrovà, 2011; Ullah et al*.*, 2013; Llopis-Torregrosa et al*.*, 2016
pZEU (AM696689)	Yeast replicon pSR1; *ScURA3*; *LacZ*; Amp^R	*Z. rouxii* episomal vector carrying *ScURA3* marker gene	Pribylova et al., 2007a
pZEN	pSR1 yeast replicon; nourseothricin resistance (ClonNAT^R); *LacZ*; Amp^R	*Z. rouxii* episomal vector derivative of pZEU; *ScURA3* marker gene was replaced with nourseothricin resistance gene *sat1* (ClonNAT^R) from YEp352-SAT1	This work
pZEG	pSR1 yeast replicon; G418 resistance (G418^R); *LacZ*; Amp^R	*Z. rouxii* episomal vector derivative of pZEU; *ScURA3* marker gene was replaced with *kanMX* cassette (G418^R) from pUG6	This work
pZCA (AM697670)	*ScARS1*/*ZrCENA*; *ScURA LacZ*; Amp^R	Centromeric plasmid carrying *Z. rouxii* centromere A	Pribylova et al., 2007b
pZCAN	*ScARS1*/*ZrCENA*; nourseothricin resistance (ClonNAT^R*)*; *LacZ*; Amp^R	*Z. rouxii* centromeric vector derivative of pZCA; *ScURA3* marker gene was replaced with nourseothricin resistance gene *sat1* (ClonNAT^R) from YEp352-SAT1	This work
pZCAG	*ScARS1*/*ZrCENA*; G418 resistance (G418^R); *LacZ*; Amp^R	*Z. rouxii* centromeric vector derivative of pZCA; *ScURA3* marker gene was replaced with *kanMX* cassette (G418^R) from pUG6	This work

pZCRE (AM697668)	pSR1 yeast replicon; *ScURA3*; *cre*, pScGAL1; Amp^R; no *LacZ*	*Z. rouxii* expression vector derivative of pZEU; *cre* gene under the *S. cerevisiae GAL1* promoter (pScGAL1)	Pribylova et al., 2007a
pGRCRE	*C. glabrata* centromere and autonomously replicating sequence (*ARS*)-based plasmid, containing *S. cerevisiae PGK1* promoter, nourseothricin resistance gene (ClonNAT^R), cre, *pScGAL1*	Derived from pCg2XpH-N: the two pHluorins are replaced with *cre* recombinase under the *S. cerevisiae GAL1* promoter; the *URA3* marker gene was replaced with a nourseothricin resistance gene (ClonNAT^R) amplified from YEp352-SAT1	This work

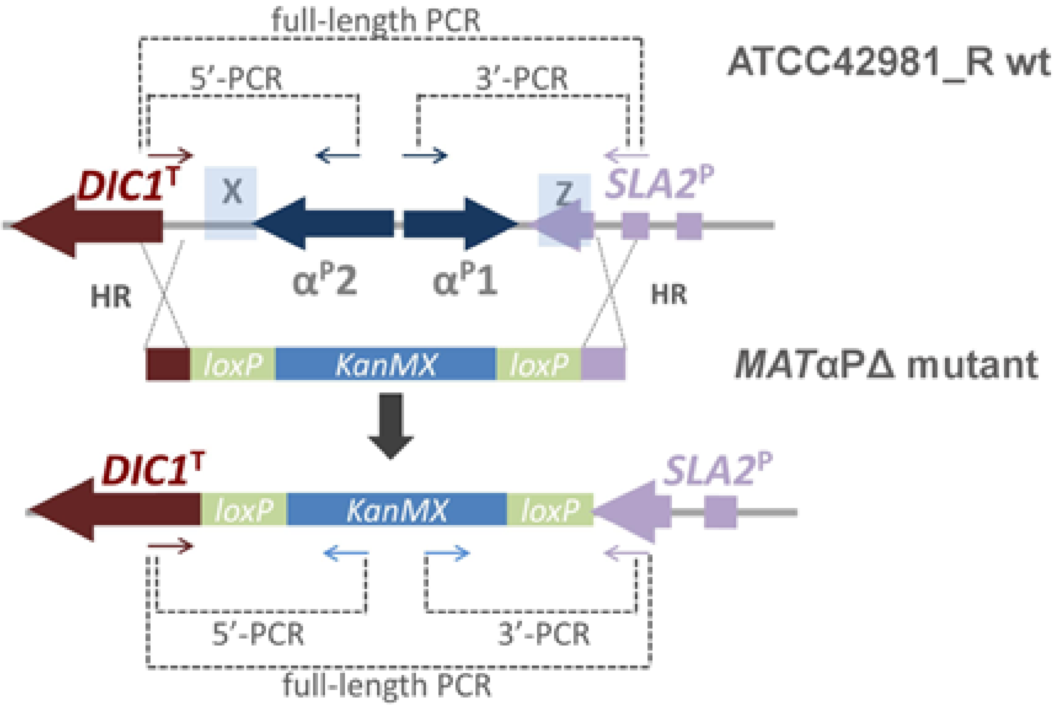
Table 6. Primer used for plasmid construction and gene disruption cassette. PCR insert containing cre gene under *S. cerevisiae GAL1* promoter is referred to as pScGAL1+cre. Abbreviation: Sc, *S. cerevisiae*.

Primer	Sequence (5'->3')	Description
pZEU/pZCA-URA3-NAT-F	GTGAGTTTAGTATACATGCATTTACTTATAATACAGTTTTatggacggtggtatgttttagtttagc	pZEU/pZCA common forward. Uppercase: region homologous to pZEU/pZCA *ScURA3;* lowercase: region homologous to YEp352-SAT1 upstream *sat1*
pZEU-URA3-NAT-R	CTTAACCCAACTGCACAGAACAAAAACCGGAAACGAAGATAAATCttaggcgtcatcctgtgctcccg	pZEU specific reverse. Uppercase: region homologous to pZEU *ScURA3*; lowercase: region homologous to YEp352-SAT1 downstream *sat1*
pZCA-URA3-NAT-R	CCCAACTGCACAGAACAAAAACATGCAGGAAACGAAGATAAATCttaggcgtcatcctgtgctcccg	pZCA specific reverse. Uppercase: region homologous to pZCA *ScURA3*; lowercase: region homologous to YEp352-SAT1 downstream *sat1*
pZEU/pZCA-URA3-kanMX-F	GTGAGTTTAGTATACATGCATTTACTTATAATACAGTTTTttcgtacgctgcaggtcgac	pZEU/pZCA common forward. Uppercase: region homologous to pZEU/pZCA *ScURA3*; lowercase: region homologous to pUG6 upstream *loxP–kanMX–loxP*
pZEU-URA3-kanMX-R	CTTAACCCAACTGCACAGAACAAAAACCGGAAACGAAGATAAATCgcataggccactagtggatctg	pZEU-specific reverse. Uppercase: region homologous to pZEU *ScURA3*; lowercase: region homologous to pUG6 downstream *kanMX*
pZCA-URA3-kanMX-R	CCCAACTGCACAGAACAAAAACATGCAGGAAACGAAGATAAATCgcataggccactagtggatctg	pZCA-specific reverse. Uppercase: region homologous to pZCA *ScURA3*; lowercase: region homologous to pUG6 downstream *kanMX*
CRE-pGRB-SacI-XbaI-F1	GCGCGCAATTAACCCTCACTAAAGGGAACAAAAGCTGGAGCTCgaattcgagctctagtacggat	Uppercase: region homologous to pCg2XpH-N upstream *Sac*I restriction site; lowercase: region homologous to pZCRE upstream pScGAL1+*cre*
CRE-pGRB-SacI-XbaI-R1	CTCTGTTTGTGTGATTTCTATGTGTACGTTATATATATAtggctaatcgccatcttcca	Uppercase: region homologous to pCg2XpH-N downstream *Xho*I restriction site; lowercase: region homologous to pZCRE downstream pScGAL1+*cre*
MATα1/2cp2-kanMX-F-80nt	CATGTTTGAACGAGTGTTTTGTTCATTGGTTTGGAATAAACAGGTCTTCGACGTTTAGCCATGTCGAGGATTTAAACGTTTGACAttcgtacgctgcaggtcgac	Uppercase: 80-bp-long region homologous to *MAT*α 5'-flanking region (between *DIC1* and X region); lowercase: sequence homology to amplify *loxP-KanMX-loxP* cassette in pUG6
MATα1/2cp2-kanMX-R-80nt	CAACCGGTAAGTGTTCTTTCAATAAGTCAGTTGTGCAATGAAGTGGCAAGTCAGTTTTTAAGCAACACACCGCACGTACCGgcataggccactagtggatctg	Uppercase: 80-bp-long region homologous to *MAT*α 3'-flanking region (between Z region and *SLA2*); lowercase: sequence homology to amplify *loxP-KanMX-loxP* cassette in pUG6

Construction of *loxP-kanMX-loxP* disruption cassette and gene deletion

The *loxP-kanMX-loxP* cassette containing the 80-bp-long target sequence homologous to the *MAT*α expression cassette (*DIC1-MAT*α*-SLA2*) was PCR-amplified from pUG6 (Güldener et al., 1996) using MATα1/2cp2-kanMX-F-80nt and MATα1/2cp2-kanMX-R-80nt (**Table 6**). After PCR product purification, 0.3 µg of deletion cassette was used to transform ATCC 42981. Transformation mixture was incubated for 2 h in YPD at 30°C, before plating on selective YPDA medium supplemented with 200 µg/ml of G418. Targeted integration of *loxP-kanMX-loxP* cassette was verified by full-length, 5'- and 3'-end diagnostic PCR analyses (**Fig. 1**), using primers detailed in **Table S1**. The resulting deletants were named according to **Table 1**.

Figure 1. Outline of the diagnostic PCR strategy used to verify the targeted integration of *loxP-kanMX-loxP* disruption cassette in ATCC 42981 genome. Full-length, 5' and 3' approaches are shown. Abbreviations: wt, wild type; cp, copy; HR, homologous recombination; *kanMX*, kanamycin resistance gene; *MAT*αcp2, *MAT*α copy 2 expressed cassette.



Plasmid loss test

ATCC 42981 *MATαΔ* transformants with pGRCRE plasmid were grown with shaking in 20 ml YPD at 28°C for 72 h. Cultures were ten-fold diluted and spread on YPD plates to obtain single colonies (approximately 200 colonies per plate). Colonies were replica-plated on YPD supplemented with 5.0 µg/ml of ClonNAT and after 3 days ClonNAT-sensitive colonies were selected and further submitted to at least two additional plasmid loss assays.

RNA extraction, cDNA synthesis and RT-PCR

RNA was extracted from ATCC 42981 wild type and deletion mutant cells cultured in YPD and harvested in stationary phase (Gietz and Schiestl, 2007; Solieri et al., 2016). RNAs were reverse transcribed using 0.5 µM oligo (dT) and RevertAid H Minus Reverse Transcriptase (Thermo Scientific, Waltham, USA) according to the manufacturer's instructions. cDNAs (25 ng) were amplified using DreamTaq polymerase and *cre* and *SOD2* gene-specific primers listed in **Table S1**.

Results

Antibiotic sensitivity tests

The first issue concealing the ATCC 42981 genetic manipulation was that this strain is prototrophic for the main amino acids used as markers in auxotrophic complementation experiments. This hampered the usage of main plasmids available for *Zygosaccharomyces* (Pribylova et al., 2007a) and the related species *C. glabrata* (Zordan et al., 2013), which contain nutritional markers instead of antibiotic-resistance markers. To avoid the search for auxotrophic mutants derived from ATCC 42981, we tested the sensitivity towards two antibiotics, namely G418 and ClonNAT. We found that ATCC 42981 cells were sensitive to G418 and ClonNAT at the minimal concentrations of 200 µg/ml and 5-10 µg/ml, respectively (data not shown). This result indicates that G418 and ClonNAT were suitable as selection agents for ATCC 42981 genetic manipulation.

Protocols optimization for *S. cerevisiae* and *Z. rouxii* transformation

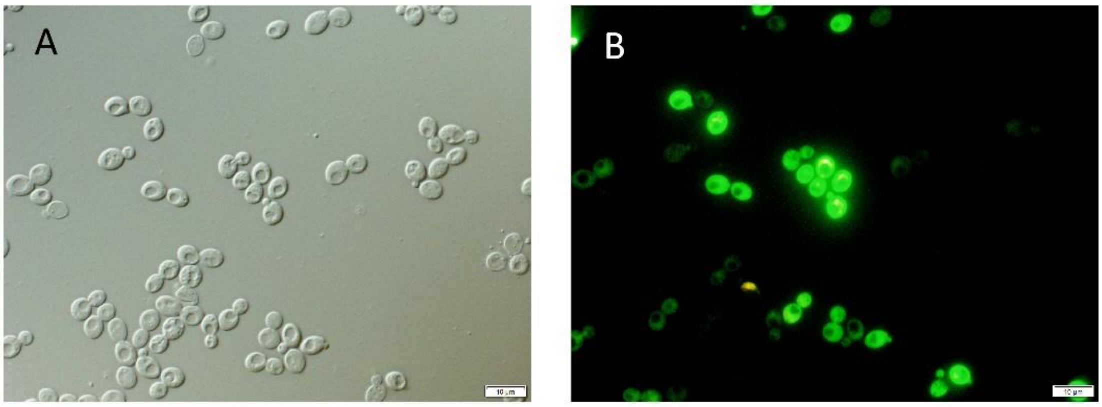
Before plasmid construction by HR, we tested both electroporation and alkali cation-based methods to transform *S. cerevisiae* strains Y13925 and BW31a using either YEp352-SAT1 or pCg2XpH-N as plasmid testers. Strain BW31a overcame Y13925 in transformation efficiency and was therefore selected for the next manipulation steps (data not shown). For each protocol the transformation efficiency was calculated (**Table 7**). Although all tested protocols resulted in high transformation efficiency, we selected the protocol n°4, which allowed a faster cell growth compared with the other methods and decreased the background of small colonies which presumably arise from abortive transformation events. The selected protocol was also positively tested to transform *S. cerevisiae* cells with pCg2XpH-N (E_t = 9520), giving fluorescent cells (data not shown).

Table 7. Comparison of transformation results in *S. cerevisiae*. N_{t0} indicates number of cells before transformation; N_t , number of transformants; µg pDNA, the quantity of plasmid DNA used; $E_{t,}$ calculated efficiency of transformation; background, the presence of false positive colonies under ClonNAT selection; growth rate cells before transformation.

Protocol n°	**N_{t0}**	**N_t**	**pDNA +(µg)**	**E_t**	**Background**	**Growth Rate**
1	2 x 10^{10}	uncountable	0.1	/	High	Low
2	2 x 10^{10}	1586	0.1	15860	High	Low
3	1 x 10^{8}	2960	0.1	29600	High	Low
4	1 x 10^{8}	91	0.1	910	Low	High

Differently from *S. cerevisiae*, *Zygosaccharomyces* cells are not as easy to transform. Furthermore, Pribylova and colleagues (2007a) observed a different response to transformation between the two most studied *Z. rouxii* strains, namely CBS 732^T and ATCC 42981, and hypothesized that a different organization in the glucan-chitin-cell wall proteins network is responsible for the higher resistance to transformation observed in ATCC 42981 compared to CBS 732^T. In order to bypass this obstacle, an electroporation-based method should be preferred over the standard lithium acetate methods to transform *Zygosaccharomyces* cells (Watanabe et al., 2010). We modified the standard electroporation-based protocol described for *Zygosaccharomyces* haploid strains by Pribylova and Sychrovà (2003), in order to improve the ATCC 42981 transformation. The transformation procedure was optimized using pCg2XpH-N, which harbours *sat1* gene encoding for ClonNAT resistance (Reuß et al., 2004) and two green fluorescent reporter genes, encoding for pH-sensitive ratiometric pHluorins (Llopis-Torregrosa et al., 2016). The ATCC 42981 transformants were positively selected on YPD plates supplemented with either 5 µg or 10 µg/ml ClonNAT. However, ClonNAT concentration of 10 µg/ml strongly reduced the transformation efficiency (data not shown). The majority of screened colonies were positive to the 5' and 3' diagnostic PCR assays (data not shown) and were intracellularly fluorescent (**Fig. 2**). Fluorescence intensity is susceptible to pH variation, as pHluorins are reporters of pH dynamics in living cells and can be tagged for organelle specificity (Ullah et al., 2013). In future ATCC 42981 cells transformed with pCg2XpH-N can be exploited to test how *Zygosaccharomyces* yeast varies intracellular pH in response to environmental cues.

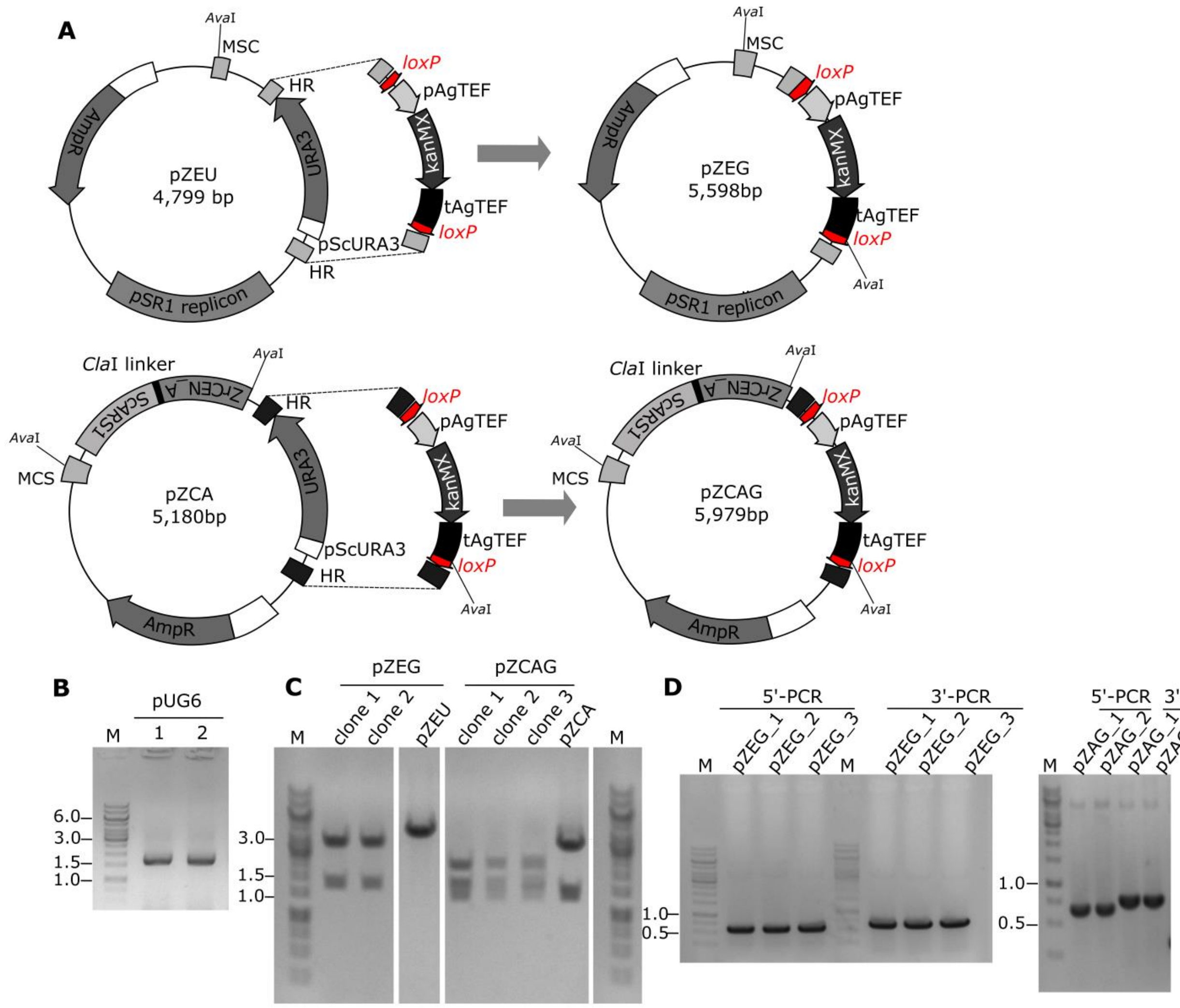
Figure 2. Modified electroporation-based protocol produces fluorescent ATCC 42981 cells when transformed with pCg2XpH-N vector. Intracellularly fluorescent ATCC 42981 transformants (B) are compared to cells microscopically observed with optical microscope under Nomarski contrast (A).



Construction of *Z. rouxii* episomal and centromeric plasmids containing G418-resistance cassette

To date, no *Z. rouxii* episomal and centromeric plasmids harbouring antibiotic selection marker has been available. To develop a first set of G418^R marked vectors, we used pZEU and pZCA as backbones with *ScURA3* auxotrophic marker. pZEU is an episomal vector containing the *Z. rouxii* pSR1 replicon as *ori* (Araki et al., 1985), while pZCA is a centromeric plasmid replicable in *Z. rouxii* owing to *Z. rouxii CEN* A sequence in combination with *S. cerevisiae ARS1* (**Fig. 3**, panel **A**). G418-resistance cassette was successfully amplified from pUG6 with the chimeric primers pZEU/pZCA-URA3-kanMX-F/pZEU-URA3-KanMX-R and pZEU/pZCA-URA3-kanMX-F/pZCA-URA3-kanMX-R, respectively (**Fig. 3**, panel **B**). The resulting 1,763 bp-long PCR amplicons contain tails homologous to *ScURA3* upstream and downstream regions in pZEU and pZCA, respectively. In both cases, PCR products encompassed the complete *loxP-kanMX-loxP* module, including two *loxP* flanking regions, pAgTEF, *kanMX* ORF from *E. coli* transposon Tn903 and tAgTEF (Wach et al., 1994; Güldener et al., 1996). *S. cerevisiae* BW31a was suitable to replace *ScURA3* in pZEU and pZCA with the corresponding *loxP-kanMX-loxP* modules by HR. The resulted plasmids, termed pZEG and pZCAG respectively, were validated by restriction analysis with endonuclease *Ava*I, diagnostic 5'- and 3'-end PCRs (**Fig. 3**, panels **C** and **D**, respectively), as well as sequencing (data not shown). The structures of pZEG and pZCAG are drawn in **Figure 3**, panel **A**. To test the functionality of the novel plasmids, *Z. rouxii* ATCC 42981 and CBS 732^T were transformed with pZEG and pCAG, and, in both cases, G418-resistant transformants were successfully isolated (data not shown).

Figure 3. Construction of episomal and centromeric vectors conferring G418 resistance. A) HR strategy was used to build episomal and centromeric vectors pZEG and pZCAG, respectively. Schematic maps of both novel dominant drug resistance marker plasmids were drawn. In both *loxP-kanMX-loxP* modules *loxP* regions were represented as red arrows, while 40-bp long pZEU and pZCA target sequences as light and dark grey blocks, respectively. B) PCR construction *loxP-kanMX-loxP* modules for *ScURA* replacement in pZEU and pZCA, respectively. Lanes 1 and 2 represent PCR products obtained with chimeric primers pZEU/pZCA-URA3-kanMX-F/pZEU-URA3-KanMX-R and pZEU/pZCA-URA3-kanMX-F/pZCA-URA3-kanMX-R, respectively. C) *Ava*I-digestion of plasmids pZEG and pZCAG and the corresponding native plasmids pZEU and pZCA. D) Validation of pZEG and pCAG with 5'- and 3'- diagnostic PCR assays. Sizes of PCR fragments (in Kb) are given on the left side of the agarose gels. For each plasmid, at least two clones were analysed. M, molecular weight marker GeneRuler 1 Kb DNA Ladder (Thermo Scientific, Waltham, MA).



Construction of *Z. rouxii* episomal and centromeric plasmids containing ClonNAT-resistance *SAT1* gene

Nourseothricin (ClonNAT) is an aminoglycoside glycopeptide (nucleoside peptide) antibiotic of the Streptothricin class which inhibits protein biosynthesis and induces miscoding. Chemically it is a mixture of Streptothricins C, D, E, F (D + F >85%) which differ for number of ß-lisine residues (n = 1-7). ClonNAT resistance is conferred by *sat* or *nat* marker genes. *nat1* and *nat2* genes are from *Streptomyces krusei* (Krügel et al., 1983), while *sat1* gene from bacterial transposon Tn 1825 (Heim et al., 1989). In *S. cerevisiae nat1* gene was joined to the *TEF* gene promoter and terminator at 5'- and 3'-end, respectively, resulting in *natMX* resistance cassettes (plasmids pAG25, pAG35 and pAG36; Goldstein et al., 1999). Similar synthetic constructs are not functional in *Z. rouxii*. In C. albicans *sat1* gene from pELSAT plamid (Garraway at al., 1997) was fused in frame with 501 bp of upstream sequences and the first 15 codons including the intron of the *C. albicans ACT1* gene. The construct was also joined to the transcription termination sequence of the *C. albicans URA3* gene, leading to the final *CaSAT1* marker (Reuß et al., 2004). *CaSAT1* was succefflully cloned in YEp352 by HR in *S. cerevisiae*, resulting in YEp352-SAT1 (Krauke and Sychrovà, 2011). During pCg2XpH-N construction *CaSAT1* from YEp352-SAT1 was shortened in the promoter and terminator sequences and found to be functional in conferring ClonNAT resistance to *C. glabarata*, which is a *Z. rouxii* closely related species (Llopis-Torregrosa et al., 2016). We firstly amplified *CaSAT1* from YEp352-SAT1, resulting in several aspecific PCR fragments (data not shown). Results from Llopis-Torregrosa et al. (2016) suggested that regulatory elements upstream and downstram *sat1* gene were not necessary to confer ClonNAT resistance. Thus, the chimeric primer pairs were designed to include *ACT1* exon 1 and the subsequent intron as well as the full-lengtht *sat1* ORF (**Table 6**). PCRs resulted in two specific 1, 407bp PCR fragments (**Fig. 4**, panels **A** and **B**). These minimal *SAT1* markers were used to succesffully replace *URA3* gene both in pZEU and in pZCA, as showed by restriction analyses and diagnostic PCRs (**Fig. 4**, panels **C** and **D**). The resulting plasmids, termed pZEN and pZCAN, were used to transform ATCC 42981. We found that a recovery step for 3h at 30°C in YPD was required to express *SAT1* gene (**Fig. 5**).

Figure 4. Construction of episomal and centromeric vectors conferring nourseothricin resistance. A) HR strategy was used to build episomal and centromeric vectors pZEN and pZCAN, respectively. Schematic maps of both novel dominant drug resistance marker plasmids were drawn. In both cases *C. albicans ACT1* exon 1 and intron are depicted as brawn block ad black line, respectively, while 40-bp long pZEU and pZCA target sequences as light and dark grey blocks, respectively. B) PCR construction of minimal *SAT1* markers for *ScURA* replacement in pZEU and pZCA, respectively. Lanes 1 and 2 represent PCR products obtained with chimeric primers pZEU/pZCA-URA3-NAT-F/pZEU-URA3-NAT-R and pZEU/pZCA-URA3-NAT-F/pZCA-URA3-NAT-R, respectively. C) *Ava*I- and *BamH*I digestion of plasmids pZEN and pZCAN and the corresponding native plasmids pZEU and pZCA. pZEU_und. indicates undigested pZEU. D) Validation of pZEN and pZCAN with 5'-end diagnostic PCR assay. The sizes of the PCR fragments (in Kb) are given on the left side of the agarose gels. For each plasmid, at least two clones were analysed.

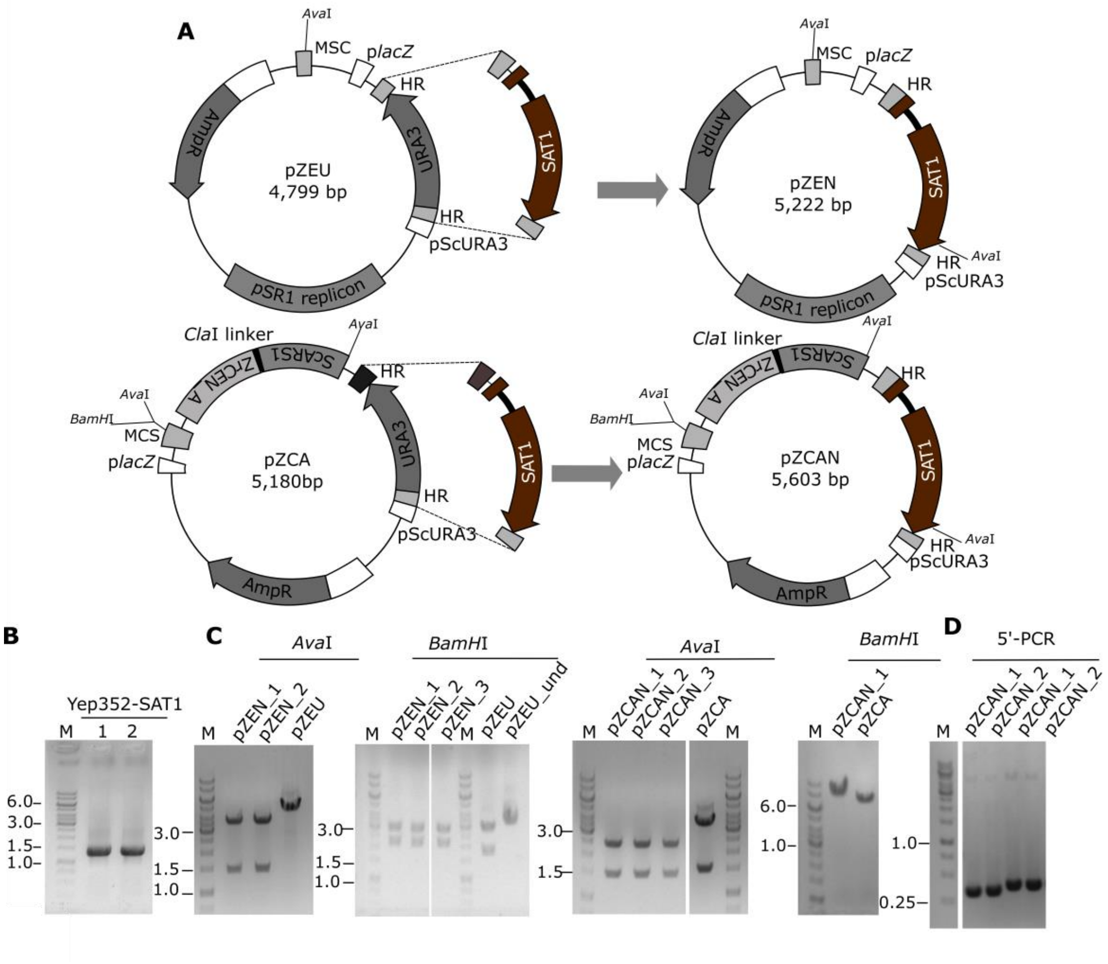
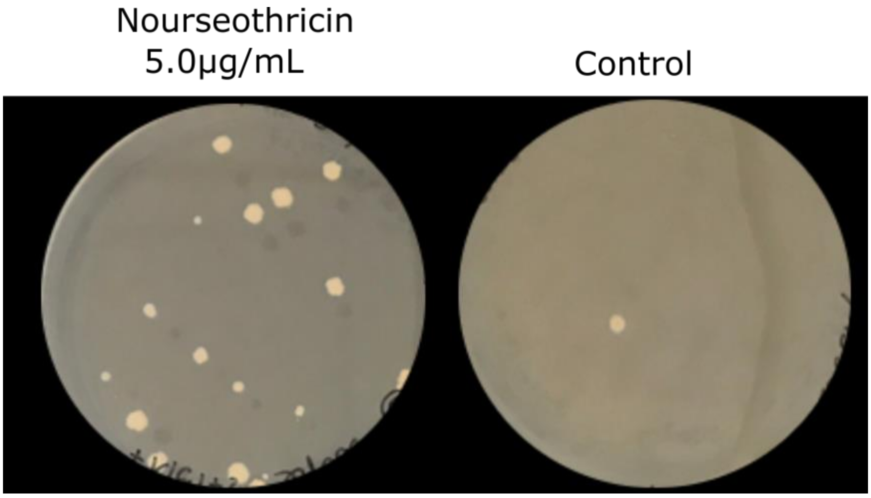


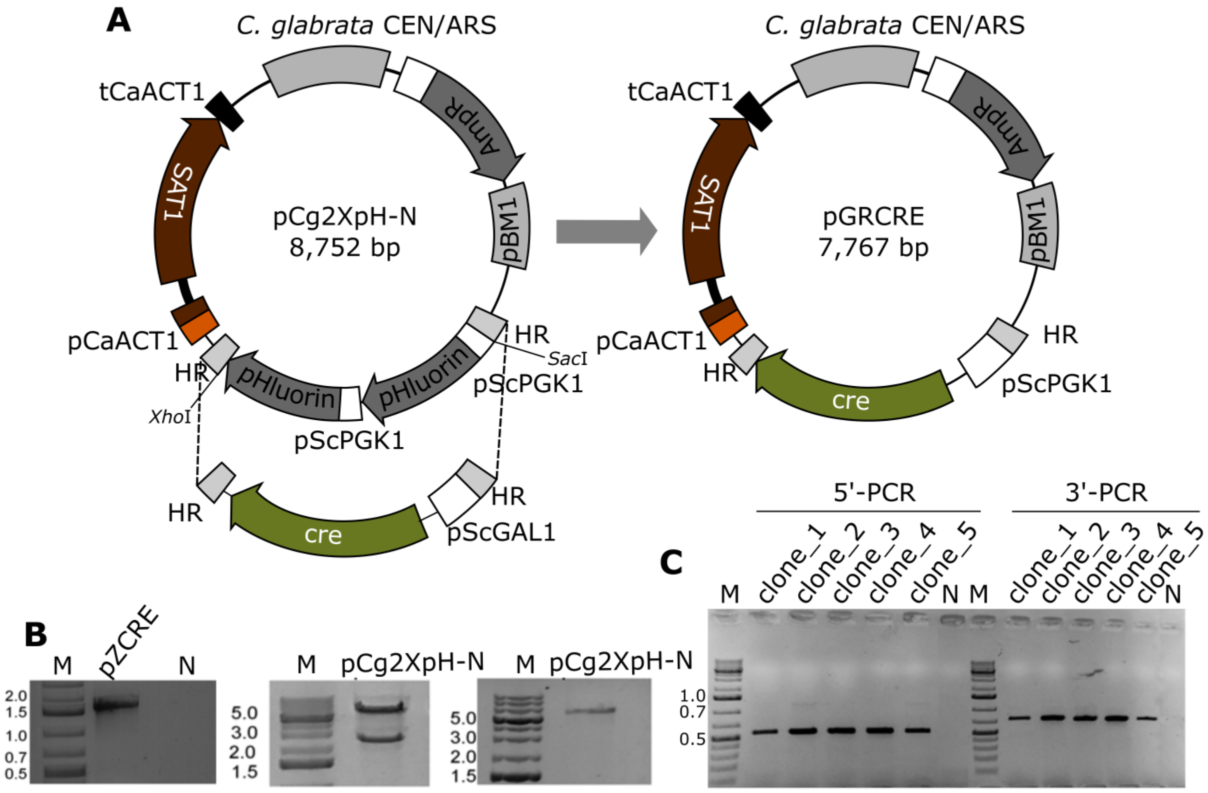
Figure 5. Selection of ATCC 42981 ClonNAT^{R} transformants obtaining by electroporation of ATCC 42981 cells with 0.1 µg pZEN.



pGRCRE construction

To construct the first *Z. rouxii* plasmid harbouring *cre* gene and ClonNAT^R as selectable marker, we cloned *cre* gene under the control of the inducible pScGAL1 promoter in pCg2XpH-N (Llopis-Torregrosa et al., 2016) by HR in *S. cerevisiae* (**Fig. 6**, panel **A**). Firstly, PCR amplification of pZCRE plasmid (Pribylova et al., 2007a) with chimeric primers CRE-pGRB-SacI-XbaI-F1 and CRE-pGRB-XhoI-XbaI-R1 resulted in a pScGAL1+*cre* PCR insert with the expected size of 1,600 bp (**Fig. 6**, panel **B**). Double digestion of pCg2XpH-N plasmid with endonucleases *Sac*I and *Xho*I resulted in two electrophoretic bands of 6,180 and 2,572 bp, respectively (**Fig. 6**, panel **B**). Fragment corresponding to pHluorin genes was discarded, while the 6,180 bp fragment corresponding to the pCg2XpH-N linearized plasmid was successfully gel-purified and used for HR procedure together with pScGAL1+*cre* (**Fig. 6**, panel **B**). The resulting novel plasmid, named pGRCRE, contained pScGAL1+*cre* module instead of double copy of pHluorin genes (**Fig. 6**, panel **A**). *S. cerevisiae* ClonNAT-resistant clones were screened by colony PCRs. Fragment orientation was checked by amplifying the 5’ and 3’ ends of the construct with one primer annealing on pCg2XpH-N backbone and the other on pScGAL1+*cre* fragment. Out of the 7 colonies screened, 5 clones were postivie to both diagnostic PCRs, indicating that pScGAL1+*cre* construct was correctly inserted into pGRCRE plasmid (**Fig. 6**, panel **C**). pDNA was extracted from BW31a ClonNAT^R clones and Sanger-sequenced in order to confirmed that the pScGAL1+*cre* PCR insert was instead of the original pHluorin genes (data not shown).

Figure 6. pGRCRE construction. A) HR strategy was used to build the centromeric vector pGRCRE containing pScGAL1+*cre* insert and ClonNAT^R selectable marker. Schematic maps of pCg2XpH-N and the novel pGRCRE plasmid were drawn. *C. albicans ACT1* exon 1 and intron in-frame fused with *sat1* gene are depicted as brawn block ad black line, respectively, while 80-bp long pCg2XpH-N target sequences as light grey blocks (HR). *C. albicans ACT1* promoter (pCaACT1) and terminator (tCaACT1) are partial compared to the corresponding regulatory elements in YEp352-SAT1 (Krauke and Sychrovà, 2011) and SAT1-flipper plasmid AY524979 (Reuß et al., 2004). B) Gel electrophoresis steps of pGRCRE construction: pScGAL1+*cre* insert was PCR-amplified from pZCRE; pCg2XpH-N was linearized with *Sac*I and *Xho*I enzymes and the pCg2XpH-N backbone without pHfluorins was gel-purified before transformation in BW31a. C) 5’- and 3’-ends diagnostic PCR assays of nourseothricin-resistant BW31a clones harbouring pGRCRE. Sizes of PCR fragments (in Kb) are given on the left side of the agarose gels. N, negative control; M, molecular weight markers GeneRuler 1 Kb or 100bp Plus DNA Ladders (Thermo Scientific, Waltham, MA).

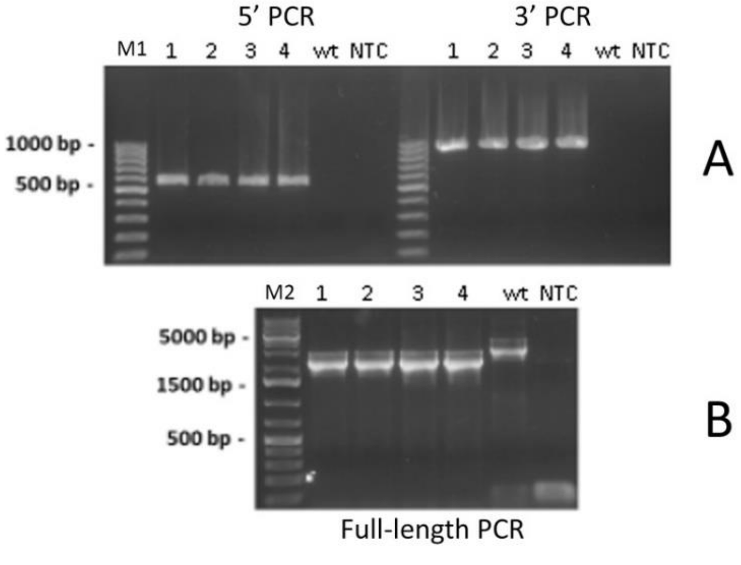


Construction of ATCC 42981 *MATαΔ* mutants

DIC1-MAT-SLA2 arrangement characterizes the expressed *MAT* cassette in all the pre-WGD species (Gordon et al., 2011). ATCC 42981 genome contains one *DIC1-MAT*α*-SLA2* locus, which was selected as target of gene disruption strategy. In order to achieve this goal, we designed one primer pair for the HR between *loxP-kanMX-loxP* and the target cassette *DIC1-MAT*α*-SLA2* (**Table 6**). *S. cerevisiae* is very prone to micro-homologous recombination and only 40 bp of target sequence homology are enough to assure high efficiency in targeting deletion (20 - nearly 100% depending on the targeted locus; Baudin et al., 1993). Pribylova and colleagues (2007a) reported that *loxP-kanMX-loxP* cassette containing 80-bp-long target sequence homology was more effective than that containing 40-bp-long target sequence homology in replacing the target sequence in *Z. rouxii* UL4 haploid strain (a derivative from CBS 732^T). Therefore, we decided to use 80 bp-long target sequence homology for target gene deletion in allodiploid ATCC 42981 too. In particular, the forward primer MATα1/2cp2-

kanMX-F-80nt had 80 bp-long upstream sequence homologous to the intergenic region between *DIC1* and *MAT*α2 copy2 genes outside the X region, while the reverse primer MATα1/2cp2-kanMX-R-80nt contained a 80 bp-long sequence homologous to *SLA2* gene downstream the Z region (**Table 6**). The resulting chimeric primers were used to PCR-amplified the *loxP-kanMX-loxP* cassette from pUG6 plasmid, giving the PCR amplicon of expected XX size (data not shown). Strain ATCC 42981 was transformed by electroporation with the PCR-generated *loxP-kanMX-loxP* cassette. Several G418^R transformants were selected on YPD supplemented with 200 µg/ml G418. Four ATCC 42981 *MAT*αΔ mutants, namely clones_6, clone_65, clone_74 and clone_177, were confirmed by 5'- and 3'-end diagnostic PCR analysis, as well as by a full-length PCR approach spaning the entire *loxP–kanMX–loxP* module (**Fig. 7**). In particular, panel B shows that PCR reactions with primers on *DIC1* and *SLA2 MAT*-flanking genes gave the expected 2,164 bp-long PCR amplicon signal in all deletants, compared to the 2,889 bp-long PCR amplicon obtained for wild type ATCC 42981.

Figure 7. Validation of ATCC 42981 *MATαΔ* mutants by 5', 3' and full-length PCR reactions. Panel A reports the 5' and 3' PCR products, which indicate the presence of the *kanMX* cassette in the deletion mutants and the absence in the ATCC 42981 wild type-strain. Panel B reports full-length *MAT*α copy 2 (2,889 bp) and *kanMX* (2,164 bp) PCR products in the wild type and the mutant cells, respectively. Numbers from 1 to 4 indicate deletion clones 6, 65, 74, and 177, respectively. Abbreviations: M1, molecular weight marker GeneRuler 100 bp DNA Ladder (Thermo Scientific, Waltham, MA); M2, molecular weight marker GeneRuler 1 Kb Plus DNA Ladder (Thermo Scientific, Waltham, MA); wt, wild type; NTC, no template control.



Recycle of *kanMX-loxP* module

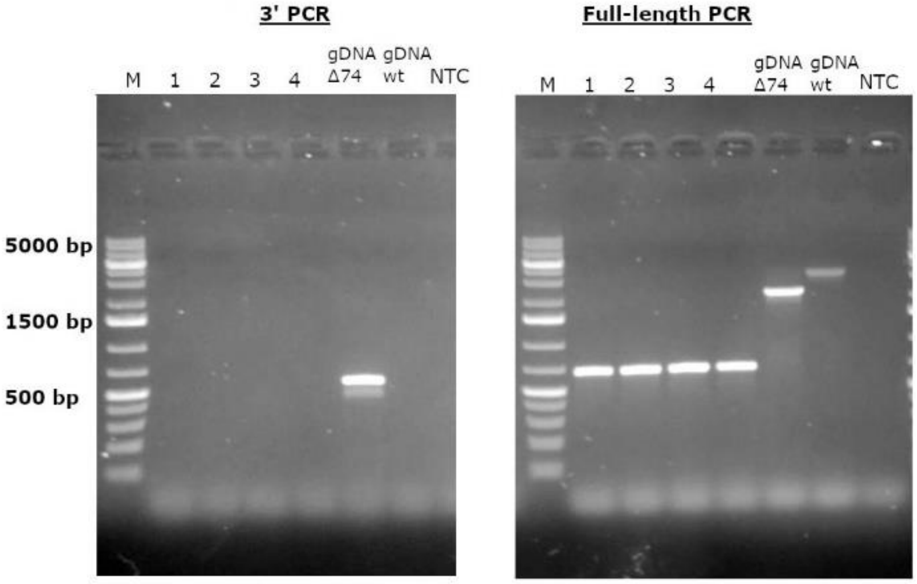
In *S. cerevisiae*, the *kanMX–loxP* module can be removed by Cre recombinase-mediated recombination between the two *loxP* sites, when cells are grown in presence of galactose, as *cre* gene is under the control of the *S. cerevisiae* pScGAL1 promoter from the pSH47 plasmid (Güldener et al., 1996). This promoter should belong to the inducible promoters, *i.e*. they enable the up-regulation of gene expression via the addition of an inducer chemical at the desired processing time. In particular, shifting the cells from glucose to galactose-containing medium induces *cre* gene expression and, consequently, *kanMX-loxP* module excision (Güldener et al., 1996).

To validate the Cre recombinase from pGRCRE for marker recycling in *Z. rouxii* protrophic strains, we transformed the previously obtained ATCC 42981 *MATαΔ* mutants with pGRCRE by electroporation. Out of four *MAT*αΔ mutants, ATCC 42981_Δ74 and ATCC 42981_Δ177 were successfully transformed with pGRCRE, giving transformants capable to grow on 5.0 µg/ml of ClonNAT. The pGRCRE presence in ATCC 42981 transformants was further demonstrated by 5' and 3'-ends diagnostic PCRs (data not shown; diagnostic primers listed in **Table S1**). However, we noticed that ATCC 42981 *MAT*αΔ cells transformed with pGRCRE were resistant to ClonNAT, but lost the G418^R phenotype (data not shown). This result suggested that, differently from *S. cerevisiae*, in strain ATCC 42981 *cre* gene was constitutively expressed under pScGAL1 promoter. According to our results, Pribylova and colleagues (2007) found that in medium with glucose pScGAL1 was only slightly repressed in comparison with medium with galactose. Galactose is also reported to be an ineffective carbon source for *Z. rouxii* strain CBS 732^T and galactose assimilation is a variable trait in this species (Kurtzman et al., 2001). Similarly, a leaky glucose repression of the *pScGAL1* promoter was also observed in *K. lactis* (Steensma and Linde, 2001).

To confirm the removal of *kanMX-loxP* module without any galactose induction, we designed two diagnostic PCRs. In 3'-end PCR approach forward primer annealed on *kanMX* cassette and reverse primer on 3' *MAT*α-flanking gene *SLA2.* We expected negative result for all the G418-sensitive clones. In full-length PCR approach, primers on *DIC1*^{T} and *SLA2*^{P} genes flanking the *loxP-kanMX-loxP* cassette should give a 700 bp-long product in clones without *kanMX-loxP* module, while it should result in a 2,166 bp-long amplicon in G418^R *MAT*αΔ cells which still contained the *kanMX* marker. Finally, a 2,889 bp-long amplicon was expected for the wild type-strain harbouring the native *MAT*α locus surrounded by *DIC1*^{T} and *SLA2*^{P}, respectively. **Figure 8** showed that ClonNAT-resistant and G418-sensitive *MAT*αΔ clones have lost the *kanMX-loxP* module. Sanger sequencing of the 3'-end

PCR product confirmed that 34 bp-long *loxP* scar remained after *kanMX-loxP* removal (data not shown). Overall these data demonstrated that pGRCRE can be successfully employed for selectable marker recycling, but the constitutive *cre* gene expression makes necessary to lose pGRCRE after marker recycling, in order to carry out further genetic manipulations.

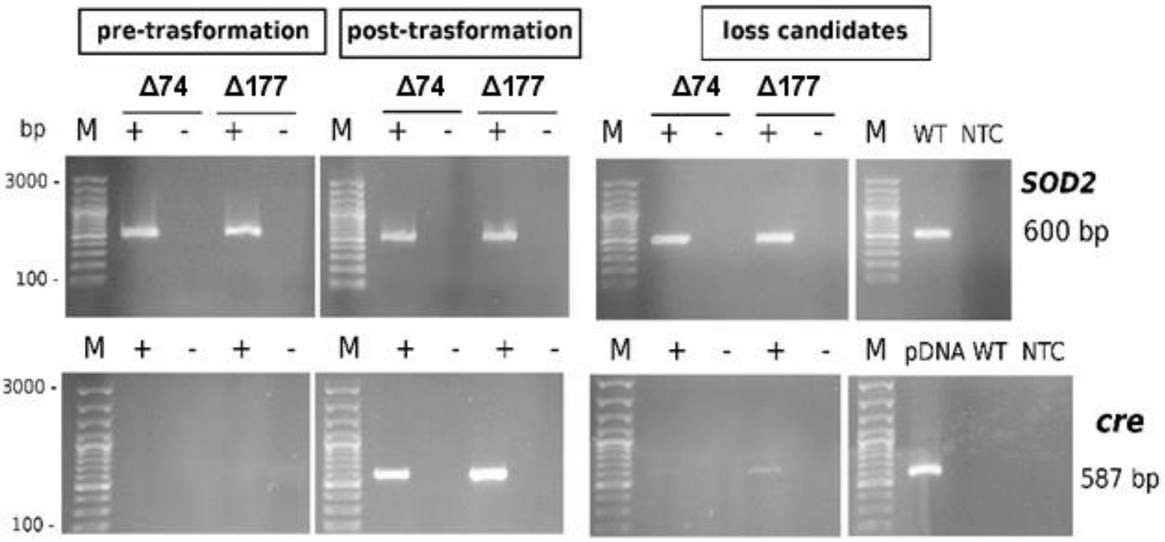
Figure 8. PCR-validation of *kanMX-loxP* removal from ATCC 42981 *MAT*αΔ clone_74. 3'-end and full-length diagnostic PCRs were performed with primer pairs listed in Table S1. Lanes from 1 to 4 represent G418-sensitive ATCC 42981 *MAT*αΔ transformants. Abbreviations: M, molecular weight marker Gene Ruler 1 Kb Plus DNA Ladder (Thermo Scientific, Waltham, MA); NTC, no template control, wt, wild type.


3' PCR
Full-length PCR
M 1 2 3 4 gDNA Δ74 gDNA wt NTC
M 1 2 3 4 gDNA Δ74 gDNA wt NTC
5000 bp
1500 bp
500 bp

pGRCRE plasmid loss assay

Clones that excised the *kanMX* marker were grown on unselective medium for several generations and then screened for ClonNAT^R phenotype by replica plating in ClonNAT-containing medium, in order to gather ClonNAT sensitive clones. After three rounds of plasmid loss assays, we obtained two ClonNAT sensitive candidates for both ATCC 42981_Δ74 and ATCC 42981_Δ177 (data not shown). RT-PCR with specific primers annealing on pScGAL1+*cre* insert validated the pGRCRE loss (**Fig. 9**). *Cre*-specific amplicons generated from cDNAs sampled before transformation with plasmid pGRCRE, after transformation and after loss plasmid assay, showed that, apart from a slight background expression, *cre* transcript signals significantly decreased after plasmid loss assay.

Figure 9. ***cre* gene expression in ATCC 42981_Δ74 and ATCC 42981_Δ177 mutants before and after pGRCRE plasmid loss assay**. Mutants resulted negative to *cre* transcripts before transformation and positive after pGRCRE transformation. After plasmid loss assay *cre* gene expression decreased. The + or - indicates cDNA synthesis reaction with or without reverse transcriptase, respectively. *SOD1* was used as housekeeping gene. gDNA and pDNA (pGRCRE) were used as positive PCR controls. Abbreviations: M, molecular weight marker Gene Ruler 100 bp Plus DNA Ladder (Thermo Scientific, Waltham, MA); WT, wild type; NTC no template control.



Concluding remarks

In the present work, we constructed four plasmids containing drug resistance dominant markers, namely pZEG, pZEN, pZCAG, pZCAN, and build a Cre/*loxP* recombination system for their recycling. These new bio-bricks will be useful for recursive genetic manipulations of industrially relevant *Z. rouxii* strains, which are generally prototrophic. The new plasmid pGRCRE harboured *cre* recombinase under pScGAL1 promoter and the *sat1* gene conferring ClonNAT^R phenotype as dominant marker. We demonstrated that this plasmid is effective in *kanMX* marker rescue and that it can be lost during recursive cultivation under unselective conditions. We also provided compelling evidence that, differently from *S. cerevisiae*, in *Z. rouxii* pScGAL1 is not leaky in glucose-containing medium, suggesting that when *cre* gene is under pScGAL1, the loss of the *loxP*-flanked marker takes place without any preliminary galactose induction. This result indicates that promoter element transferability between *S. cerevisiae* and *Z. rouxii* is not as high as occurs in other non-conventional yeasts, and suggests that appropriate inducible promoters should be specifically selected for *Z. rouxii*.

Author Contributions

SC and LS contributed to the conception and design of the study. MB conducted the experiments described in this work. HS and MD contributed to the study by providing plasmids, reagents and materials. MB wrote the manuscript and LS contributed to draft revision.

CHAPTER 7: EVALUATION OF EXTRACTION METHODS FOR HIGH-QUALITY DNA EXTRACTION SUITABLE FOR NANOPORE SEQUENCING AND *DE NOVO* HYBRID ASSEMBLY OF *ZYGOSACCHAROMYCES ROUXII* ATCC 42981 AND *ZYGOSACCHAROMYCES SAPAE* ATB301^{T} GENOMES

Bizzarri M., Cassanelli S., Leszek L. P, Gawor J., Gromadka R., and Solieri L. Draft genome sequences of the highly halotolerant strain Zygosaccharomyces rouxii ATCC 42981 and the novel allodiploid strain Zygosaccharomyces sapae ATB301^{T} obtained using the MinION platform. (2018). Microbiology Resource Announcements. 7(4), e00874-18. doi: 10.1128/MRA.00874-18.

Abstract

Over the recent years, third generation sequencing technologies (3GS) have revolutionized scientific research with their applications to high-throughput analysis of biological systems. Among them, long-read MinION sequencing from Oxford Nanopore Technology (ONT) is transforming our ability to *de novo* sequencing highly complex genomes, greatly improving assembly contiguity, since the long reads expand into problematic or repetitive regions peculiar of highly heterozygous genomes. Realising MinION full potential requires high quality, pure and intact high molecular weight (HMW) genomic DNA (gDNA) from the organisms of interest. Here, we established a workflow for MinION genome sequencing of two non-conventional allodiploid yeasts, namely *Zygosaccharomyces rouxii* ATCC 42981 and *Zygosaccharomyces sapae* ABT301^{T}, which are important halotolerant and osmotolerant biocatalyzers used in food fermentation and spoilage agents. In particular, we evaluated four different DNA extraction procedures to determine the most effective one for extracting HMW gDNAs from these yeasts. Moreover, we compared Covaris g-tube and BluePippin electrophoresis as DNA shearing and size selection systems for obtaining a 20 Kb template preparation. Following the workflow illustrated, we obtained >6 Gb of long reads sequencing data for both strains, with a mean read length of 26 Kb and 28 Kb as well as a read length N50 of 33 Kb and 38 Kb for ATCC 42981 and ABT301^{T}, respectively. We envision that our workflow for establishing MinION library preparation and sequencing, including the illustration of potential strengths and weaknesses, will be useful to others who plan to set-up long-read sequencing of challenging allodiploid genomes. Finally, we illustrated the flow-chart of the hybrid assembly approach adopted for both target strains that combines short reads from Illumina platform and long reads from MinION sequencing. Sequencing results revealed genome sizes of 20.9 and 24.7 megabase (Mb) for ATCC 42981 and ABT301^{T}, respectively. We are confident that this information will be useful for deciphering the genetics of hybrid adaptation to high salt and sugar concentrations in non-conventional yeasts.

Keywords: *Zygosaccharomyces*, draft genome announcement, MinION, nanopore sequencing, hybrid assembly, heterozygosity, gDNA extraction, method evaluation.

Introduction

The revolution of genome sequencing is continuing after the next generation sequencing (NGS) technology. The third-generation sequencing (3GS) technology, led by Pacific Biosciences (PacBio), is progressing rapidly with the release in 2014 by Oxford Nanopore Technologies (ONT), of the MinION, the first commercial sequencer using Nanopore technology (Giordano et al., 2017). MinION identifies DNA bases by measuring the changes in electrical conductivity generated as DNA strands pass through an engineered pore across a chemical gradient (Jain et al., 2016). This sequencing platform is characterized by portability, affordability and great speed in data production and it is capable of producing long sequencing reads with average fragment lengths of over 10,000 base-pairs and maximum lengths reaching 100,000 base-pairs. This technology can sequence DNA fragments of varied lengths, from a few hundred bases to over a Mb, which compares favourably to sequencing by synthesis (e.g. Illumina), which is limited to hundreds of bases (Leggett and Clark, 2017). Compared to NGS short reads, MinION long reads have a number of important applications. They greatly improve assembly quality and quantity, since long fragments, despite having higher error rates at the base level, are able to expand into problematic or repetitive regions that are particularly abundant in diploid genomes, in which the extent of heterozygosity can vary dramatically across chromosomal regions, leading to more fragmented assemblies. MinION library preparation is similar to that for other NGS applications, requiring DNA shearing, end repair, adaptor ligation and size selection. Library preparation takes about half a day and is of comparable complexity and cost to library preparation for other platforms. More recently, improvements in the protein pore (a laboratory-evolved *Escherichia coli* CsgG mutant named R9.4), library preparation techniques (1D ligation and 1D rapid), sequencing speed (450 bases/s), and control software allowed the usage of Nanopore sequence data, in combination with other sequencing technologies, for assemblies of eukaryotic genomes, including fish (Read et al., 2017), yeasts (Istace et al., 2017; Jansen et al., 2017), fungi (Dutreux et al., 2018), and human (Jain et al., 2018). In short, Nanopore sequencing solves the technical challenges of reading long DNA fragments, while still having room for improvement in terms of accuracy. One of the main remaining challenges is to extract highly concentrated, pure and intact long gDNA fragments from the target organisms that have to be sequenced. This democratization of sequencing forces every laboratory to establish the sequencing platform and, more important, new DNA extraction and library preparation procedures. This can be challenging and time consuming.

Here, we present the workflow we applied to set up MinION sequencing using *Zygosaccharomyces* yeasts as a case study. They find relevant applications in food spoilage and fermentation (Dakal et al., 2014) and exhibit high diversity in response to high solute concentration, tendency to hybridization and ectopic recombination at the mating-type loci, leading to ploidy and karyotype variation (Watanabe et al., 2013; Bizzarri et al., 2016). *Z. rouxii* ATCC 42981 is an allodiploid strain, isolated from Japanese miso, which grows at NaCl and dextrose concentrations up to 3.0 M and 70% w/v, respectively (Solieri et al., 2014a). *Zygosaccharomyces sapae* represents a novel species, firstly described in high-sugar traditional balsamic vinegar (TBV), for which ABT301^T (= CBS 12607^T = MUCL 54092^T = UMCC 152^T) is the type-strain (Solieri et al., 2013b). ABT301^T is a sugar-resistant and slow-growing strain more sensitive to salt than ATCC 42981. Under standard conditions, ATCC 42981 produces more glycerol than ABT301^T and better retains it into the cell under salt stress (Solieri et al., 2016). ATCC 42981 was supposed to evolve by hybridization between two divergent parents (Bizzarri et al., 2016; James et al., 2005; Gordon and Wolfe, 2008; Watanabe et al., 2017), while no evidences about origin are available for ABT301^T.

The objective of this work was to sequence and assembly the genome of the allodiploid strain ATCC 42981 and the *Z. sapae* strain ABT301^T by exploiting a *de novo* hybrid strategy, which combined

MinION long and Illumina short reads. Preliminarily, we reported reliable and repeatable ways of measuring DNA purity and integrity to optimize input for the MinION sequencer.

In particular, we evaluated four different methods for HMW gDNA extraction from *Zygosaccharomyces* yeast cells, which are reported to have a different cell wall composition compared to the model species *Saccharomyces cerevisiae* (Pribylova and Sychrovà, 2003). Moreover, we discussed important consideration about the extraction methods, showing that small alterations or modifications in sample preparation protocols can dramatically alter DNA fragment lengths, and compared size selection methods using electrophoresis versus DNA shearing. Overall, we can conclude that our workflow is applicable well beyond the case of study presented here.

Materials and Methods

Yeast cultures

Yeast cells were grown on YPD (1% yeast extract, 2% peptone, 2% glucose) medium. Single-colony isolates were obtained from the Unimore Microbial Culture Collection (UMCC) of the University of Modena and Reggio Emilia, Italy. *Z. rouxii* ATCC 42981 is our in-house stock culture that comes from the ATCC culture collection, while *Z. sapae* ABT301^T was isolated from a TBV sample in May-June 2004 (Solieri et al., 2006).

DNA extraction methods

Total genomic DNA (gDNA) was isolated from stationary phase culture (~12.0 OD$_{600)}$ derived from a single colony. Lyticase from *Arthrobacter luteus*, activity ≥33 kat/mg protein (Sigma, cat. no. L2424) and Zymolyase 20T, activity 1.7 kat/mg protein (Seikagaku America, cat. no. 120491-1) were used as lysing enzymes. Genomic DNA for Illumina sequencing platform was extracted according to Hoffman and Winston (1987), which is a standard phenol-chloroform based extraction protocol containing mechanical cell lysis with acid-washed glass beads (Sigma, G8772). For long-read sequencing with MinION platform, the following extraction methods were tested: (1) Hoffman and Winston (1987) phenol-chloroform based with beat beads lysis (PC_B); (2) Hoffman and Winston (1987) modified with enzymatic cell lysis (FC_L); (3) Wizard® Genomic DNA Purification Kit

(Promega, Madison, WI, USA) according to the manufacturer's instructions (W); (4) a slightly modified Teeny Prep protocol (Boeke et al., 1985) detailed in **Table 1** (TeP_L). For each extraction method and for each strain gDNA was extracted in four replica. Genomic DNA was eluted in 50 µl with 10mM Tris-HCl pH 8.0.

Table 1. Modified Teeny Prep protocol tested to isolate gDNA from yeast samples. Protocol was modified from Boeke et al. (1985). Main modifications are in red. Abbreviations: conc., concentration; Rxn, reaction.

Step	**Description**		
Genomic DNA isolation			
1	For each sample, use a total of 12 OD_{600} to prepare gDNA		
2	Add 1 ml of 1M Sorbitol/0.1M EDTA to the cells		
3	Spin at top speed for 10 min. to collect pellet		
Cell lysis			
1	Resuspend pellet in 300 µl of 1M Sorbitol/0.1M EDTA+14 mM 2-Mercaptoethanol (b-ME)+0.5 mg/ml Lyticase. The mixture is prepared as follows:		
	Stock reagent	Volume (µl)	Final conc.
	1M Sorbitol/0.1M EDTA	300	~1M Sorbitol/0.1M EDTA
	14.3M b-ME	0,35	14mM b-ME
	8 mg/ml Lyticase (Sigma, cat. no. L2424)	19	300U/µl
2	Incubate at 37°C for 1h for Lyticase treatment. *Spheroplasted yeast cells look much darker at the microscope than unspheroplasted yeast cells.*		
3	Spin at top speed for 30 sec. Resuspend in 400 µl of 1M TE pH 8.0. *If spheroplasting worked, the pellet is very steacky and difficult to resuspend.*		
4	Add 100 µl of lysis buffer to each tube. Mix by inversion ten times. The lysis buffer is prepared as follows:		
Lysis Buffer	Volume for 26 Rxn (ml)	Volume per Rxn (µl)	Final conc. including 400 µl TE
0.5M EDTA pH 8.0	1,5	55,5	55 mM
2M Tris-HCl pH 8.0	0,6	22,2	89 mM
10% SDS	0,6	22,2	0.44%
5	Incubate at 65°C for 30 min.		
6	Add 100 µl of 5M potassium acetate pH 4.8, mix well and incubate at 4°C for 1h		
7	Before centrifuging, add 50 µl of chloroform to each reaction. Mix well.		
8	Spin at top speed for 15 min. at 4°C. Transfer supernatant (~600 µl) to a new tube		
9	Add equal volume (~600 µl) of 100% ethanol. Spin at top speed for 10 min. at 4°C		
10	Wash in 70% ethanol. Incubate at RT for 5 min. Resuspend in 400 µl of 1M TE buffer pH 8.0		

RNase treatment and ethanol precipitation of genomic DNA	
1	Add 2,5 µl of 10 mg/ml RNase A (Sigma). Incubate at 37°C for 30 min. and overnight at 4°C
2	Extract with equal volume (400 µl) of phenol:chloroform:isoamyl alchol (25:24:1)
3	Extract with equal volume (400 µl) of chloroform
4	Precipitate with 120 µl of 7.5M ammonium acetate and 1 ml of 100% ethanol
5	Set at -20°C for 1h
6	Spin at top speed for 15 min. at 4°C. Wash in 70% ethanol. Air dry
7	Resuspend in 50 µl of 10mM Tris-HCl pH 8.0
8	Determine the concentration of gDNA
9	Resolve 2-3 µl of gDNA on a 0.4% agarose gel to determine the quality of the preparation. *The entire gDNA preparation should run as one discrete band larger than 20 Kb in size*

Evaluation of genomic DNA extraction methods

To compare the different extraction protocols described above the quantity and quality of gDNA were evaluated using spectrophotometrical, fluorimetrical and electrophoretical methods. DNA quantity and quality was measured by reading the whole absorption spectrum (220-750 nm) with NanoDrop ND-1000 device (Thermo Scientific, Waltham, MA) and calculating DNA concentration and absorbance ratio at both 260/280 and 230/260 nm. Quantity of gDNA was also assessed using Qubit®3.0 double stranded DNA (dsDNA) BR and HS Assay Kits (Invitrogen, Life technologies), according to the manufacturer's instructions. Qubit fluorimeter calculates concentration based on the fluorescence of a dye, which binds to dsDNA (Simbolo et al., 2013). This is a quantification system relying on dyes that produce fluoresce only when bound to specific molecules, such as dsDNA, ssDNA or RNA. DNA integrity and size were assessed by conventional and pulse field gel electrophoresis (PFGE). In the first case, 1 µl of gDNA sample was loaded in a 0.4% agarose gel containing 0.5% ethidium bromide and prepared in 0.5X TBE buffer, and run at 80V for 4 h before visualization under UV light. In PFGE, DNA samples (4 µl) embedded in agarose plugs were separated on a 1.0% agarose gel in 0.5X TBE buffer chilled to 14°C on a Bio-Rad CHEF DR-III system. The DNA was separated in four subsequent 14 h runs at 6V/cm, using a 120° angle and a constant switching time of 1-6 sec. The gel was afterwards stained with ethidium bromide and imaged.

DNA shearing and target size selection

We evaluated two DNA shearing and size-selection tools for MinION library preparation. Firstly, gDNAs were sheared to 20-Kb fragments using the Covaris g-tube, which uses centrifugal force to pass the DNA sample through a finely engineered ruby shearing orifice, and then captures the fragmented DNA sample in an integrated collection chamber. In particular, 4 µg of gDNA in 150 µL of deionized water was loaded into the Covaris g-tube and spin twice at 6,000 rpm in an Eppendorf 5424 for 1.5 min. To quality check DNA shearing, 5.0 µl and 2.0 µl of sheared and unsheared gDNA samples were loaded and for standard gel electrophoresis, while 5.0 µl and 0.5 µl for PFGE, respectively. Alternatively, gDNA was size-selected using the high-pass BluePippin size selection system (Sage Science, Beverly, MA, USA), that features an alternating field power supply capable of size-fractionating DNA fragments as large as 50 Kb. Fragments ≥ 20 Kb were selected for both ATCC 42981 and ABT301^T strains. Each sample was prepared according to the BluePippin gel cassette manufacturer's protocol (PAC30KB 30-40 Kb 0.75% DF Marker U1 high-pass 30-40 Kb v3). In particular, for each strain, we totally loaded 20 µg of each gDNA divided in four wells. The concentration of gDNA recovered from BluePippin was measured with Qubit 3.0 dsDNA HS assay kit (ThermoFisher, Wilmington, DE, USA). Finally, the recovered gDNA was further concentrated using Amicon® Ultra-4 10K centrifugal filter devices (Merck Millipore Ltd.) by centrifugation for 10 min at 5,000 rpm in the high-speed Eppendorf centrifuge 5804R, and quantified by Qubit 3.0 dsDNA BR assay kit (ThermoFisher, Wilmington, DE, USA).

Illumina library preparation and sequencing

llumina libraries were prepared with an average insert size of ~600 bp and sequenced in paired-end mode on MiSeq instrument using v3 600-cycle chemistry kit.

MinION 1D sequencing library preparation

MinION libraries were constructed using 1D Ligation Sequencing Kit (SQK-LSK108), including the NEBNext FFPE DNA repair step, with the modifications reported in the One-Pot ligation manual (dx.doi.org/10.17504/protocols.io.k9acz2e). In particular, manufacturer-recommended library preparations involving DNA repair and end-prep are optimized for 0.2 pmol of input DNA with an average fragment size of 8 Kb, which in turn requires 1 µg of double-stranded DNA. The DNA mass for 0.2 pmol was calculated using the Promega Biomath calculator (http://www.promega.com/a/apps/biomath/) by setting the average fragment range to 20 Kb. The ideal output of long fragments was ~2.64 µg.

MinION flow cell preparation and sample loading

Libraries were loaded onto an R9.0 flow cell (FLO-MIN107, ONT) according to manufacturer's instructions. All sequencing runs were performed on MinION Mk1b devices for 48 h and, when possible, sequencing time was extended to 72 h. All sequencing runs used MinKNOW software (version 1.2.8). All flow cells used for sequencing underwent the standard MinION Platform QC for analysis of overall quality and number of functional pores.

Basecalling and ssembly assessment

The fast5 reads from 1D MinION runs were base-called using ONT's Albacore v2.1.7 basecaller that converts the raw signal data from the MinION into DNA sequence data in fastq format. Long reads were quality trimmed with PoreChop v0.2.1 (https://github.com/rrwick/Porechop) and error-corrected with Canu v1.7 (Koren et al., 2017). Quast was used to assess quality of genome assemblies (Gurevich et al., 2013).

Results and Discussion

Genomic DNA extraction methods comparison

The impact of DNA molecule length for 3GS, appears to be crucial, in order to assure long reads sequencing. The first step of this work was to optimize extraction protocol to yield highly intact and high purity DNA suitable for MinION sequencing. We evaluated the suitability of four extraction methods, namely PC_B, PC_L, W, and TeP_L in HMW gDNA recovery. These protocols mainly differed in the method (mechanical or enzymatic) used for cell lysis. Zymolyase and Lyticase are the main enzymes used to generate yeast spheroplasts. In literature, only few studies investigated the susceptibility of *Z. rouxii* cell wall to these lysing enzymes. Pribylova and colleagues (2007) found that Zymolyase was more aggressive in lysing *Z. rouxii* type-strain CBS 732^{T} and the allodiploid ATCC 42981 than Lyticase, reducing the spheroplast survival compared to Lyticase. In particular, ATCC 42981 strain was more sensitive to Zymolyase than CBS 732^{T}. Moreover, Solieri et al. (2008) also observed that Lyticase was the more efficient than Zymolyase to generate *Z. sapae* spheroplasts. Based on these observations, in TeP_L protocol, we incubated ATCC 42981 and ABT301^{T} cells with Lyticase for 1 h, instead of adopting treatment with Zymolyase for 30 min conventionally used for *S. cerevisiae* spheroplast generation. Increasing the incubation time allowed us to lyse the ATCC 42981cell wall more efficiently since it is more rigid due to a higher content of chitin and cell wall polymer compared to CBS 732^{T}.

To compare the different tested extraction methods, we firstly evaluated DNA quality and quantity by NanoDrop and Qubit 3.0®, which uses fluorochromes specifically binding dsDNA. For the NanoDrop spectrophotometric measurement, a descriptive statistics relative to all four methods is summarized in **Table 2**. A low 260/280 nm ratio is indicative of contamination with proteins, which could inhibit downstream applications and hamper DNA-banking (Wilson, 1997). A low 260/230 nm ratio is indicative of contamination with organic compounds, like phenol or guanidine, which negatively affect downstream MinION sequencing (Salonen et al., 2010). According to the 260/280 nm absorbance ratio values, PC_B and PC_L protocols did not extract pure DNA suitable for Nanopore sequencing. This result could be attributed to the organic compounds, such as phenol-chloroform used for the extraction. Although TeP_L method also included the use of phenol-chloroform, it gave more satisfactory spectrophotometric measurements than PC_B and PC_L. This could be due to the two precipitation steps included in TeP_L, that allow the salting-out of proteins with 5M potassium acetate and gDNA precipitation with 7.5M ammonium acetate and ethanol. Overall, these results indicate that TeP_L protocol yielded the purest DNA for both *Zygosaccharomyces* strains, followed by W, FC_L, and FC_B, respectively.

Table 2. Descriptive statistics of the Nanodrop measurements used to evaluate the respective purity of DNA extracted with four methods. Abbreviation: SD, standard deviation.

Method	ATCC 42981						ABT301^T					
	OD 260/280			OD 260/230			OD 260/280			OD 260/230		
	Median	Range	SD	Median	Range	SD	Median	Range	SD	Median	Range	SD
FC_B	1.74	1.34-2.08	0.37	2.11	1.53-2.40	0.44	2.03	2.01-2.12	0.06	2.17	2.08-2.30	0.11
FC_L	1.8	1.34-2.07	0.37	2.02	1.53-2.03	0.39	2.03	1.92-2.07	0.08	2.17	2.16-2.30	0.08
W	1.81	1.81-2.12	0.21	1.63	1.09-2.45	0.68	2.04	1.91-2.07	0.09	1.52	1.52-2.26	0.43
TeP_L	2.09	1.98-2.10	0.05	2.45	1.76-2.50	0.29	2.09	2.04-2.12	0.03	2.19	2.12-2.35	0.09

Concentration and total yield of extracted gDNAs were also assessed using the Qubit Platform v3.0, which is the method of choice for accurate estimation of DNA quantity for 3GS techniques (Simbolo et al., 2013). Generally Qubit DNA estimations were lower than Nanodrop, but more reliable for dsDNA quantification since they are minimally affected by interference from RNA, protein, single stranded DNA (primers) or other common contaminants (O'Neill et al., 2011). According to these values, FC_L and TeP_L methods provided enough gDNAs for *de novo* Nanopore sequencing (**Table 3**).

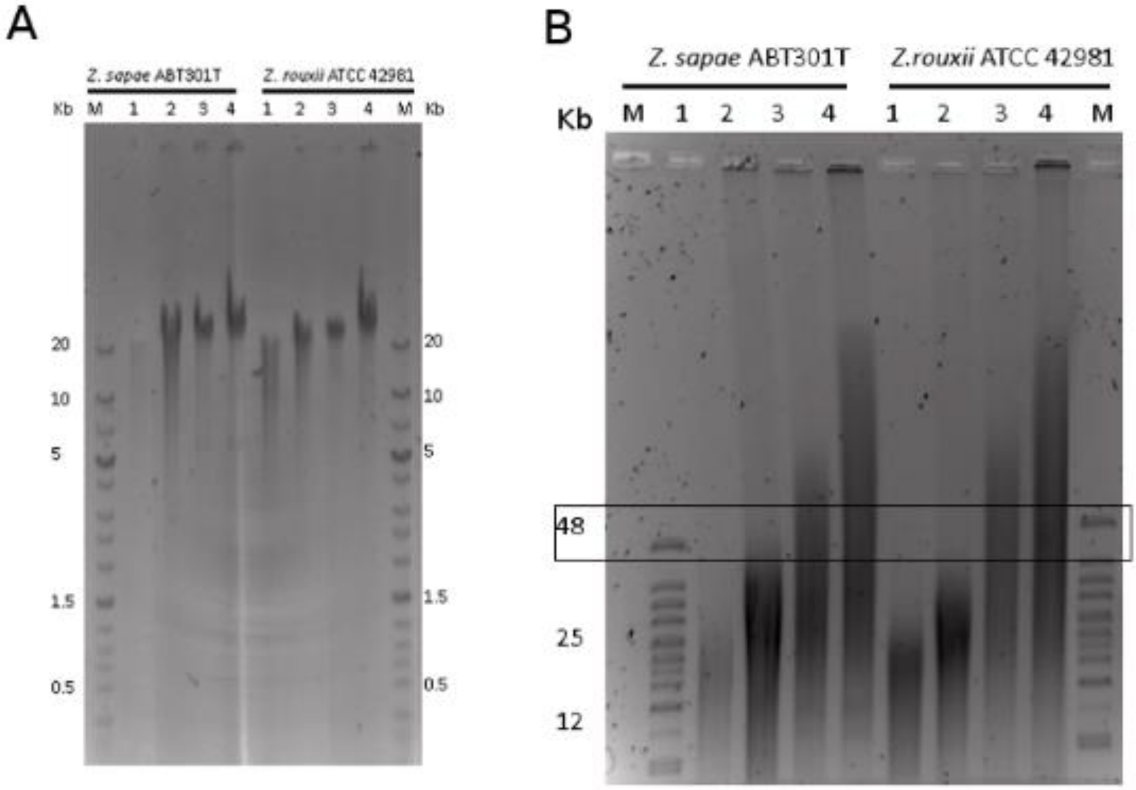
Table 3. Total DNA yields measured with Qubit assay. Concentration measured for one representative replica for each extraction method was reported. For each protocol, total DNA yield extracted with each protocol was calculated by multiplying DNA concentration measured with the Qubit platform with final elution volume.

Method	DNA concentration (ng/µl)	
	ATCC 42981	ABT301^T
FC_B	70	19
FC_L	76	179
W	42	80
TeP_L	117	104

Finally, DNA integrity and size were assessed by agarose gel electrophoresis and PFGE (**Fig. 1,** panel **A** and **B**), respectively. DNA extracted with FC_B protocol showed an evident smear for both strains (**Fig. 1,** panel **A**), suggesting that mechanical cell disruption with beat beading reduces gDNA integrity and decreases the length of fragments. Although PFGE showed that all the DNA samples are quite fragmented, TeP_L protocol assured higher DNA integrity compared to the other protocols, with the majority of DNA fragments in the 48-38 Kb size range (**Fig. 1,** panel **B**).

In conclusion, the evidences supported TeP_L as the most suitable method for yielding HMW gDNA for Nanopore sequencing.

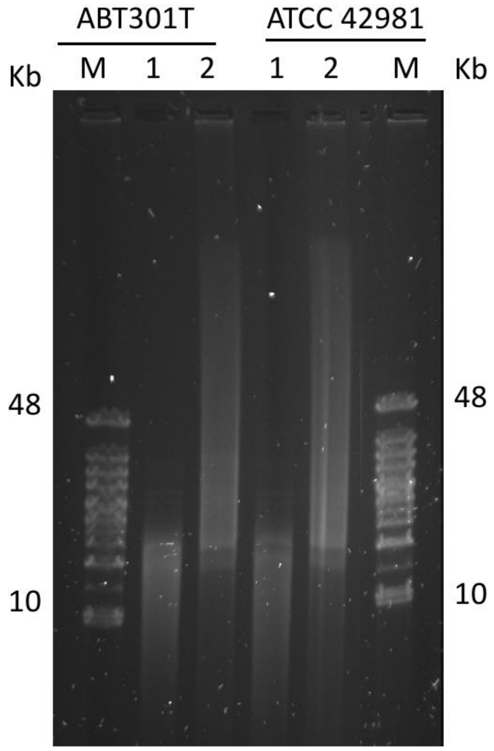
Figure 1.Illustration of the impact on DNA extraction procedures on DNA fragment length. Representative results from standard gel electrophoresis (**A**) and PFGE analysis (**B**) of gDNAs from ABT301^{T} and ATCC 42981 samples extracted by four methods. Lanes from 1 to 4 represent FC_B, FC_L, W, and TeP_L, respectively. Abbreviation: M, molecular weight markers ZipRuler Express DNA Ladder Set (**A**) and Lambda Mix Marker, 19 (**B**) (ThermoFisher, Wilmington, DE, USA).



DNA shearing and size selection

We next assessed the effect of DNA shearing and gel-based size-selection procedures on read length distribution. We defined the optimal gDNA fragment length for MinION platform input of 20 Kb. The Covaris g-tube and the Sage BluePippin system provide a simplified and complementary workflow for the optimization of HMW gDNA fragmentation and size-selection. In the majority of protocols, g-tube and BluePippin systems were combined for producing tight DNA fragment distributions in the 8-12 Kb size range (Wang et al., 2015a). When gDNA sample extracted with TeP_L protocol was shared by Covaris g-tube, DNA fragments shifted to smaller size values compared with unsheared DNA (**Fig. 2**). Therefore, we decided not to perform DNA shearing. The negative result could be due to the low DNA purity because the opening of g-TUBEs can be blocked by particles, resulting in inconsistent performance and occasionally considerable sample loss. Additionally, we tested the effect of removing DNA fragments below 20 Kb by size selection using the BluePippin system in the high-pass mode (20-80 Kb range). We observed a sample recovery of app. 20% and standard gel electrophoresis confirmed DNA fragment distribution around 20 Kb (data not shown). Therefore, we decided to use BluePippin size-selected DNAs for downstream sequencing, since it allowed us to perform target size selection with improved accuracy and sample recovery.

Figure 2. The impact of DNA shearing by Covaris g-tube on DNA fragment distribution. DNA fragmentation was assessed by PFGE. Lanes 1 and 2 represent Covaris g-tube sheared and unsheared gDNAs, respectively. Abbreviation: M, molecular weight marker Lambda Mix Marker, 19 (ThermoFisher, Wilmington, DE, USA).



Illumina sequencing and MinION real time run quality checks

The first NGS sequencing using Illumina platform generated a total of 2,234,027 and 2,649,084 short paired-end reads for ATCC 42981 and ABT301^T, respectively.

Before genome assembling, FASTQ files were preprocessed for quality control and cleaning (adapters cutting and low quality base called trimming) by fastp tool (v0.19.5) using default parameters (Chen et al., 2018). **Table 4** reported a general summary of the Illumina sequencing metrics for both *Zygosaccharomyces* strains. An example of fastp output about ATCC 42981 quality score per base position before and after filtering is reported in **Figure 3**.

MiSeq sequencer produced good quality short reads for both genomes, since more than 82% of the reads had a Q20 quality score. The short reads preprocessing further improved the base score quality (Q20>85%) with a negligible loss of sequence amount (more than 88% of the reads passed the quality filtering).

MinION (ONT) sequencing, on the other hand, produced a total of 260,559 and 226,374 raw long reads for ATCC 42981 and ABT301^{T}, respectively.

The software MinKNOW (ONT) makes it possible to perform a real time monitoring during the MinION sequencing run. Interpreting the pore signal statistics and the length graph during the first two hours of sequencing gave us a clear idea if the run should be continued or stopped (Schalamun et al., 2018). According to pore occupancy detected by MinKNOW software, we stopped and re-started the run when the value decreased below 70%. Usually, low throughput runs are due to insufficient DNA molecules being ligated to sequencing adapters during library preparation (Schalamun et al., 2018). To ensure sufficient adapter ligated DNA, we started library preparation with at least 2,5 ug of high quality size selected starting DNA. An example of real time analysis of ATCC 42981 MinION sequencing run after 19 hours via the MinKNOW software is presented in **Figure 4**, which showed a satisfactory length distribution with a high concentration of long reads in the range from 11 Kb to 22 Kb. This result was consistent with the size selection range setted with BluePippin system (**Fig. 2**).

Table 4. ATCC 42981 and ABT301^{T} sequencing metrics summary. The comprehensive information on quality-profiling results, with both pre-filtering and post-filtering data, is provided by fastp web tool (Chen et al., 2018). Abbreviations: M, million; G, giga.

General	**ATCC 42981**	**ABT301^{T}**
Sequencing	paired end (301 cycles + 301 cycles)	paired end (301 cycles + 301 cycles)
Mean length before filtering	300bp, 300bp	300bp, 300bp
Mean length after filtering	299bp, 299bp	300bp, 300bp
Duplicaton rate	0.85%	0.66%
Insert size peak	506	508
Before filtering		
Total reads	4.468054 M	5.298168 M
Total bases	1.343151 G	1.592764 G
Q20 bases	1.106055 G (82.34%)	1.369078 G (85.96%)
Q30 bases	907.556471 M (67.57%)	1.157899 G (72.69%)

GC content	40.09%	40.17%
After filtering		
Total reads	3.933094 M	4.952802 M
Total bases	1.176180 G	1.486096 G
Q20 bases	1.005206 G (85.463653%)	1.304287 G (87.76%)
Q30 bases	840.319082 M (71.44%)	1.114287 G (74.98%)
GC content	39.90%	40.04%

Figure 3. Quality score across all positions for ATCC 42981 Illumina sequencing provided by fastp (v0.19.5). Panel **A** depicts the quality score across all positions of read1 before filtering, while panel **B** reports how quality score changes after filtering.

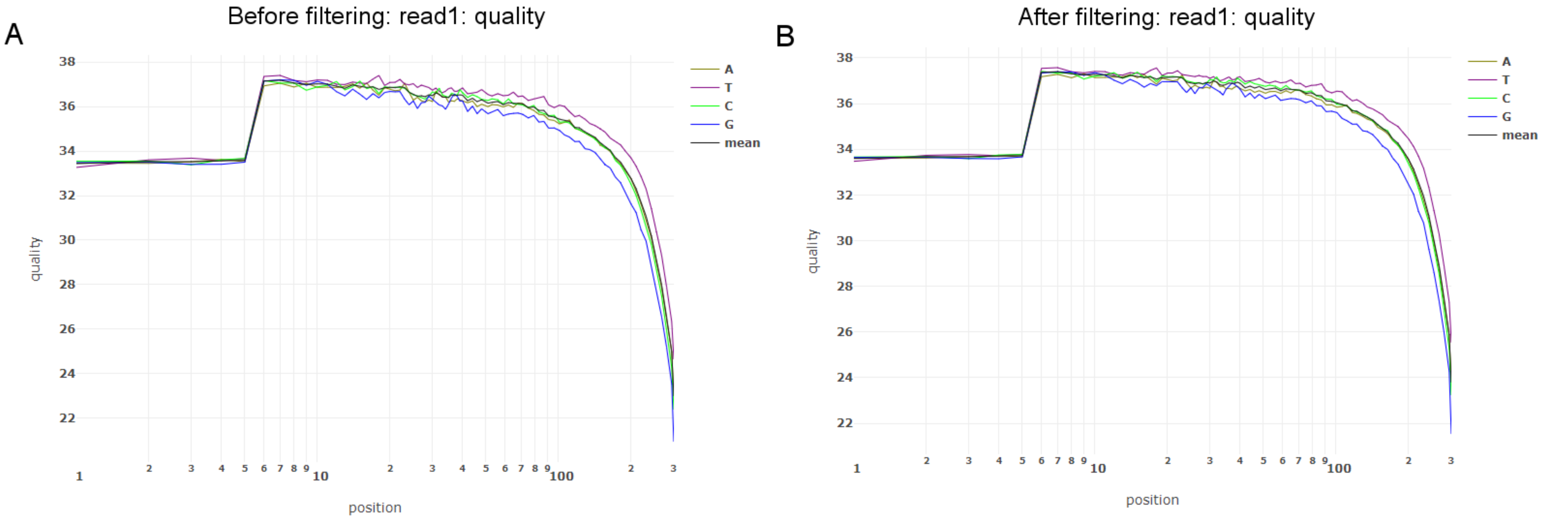
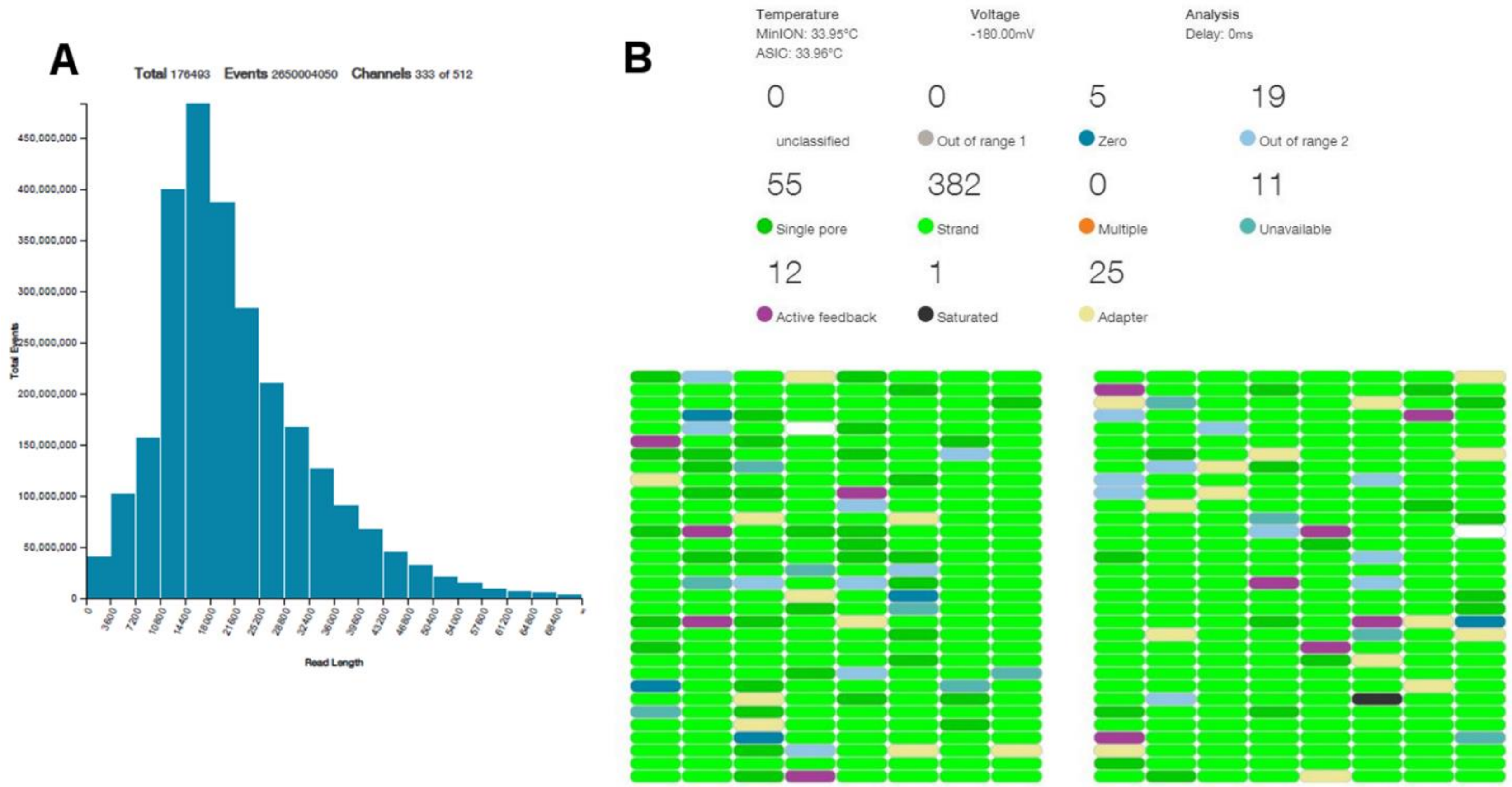


Figure 4. Real time analysis of ATCC 42981 sequencing run via the MinKNOW graphical user interface. Panel **A** illustrates an excellent read length distribution (11-21 Kb) after 19 hours and 14 minutes of sequencing. Panel **B** shows a satisfactory sequencing run with good pore occupancy (light green) and most pores still available for sequencing.



Temperature
MinION: 33.95°C
ASIC: 33.96°C

Voltage
-180.00mV

Analysis
Delay: 0ms

0 unclassified	0 ● Out of range 1	5 ● Zero	19 ● Out of range 2
55 ● Single pore	382 ● Strand	0 ● Multiple	11 ● Unavailable
12 ● Active feedback	1 ● Saturated	25 ● Adapter	

After running the two flowcells, we used ONT's Albacore v2.1.7 to produce sequencing_summary.txt files, which contained a summary of every sequencing read that was used as input file to produce some statistics analysis of the runs. To analyse sequencing data, we used NanoPack (https://github.com/wdecoster/nanopack), a set of tools developed for visualization and processing of long read sequencing data from ONT and Pacific Biosciences (De Coster et al., 2017). In particular, NanoStat was used to produce a comprehensive statistical data summary of ATCC 42981 and ABT301^{T} sequencing runs (**Table 5**) and NanoPlot web tool (http://nanoplot.bioinf.be/) allowed us to visualize long read sequencing data by generating informative QC graphs displaying multiple aspects of the running (**Figs. 5** and **6**).

Table 5. NanoStat general statistics output generated for sequenced yeast strains.

ATCC 42981		**ABT301^{T}**	
Active channels:	342	Active channels:	496
Mean read length (bp):	26,527.3	Mean read length (bp):	28,014.1
Number of reads:	260,559	Number of reads:	226,374
Mean read quality:	9.6	Mean read quality:	9.3
Median read length (bp):	25,210.0	Median read length (bp):	26,970.0
Median read quality:	9.9	Median read quality:	9.7
Read length N50:	33,563	Read length N50:	38,024
Total bases:	6911916550	Total bases:	6341657185

For both sequenced strains, we were able to obtain output long reads of exceptional mean length (**Figs. 5** and **6**, panels **A**) and characterized by an uncommon high quality level. In particular, the bivariate plots comparing log transformed read lengths with their mean quality score (**Figs. 5** and **6**, panels **B**) showed that the majority of reads can be identified at lengths of 50 Kb and quality scores of 11 by the color intensity of the hexagonal bins, with a subgroup of low quality short reads. The heat maps of the physical layout of the MinION flow cell (**Figs. 5** and **6**, panels **C**) showed that most of the nanopores started to be saturated after 24 hours running.

Figure 5. *Z. rouxii* ATCC 42981 plots provided by NanoPlot. (**A**) NanoPlot histogram of log transformed read lengths in bins of 100 bp, the read length N50 value is showed (**B**) Bivariate plot of log transformed read length against base call quality with hexagonal bins and marginal histograms (**C**) Flow cell activity heat map showing number of reads per channel.

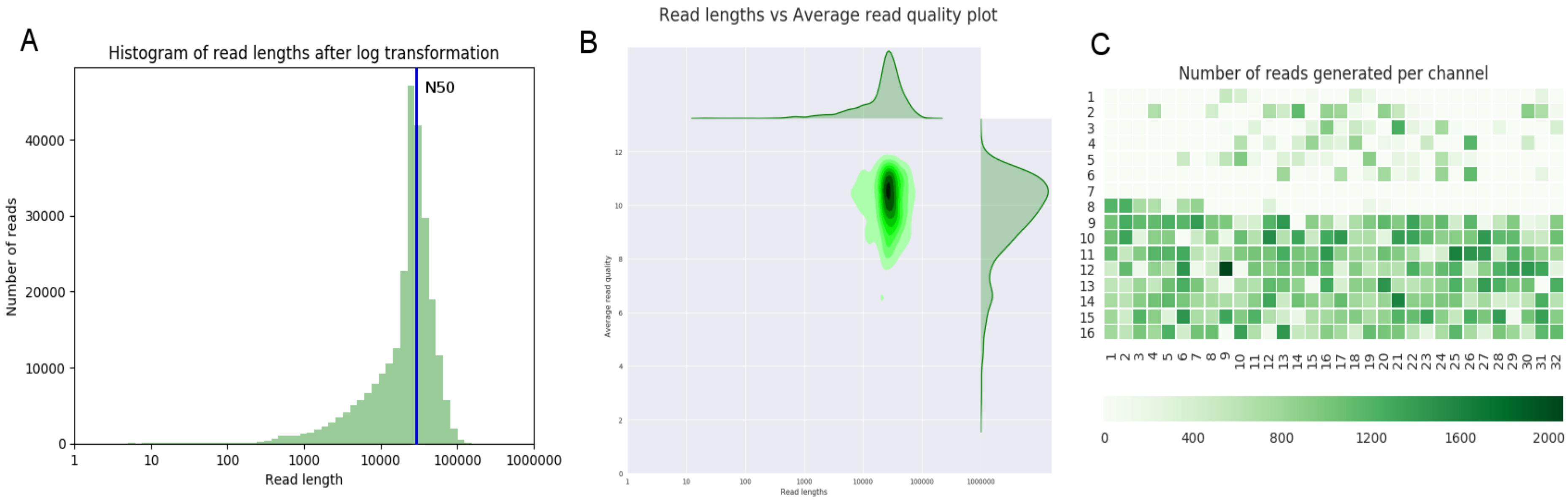
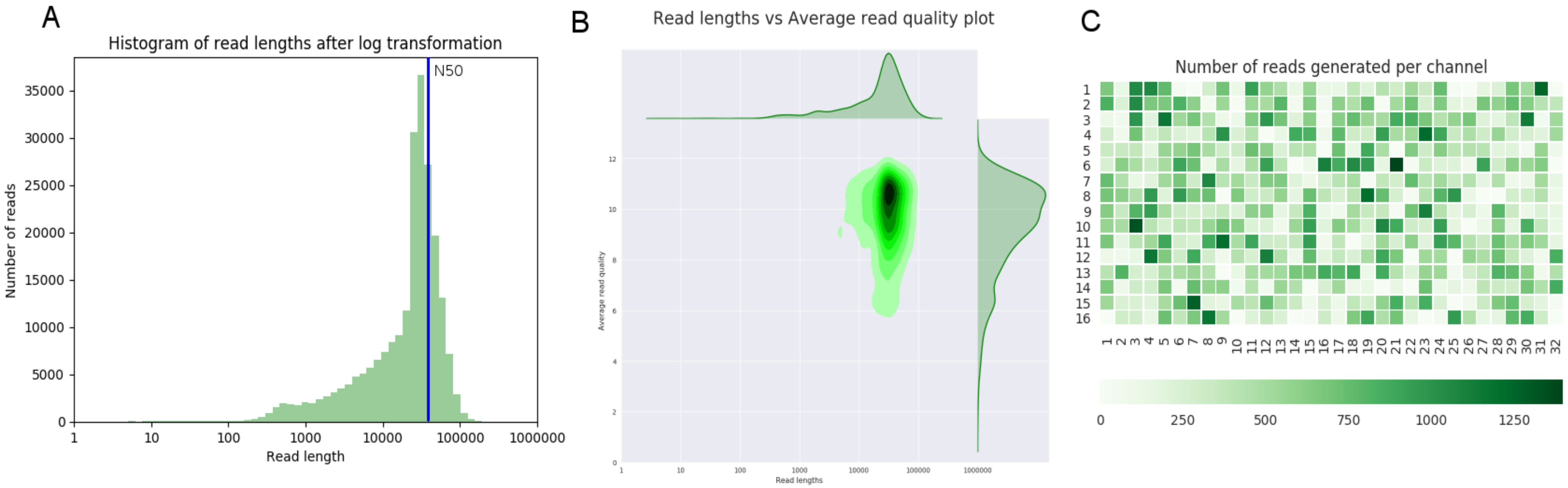


Figure 6. *Z. sapae* ABT301^T plots provided by NanoPlot. (**A**) NanoPlot histogram of log transformed read lengths in bins of 100 bp, the read length N50 value is showed (**B**) Bivariate plot of log transformed read length against base call quality with hexagonal bins and marginal histograms (**C**) Flow cell activity heat map showing number of reads per channel.



Hybrid genome assemblies

ATCC 42981 and ABT301^T genomes are allodiploid, combining haplotypes from unknown parents. High heterozygosity and span repetitive regions represent the greatest technical challenges during the assembly of complex non-haploid genomes (Treangen and Salzberg, 2012; Del Angel et al., 2018). The recently released assembler DBG2OLC allows to efficiently assembly large genomes using long erroneous reads such as MinION (Ye et al., 2016). This hybrid assembly approach reduces both NGS and 3GS sequencing requirements, since Illumina and MinION reads are combined compensating for each other and, therefore, reducing the high error rates of long erroneous reads. As depicted in **Fig. 7**, the hybrid strategy used to *de novo* assemble the ATCC 42981 and ABT301^T genomes exploited DBG2OLC (Bizzarri et al., 2018). In particular, Platanus v1.2.4 (Kajitani et al., 2014) was used to assemble initial Illumina contigs, which were subsequently scaffolded with corrected MinION reads using DBG2OLC (Ye et al., 2016). Finally, scaffolds were polished with long reads using Racon v1.2.0 (Vaser et al., 2017), with short reads using Pilon v1.22 (Walker et al., 2014), and reduced using Redundans v.014 (Pryszcz et al., 2016).

A preliminary gene prediction and functional annotation was performed by homology to the closest haploid relative *Z. rouxii* CBS 732^T (Souciet et al., 2009) using Exonerate v2.2.0 (http://www.ebi.ac.uk/~guy/exonerate/) (Slater and Birney, 2005). Assembly completeness was assessed by Benchmarking Universal Single-Copy Orthologs (BUSCO) v3.0.2 (Simão et al., 2012) using saccharomycetales_odb9 data set.

Comparison with haploid type-strain CBS 732^T showed that the ATCC 42981 and ABT301^T assembled genomes had a 2.14 and 2.53 times larger assembly size and contained a 2.11 and 2.46 times higher number of protein-coding genes, respectively (**Table 6**). For both genomes, we dissected three haplotypes. The first haplotype was identical to that of the reference haploid type-strain CBS 732^T (identity cutoff, 0.92), therefore we named it haplotype T according to Gordon and Wolfe (2008). The second one resembled *Z. pseudorouxii* NCY3042 and therefore we called it haplotype P. According to Gordon and Wolfe (2008), these two parental genomes are quite divergent (averaging about 85% nucleotide sequence identity) but two lines of evidence indicate that the hybridization event that formed allodiploid ATCC 42 981 happened relatively recently. Unfortunately, no data are available regarding the origin of the new species *Z. sapae* ABT301^T (Solieri et al., 2013b). We found an additional third haplotype, named N (new) in both sequenced strains.

We successfully separated ATCC 42981 and ABT301^T contigs/chromosomes into parental haplotypes using the reference haploid genome of CBS 732^T and the script fasta2split.py included in Redundans pipeline (Pryszcz et al., 2016). In ATCC 42981, 9,967,252 contigs were assigned to haplotype T (39.08% GC) and 10,942,807 contigs were splitted in 9,967,959 from haplotype P (40.40% GC) and 974,848 from haplotype N (37.75% GC). In ABT301^T genome 8,851,027 contigs derived from haplotype T (39.26% GC) and 15,874,077 from haplotype P. Interestingly, ABT301^T haplotype P can be further split into 7,729,369 contigs from haplotype P itself (40.26% GC) and 8,144,708 from haplotype N (39.25% GC).

These data suggest a recursive hybridization model for both sequenced strains, in which they derived from at least two hybridization events between three divergent parental genomes (Dakal et al., 2016). Unfortunately, the recovery of the third new haplotype was particularly challenging for ATCC 42981 since parts of that are still missing from the current assembly, revealing that it probably lost most of the haplotype N following rearrangement events during evolution.

Figure 7. Box flow-chart that outlines the strategy used to perform ATCC 42981 and ABT301^T genomes hybrid assembly. Abbreviation: hap, haplotype.

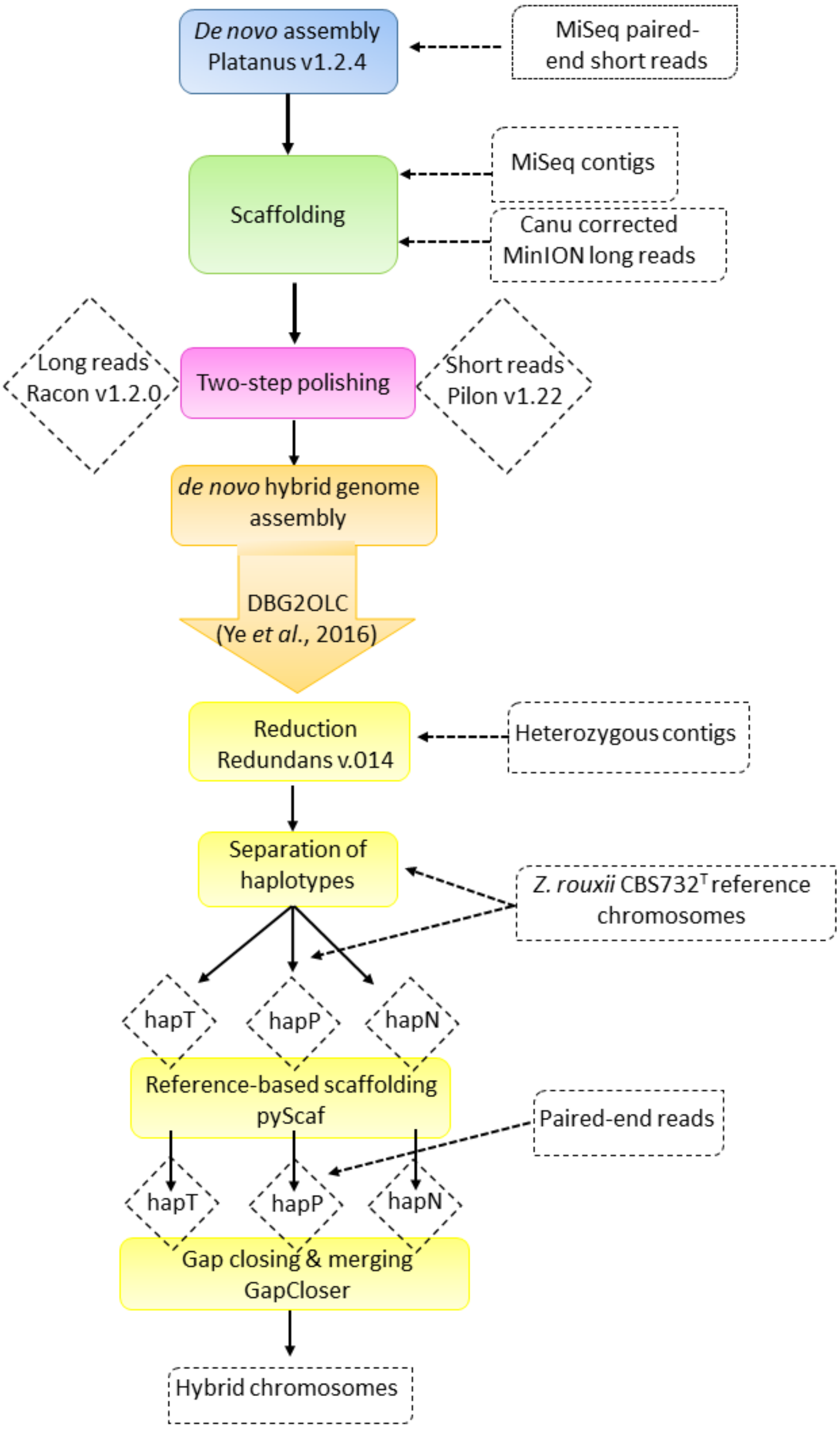


Table 6. Assembly metrics and annotation completeness by BUSCO using universal fungal genes (fungi_odb9) dataset.

Features	**Strains**		
	CBS 732^T	**ABT301^T**	**ATCC 42981**
Assembly size (bp)	9,764,635	24,741,993	20,910,059
No. of scaffolds	7	52	33
G+C content (%)	39.13	39.57	39.65
N_{50} contig size (bp)	1,496,342	1,409,619	1,393,912
N_{90} contig size (bp)	1,114,666	146,869	400,395
Gaps	1,269	0	0
Longest scaffold (bp)	1,865,392	1,913,612	1,903,919
No. of genes	4,991	12,300	10,524
BUSCO complete genes	285 (98.28%)	282 (97.24%)	285 (98.28%)
BUSCO duplicated genes	0 (0%)	240 (85.11%)	264 (92.63%)

Concluding remarks

In this work, *Z. rouxii* hybrid ATCC 42981 and the allodiploid *Z. sapae* ABT301^T strains were used as study models to establish a MinION (ONT) genome sequencing workflow, which could be generalized to other allodiploid or highly heterozygous species. In particular, four extraction protocols were examined for their effectiveness and efficiency in extracting and purifying HMW gDNA, suitable for *de novo* 3GS with the recently released MinION sequencing platform.

In general, DNA requirements for Nanopore sequencing cannot easily be met with the use of most of the commercially available extraction kits. In our case of study, we chose a quantity of, at least, 2.6 µg of 20 Kb fragmented gDNA as input. The most common problem with blood, animal tissue, plant tissue, yeast, Gram-positive and Gram-negative bacteria extraction kits, is that, in many cases, less than the required concentrated DNA is extracted (<50 ng/µl). Three tested protocols, namely PC_B, PC_L and TeP_L, were able to provide very high quantities of pure gDNA at a relatively low cost compared to the W protocol, based on the commercial kit Wizard® Genomic DNA Purification Kit (Promega, Madison, WI, USA). The protocol PC_B represents the standard and consolidate Winston and Hoffman (1987) method for yeast gDNA extraction, which provides high DNA amount, but with low purity due to the presence of just one precipitation step after phenol-chloroform extraction. Moreover, it is not suitable for Nanopore sequencing purpose since it generates highly fragmented DNA after the mechanical lysis of yeast cells with silica beads. When this mechanical lysis was replaced with a softer enzymatic treatment with Lyticase, the resulting DNA was less fragmented but still not enough pure. The TeP_L protocol appeared to be the most efficient DNA extraction method, capable to provide high gDNA yields with better quality and integrity compared with the other ones. In particular, DNA fragmentation overcame the initial desired score of 20 Kb for both *Zygosaccharomyces* strains.

This method of choice is relatively low cost without being excessively laborious, thus it could be appropriate for whole-genome large-scale applications. As 3GS applications will become routine in the near future, the TeP_L method might provide very handy solutions. In addition, extracting quite higher gDNA quantities than needed for present applications and creating DNA banks can be very useful for follow-up or parallel research projects.

Since the TeP_L method produces poorly fragmented DNAs, we found that BluePippin electrophoresis target size selection overcame Covaris g-tube DNA shearing method to provide suitable HMW DNA sample for 1D library preparation. Therefore, we included BluePippin electrophoresis downstream TeP_L DNA extraction and QC steps in our workflow.

Another very challenging question in ONT sequencing is the high error rate of MinION long reads, which makes assembly imprecise and computationally expensive, since repetitive regions are often longer than the read length. On the other hand, long reads could be very useful in resolving high levels of heterozygosity and span repetitive regions, which represented the greatest technical challenges during the previously Illumina assembly of complex *Zygosaccharomyces* non-haploid genomes. In the case of the allodiploid ATCC 42981, distinguishing between homeologous sequences is further challenging as only the haploid type-strain *Z. rouxii* CBS 732^T genome is available to guide homologous scaffold assembly. The high error rate associated to long Nanopore fragments made necessary to polish MinION reads with Illumina-derived reads, resulting into DBG2OLC-driven hybrid *de novo* genome assembly.

Overall, the reported ATCC 42981 and ABT301^T assemblies, even if they represent just an initial step, will assist us in deciphering how hybridization, followed by functional genome stabilization, may offer a rapid speciation and adaptation strategy in non-conventional yeasts.

Data availability

The BioProject has been deposited in GenBank under number PRJEB26771. All sequencing reads of *Z. rouxii* ATCC 42981 and *Z. sapae* ABT301^T have been deposited at EMBL/GenBank under the accession numbers UEMZ01000001 to UEMZ01000033 and UEGL01000001 to UEGL01000052, respectively.

Acknowledgements

We acknowledge Prof. Paolo Giudici for his valuable support and Luciana De Vero PhD for maintaining UMCC strains. We are grateful to Jan Gawor for the technical support in MinION library preparation and sequencing (Robert Gromadka's group, Institute of Biochemestry and Biophysics, Warsaw, Poland). We thank Leszek P. Pryszcz for the realisation of the hybrid assembly.
L.P.P. is funded by National Science Centre (Poland) Polonez-1 framework from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 665778.

Author Contributions

Conceived and designed the experiments: SC, LS, MB and JG. Performed the experiments: MB and JG. Analyzed the data: LP, JG and MB. Contributed to reagents/materials/analysis tools: SC and RG. Wrote the manuscript: MB. Contributed to manuscript revision: LS and SC.

CHAPTER 8: RESOLVING ATCC 42981 *MAT* LOCI BY NANOPORE SEQUENCING AND SYNTHETIC BIOLOGY

Bizzarri M., Cassanelli S., Bartolini L., Pryszcz P. L., Dušková M., Sychrová H., Solieri L. Interplay of chimeric mating-type loci impairs fertility rescue and accounts for intra-strain variability in Zygosaccharomyces rouxii interspecies hybrid ATCC42981. Paper accepted for publication in Frontiers in Genetics.

Abstract

The pre-whole genome duplication (WGD) *Zygosaccharomyces* clade comprises several allodiploid strain/species with industrially interesting traits. The salt-tolerant yeast ATCC42981 is a sterile and allodiploid strain which contains two subgenomes, one of them resembling the haploid parental species *Z. rouxii*. Recently, different mating-type-like (*MTL*) loci repertoires were reported for ATCC42981 and the Japanese strain JCM22060, which are considered two stocks of the same strain. *MTL* reconstruction by direct sequencing approach is challenging due to gene redundancy, structure complexities, and allodiploid nature of ATCC42981. Here, DBG2OLC and MaSuRCA hybrid *de novo* assemblies of ONT and Illumina reads were combined with *in vitro* long PCR to definitively solve these incongruences. ATCC42981 exhibits several chimeric *MTL* loci resulting from reciprocal translocation between parental haplotypes and retains two *MAT***a**/*MAT*α expression loci, in contrast to *MAT*α in JCM22060. Consistently to these reconstructions, JCM22060, but not ATCC42981, undergoes mating and meiosis. To ascertain whether the damage of one allele at the *MAT* locus regains the complete sexual cycle in ATCC42981, we removed the *MAT*α expressed locus by gene deletion. The resulting *MAT***a**/- hemizygous mutants did not show any evidence of sporulation, as well as self- and out-crossing fertility, probably because incomplete silencing at the chimeric *HML*α cassette masks the loss of heterozygosity at the *MAT* locus. We also found that *MAT*α deletion switched off **a**2 transcription, an activator of **a**-specific genes in pre-WGD species. These findings

suggest that regulatory scheme of cell identity needs to be further investigated in *Z. rouxii* protoploid yeast.

Keywords: mating-type, MinION, sexual cycle, *Zygosaccharomyces*, chimeric loci, interspecies hybridization, yeast.

Introduction

Polyploidization, a state resulting from doubling of a genome within a species (autopolyploidy) or the merging between different species (allopolyploidy) (Campbell et al., 2016), is an important evolutionary force which shapes eukaryotic genomes (Albertin and Marullo, 2012), triggers speciation, and can result in phenotypic changes driving adaptation (Ohno, 1070). A whole-genome duplication (WGD) event occurred approximately 100–200 Mya in the common ancestor of 6 yeast genera in the family Saccharomycetaceae, including *Saccharomyces cerevisiae* (as reviewed by Wolfe et al., 2015). WGD was recently proposed to be a direct consequence of an ancient hybridization between two ancestral species (Marcet-Huben and Gabaldón, 2015), followed by genome doubling of initially sterile hybrid to regain fertility, ie the ability to undergo meiosis and produce viable spore (Wolfe, 2015).

Different mechanisms can contribute to hybrid infertility, such as chromosomal missegregation caused by meiosis I non-disjunction (Boynton et al., 2018), chromosomal rearrangements (Liti et al., 2006; Rajeh et al., 2018), and Dobzhansky–Muller gene incompatibilities either between nuclear genes (Bizzarri et al., 2016) or between mitochondrial and nuclear genes (Lee et al., 2008). Specialized loci, called the mating-type (*MAT*)-like (*MTL*) cassettes, regulate mating between haploid cells with opposite *MAT***a** and *MAT*α idiomorphs, as well as meiosis in diploid **a**/α cells. In diplontic yeast *S. cerevisiae MAT* locus on chromosome III contains either the **a**1 or the α1 and α2 genes in Y**a** and Yα segments, respectively, surrounded by X and Z regions at the left and right sides. In haploid α cells, α1 activates the α-specific genes (αsgs), while α2 represses a cohort of **a**-specific genes (**a**sgs), which **a** cells transcribe by default (Haber, 2012). Finally, diploid **a**/α cells are meiosis but not mating-competent, because the **a**1-α2 heterodimer positively regulates *IME1* (Inducer of Meiosis) expression and represses the expression of haploid-specific genes (hsg) required for mating responses. *S. cerevisiae* cells also have extra copies of *MAT* genes at the *HMR***a** and *HML*α loci located close to telomeres of chromosome III and silenced by a combination of the Sir1–4 proteins (Hickman et al., 2011). These extra copies serve as donors during the mating-type switching, which enables *MAT***a** cells to convert into *MAT*α cells, or vice versa, and to mate each other. This autodiploidization event is triggered by a site-specific endonuclease called HO, which induces double-strand break at Z region of the *MAT* locus. In *Saccharomyces* interspecies hybrids, experimental deletion of one *MAT* locus or elimination of the entire chromosome carrying one *MAT* locus yielded fertile allotetraploids (Greig et al., 2002; Pfliegler et al., 2012; Karanyicz et al., 2017). More recently, the *MAT* locus damage was proposed to be the most plausible evolutionary route, which enables natural interspecies hybrids of the *Zygosaccharomyces bailii* complex to rescue mating and meiosis (Ortiz-Merino et al., 2017; Braun-Galleani et al., 2018).

In the Saccharomycetaceae lineage, *Zygosaccharomyces rouxii* stands on the crossroad where different and relevant evolutionary events take their way (Dujon and Louis, 2017). This evolutionary route involves ancient allopolyploidization between two parental lineages, one of which was close to *Z. rouxii* and *Torulaspora delbrueckii* (ZT) clade (Marcet-Huben and Gabaldón, 2015). *Z. rouxii* represents the early branching species before WGD that recruits HO from a LAGLIDADG intein to catalyze the first step of mating-type switching (Fabre et al., 2005). Furthermore, *Z. rouxii* exhibits the triplication of *MTL* loci, which is a genomic landmark of the Saccharomycetaceae family, but, in contrast to *S. cerevisiae*, it lacks of *MAT-HMR* linkage. Whereas the route of αsg regulation appears to be conserved, the regulatory circuit of **a**sgs has been extensively rewired across the Saccharomycotina clade. Instead of the negative regulatory circuit widespread in post-WGD species, several pre-WGD species activate **a**sgs by an HMG-domain protein (**a**2) that is encoded by *MAT***a** (Tsong et al., 2003). Conventionally, *Z. rouxii* displays haplontic life style, where heterothallic haploid cells with opposite mating-type mate each other or, alternatively, homothallic haploid cells switch mating-type and subsequently undergo mating between mother and daughter cells. In both cases, the transient diploid zygote should sporulate to restore the haploid state. Alternatively, stable allodiploid strains arose from mating between divergent haploid parents. One parental haplotype (called T-subgenome) resembles *Z. rouxii* and was 15% different from the other parental haplotype (called P-subgenome) (Gordon and Wolfe, 2008; Bizzarri et al., 2016; Watanabe et al., 2017; Bizzarri et al., 2018).

Both haploid and allodiploid strains show highly variable gene arrangements around *MTL*, suggesting that these loci are recombination hotspot during error-prone mating-type switching events (Watanabe et al., 2013; Solieri et al., 2014). Structural rearrangements are so rampant in these regions that different stock cultures of the same haploid (Watanabe et al., 2013) or allodiploid (Bizzarri et al., 2016; Watananbe et al., 2017) strains can display distinct *MTL* repertoires. For instance, differences in *MTL* loci were recently found between two sub-cultures of the allodiploid strain ATCC42981. In our previous work, we found 7 *MTL* loci in in-house stock of ATCC42981 (termed ATCC42981_R for convenience) (Bizzarri et al., 2016), while Watanabe et al. (2017) detected 6 *MTL* loci in strain JCM22060, the Japanese stock of ATCC42981. Ectopic recombination between *MTL*-flanking regions from divergent parental haplotypes yields chimeric arrangements hardly to resolve both by targeted long PCR approaches (Bizzarri et al., 2016) and by genome sequencing technologies based on short reads (Watanabe et al., 2017).

In 2014, the MinION sequencing device (Oxford Nanopore Technology, ONT) was released and initially exploited to sequence and assemble PCR products or microbial genomes (Jain et al., 2016). Recent improvements in protein pore (a laboratory-evolved *Escherichia coli* CsgG mutant named R9.4), library preparation techniques (1D ligation and 1D rapid), sequencing speed (450 bases/s), and control software enabled the usage of Nanopore sequence data, in combination with other sequencing technologies, for assembling eukaryotic genomes including yeasts, nematodes and human (Istace et al., 2017; Yue et al., 2017; Jansen et al., 2017; Jain et al., 2018). The main advantage of ONT is that reads can reach tens of kilobases (Jain et al., 2016), making more easy to resolve repeat regions and to detect structural variation. Recently, the genome of allodiploid strain ATCC42981_R was sequenced and assembled through a *de novo* hybrid strategy which combined MinION long and Illumina short reads (Bizzarri et al., 2018).

Here, we took advantage from the newly released genome of ATCC42981_R (Bizzarri et al., 2018), in order to resolve incongruences in the highly dynamic *MTL* loci. Furthermore, we deleted the expressed *MAT*α^{P} locus in ATCC42981_R to test whether the loss of *MAT* heterozygosity can induce genome doubling and rescue fertility in allodiploid cells of the ZT clade.

Materials and Methods

Strains, plasmids, and culture conditions

Yeast strains and plasmids used in this study are listed in **Table 1**. Yeast cells were routinely propagated at 28°C in YPD (1% yeast extract, 2% peptone, 2% glucose) medium with 1.5% agar when necessary. Stock cultures were stored at -80°C with glycerol at final concentration of 25% (v/v) for long-term preservation. For sporulation and mating assays, MEA (5% malt extract, 2% agar) with and without 6% NaCl, and YM (0.3% yeast extract, 0.5% peptone, 0.3% malt extract, 1% dextrose, 1.5% agar) media were used. *Zygosaccharomyces parabailii* strain G21C was used as control for conjugated asci formation after growth on MEA medium. When required, YPD medium was supplemented with G418 (100 mg/ml, MP Biochemicals, Germany) to the final concentration of 200 µg/ml.

Table 1. Yeast strains and plasmid used in this work. ATCC 42981_R represents in-house stock culture of strain ATCC 42981. Other codes indicate the name of strains in other culture collections. Genotype reports Y sequence from the putatively expression active mating-type locus (*MAT*). T and P superscripts indicate Y**a** or Yα sequences from T- or P-subgenomes, respectively. Abbreviations: na, not available.

Strains	Other codes	Relevant characteristics	References
Z. parabailii			
G21C	na	Isolated from balsamic glaze	this work
***Z. rouxii* strain**			
ATCC 42981_R	JCM22060	*MAT***a**^{T}/*MAT*α^{P}	Solieri et al., 2008; Bizzarri et al., 2017

CBS4837	NCYC1682;
	NBRC1876
		*MAT*α^{P}	Solieri et al., 2008;
			Sato et al., 2017;
			Watanabe et al., 2017

CBS4838	NBRC1877	*MAT***a**^{P}	Solieri et al., 2008; Watanabe et al., 2017
Transformants			
ATCC 42981 *MAT*αΔ clone_6	na	*MAT*α^{P} *Δ::loxP-KanMX-loxP*; *MAT***a**^{T}	this work
ATCC 42981 *MAT*αΔ clone_65	na	*MAT*α^{P} *Δ::loxP-KanMX-loxP*; *MAT***a**^{T}	this work
ATCC 42981 *MAT*αΔ clone_74	na	*MAT*α^{P} *Δ::loxP-KanMX-loxP*; *MAT***a**^{T}	this work
ATCC 42981 *MAT*αΔ clone_177	na	*MAT*α^{P} *Δ::loxP-KanMX-loxP*; *MAT***a**^{T}	this work
Plasmids			
pUG6		*loxP-KanMX-loxP* module	Güldener et al., 1996

DNA manipulations

DNA manipulations were performed according to standard protocols (Sambrook et al., 1989). Genomic DNA from yeast cells were isolated according to Hoffman and Winston (1987), plasmid DNA from *E. coli* was isolated using the GenEluteTM Plasmid Miniprep Kit (Sigma). DNA quantity and quality were evaluated electrophoretically and spectrophotometrically using a NanoDrop ND-1000 device (Thermo Scientific, Waltham, MA). Zymoclean™ Gel DNA Recovery and DNA Clean & Concentrator™-5 Kits (Zymo Research, Orange, CA, USA) were used for the isolation of DNA fragments from agarose gels and for PCR amplicons purification, respectively. Long PCR amplifications were carried out with rTAQ DNA polymerase (Takara Bio, Shiga, Japan) according to manufacturer's instructions. For colony PCR 1 µl of DNA extracted with lithium acetate-SDS method (Lõoke et al., 2011) was amplified with DreamTaq polymerase (Thermo Scientific, Waltham, MA) according to the manufacturer's instructions in 20 µl reaction volume. All PCR amplifications were carried out in a T100 Thermalcycler (Bio-Rad). All primers used in this study are listed in **Supplementary Table S1**.

Genome re-assembly

Hybrid assembly of ATCC 42981_R genome from Oxford Nanopore and Illumina reads was released to the European Nucleotide Archive under the accession number PRJEB26771 (Bizzarri et al., 2018). In the deposited assembly, Platanus contigs were scaffolded into 33 scaffolds with corrected MinION reads using DBG2OLC (Ye et al., 2016). These scaffolds were submitted to two-step polishing with long reads using Racon v1.2.0 (Vaser et al., 2017) and with short reads using Pilon v1.22 (Walker et al., 2014), and, finally, reduced using Redundans v.014 (Pryszcz et al., 2016). Here, both long and short reads were assembled jointly with the alternative assembly algorithm Maryland Super-Read Celera Assembler v.3.2.2 (MaSuRCA) (Zimin et al., 2017) with default settings. Gene identification and annotation were carried out through the Yeast Genome Annotation Pipeline (YGAP) (http://wolfe.ucd.ie/annotation/) without frameshift correction (Proux-Wéra et al., 2012). Assembly completeness was assessed by Benchmarking Universal Single-Copy Orthologs (BUSCO) v3.0.2 (Simão et al., 2012) using saccharomycetales_odb9 data set.

MTL loci search and Sanger-based validation

Search for *MTL* loci on scaffolds generated by DBG2OLC and MaSuRCA hybrid assemblies was carried out with a custom BLAST server built using the Sequenceserver software package (Pryiam et al., 2015). Y**a** and Yα sequences and *MTL* flanking genes from haploid reference genome of *Z. rouxii* CBS 732^T (Souciet et al., 2009) were used as queries.

The *in silico MTL* arrangements were *in vitro* validated by PCR and Sanger sequencing. Specific primer sets were built on *MTL* locus flanking regions outside of the X and Z regions (**Table S1**). For putatively active *MAT*α^P cassette, walking strategy was adopted to cover ~ 1 Kb downstream and upstream Yα (Wang et al., 2011). According to the nomenclature adopted by Watanabe et al., (2017), *MTL* and flanking genes were marked with T and P superscripts when they shared >99% identity with *Z. rouxii* CBS 732^T or with P-subgenome from allodiploid NBRC110957 (Watanabe et al., 2017), respectively. N superscript was used to identify gene variants divergent from both T and P counterparts (identity % lower than 99). The 5' *MTL*-flanking gene ZYRO0F18524g was named as *CHA1*$_L$ for brevity. Sequences were aligned with Clustal Omega (Sievers and Higgins, 2014) and viewed using Jalview (Waterhouse et al., 2009). Neighbor-joining tree was built using MEGA v.6 software (Tamura et al., 2013).

Deletion cassettes construction and yeast transformation

Deletion of the active *MAT*α locus from P-subgenome (abbreviated as *MAT*α^P) was performed with the reusable *loxP-kanMX-loxP* cassette as described previously (Güldener et al., 1996). The MATα1/2cp2-kanMX-F-80nt and MATα1/2cp2-kanMX-R-80nt primers contained ~80 bp homology sequences outside the X ad Z regions of *MAT*α^P locus, respectively, and were used to amplify the *kanMX* deletion cassette from pUG6. After purification, the resulting PCR product was used to transform *Z. rouxii* cells by electroporation with a modified protocol from Pribylova and Sychrova (2003) Briefly, cells were grown (28°C; 180 rpm) in 80 ml of YPD medium supplemented with 300 mM NaCl until the exponential phase (corresponding to OD$_{600nm}$ of 0.7-0.8). After washing with ddH$_2$O, cells were resuspended into 16 ml of TE buffer (Tris-hydrochloride buffer, pH 8.0, containing 1.0 mM EDTA) supplemented with 25 mM dithiothreitol and 20 mM LiAc, and incubated at 30°C for

30 min with gently shaking. Cells were centrifuged at 6,000 *g* for 5 min at 4°C, and washed twice by resuspension in 20 mL of ice-cold 1M sorbitol. Finally, cells were washed in 5 ml of ice-cold 1M sorbitol and resuspended in 800 µl of ice-cold 1M sorbitol. One hundred microliter of competent cell suspension were transferred into a pre-chilled 2-mm electroporation cuvette (Molecular Bioproducts Inc., San Diego, CA, USA) and 1 µg of *loxP-kanMX-loxP* deletion cassette was added before the electroporation at 2250 V/cm for 5 ms (Eporator, Eppendorf, Germany). Immediately after electroporation, 100 µl of ice-cold 1M sorbitol were added to electroporation mixture. Before plating on selective YPDA medium supplemented with G418, the transformation mixtures were incubated for 2 h in 5 ml of YPD at 30°C. In G418^R clones, targeted gene disruption was confirmed by full-length, 5'-, and 3'-end diagnostic PCRs (**Fig. S1**).

RNA extraction, cDNA synthesis and RT-PCR

RNA was extracted from ATCC 42981 wild type and deletion mutants cultured in YPD and harvested at stationary phase, as previously reported (Solieri et al., 2016). RNAs were reverse transcribed using 0.5 µM oligo (dT) and RevertAid H Minus Reverse Transcriptase (Thermo Scientific, Waltham, USA) according to the manufacturer's instructions. cDNAs (25 ng) were amplified using DreamTaq polymerase with primers specific for different variants of *MAT***a**1, *MAT*α1 and *MAT*α2 genes, as well as for T and P variants of **a**sg *AGA2*, *STE2*, and *STE6* (**Table S1**).

Results

Inventory of ATCC42981_R *MTL* cassettes

To unambiguously characterize *MTL* loci in our stock culture, we exploited the new available ATCC 42981_R draft genome (Bizzarri et al., 2018). This draft genome deals with the hybrid DBG2OLC assembly of MinION ultra-long and Illumina MiSeq short reads to resolve high heterozygosity and span repetitive regions, which represent the greatest technical challenges during the assembly of complex non-haploid genomes (Treangen and Salzberg, 2011; Del Angel et al., 2018). Custom BLAST searches using SequenceServer identified 6 scaffolds harboring 8 *MTL* loci (2 *MTL*α^T, 4 *MTL*α^P, and

2 *MTL*a) (**Table 2**). As this pattern matched only partially either with our previous results or with the JCM22060 set of *MTL* loci, we took into account the possibility of misassembled loci, mainly considering that reference genomes from the parental haplotype P is not available for ATCC 42981_R. This problem is even more burdensome for *MTL* loci, which contain the long non-tandem repeated X and Z sequences enriched in homopolymeric stretches. To circumvent these caveats, we validated the *MTL* cassettes found in DBG2OLC assembly *in silico* by using the alternative assembler MaSuRCA, as well as *in vitro* by direct PCR and Sanger sequencing. With appropriate caution, agreement between the assemblies – which are completely independent in assembly algorithms – and among assemblies and Sanger sequencing can confirm the integrity of *MTL* cassettes.

Table 2. Overview of the *MTL* cassettes confirmed by hybrid *de novo* genome assemblies and PCR approach. *MTL* cassettes were found by BLAST searching Y**a** and Yα coding DNA sequences from *Z. rouxii* CBS 732^T reference genome against DBG2OLC and MaSuRCA assemblies and then they were validated by long PCR and Sanger sequencing. Gray shadow indicates *MTL* cassettes found in both the assemblies. JCM66020 *MTL* cassettes were described based on flanking genes according to nomenclature reported by Watanabe et al., (2017). Briefly, numbers 1 to 6 indicate 5'-flanking genes *DIC1*^{T}, *CHA1*$_L^T$, *CHA1*^{T}, *DIC1*^{P}, *CHA1*$_L^P$, and *CHA1*^{P}, respectively. Capital letters A to F indicate 3'-flanking genes *SLA2*^{T}, ZYRO0F18634g^T, ZYRO0C18392g^T, *SLA2*^{P}, ZYRO0F18634g^P and ZYRO0C18392g^P, respectively. NBRC110957 *MTL* cassettes were derived from BioProjects PRJDB4974 (Watanabe et al., 2017). Abbreviation: r.c, reverse complement.

Assembler	Cassette	Scaffolds	Coordinates	PCR	JCM22060
DBG2OLC	**Yα**^{T}				
	DIC1^{P}-*MTL*α^{T}-ZYRO0F18634g^{T}	UEMZ01000028.1	45,980…56,093	+	4B
	CHA1$_{L}^{T}$-*MTL*α^{T}-ZYRO0F18634g^{T}	UEMZ01000013.1	263,261…275,557	+	-
	Yα^{P}				
	DIC1^{T}-*MTL*α^{P}-*SLA2*^{P}	UEMZ01000013.1	35,683…40,522	+	1D
	CHA1$_{L}^{T}$-*MTL*α^{P}-*SLA2*^{P}	UEMZ01000003.1	11,848…18,890 (r.c)	+	2D
	CHA1$_{L}^{P}$-*MTL*α^{P}-ZYRO0F18634g^{P}	UEMZ01000003.1	241,988…250,941 (r.c.)	+	5E
	DIC1^{T}-*MTL*α^{P}-*SLA2*$^{\mathbf{N}}$	UEMZ01000007.1	1,444,839…1,449,671 (r.c.)	-	-
	Ya				
	DIC1$^{\mathbf{N}}$-*MTL***a**$^{\mathbf{N}}$-*SLA2*^{T}	UEMZ01000008.1	1,427,380…1,431,846	+	-
	CHA1^{T}-*MTL***a**^{T}-ZYRO0C18392g^{T}	UEMZ01000015.1	1,296,432…1,304,606 (r.c.)	+	3C
MaSuRCA	**Yα**^{T}				
	DIC1^{P}-*MTL*α^{T}-ZYRO0F18634g^{T}	jcf7180000000244	822,883…832,993 (r.c.)	+	4B
	Yα^{P}				
	DIC1^{T}-*MTL*α^{P}-*SLA2*^{P}	jcf7180000000243	2,467,548…2,467,412 (r.c.)	+	1D
	CHA1$_{L}^{T}$-*MTL*α^{P}-*SLA2*^{P}	jcf7180000000243	2,697,258…2,697,258 (r.c.)	+	2D
	CHA1$_{L}^{P}$-*MTL*α^{P}-ZYRO0F18634g^{P}	jcf7180000000243	2,920,342…2,929,342 (r.c.)	+	5E
	Ya				
	CHA1^{P}-*MTL***a**^{P}-ZYRO0C18392g^{P}	jcf7180000000321	1,284,293…1,291,457 (r.c.)	+	6F
	CHA1^{T}-*MTL***a**^{T}-ZYRO0C18392g^{T}	jcf7180000000315	1,417,152…1,425,321	+	3C

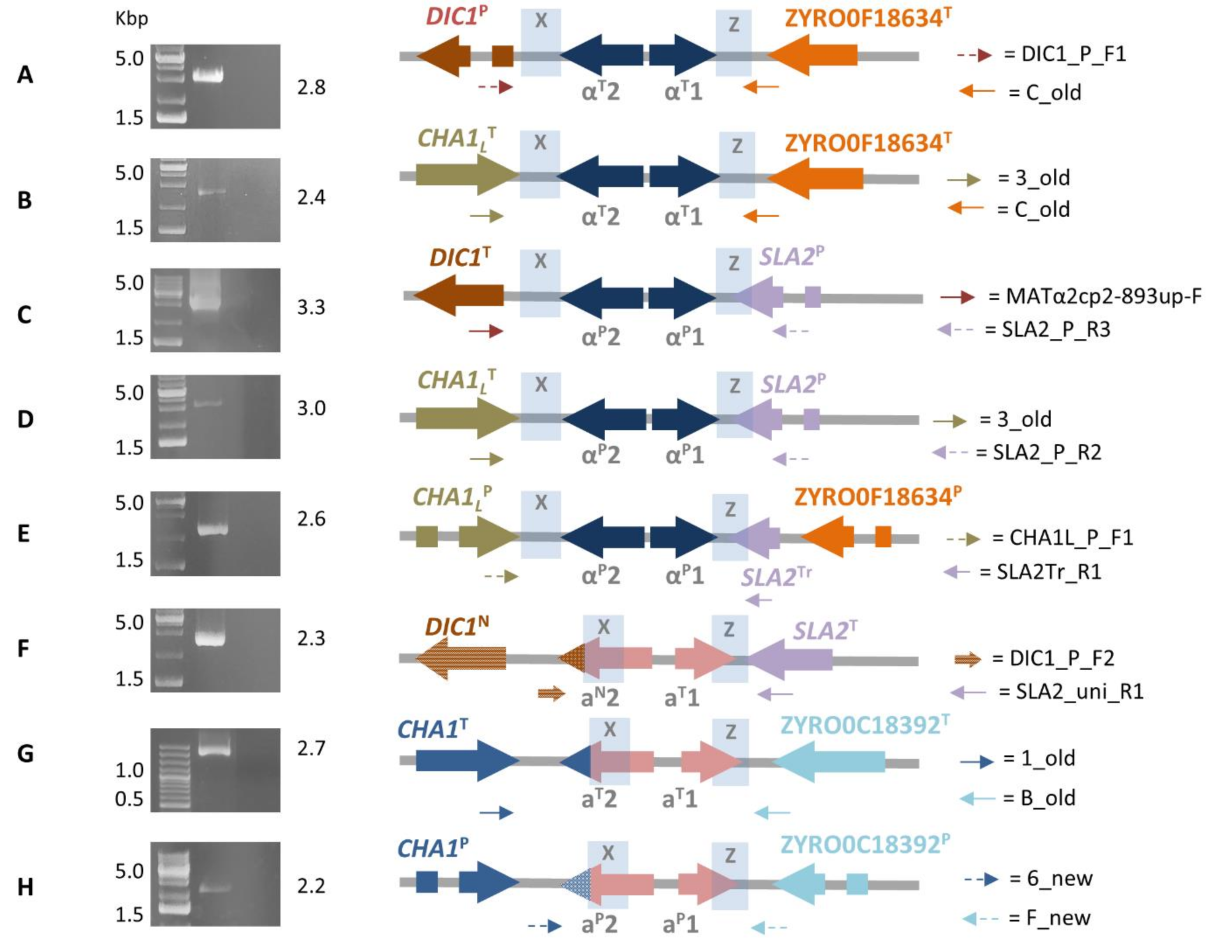
MaSuRCA assembly resulted in an assembled genome size of 21.09 Mb distributed across 59 scaffolds with N_{50} of 1.34 Mb (**Table 3**). In our previous analysis, 10,524 predicted genes were estimated by Exonerate (http://www.ebi.ac.uk/~guy/exonerate/) (Bizzarri et al., 2018). Here, gene number was re-calculated both for DBG2OLC and MaSuRCA assemblies using YGAP software. YGAP was developed for fungal genomes and exploits the existence of a large number of gene sequences ("pillars") conserved among fungal species and maintained in the Yeast Gene Order Browser (YGOB) database (as well as the syntenic arrangement of coding regions among a large number of fungi. Based on this analysis, DBG2OLC and MaSuRCA displayed roughly the same number of predicted genes (**Table 3**). Moreover, BUSCO 3.0 toolkit was employed to assess genome based on the conservation of a curated set of Saccharomycetales lineage-specific single-copy orthologs. The analysis revealed a high degree of completeness in both assemblies (>98.0%), even if MaSuRCA was able to retrieve more duplicated orthologs than DBG2OLC.

Table 3. Assembly metrics and annotation completeness obtained by using BUSCO universal fungal genes (saccharomycetales_odb9) data set.

Feature	**Assembler**	
	DBG2OLC	**MaSuRCA**
Assembly size (bp)	20,910,059	21,093,102
No. of scaffolds	33	59
G+C content (%)	39.65	39.95
N_{50} contig size (bp)	1,393,912	1,337,761
N_{90} contig size (bp)	400,395	638,558
Gaps	0	0
Longest scaffold (bp)	1,903,919	2,966,114
No of genes	10,678	10,362
BUSCO complete genes	1,687 (98.6%)	1,692 (98.9%)
BUSCO duplicated genes	1,491 (87.1%)	1,582 (92.5%)

Four MaSuRCA scaffolds contained 6 *MTL* loci identical to those previously found in JCM22060 (**Table 2**; **Table S2**). Like in DBG2OLC, *MTL* cassettes are especially present at the MaSuRCA scaffold edges, confirming difficulties in scaffolding over repeated X and Z sequences shared among multiple and partially divergent *MTL*-flanking regions. Five MaSuRCA supported *MTL* cassettes were also found in DBG2OLC assembly, while one was specific for MaSuRCA assembly and three were specific for DBG2OLC assembly. Direct *in vitro* PCR validated 8 *MTL* arrangements (**Table 2**; **Fig. 1**). Moreover, MaSuRCA consensus sequences were often more consistent with Sanger sequencing compared to DBG2OLC. Probably, this discrepancy resulted from a more aggressive DBG2OLC approach enabled to reduce the genome fragmentation, but at the price of local assembling accuracy.

Figure 1. Final ATCC 42981_R *MAT*-like cassettes organisation resulting from DBG2OLC and MaSuRCA assemblies. Picture depicts full-length PCR results (left) and the corresponding inferred *MTL* arrangements (right). Genes from T and P-subgenomes are marked with T and P superscripts, respectively. *SLA2* truncated variant is marked with Tr superscript, while *DIC1* and *MAT***a**2 new variants with N superscript. Blue shading denotes the X and Z regions. Dot arrows indicate P-variants of *MAT* flanking genes, while full arrows stand for T-variants. Variable *MAT***a**2 3'-end tags are indicated as coloured tips. Capital letters from A to H refer to primer pairs listed in **Table S1**. Numbers on the right of gel electrophoresis images indicate PCR product length in Kb.



***MTL*α^{P} cassettes**

Congruently with our previous data (Bizzarri et al., 2016), the DBG2OLC assembly of ATCC 42981_R genome supported the cassettes *DIC1*^{T}-*MTL*α^{P}-*SLA2*^{P} and *CHA1*$_{L}^{T}$-*MTLα*^{P}-*SLA2*^{P}*,* which were also present in JCM22060 (Watanabe et al., 2017). MaSuRCA assembly and PCR approach confirmed these arrangements. Pairwise comparisons showed that *DIC1*^{T} and *CHA1*$_{L}^{T}$ were 100% identical to the *Z. rouxii* CBS 732^{T} counterparts. In cassette *DIC1*^{T}-*MTL*α^{P}-*SLA2*^{P}, the 3'-flanking gene *SLA2*^{P} diverged from CBS 732^{T} counterpart (83.65% identity), and resembled *SLA2* found in allodiploids NBRC110957 and NBRC1876 (99.58% identity) (Sato et al., 2017; Watanabe et al., 2017). In *CHA1*$_{L}^{T}$-*MTLα*^{P}-*SLA2*^{P} cassette, DBG2OLC assembly reported mismatches compared to *SLA2*^{P} in NBRC110957 (93.12% identity), which were not supported by MaSuRCA. Sanger sequencing did not support the consensus sequence built with DBG2OLC assembly and confirmed the accuracy of MaSuRCA assembling (**Fig. S2**).

According to the model of T- and P-subgenomes, *DIC1*^{T}-*MTL*α^{P}-*SLA2*^{P} and *CHA1*$_{L}^{T}$-*MTLα*^{P}-*SLA2*^{P} should be chimeric cassettes with α^{P} mating-type, arisen from rearrangements involving the X regions. Intriguingly, NBRC110957 also contains a similar chimeric arrangement with a Ya^{P} sequence (Watanabe et al., 2017; Supplementary Table S2), suggesting that recombination could be frequent upstream the Y sequence. Thus, the X region could represent specific 'fragile' chromosomal location susceptible to double strand breakage (DSB). Recombinant sites at the *MAT* locus were documented in several *Saccharomyces* lager yeasts (Bond et al., 2004; Hewitt et al., 2014). Breakpoints frequently occurred at the right of the *MAT* locus resulting in hybrid *S. cerevisiae*-*S. eubayanus* chromosomes III. These chromosomes contain *S. eubayanus* sequences in the W region and *S. cerevisiae* in the Y region hitch-hiking downstream genes or *viceversa* (Monerawela and Bond, 2017). In lager yeast Ws34/70 a possible location for the recombination event is a 9-bp insertion in the *S. eubayanus* X region compared to *S. cerevisiae*. We found a similar indel between X regions of *DIC1* variants found in ATCC 42981_R (**Fig. S3**).

Novel sets of P-subgenome-specific primers confirmed an additional *MTL*α^{P} locus (*CHA1*$_{L}^{P}$-*MTL*α^{P}-ZYRO0F18634g^{P}) which escaped our previous reconstruction (Bizzarri et al., 2016). Based on Watanabe *et al.,* (2017), this locus should be a cryptic *HML* cassette, which did not affect the proper cell identity. This cassette had a truncated *SLA2* sequence downstream the Z region, confirming DNA erosion on the right side of *MAT* locus (Gordon et al., 2011). Interestingly, in both DBG2OLC and MaSuRCA assemblies this cassette is linked to *CHA1*$_{L}^{T}$-*MTL*α^{P}-*SLA2*^{P} on the same scaffold (**Fig. S4**).

***MTL*α^{T} cassettes**

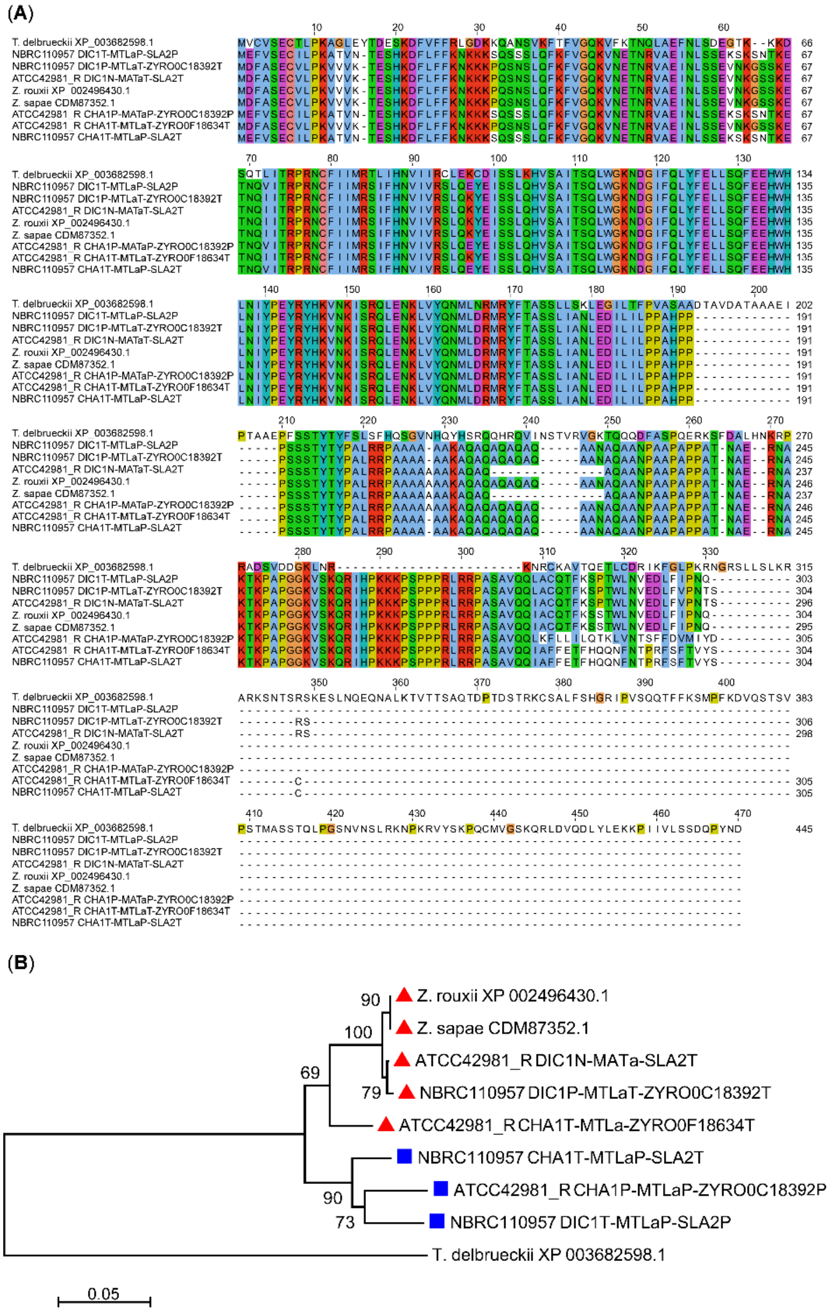
DBG2OLC and MaSuRCA assemblies failed to congruently reconstruct *MTL*α^{T} loci (**Table 2**). DBG2OLC scaffold UEMZ01000013.1 contains *CHA1*$_{L}^{T}$-*MTLα*^{T}-ZYRO0F18634g^{T} linked to the chimeric cassette *DIC1*^{T}-*MTL*α^{P}-*SLA2*^{P}*,* while another *MTL*α^{T} locus *(DIC1*^{P}-*MTL*α^{T}-ZYRO0F18634g^{T}) lies on the scaffold UEMZ01000028.1. MaSuRCA assembly did not report *CHA1*$_{L}^{T}$-*MTLα*^{T}-ZYRO0F18634g^{T}, but only *DIC1*^{P}-*MTL*α^{T}-ZYRO0F18634g^{T} on scaffold jcf7180000000244. *DIC1*^{T}-*MTL*α^{P}-*SLA2*^{P} was linked to *CHA1*$_{L}^{T}$-*MTLα*^{P}-*SLA2*^{P} and *CHA1*$_{L}^{P}$-*MTL*α^{P}-ZYRO0F18634g^{P} (scaffold jcf7180000000243) (**Fig. S4**). PCR approach supported both *MTL*α^{T} cassettes from DBG2OLC assembly, while scaffold comparison suggests that MaSuRCA collapsed the *CHA1*$_{L}^{T}$ flanking regions into a single locus (**Fig. S4**).

***MTL*a cassettes**

Blast search against the DBG2OLC assembly revealed two *MTL*a cassettes. The arrangement *CHA1*^{T}-*MTL*a^{T}-ZYRO0C18392g^{T} was supported by MaSuRCA and PCR approach, and was also congruent with our previous reconstruction (Bizzarri et al., 2016) and with JCM22060 (Watanabe et al., 2017) (**Table S2**). The second *MTL*a locus resolved by DBG2OLC, *DIC1*^{N}-*MTL*a^{T}-*SLA2*^{T}, contained a^{T}1 and a novel a^{N}2 gene variant (indicated with N superscript) which was 97.99% identical to *MAT*a2 from NBRC110957 *DIC*1^{P}-*MTL*a^{T}-ZYRO0C18392^{T} cassette (**Fig. 2**). PCR approach and Sanger sequencing demonstrated that this cassette really exists in ATCC 42981_R genome, even if it was missing both in MaSuRCA assembly and in JCM22060. Like in case of *SLA2*^{P} from *CHA1*$_{L}^{T}$-*MTLα*^{P}-*SLA2*^{P}, DBG2OLC *MAT*a2 sequence showed some indels in homopolymeric stretches compared to the Sanger-sequence data (98.54% pairwise identity), resulting in a prematurely interrupted ORF (data not shown). The neighbor genes at the 5' and 3' sides were a novel *DIC1* variant (named *DIC1*^{N}) and the *SLA2*^{T} gene, respectively. Noteworthy, the *DIC1-MAT-SLA2* arrangement is retained around the transcriptionally active *MAT* loci in almost all the pre-WGD species (Gordon et al., 2011). Therefore *DIC1*^{N}-*MTL*a^{N}-*SLA2*^{T} cassette could be a good candidate to be the active *MAT*a cassette in ATCC 42981_R.

Finally, PCR approach with haplotype P-specific primers identified a third *MTL***a** locus (*CHA1*^{P}-*MTL***a**^{P}-ZYRO0C18392g^{P}) which was present in JCM22060 (**Table S2**) and in MaSuRCA assembly (**Table 2**). Blast search for *CHA1*^{P} gene revealed that DBG2OLC assembler did not extend scaffold UEMZ01000005.1 beyond this gene.

Figure 2. Multiple sequence alignment and phylogenetic analysis of MATa2 proteins. Panel (**A**) depicts the alignment involving 9 MATa2 amino acid sequences. The amino acid identities were coloured according to Clustal Omega colour scheme (Sievers and Higgins, 2014). In panel (**B**) dendrogram was inferred using the Neighbor-Joining method. The percentages of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches, when ≥ 50%. The evolutionary distances were computed using the p-distance method and are in the units of the number of amino acid differences per site. All positions containing gaps and missing data were eliminated. Red triangles and blue squares marked T and P variants.


(A)
T. delbrueckii XP_003682598.1
NBRC110957 DIC1T-MTLaP-SLA2P
NBRC110957 DIC1P-MTLaT-ZYRO0C18392T
ATCC42981_R DIC1N-MATaT-SLA2T
Z. rouxii XP_002496430.1
Z. sapae CDM87352.1
ATCC42981_R CHA1P-MATaP-ZYRO0C18392P
ATCC42981_R CHA1T-MTLaT-ZYRO0F18634T
NBRC110957 CHA1T-MTLaP-SLA2T
(B)
90
Z. rouxii XP 002496430.1
100
Z. sapae CDM87352.1
ATCC42981_R DIC1N-MATa-SLA2T
69
79
NBRC110957 DIC1P-MTLaT-ZYRO0C18392T
ATCC42981_R CHA1T-MTLa-ZYRO0F18634T
NBRC110957 CHA1T-MTLaP-SLA2T
ATCC42981_R CHA1P-MTLaP-ZYRO0C18392P
90
73
NBRC110957 DIC1T-MTLaP-SLA2P
T. delbrueckii XP 003682598.1
0.05

Reconstruction of *MTL* structure

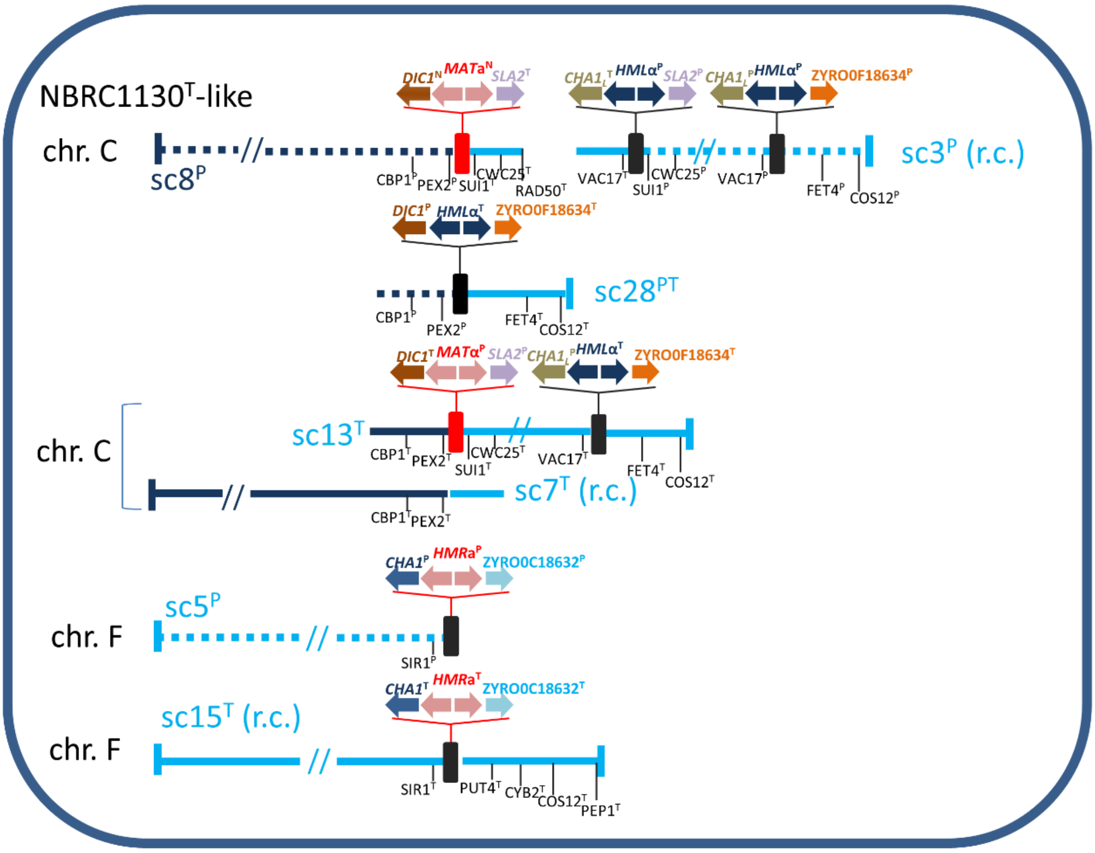
Analysis of regions around *MTL* loci assisted us to reconstruct the putative *MTL* structure in ATCC42981_R. NBRC1130^{T} culture retains ancestral *MTL* arrangement compared to CBS732^{T} (Watanabe et al., 2013) and was used as reference strain. In this strain, chromosome C contains *MAT* and *HML* loci flanked by sets of genes which were also conserved around ATCC42981_R *MTL* cassettes. In particular, *MAT* locus was flanked on the left by *PEX2* and *CBP1* and on the right by *SUI1* and *CWC25*, while *HML* cassette was flanked by *VAC17* at the left side and by *FET4* and *COS12* at the right side (**Fig. 3**). Blast analysis of DBG2OLC scaffold UEMZ01000008.1 indicated that the first 1,427,380 bp was almost collinear to NBRC1130 chromosome C and that genes upstream and downstream the *MAT***a**$^{\mathrm{N}}$ cassette were P and T-type, respectively. Congruently, *MAT***a**$^{\mathrm{N}}$ cassette retained the synteny with *PEX2*$^{\mathrm{P}}$ and *CBP1*$^{\mathrm{P}}$ at 5'- and *SUI1*$^{\mathrm{T}}$ and *CWC25*$^{\mathrm{T}}$ at 3'-end (2). However, 3-end side was interrupted at *RAD50*$^{\mathrm{T}}$. Scaffold UEMZ01000003.1 (rc) linked *CHA1*$_{\mathrm{L}}^{\mathrm{T}}$-*MTLα*$^{\mathrm{P}}$-*SLA2*$^{\mathrm{P}}$ and *CHA1*$_{L}^{\mathrm{P}}$-*MTL*α$^{\mathrm{P}}$-ZYRO0F18634g$^{\mathrm{P}}$ cassettes (**Fig. 3**). Reciprocal translocation between chromosomes C from T and P haplotypes led to a similar arrangement in CBS4837 (Watanabe et al., 2017). As result, in CBS4837 the *MAT*α$^{\mathrm{P}}$ expression cassette is linked to *CHA1*$_{\mathrm{L}}^{\mathrm{T}}$-*MTLα*$^{\mathrm{P}}$-*SLA2*$^{\mathrm{P}}$ and *CHA1*$_{L}^{\mathrm{P}}$-*MTL*α$^{\mathrm{P}}$-ZYRO0F18634g$^{\mathrm{P}}$. In ATCC42981_R, flanking gene analysis also supported a linkage between *MAT***a**$^{\mathrm{N}}$ and *CHA1*$_{\mathrm{L}}^{\mathrm{T}}$-*MTLα*$^{\mathrm{P}}$-*SLA2*$^{\mathrm{P}}$/*CHA1*$_{L}^{\mathrm{P}}$-*MTL*α$^{\mathrm{P}}$-ZYRO0F18634g$^{\mathrm{P}}$ cassettes, suggesting that scaffolds UEMZ01000008.1 and UEMZ01000013.1 contributed to the chimeric chromosome C. Like in CBS4837, this chromosome C could arise from a reciprocal translocation between two ancestor T and P chromosomes C (Watanabe et al., 2017).

Scaffold UEMZ01000028.1 was chimeric with P-type (*PEX2* and *CBP1*) and T-type (*FET4* and *COS12*) genes upstream and downstream the cassette *DIC1*$^{\mathrm{P}}$-*MTL*α$^{\mathrm{T}}$-ZYRO0F18634g$^{\mathrm{T}}$ (**Fig. 3**). The loss of gene block between *MAT* and *HML* cassettes suggested that a deletion between *MAT* and *HML* cassettes led to this arrangement, similar to that described in strain NBRC0686 (Watanabe et al., 2013; **Fig. S5**). Alternatively, in CBS4837 a similar arrangement resulted from reciprocal translocation leading to chimeric chromosome C (Watanabe et al., 2017).

DBG2OLC scaffold UEMZ01000013.1 exhibited T-type flanking genes around *DIC1*$^{\mathrm{T}}$-*MTL*α$^{\mathrm{P}}$-*SLA2*$^{\mathrm{P}}$ and *CHA1*$_{L}^{\mathrm{T}}$-*MTL*α$^{\mathrm{T}}$-ZYRO0F18634g$^{\mathrm{T}}$. Overlapping region with scaffold UEMZ01000007.1 suggested that scaffolds UEMZ01000013.1 and UEMZ01000007.1 could contribute to the T-type chromosome C in ATCC42981_R (**Fig. 3**).

NBRC1130$^{\mathrm{T}}$ strain has the *HMR***a** locus on chromosome F. *SIR1* and a set of genes including *PUT4*, *CYB2*, *COS12*, and *PEP1* are upstream and downstream to *HMR***a**, respectively (**Fig. S5**). ATCC42981_R DBG2OLC assembly exhibited two scaffolds retaining this synteny, namely 5 and 15 (rc). Scaffold 5 contained P-type genes, including *SIR1*$^{\mathrm{P}}$ (**Fig. 3**). Unfortunately, DBG2OLC assembler interrupted scaffold 5 after *CHA1*$^{\mathrm{P}}$. However, MaSuRCA assembly retained *PUT4*$^{\mathrm{P}}$, *CYB2*$^{\mathrm{P}}$, *COS12*$^{\mathrm{P}}$, and *PEP1*$^{\mathrm{P}}$ downstream to *HMR***a**, suggesting that ATCC42981_R has a P-type chromosome F collinear to NBRC1130 chromosome F. Syntenic relationships and Blast analysis supported scaffold UEMZ01000015.1 as the T-type version of NBRC1130$^{\mathrm{T}}$ chromosome F (**Fig. S5**).

Figure 3. **Inferred gene organization around the *MTL* loci in ATCC42981_R.** Scaffold (sc) numbers referred to the DBG2OLC genome assembly deposited in European Nucleotide Archive under accession number PRJEB26771 (Bizzarri et al., 2018); for brevity each scaffold is identified by the last number of ENA code (*i.e*. UEMZ01000028.1 in short sc28). Solid and dotted lines referred to T and P-subgenomes, respectively. Genes from T and P-subgenomes are marked with T and P superscripts, respectively, while *DIC1* and *MAT***a**2 new variants with N superscript. Red and black rectangles defined *MAT* and *HML*/*HMR* loci, respectively. Scaffold lengths are not in scale. Abbreviation: r.c., reverse complement.


NBRC1130^T-like
chr. C
sc8^P
DIC1^N *MAT*a^N *SLA2*^{T}
CBP1^P PEX2^P SUI1^T CWC25^T RAD50^T
CHA1$_L$^{T} *HML*α^P *SLA2*^{P} *CHA1*$_L$^{P} *HML*α^P ZYRO0F18634^P
VAC17^T SUI1^P CWC25^P VAC17^P FET4^P COS12^P
sc3^P (r.c.)
DIC1^P *HML*α^T ZYRO0F18634^T
sc28^{PT}
CBP1^P PEX2^P FET4^T COS12^T
DIC1^T *MAT*α^P *SLA2*^{P} *CHA1*$_L$^{P} *HML*α^T ZYRO0F18634^T
chr. C
sc13^T
CBP1^T PEX2^T SUI1^T CWC25^T VAC17^T FET4^T COS12^T
sc7^T (r.c.)
CBP1^T PEX2^T
CHA1^P *HMR*a^P ZYRO0C18632^P
sc5^P
chr. F
SIR1^P
CHA1^T *HMR*a^T ZYRO0C18632^T
sc15^T (r.c.)
chr. F
SIR1^T PUT4^T CYB2^T COS12^T PEP1^T

Disclosing the true cell identity

Watanabe et al., (2017) identified two *MTL* patterns: strains with pattern A, such as NBRC110957, exhibit two active *MAT* loci, namely *DIC1*^{T}-*MAT*^{P}-*SLA2*^{P} and *CHA1*^{T}-*MTL*^{P}-*SLA2*^{T}, while strains with pattern B have *DIC1*^{T}-*MAT*^{P}-*SLA2*^{P} as active *MAT* locus, even if they also actively transcribed genes from *CHA1*$_L$^{T}-*MTL*^{P}-*SLA2*^{P}. JCM66020 belonged to this last group, exhibited a *MAT*α^P idiomorph and, congruently, mates only the tester strain **a** (CBS4838). Conversely, ATCC 42981_R displays another pattern of putatively active *MAT* loci, namely, *DIC1*^{T}-*MAT*α^P-*SLA2*^{P} and *DIC1*^{N}-*MAT*a^N-*SLA2*^{T}, in addition to the *CHA1*$_L$^{T}-*MTL*^{P}-*SLA2*^{P} cassette. RT-PCR analysis confirmed that α^{P}1, α^{P}2, **a**^{N}2 and **a**^{T}1 genes were expressed, while **a**^{P}1 gene encoded by *CHA1*^{P}-*MTL*^{P}-ZYRO0C18392g^{P} cassette was silent (**Fig. 4**). Interestingly, **a**^{T}1-specific RT-PCR resulted in two PCR amplicons compatible with alternative spliced intronic sequence.

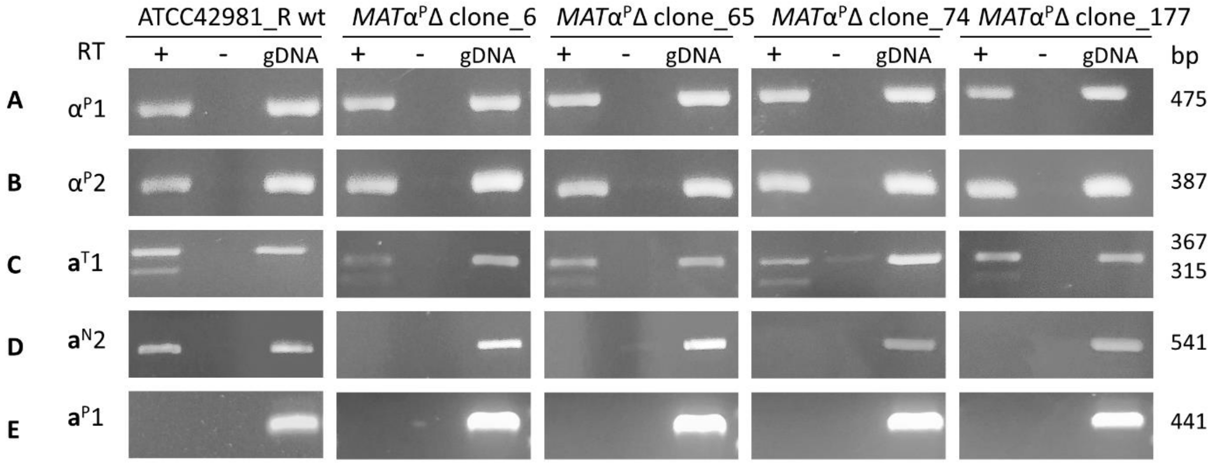
Genome comparison with other pre-WGD yeasts indicates that *HML*α silent cassettes are generally 5'-flanked by *CHA1*$_{L}$ (Gordon et al., 2011). Conversely, strains with pattern B actively transcribed *MTL* genes from *CHA1*$_{L}^{T}$-*MTL*-*SLA2*^{P} cassette without that these transcripts affect cell identity (Watanabe et al., 2017). This is evident for strain CBS4837, where genes encoding opposite α^{P} and **a**^{P} idiomorphs are both expressed by *DIC1*^{T}-*MAT*^{P}-*SLA*2^{P} and *CHA1*$_{L}^{T}$-*MTL*^{P}-*SLA2*^{P} cassettes, respectively. In JCM22060 (encoding α^{P} genes at both these loci), outcross experiment with CBS4837 and gamete segregation support that cell identity was determined by *DIC1*^{T}-*MAT*^{P}-*SLA*2^{P} cassette. To establish which cassette contributes to cell identity in ATCC 42981_R, we deleted α^{P} idiomorph genes by replacing all segment including α^{P}1, α^{P}2 encoding genes and intergenic region from *DIC1*^{T}-*MAT*α^{P}-*SLA*2^{P} with *loxP-kanMX-loxP* module. From approximatively 300 colonies we obtained four G418-resistant *MAT*α^{P}Δ candidates. PCR approaches showed that these clones contained *loxP-kanMX-loxP* surrounded by *DIC1*^{T} *SLA*2^{P} instead of *MAT*α^{P} locus (**Fig. S1**).

Gene deletion of *DIC1*^{T}-*MAT*α^{P}-*SLA*2^{P} cassette should abolish the heterozygosity at the *MAT***a**/α active loci and results in an allodiploid partially resembling a haploid cell with a mating-type. Conversely, we still observed α^{P}1 and α^{P}2 gene expression in ATCC 42981_R *MAT*α^{P}Δ transformants (**Fig. 4**). These mRNAs could be only transcribed by the not completely silenced cassettes *CHA1*$_{L}^{T}$-*MTL*α^{P}-*SL21*^{P} or by *DIC1*^{P}-*MTL*α^{P}-ZYRO0F18634g^{P}.

Since allodiploid lacking one *MAT* active locus might behave like haploid with opposite mating-type, we expected to detect both a1 and a2 transcripts in ATCC 42981_R *MAT*α^{P}Δ mutants. In some haploid pre-WGD species, **a**2 gene encodes a transcription activator of **a**sg, while **a**1 should not affect **a**sg in **a** cells (Tsong et al., 2003, 2006; Baker et al., 2012). Unexpectedly, RT-PCR showed that *MAT*α^{P} deletion switched off **a**2 but not **a**1 gene expression (**Fig. 4**). By contrast, ATCC 42981_R wild type both transcribed both **a**1 and **a**2 genes. Preliminary end-point RT-PCRs showed that the **a**sg *AGA2*, *STE6* and *STE2* are transcriptionally active in both ATCC 42981_R and *MAT*α^{P}Δ clone_74 (data not shown).

Figure 4. Gene expression at the *MAT* loci of ATCC 42981_R wild type and *MAT*α^{P}Δ deletion mutants. cDNA was amplified from total RNA extracted from stationary growing cells. Plus or minus indicates with or without reverse transcriptase in cDNA synthesis reaction, respectively. Gene

variants from T and P-subgenomes are indicated with T and P superscripts, while divergent *MAT***a**2 gene variant from *DIC1*$^{\mathbf{N}}$-*MAT***a**$^{\mathbf{N}}$-*SLA2*^{T} cassette with N superscript. Capital letters from A to E refer to primer sets listed in **Table S1**. Abbreviations: wt, wild type; gDNA, genomic DNA.

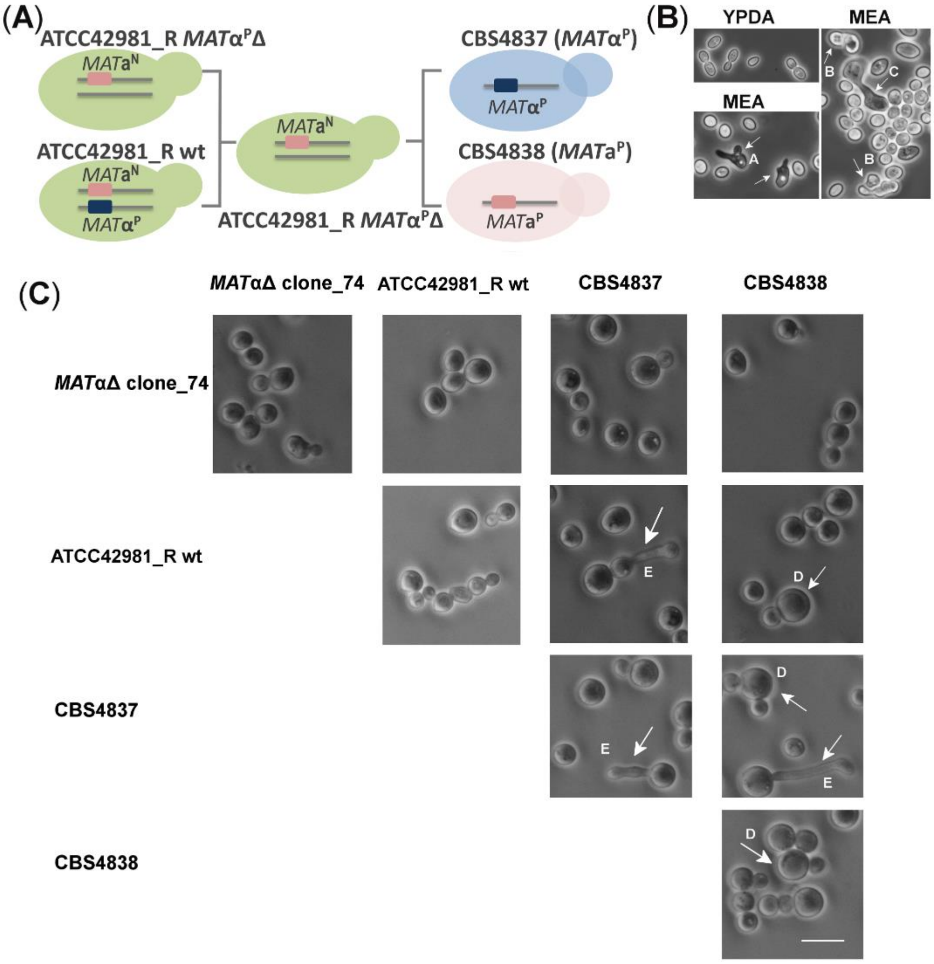


Mating and sporulation competence assays

To test whether the *MAT*α^{P} deletion rescues the mating competence in ATCC42981_R, we carried out self- and out-cross fertility assays of the wild type strain and the *MAT*α^{P}Δ transformants as monoculture or in mixture with CBS4837 (α) or CBS4838 (**a**) mating testers, respectively. If *MAT*α^{P}Δ transformants behave as homothallic haploid, they should produce shmoo and conjugated asci as monoculture, while, if they are like heterothallic haploid, they should mate and sporulate in mixture with CBS4837 or CBS4838. We used three media reported in literature to promote zygote formation and conjugated asci of *Zygosaccharomyces* cells, as proved for *Z. parabailii* G21C (**Fig. 5**). In particular, 5-6% NaCl addition was reported to increase sporulation occurrence (Mori and Onishi, 1967). Like the wild type strain, *MAT*α^{P}Δ mutants did not show any evidence of conjugative bridge and/or conjugative asci either as monoculture or in mixture with the mating testers (**Fig. 5**). The composition of three test media did not affect the inability to mate or to undergo meiosis. Overall, these evidences indicate that the deletion of active *MAT*α^{P} locus did not make ATCC42981_R cells phenotypically heterothallic or homothallic haploids.

Figure 5. Self- and out-cross fertility assays of ATCC 42981_R wild type and *MAT*α^{P}Δ deletion mutant. Panel (**A**) shows the scheme of self- and out-cross fertility assays involving ATCC 42981_R

wild type and *MAT*αΔ clone_74, while panel (**B**) depicts representative phase contrast microscopic images of control strain *Zygosaccharomyces parabailii* G21C grown in YPDA and MEA media. Panel (**C**) shows selected phase contrast microscopic images of representative ATCC 42981_R *MAT*αΔ clone_74 grown as monoculture and in mixture with ATCC 42981_R wild type, CBS4837 (α) and CBS4838 (**a**). Phase contrast microscopic images of ATCC 42981_R wild type, CBS4837 and CBS4838 monocultures were reported for comparative purposes. The scale bar represents 11 µm and was reported in one picture for brevity. Capital letters from A to E indicate abortive shmoo, conjugated asci with ascospores, conjugative bridge, giant cells, and abnormal conjugative tube, respectively. All cells were photographed after 7 days on incubation on MEA medium at 27°C. Abbreviation: wt, wild type.



Discussion

Our study is the first to combine the Nanopore whole-genome sequencing to conventional PCR-based methods in order to survey *MTL* loci in a *Z. rouxii* allodiploid genome. This yeast is particularly prone to outbreeding and provides a particularly appealing platform to study genome re-shaping after the merging of two parental subgenomes. Recombination and introgression between the two subgenomes have been rampant in hybrid yeasts, resulting in loss of heterozygosity and gradual genome reduction (Sipiczki, 2008). In *Z. rouxii MTL* loci markedly contribute to this genomic plasticity (Watanabe et al., 2013; Solieri et al., 2014b). As consequence, this species frequently undergoes chromosomal translocations at the *MTL* loci, which make hard the understanding of true cell identity. For example, haploid *Z. rouxii* strain CBS 732^T switched mating-type at the *CHA1-MAT-SLA2* locus (Bizzarri et al., 2018), suggesting that *CHA1* gene flanks the actively transcribed *MAT* locus instead of *DIC1*. Several combinations of different flanking gene variants and distinct idiomorph encoding genes make challenging and laborious to resolve the complex genetic *MTL* architecture by PCR targeted approaches. For these reasons, we generated a high-quality genome assembly in order to dissect complex rearrangements at the *MTL* loci that were not fully resolvable from the earlier survey based only on long-range PCR amplification (Bizzarri et al., 2016). One of the major advantages of the ONT is the possibility of sequencing very long DNA fragments, which span the entire *MTL* cassettes. This strategy assures to accurately reconstruct gene order around different *MTL*s. On the other hand, using noisy ultra-long reads for self-correction and assembling of highly heterozygous genomes can affect the consensus sequence accuracy and the parental haplotypes sorting. In the case of ATCC 42981_R, distinguishing between homeologous sequences is further challenging as only the *Z. rouxii* parental genome is available to guide homeologous scaffold assembly. Error rate made necessary to polish MinION reads with Illumina-derived reads, resulting into DBG2OLC-driven hybrid *de novo* genome assembly (Bizzarri et al., 2018). However, our result showed that a single “best assembler” does not exist to resolve highly heterozygous and highly repeated *MTL* regions. DBG2OLC assembly suffers from poor performance in certain sequence contexts such as regions with low coverage or regions that contain short repeats. Besides, the new assembly generated with MaSuRCA showed higher accuracy compared to DBG2OLC, but loses some *MTL* cassettes. Combination of both assemblies facilitated the elucidation of complex rearrangements around *MTL* loci in ATCC 42981_R and allowed the capture of *MTL* cassettes divergent from the expected T- and P-subgenomes. As bottom-end validation step, Sanger sequencing was used to discard artificial *MTL* arrangements arisen from flawed contig assemblies. This integrated strategy should allow the resolution of controversies over the extent of *MTL* loci in

ATCC 42981_R genome derived from the analysis of the Japanese stock JCM22060 (Watanabe et al., 2017).

Reconstruction of *MTL* structure indicates that ATCC42981_R resembles CBS4837 for the exception of an additional scaffold containing *DIC1*^{T}-*MTL*α^{P}-*SLA2*^{P} linked to *CHA1*$_{L}^{T}$-*MTL*α^{T}-ZYRO0F18634g^{T} (**Fig. 3**). This assessment was congruent with previous PFGE-Southern blotting which showed two signals for *MAT*α-specific probe (Bizzarri et al., 2016).

The most significant difference between ATCC 42981_R and JCM22060 is that ATCC 42981_R harbors the transcriptionally active *MAT***a**^{N} cassette in addition to the expected *MAT*α^{P}. Differently from *Z. parabailii* (Ortiz-Merino et al., 2017), *MAT***a**^{N} cassette of ATCC 42981_R contains *MAT***a**1 gene. This means that *Z. rouxii* retains the ancestral regulatory circuit based on **a**1–α2 heterodimer as diploid cell sensor (Booth et al., 2010). Watanabe et al., (2017) showed that strain JCM22060, which contains only *MAT*α^{P}, mates the tester strain **a** in a medium containing Shoyu-koji extract. By contrast, we did not find any evidence of meiosis or mating in ATCC 42981_R (Bizzarri et al., 2016), when grown on the media reported in literature to promote *Z. rouxii* mating and sporulation (James and Stratford, 2011). Watanabe et al., (2017) argued that difference in medium composition could account for the phenotypic discrepancy between ATCC 42981_R and the sister stock JCM22060. As the Shoyu-koji extract is difficult to gain in western countries, we cannot rule out this hypothesis. Otherwise, heterozygosity at the *MAT* locus could significantly contribute to the allodiploid infertility. In particular, the hybrid heterodimer with divergent **a**1 and α2 subunits brings the cell in an 'haploid-diploid intermediate' functional state which hamper the meiosis commitment and the responsiveness to mating stimuli (Bizzarri et al., 2016).

In *Saccharomyces* clade, experimental deletion of one *MAT* locus leads to allotetraploids suitable to undergo meisosis (Greig et al., 2002; Pfliegler et al., 2012). Based on this consideration, *Z. parabailii* and *Z. pseudobailii* hybrid strains ATCC60483 and MT15 were recently supposed to be fertile due to the accidental breakage of 1 of the 2 homeologous copies of the *MAT* locus (Ortiz-Merino et al., 2017; Braun-Galleani et al., 2018). A prediction of this model is that artificial deletion of one *MAT* locus in *Zygosaccharomyces* cells should override the arrest in mating commitment. In our model, ATCC 42981_R cells did not behave as haploids with idiomorph **a**, when the *MAT*α^{P} locus was deleted. This suggests that mechanism underpinning the cell identity in *Z. rouxii* could be different from those involved in cell identity regulation of the sister species *Z. parabailii* and *Z. pseudobailii*. Compared to ATCC60483 and MT15, ATCC 42981_R exhibits more complex sex determining loci, as it harbors opposite idiomorphs at the *MAT* loci and retains the ancestral **a**1-α2 circuit which prevents mating competence in allodiploids.

Gene deletion of transcriptionally active *MAT*α^P locus did not rescue the ability to produce conjugated asci in ATCC 42981_R, while the persistence of α1 and α2 transcripts suggests that silencing of *HML*α was leaky in ATCC 42981_R. Consequently, α^P genes either from *CHA1*$_L^T$-*MTL*α^P-*SLA2*^{P} or *CHA1*$_L^P$-*MTL*α^P-ZYRO0F18634g^P are transcriptionally active in *MAT*α^{P}Δ mutants. Strain NBRC110957, which does not have the *CHA1*$_L^T$-*MTL*α^P-*SLA2*^{P} cassette, uses *CHA1*$_L^P$-*MTL*α^P-ZYRO0F18634g^P as donor during switching from **a**^{P} to α^P (Watanabe et al., 2017). This suggests that *CHA1*$_L^P$-*MTL*α^P-ZYRO0F18634g^P cassette is most likely silenced and that α^P could be expressed by the *CHA1*$_L^T$-*MTL*α^P-*SLA2*^{P} in ATCC 42981_R. To support this, in strain CBS4837 the *CHA1*$_L^T$-*MTL***a**^{P}-*SLA2*^{P} cassette is actively transcribed. These findings make less probable the alternative hypothesis that *MAT*α^P deletion induces *HML*α cassette de-silencing. Abnormal expression of cryptic *HMR*/*HML* loci has been described in *Vanderwaltozyma polyspora*, the *Z. rouxii* closest relative that branched after the WGD (Roberts and Van der Walt, 1959). Consequently, *V. polyspora* haploid cells behave as **a**/α diploid and appear mating-incompetent for many generations only to subsequently restore silencing. Significantly, *V. polyspora* lacks of Sir1 histone deacetilase, which mediates the *HM* loci silencing in *S. cerevisiae* together with the SIR complex (Sir2/Sir3/Sir4). In *S. cerevisiae* failure to recruit Sir1 is thought to account for the instability of subtelomeric silencing relative to *HM* loci (Chien et al., 1993). Like *V. polyspora*, *C. glabrata* is another species close to *Z. rouxii*, which lacks of a *SIR1* ortholog (Gábaldon et al., 2013). The defection of a complete silencing system lead to the expression of *MAT***a** gene in *C. glabrata MAT*α cells (Muller et al., 2008) and makes *HML* more prone to HO cleavage at the Y/Z junctions (Boisnard et al., 2015). *Z. rouxii* has the archetypal member of the *SIR1* family, *KOS3* (*K*in *o*f *S*ir1 3) (Gallagher et al., 2009). In pre-WGD species *Torulaspora delbrueckii KOS3* located ~1 Kb away from the copy of *HMR* and plays a key role in *HML*/*HMR* silencing (Ellahi and Rine, 2016). Strikingly, in ATCC 42981_R two *KOS3* copies, *KOS3*^{T} and *KOS3*^{P}, were found upstream the *CHA1*^{T} and *CHA1*^{P} genes flanking *HMR***a**^{T} and *HMR***a**^{P} loci, respectively. In addition, Sir1 and the components of SIR complex have been reported to rapidly evolve in the Saccharomycetaceae species. This could potentially jeopardize the efficiency of the silencing machinery in interspecific hybrids. For example, Sir1, Sir4 and the *cis*-acting silencer sequences are incompatible in *S. cerevisiae* x *S. uvarum* hybrids (Zill et al., 2012; Zill et al., 2010). In ATCC 42981_R, heterochromatin formation across silent loci could be less effective due to the incompatibility in the silencing machinery between the T- and P-subgenomes. Watanabe et al., (2017) suggest that chimeric *MTL* cassettes could be particularly prone to perturbation of cell-to-cell gene expression due to incompatibility among elements of silencing machine. This could produce allodiploid cells undergoing epigenetic silencing at one of *MAT* loci. According to this model, only inappropriate silencing of *MAT* locus should lead to fertility restoration. Our results suggest that the lapses in *HML* silencing mask the loss of heterozygosity at the *MAT* locus induced by gene deletion in ATCC 42981_R.

Strikingly*, MAT*α^{P} locus deletion also revealed that the depletion of α^{P}1 and α^{P}2 genes switched off the **a**2 but not the **a**1 gene transcription. Moreover in both deleted and wild type-strains two **a**1 alternative spliced isoforms, one of them compatible with the retention of first intron. In *S. cerevisiae* exon-intron structure is conserved and the retention of first intron resulted in a functional **a**1 transcriptional factor that prevents mating (Ner and Smith, 1989). Since α1 activates the αsgs in the ancestral circuit of yeast cell identity (Baker et al., 2011), we rule out the possibility that α1 is involved in **a**2 gene repression. In *S. cerevisiae*, α2 represses **a**sg by binding **a**sg *cis*-regulatory sequences cooperatively with a MADS-box transcription regulator, Mcm1 (Tsong et al., 2003). *Z. rouxii*, which branched from the *S. cerevisiae* lineage prior to the loss of the **a**2 gene, should maintain both the **a**2 activation and the α2 repression of **a**sgs (Tsong et al., 2006; Baker et al., 2012). In *Lachancea kluyveri* haploid cells, α2 deletion induces the transcription of the **a**sgs *AGA1* and *AGA2,* while **a**2 deletion decreases the **a**sg transcript levels (Baker et al., 2012). However, to the best of our knowledge, no evidence has been provided until now about the consequences of α2 gene deletion in diploid cells which retain **a**2 gene. As **a**1 is still expressed in *Z. rouxii MAT*αΔ/*MAT***a** hemizygous cells, we speculate that **a**2 silencing could be a promoter-driven event directly or indirectly regulated by α2. Furthermore, in our *MAT*αΔ/*MAT***a** model, the **a**gs were expressed even when **a**2 was switched off by the *MAT*α2 deletion, supporting the existence of an **a**sgs hybrid regulatory network in *Z. rouxii*.

Concluding remarks

This study recharges the pattern of *MTL* loci in the allodiploid strain ATCC 42981_R. By taking advantage from ONT technology, we captured a novel *MAT***a** cassette which did not correspond to the expected T and P counterparts, providing preliminary evidences that a third haplotype contributes to this genome. The differences between ATCC 42981_R and JCM22060 support that

*MTL*s are a root source of genetic variation, leading to novel chimeric *MTL* cassettes, to different cell identities and, consequently, to distinct phenotypic behaviors. While further researches are required to investigate mechanisms responsible of this extensive *MTL* reshaping, our results confirm that these yeast stocks are genetically unstable (Watanabe et al., 2013; Bizzarri et al., 2018). We also demonstrated how *HMR*/*HML* silencing is crucial to establish the cell identity, as leakage in *HML* silencing prevents allodiploid cells deleted for the *MAT*α^{P} locus, to behave like haploids. How allodiploid cell modulates **a**2 expression via α2 transcriptional factor represents an unexplored regulatory circuit that has to be investigated in future.

Funding

LS was partially supported by the Italian Ministry of Education, University and Research (MIUR), within the framework of the Italian National Grant for Fundamental Research (FFABR 2017). The work of HS group was supported by the Ministry of Education, Youth and Sports of CR (MEYS) within the LQ1604 National Sustainability Program II (Project BIOCEV-FAR) and by the project “BIOCEV” (CZ.1.05/1.1.00/02.0109).

Acknowledgments

We are grateful to Prof. Paolo Giudici for his valuable comments and to Marcello Benevelli for help in mating assays.

Author Contributions

SC and LS contributed conception and design of the study. MB conducted the experiments described in this study. LB contributed to *in vitro* PCR validation and **a**sg expression; HS and MD contributed to deletion mutant construction. SC and LPP performed bioinformatic analysis of the whole genome sequence data. LS wrote the manuscript, SC and MB contributed to draft revision. All authors read and approved the final manuscript.

CHAPTER 9: ATCC 42981 GENE PREDICTION AND FUNCTIONAL ANNOTATION

Introduction

Interspecies hybridization between two haploid parental species, which results in sterile allodiploids, may represent the first step in yeast speciation when genome duplication allows the recovery of fertility and mating competence (Marcet-Houben and Gabaldòn, 2015).

Previous studies suggested that inside the *Zygosaccharomyces rouxii* clade, the ATCC 42981 strain was an allodiploid formed after hybridization between two strains of *Z. rouxii* and *Z. pseudorouxii*, since ribosomal RNA gene sets and several *Z. rouxii* genes responsive for the peculiar osmo- and halo-tolerance (sodium membrane pumps and enzymes involved in glycerol production and accumulation) were duplicated (Gordon and Wolfe, 2008). Random sequencing reads from a genomic plasmid library showed a limited nucleotide divergence between the putative parental genomes (4%-15%). Nevertheless, fertility and mating assays revealed that ATCC 42981 was unable to sporulate (Bizzarri et al., 2016). Different cytological and molecular mechanisms could be responsible for this phenotypic inability, such as homeomologous chromosomes miss pairing and the lack of interaction or the malfunctioning of interacting alleles derived from divergent genomes (Bateson-Dobzhansky-Muller model) (Liti et al., 2006; Morales and Dujon, 2012).

Inferring genome-wide incompatibility at the molecular level requires the availability of ATCC 42981 whole genome sequence. The plasmid genome library covered just the 0.5% of ATCC 42981 genome size that was previously estimated to be around 21.9 ± 0.2 Mb by flow cytometry (Solieri et al., 2008).

The recovery of parental haplotypes from heterozygous diploid genomes is a challenging task by standard assembly pipelines, which use short-read sequencers output. The collapsing of homozygous regions and multiple contig assignments of heterozygous sequences often result in highly fragmented assembly. In the case of ATCC 42981 genome sequencing, we tried to circumvent this caveat adopting a hybrid approach that uses high quality MiSeq short reads to correct errors in the MinION long reads and an assembling pipeline specifically designed for resolving parental haplotypes in hybrid genomes (Ye et al., 2016; Pryszc and Gabaldón, 2016; Bizzarri et al., 2018). Frequently post-hybridization recombination events and loss of heterozygosity occurred after interspecific mating. We took advantage of the released annotated genome of *Z. rouxii* haploid type-strain CBS 732^T, along with syntheny conservation in pre- and post-WGD species to reconstruct the parental contribution to ATCC 42981 gene repertoire. To this aim, we integrated two annotation pipelines, which exploited *ab initio* as well as evidence-based approaches for gene model prediction and functional annotation.

Materials and Methods

Gene prediction and functional annotation

To perform gene prediction and annotation we used two alternative pipelines: the Yeast Genome Annotation Pipeline (YGAP) (Proux-Wéra et al., 2012) and Funannotate pipeline (https://github.com/nextgenusfs/funannotate) (Palmer, 2016). Pipeline annotations flow-chart is outlined in **Figure 1**. Default parameters were set for YGAP, except for the status of the species, which has been switched from pre- to post-WGD species with no frameshift correction.

Proteome from *Z. rouxii* CBS 732^T and UniProt fungi dataset (uniprot.sprot.fungi.dat.gz) (The UniProt Consortium, 2012) provided Evidence-based prediction for the Funannotate pipeline.

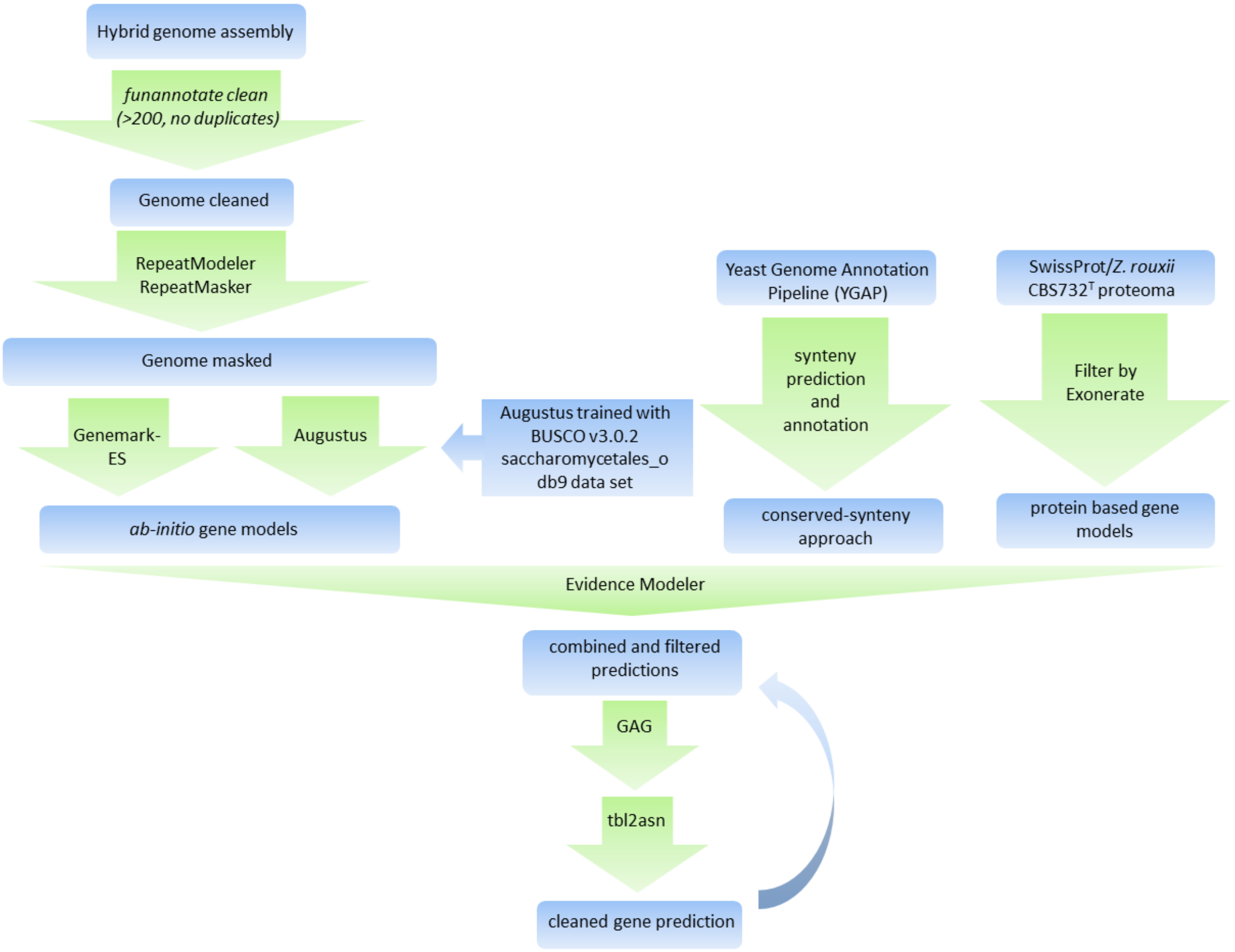
Ab initio gene predictor Augustus (Stanke et al., 2006) was trained with BUSCO2 orthologs for Saccharomycetales (saccharomycetales_odb9.tar.gz) (Simão et al., 2012).

Functional annotation of predicted genes was performed including in Funannotate pipeline InterProScan 5 (Jones et al., 2014), antiSMASH 3.0 (Weber et al., 2015), Phobius (Käll et al., 2005) and EggNOG (Cepas et al., 2016).

Non-coding RNAs (ncRNA) were inferred with tRNAscan-SE (Lowe and Eddy, 1997) for transfer RNA (tRNA) genes and with Blastn for ribosomal RNA (rRNA) genes using *Z. rouxii* CBS 732^T rRNA genes as query (XR_002648416, XR_002648417, XR_002648418, XR_002648419, XR_002648420, XR_002648422, XR_002648423, and XR_002648424). Mitochondrial DNA (mtDNA) was annotated by Blastn using *Z. rouxii* CBS 732^T *COXII* gene (AF442248.1) and full-length *Zygosaccharomyces mellis* strain Y-12628 mitochondrial genome (KU920675).

Genome annotation metrics was produced by COGNATE v1.01 (Wilbrandt et al., 2017).

Figure 1. Flow-chart of the pipelines used for ATCC 42981 gene prediction and genome annotation.


Hybrid genome assembly
funannotate clean (>200, no duplicates)
Genome cleaned
RepeatModeler
RepeatMasker
Genome masked
Genemark-ES
Augustus
Augustus trained with BUSCO v3.0.2 saccharomycetales_odb9 data set
ab-initio gene models
Yeast Genome Annotation Pipeline (YGAP)
synteny prediction and annotation
conserved-synteny approach
SwissProt/*Z. rouxii* CBS732^T proteoma
Filter by Exonerate
protein based gene models
Evidence Modeler
combined and filtered predictions
GAG
tbl2asn
cleaned gene prediction

Orthologous cluster and GO terms annotation

Annotation datasets generated by both pipelines were exploited to identify orthologous clusters between *Z. rouxii* type-strain CBS 732^T and the allodiploid *Z. rouxii* ATCC 42981. Proteome from CBS 732^T genome project (PRJNA39573) was used as reference. OrthoVenn web server provided cluster analysis among putative orthologs (with default *E*-value and Inflation value) (Wang et al., 2015b).

OrthoVenn inferred putative functions of orthologous clusters in Venn diagram in overlapping regions using fungi pre-computed GO Slim sets.

KEGG pathway mapping

KEGG Orthology (KO) for the five main categories (metabolism, genetic information processing, environmental information processing, cellular process and human disease) was determined exploiting the YGAP annotation dataset, which was more conservative than Funannotate annotation dataset (Ogata et al., 1999). *Z. rouxii* CBS 732^T orthologs were retrieved from YGAP annotation pipeline, kindly modified by Ortiz-Merino and colleagues (UCD Conway Institute, School of Medicine, University College Dublin, Ireland).

The associated KO terms were provided by BlastKOALA tool and mapped in pathways by KEGG Mapper - Search Pathways (Kanehisa et al., 2015).

Gene products mapped on yeast meiosis pathway (ID: 04113) were manually curated by directly blasting both ATCC 42981 annotations against *Z. rouxii* CBS 732^T orthologs and amino acid identity was calculated along with CBS 732^T : ATCC 42981 gene copy number ratio. To recover *IME1* gene product missing in CBS 732^T KO annotation, we performed blasting against *S. cerevisiae* proteome.

Syntheny map reconstruction

Syntheny block reconstruction between the reference CBS 732^T chromosomes (genome project: PRJNA39573) and ATCC 42981 scaffolds was performed by SynChro (Drillon et al., 2014). SynChro algorithm identifies the syntheny blocks relaying on a series of anchored Reciprocal Best Hits (RBH) between putative orthologous, which co-localize along chromosomes and scaffolds in the two compared annotation datasets. For CBS 732^T the GenBank deposited proteome was compared with annotated proteins in ATCC 42981, considering the two annotation pipelines separately. SynChro was run with the highest syntheny block stringency by setting the Δ parameter to 2.

Colinearity among ATCC 42981 scaffolds was reconstructed using putative homeologs identified by YGAP pipeline: gene products annotated as similar to the same CBS 732^T orthologs were considered homeologs. Relationships among ATCC 42981 scaffolds were represented using shinyCircos tool (Yu and Yao, 2018).

Results

Gene prediction and functional annotation

YGAP and Funannotate pipelines resulted in an equivalent number of protein-coding genes: 11,031 and 11,117, respectively. However, the detailed prediction metrics showed a different distribution in genomic features between the two pipelines (**Tables 1, 2**). Generally, YGAP reported a wider distribution of the genomic features compared to Funannotate. The average number of transcript and CDS data per scaffold were similar (YGAP vs Funannoate: 334.27 vs 318.48 transcript/scaffold and 670.21/410.72 CDS/scaffold). Transcripts predicted by Funannotate were, on average, longer (1334.7 vs 1500.56). However, the mean individual length of CDS, exon and intron were higher in genes predicted by YGAP (**Tables 1, 2**). The most evident discrepancy affected the mean protein length (888.21 vs 488.45).

Among ncRNAs we annotated one complete set of rDNA genes (5S, 5.8S, 18S and 26S) on scaffold 05, while the second rDNA set on scaffold 22 lacked of the gene coding for 5S ribosomal subunit. According to Gordon and Wolfe (2008), ATCC 42981 cloned rDNA from T (AM943656) and P (AM943657) subgenomes differed for a single nucleotide polymorphism (SNP) in 5.8S and some deletions in ITS1 and ITS2. Genotyping of the two assembled rDNA attributed rDNA set on scaffold 5 to T subgenome and rDNA set harboured by scaffold 22 to P subgenome.

Based on YGAP annotation, we counted a total of 635 tRNAs; 14 of them were annotated as no true tRNAs by the program tRNAscan-SE, owing to the lacking of a proper predicted secondary structure. Each ATCC 42981 scaffold contained tRNA, with the exception of scaffold 30 (tRNA count range per scaffold goes from 65 to 1). The scaffold with the highest tRNA denssity (*i.e.* tRNA counts/scaffold length) was scaffold 29 with 59.7 tRNA/100,000 bp. While, scaffold 02 contained the lowest number of tRNA (0.7 tRNA/100,000 bp). tRNA codons for Lysine (tK-cuu) and Glycine (tG-gcc) residues are the most represented (36 and 40, respectively). tRNA codons for Isoleucine (tI-uau), Leucine (tL-gag) and Arginine (tR-ccg) are the less abundant.

Concerning mitochondrial DNA (mtDNA), we annotated *Z. mellis* mtDNA on ATCC 42981 scaffold 32 (34,123..28,690) with a query coverage of 23.0% and an identity of 96.0%. The annotated mtDNA spanned from the 3’ end of exon 1 to 5’ end of exon 4 of *COXI* gene. When we blasted the only mtDNA available in GenBank for *Z. rouxii* CBS 732^T (*COXII* gene), we annotated three copies: two of them on scaffold 29 (12056..12640; 45483..46067) with 100.0% identity and coverage, and the third one was on scaffold 32 (4579...3995) with 98.1% identity and 100.0% coverage.

Table 1. YGAP ATCC 42981 genome functional annotation metrics summary. The open-source command-line tool COGNATE v1.01 (Wilbrandt et al., 2017) produced the detailed description of the respective gene and genome structure parameters. Individual lengths of genome features are in nucleotide (nt) and aminoacids (aa). Abbreviations: st. dv., standard deviation; scaff, scaffold; wo, without.

Genome feature	Parameter	Minimum	Maximum	Media	Median	St. dv.
Transcript data per scaffold	Transcript count per scaffold	2.00	977.00	334.27	206.00	321.56
	Transcript coverage (added transcript length/scaff lenght	0.01	0.75	0.53	0.69	0.26
CDS data per scaffold	CDS count per scaffold	4.00	1962.00	670.21	414.00	644.84
Exon data per scaffold	Exon count per scaffold	2.00	985.00	335.94	208.00	323.28
Individual lengths	Scaffold	8827.00	1903919.00	633638.15	400395.00	597839.03
	Transcript (genomic)	57.00	14841.00	1334.67	1071.00	1066.50
	Protein	38.00	9894.00	888.21	712.00	711.20
	CDS	17.00	14841.00	1321.78	1059.00	1067.09
	Exon	17.00	14841.00	1320.10	1059.00	1067.00
	Intron	55.00	552.00	146.31	83.00	133.22
Median lengths per transcript	CDS	17.00	14841.00	1330.69	1068.00	1068.08
	Exon	55.50	14841.00	1331.90	1068.00	1066.97
	Intron	27.50	276.00	73.95	41.75	66.97
Individual GC content	Transcript	13.61	61.98	41.11	40.74	3.47
	Transcript (wo ambiguity)	13.61	61.98	41.11	40.74	3.47
	CDS	13.61	61.98	41.13	40.76	3.52
	CDS (wo ambiguity)	13.61	61.98	41.13	40.76	3.52
	Exon	13.61	61.98	41.13	40.76	3.52
	Exon (wo ambiguity)	13.61	61.98	41.13	40.76	3.52
	Intron	24.66	51.88	34.72	34.23	5.63
	Intron (wo ambiguity)	24.66	51.88	34.72	34.23	5.63

Table 2. Funannotate ATCC 42981 genome functional annotation metrics summary. The open-source command-line tool COGNATE v1.01 (Wilbrandt et al., 2017) produced the detailed description of the respective gene and genome structure parameters. Individual lengths of genome features are in nucleotide (nt) and aminoacids (aa). Abbreviations: st. dv., standard deviation; scaff, scaffold; wo, without.

Genome feature	**Parameter**	**Minimum**	**Maximum**	**Media**	**Median**	**St. dv.**
Transcript data per scaffold	Transcript count per scaffold	2.00	944.00	318.48	202.00	307.05
	Transcript coverage (added transcript length/scaff lenght	0.06	0.78	0.69	0.75	0.17
CDS data per scaffold	CDS count per scaffold	7.00	1093.00	410.73	280.00	329.48
Exon data per scaffold	Exon count per scaffold	7.00	1093.00	410.73	280.00	329.48
Individual lengths	Scaffold	8827.00	1903919.00	633638.15	400395.00	597839.03
	Transcript (genomic)	153.00	14781.00	1500.58	1212.00	1144.26
	Protein	50.00	4926.00	488.45	398.00	368.21
	CDS	3.00	14781.00	1138.59	882.00	1042.78
	Exon	3.00	14781.00	1138.59	882.00	1042.78
	Intron	18.00	2932.00	113.26	54.00	216.78
Median lengths per transcript	CDS	22.00	14781.00	1317.14	1062.00	1028.73
	Exon	22.00	14781.00	1317.14	1062.00	1028.73
	Intron	9.00	847.00	48.86	36.00	55.58
Individual GC content	Transcript	8.21	59.28	40.93	40.58	3.26
	Transcript (wo ambiguity)	8.21	59.28	40.93	40.58	3.26
	CDS	0.00	75.00	40.88	40.56	4.05
	CDS (wo ambiguity)	0.00	75.00	40.88	40.56	4.05
	Exon	0.00	75.00	40.88	40.56	4.05
	Exon (wo ambiguity)	0.00	75.00	40.88	40.56	4.05
	Intron	9.52	73.33	39.54	38.78	7.88
	Intron (wo ambiguity)	9.52	73.33	39.54	38.78	7.88

Orthologous cluster and GO terms annotation

OrthoVenn identified 4813 orthologous clusters (including 17,246 sequences) shared between *Z. rouxii* CBS 732^T and ATCC 42981, considering both annotation methods (**Fig. 2**). The clustered sequences covered 97.1%, 56.53% and 59.77% of the input protein sequences from CBS 732^T, ATCC 42981 annotated with YGAP and ATCC 42981 annotated with Funannotate, respectively.

A minimal fraction of annotated genes clustered separately for YGAP (0.62%) and for Funannotate (0.57%). Five gene clusters did not find the corresponding putative orthologous in ATCC 42981, regardless of the annotation approach. Best reciprocal blast between CBS 732^T proteome and YGAP ATCC 42981 *in silico* annotated gene products validated four out of five independently OrthoVenn clustered proteins. OrthoVenn, using UniProt database, annotated 92.66% of orthologous clusters (4,460). Top 20 GO terms, divided in biological processes, molecular function and cellular component, are represented in **Figures 3, 4** and **5**.

Figure 2. Venn diagram showing the distribution of shared gene families (orthologous clusters) among *Z. rouxii* CBS 732^T and ATCC 42981, considering YGAP and Funannonate gene prediction methods.

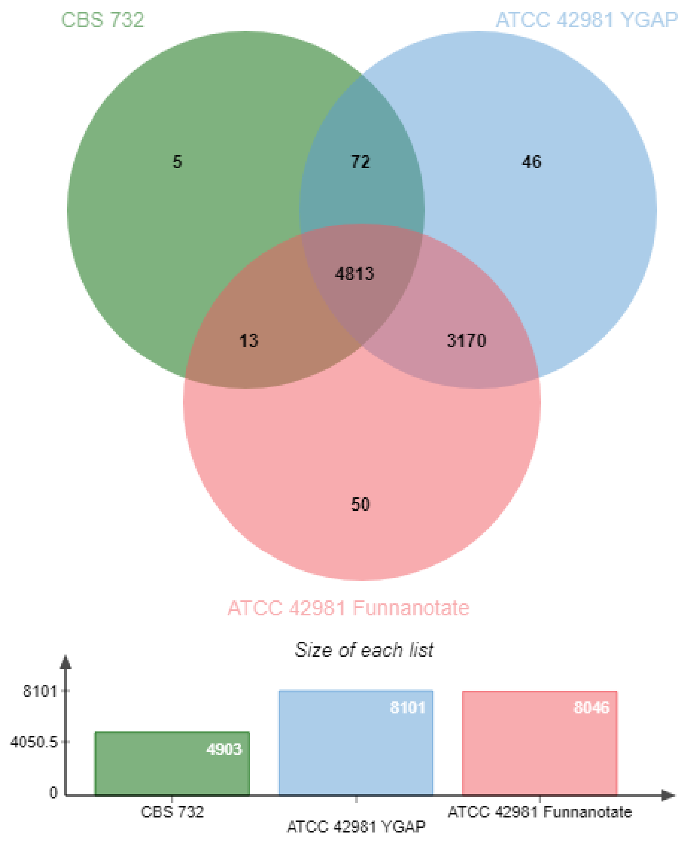


Figure 3. Top 20 slimmed biological process GO. OrthoVenn assigned the GOSlim terms for biological process category to the corresponding orthologous cluster. Orthologous cluster annotation was made for the haploid reference type-strain *Z. rouxii* CBS 732^{T}, ATCC 42981 genome annotated with YGAP (Proux-Wèira, 2012) and for ATCC 42981 genome annotated with Funannotate pipeline (Palmer, 2016).

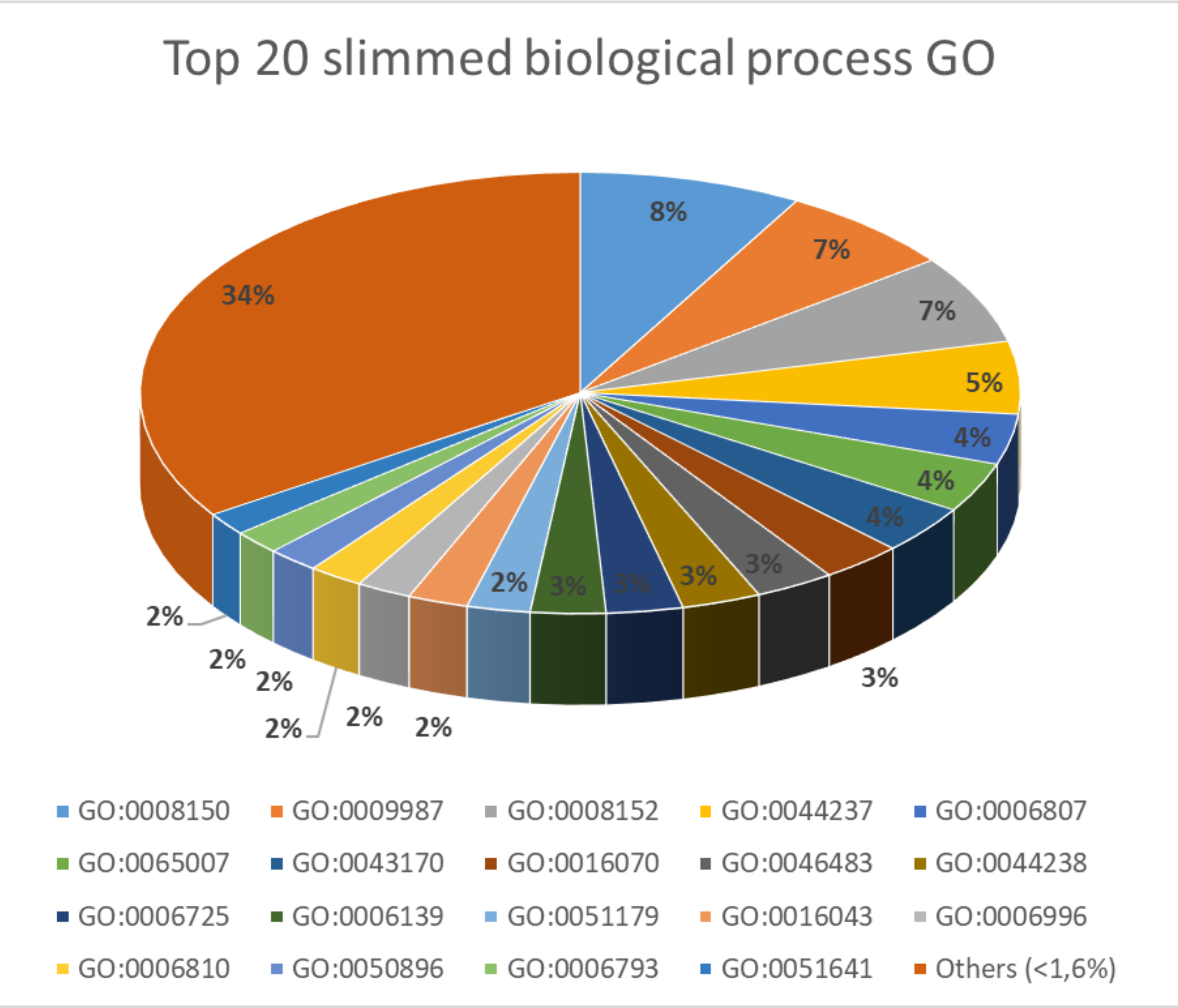


Figure 4. Top 20 slimmed cellular component GO. OrthoVenn assigned the GOSlim terms for the cellular component category to the corresponding orthologous cluster. Orthologous cluster annotation was made for the haploid reference type-strain *Z. rouxii* CBS 732^{T}, ATCC 42981 genome annotated with YGAP (Proux-Wèira, 2012) and for ATCC 42981 genome annotated with Funannotate pipeline (Palmer, 2016).

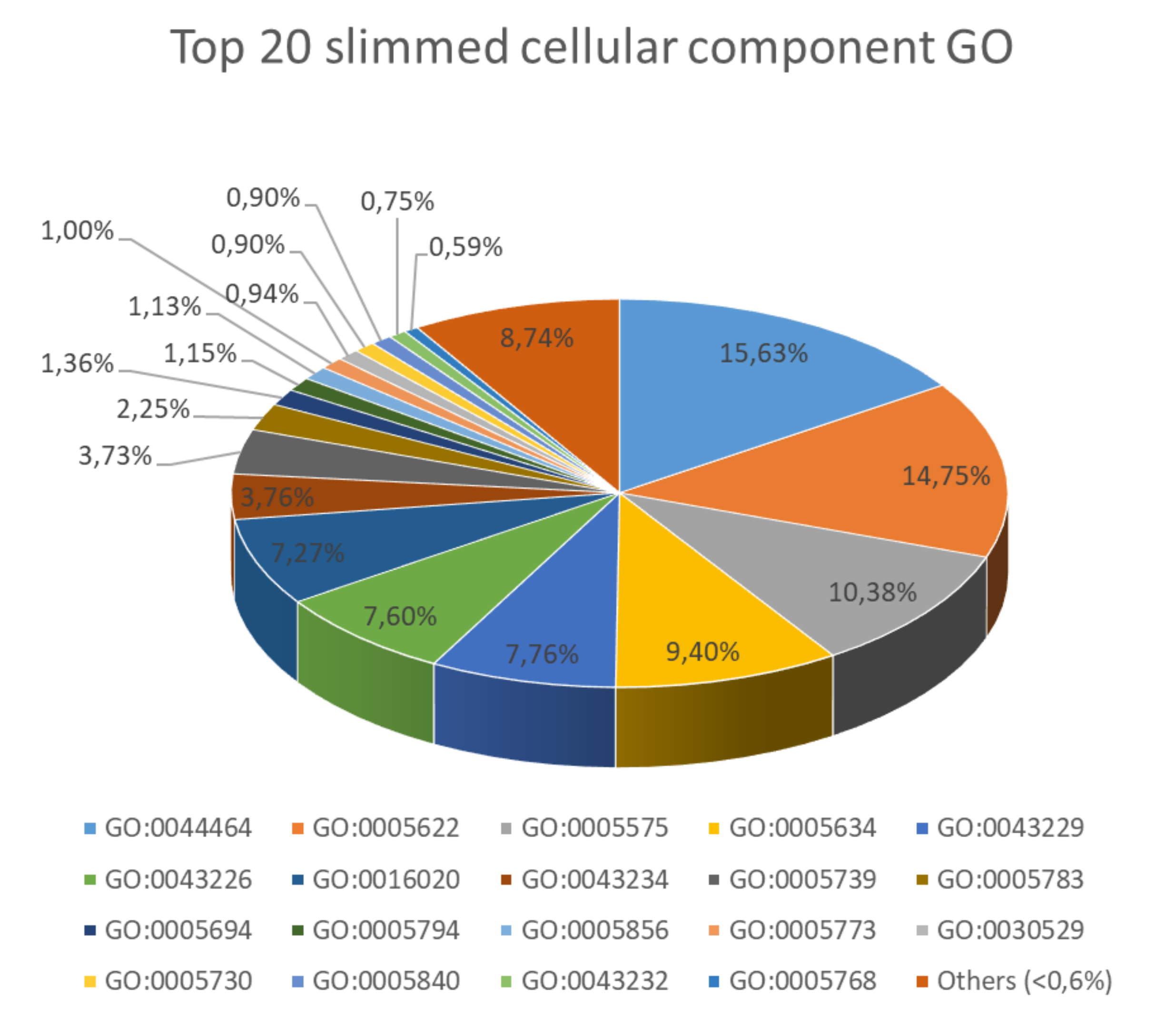
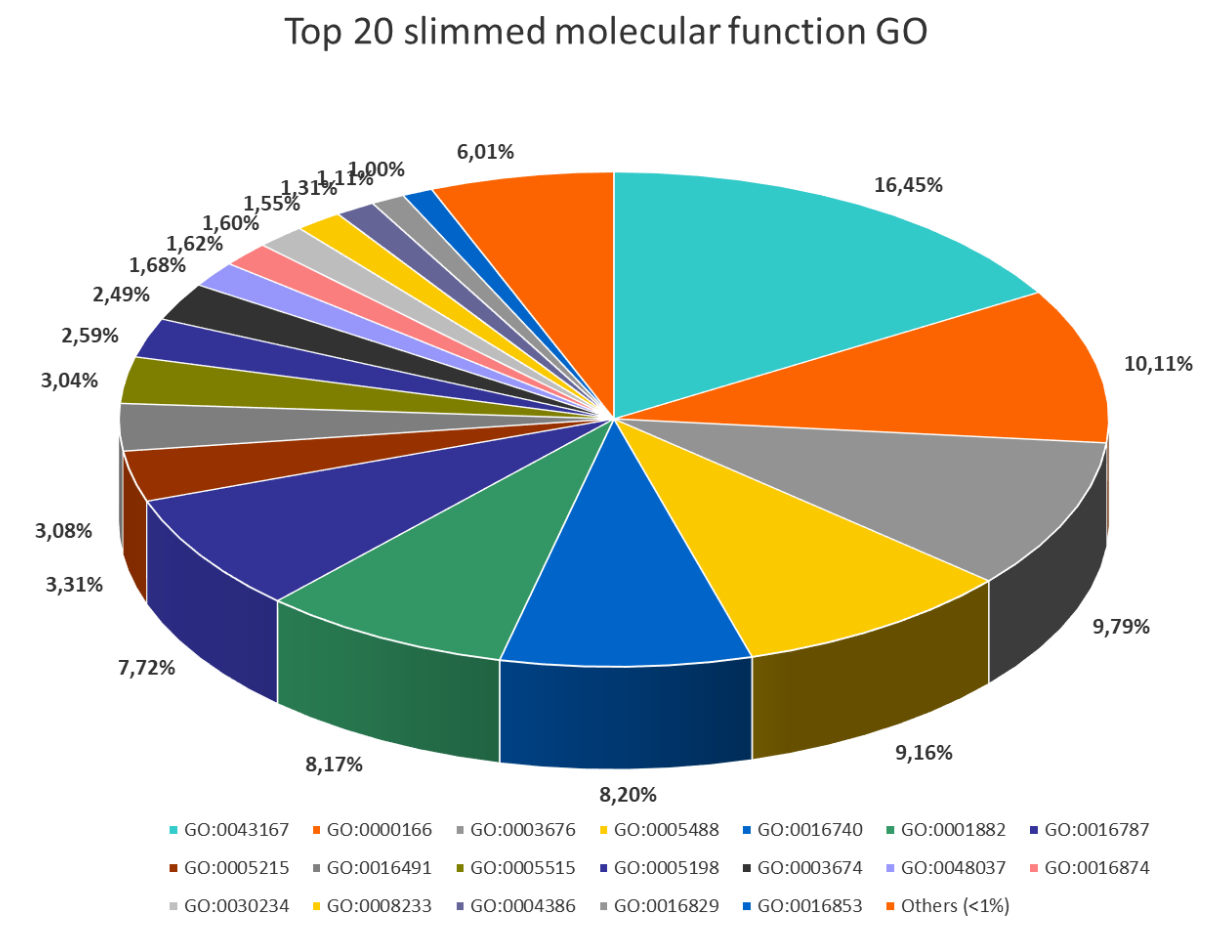


Figure 5. Top 20 slimmed molecular function GO. OrthoVenn assigned the GOSlim terms for molecular function category to the corresponding orthologous cluster. Orthologous cluster annotation was made for the haploid reference type-strain *Z. rouxii* CBS 732^{T}, ATCC 42981 genome annotated with YGAP (Proux-Wèira, 2012) and for ATCC 42981 genome annotated with Funannotate pipeline (Palmer, 2016).

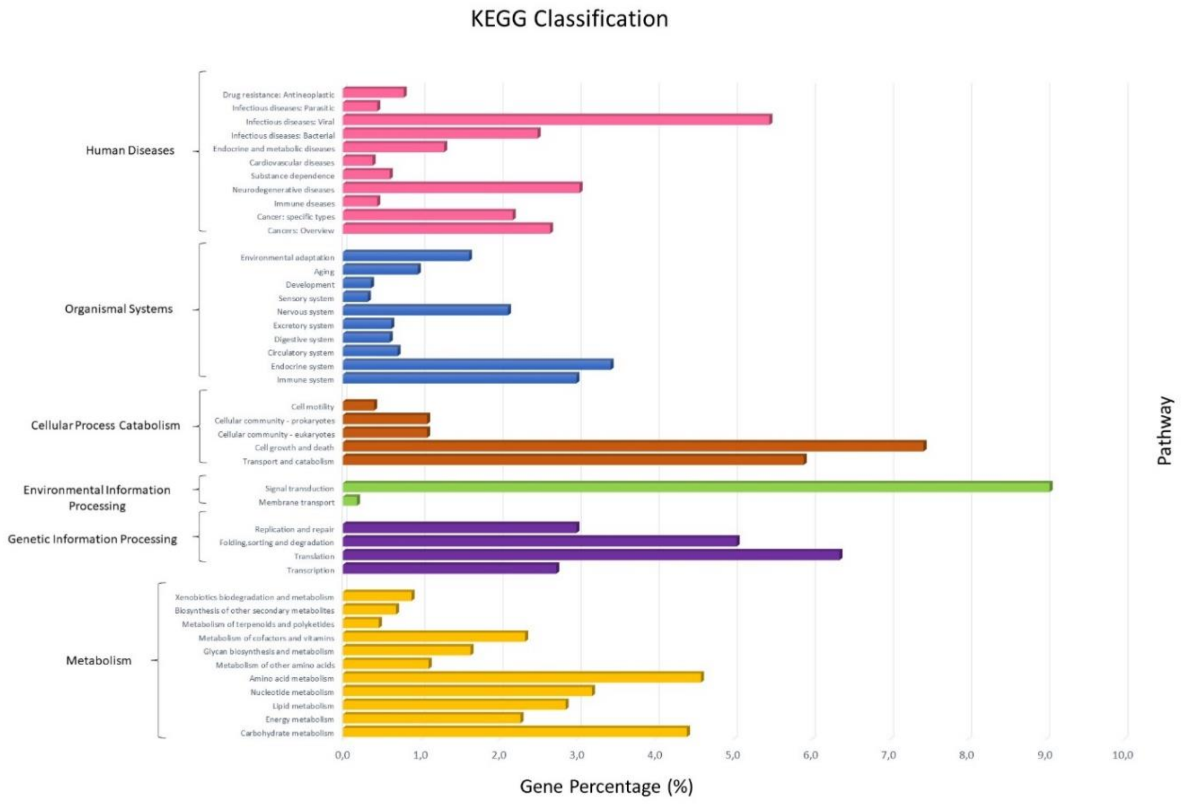


KEGG pathway mapping

The annotated sequences coding for proteins by YGAP pipeline were mapped on reference canonical pathways contained in KEGG database. Out of 10,821 putative proteins, 6,364 (58.8 %) were annotated on putative *Z. rouxii* CBS 732^{T} orthologs and 4,301 (39.7%) were assigned to KO terms.

Among the main functional categories (excluding human diseases), KEGG Orthology analysis showed that the most represented one was metabolism category (1,223), followed by genetic information processing (859) (**Fig. 6**). In metabolism category, the amino acid metabolism associated KO terms were prevalent, followed by carbohydrate (221) and lipid metabolisms (143).

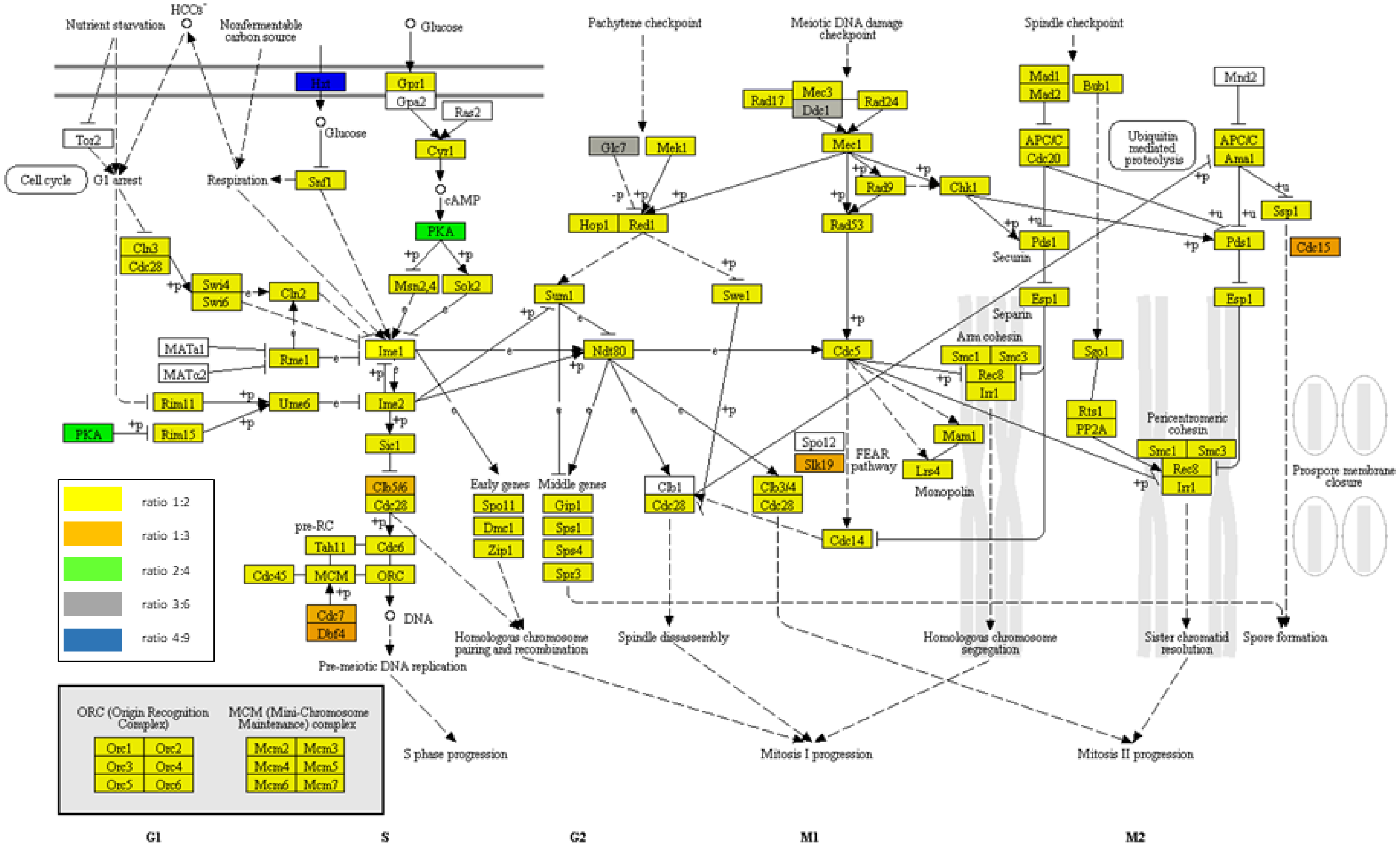
Figure 6. KEGG metabolism pathway categories assigned to ATCC 42981 YGAP annotated gene products.



We focused on the analysis of yeast meiosis pathway and all the 79 KO terms previously mapped in *Z. rouxii* CBS 732^T were successfully mapped also in ATCC 42981 (**Fig. 7**). These 79 KO terms did not include *MAT* loci and *IME1*. Therefore, we manually included the Ime1 putative protein (K12764) that was not mapped in CBS 732^T meiosis pathway. KO terms were assigned to a total of 205 protein sequences and the majority of them was in a 1:2 CBS 732^T: ATCC 42981 gene copy ratio.

In this case, protein identity ranged from 72.4% to 100.0%. A few CBS 732^T gene products had more than two corresponding orthologs in ATCC 42981. In particular, Cdc15, Slk19, Cdc7 and Dbf4 had 1:3 gene copy ratio; PKA had 2:4 gene copy ratio, Glc7and Dcd1 had 3:6 gene copy ratio and Hxt had 4:9 gene copy ratio. In this case, protein identity ranged from 79.0% to 100.0% (**Fig. 7**).

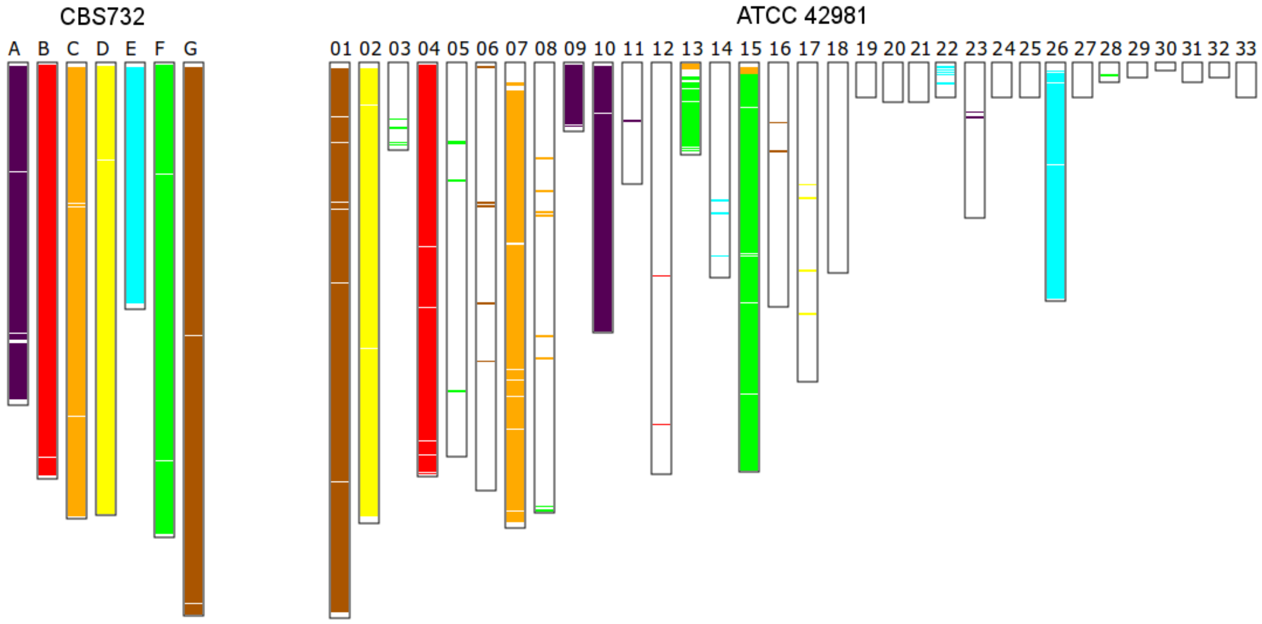
Figure 7. Graphical diagram representing the KEGG meiosis pathway in yeast. Coloured boxes represent putative gene products in ATCC 42981 annotated in KEGG pathway. Different colours indicate CBS 732^T : ATCC 42981 different gene copy ratios.



Syntheny map reconstruction

SynChro exploited the RBH identification to map the ATCC 42981 scaffolds corresponding to the T subgenome to CBS 732^T chromosomes (from A to G). Both ATCC 42981 annotation datasets from YGAP and Funannotate pipelines resulted in the same chromosomal painting representation. For brevity, **Figure 8** only shows the Funannotate scaffolds. Ten scaffolds, out of a total of 33, were assigned to T subgenome. Among them, four scaffolds (numbered 01, 02, 04 and 07) covered almost the corresponding CBS 732^T whole chromosomes, while each of *Z. rouxii* chromosomes A, E and F were fragmented into two ATCC 42981 scaffolds. The almost complete reconstruction of splitted CBS 732^T chromosomes was achieved by manual inspection of ATCC 42981 scaffold reciprocal orientation.

Figure 8. Chromosomal painting representation of ATCC 42981 scaffolds mapped on reference *Z. rouxii* CBS 732^T chromosomes. CBS 732^T chromosomes are indicated with capital letters from A to G, while ATCC 42981 scaffolds are numbered from 01 to 33. Colours are in accordance to syntheny blocks determined by RBH between the two proteome datasets.



The syntheny map showed that inversions affected the ends of some scaffolds, due to putative chromosomal rearrangements and/or to miss-scaffolding in ATCC 42981 (**Fig. 9**).

In particular, three inversion events are evident at the 5' end of scaffold 07 (corresponding to chromosome C), two at the 3' end of scaffold 02 (corresponding to chromosome D), two at the 5' end of scaffold 15 (corresponding to chromosome F), two at the 3' end of scaffold 13 (corresponding to chromosome F), and finally one inversion at the 3' end of scaffold 01 (corresponding to chromosome G). All the inversions involved non RBH linked orthologs.

ATCC 42981 homeomologs links representation showed that most of the genome assembly was duplicated, suggesting a hybrid origin (**Fig. 10**). Most of the ATCC 42981 scaffolds showed collinearity with at least one scaffold that SynChro mapped on CBS 732^{T} chromosomes (T subgenome), that are coloured in **Figure 10**. Only short sections at the end of scaffolds 06 and 16 (P subgenome) displayed collinearity with scaffolds 09 and 10 (T subgenome), suggesting miss-assemblies. ATCC 42981 shortest scaffolds (≤ 141.689 bp) lacked of or had a few homeologous (scaffolds 19, 20, 21, 24, 25, 27, 28, 29, 30, 31, 32 and 33), except of scaffold 22.

Figure 9. Detailed syntheny map among syntheny blocks with all homology relationships and syntheny breakpoints between CBS 732^T chromosomes (C, D, F and G) and ATCC 42981 scaffolds. In green are represented links between two RBH genes and in red between orthologs non RBH linked.

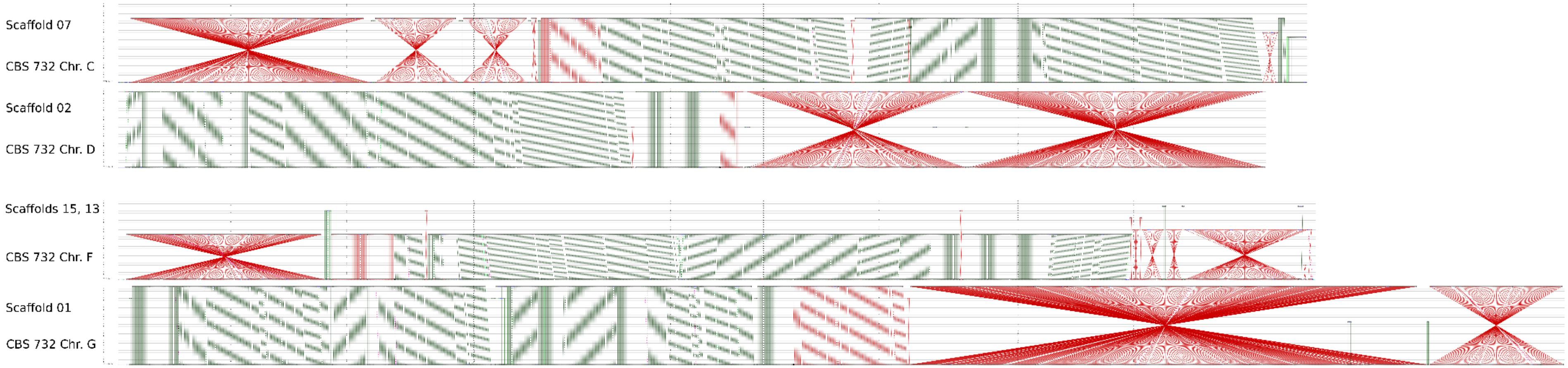
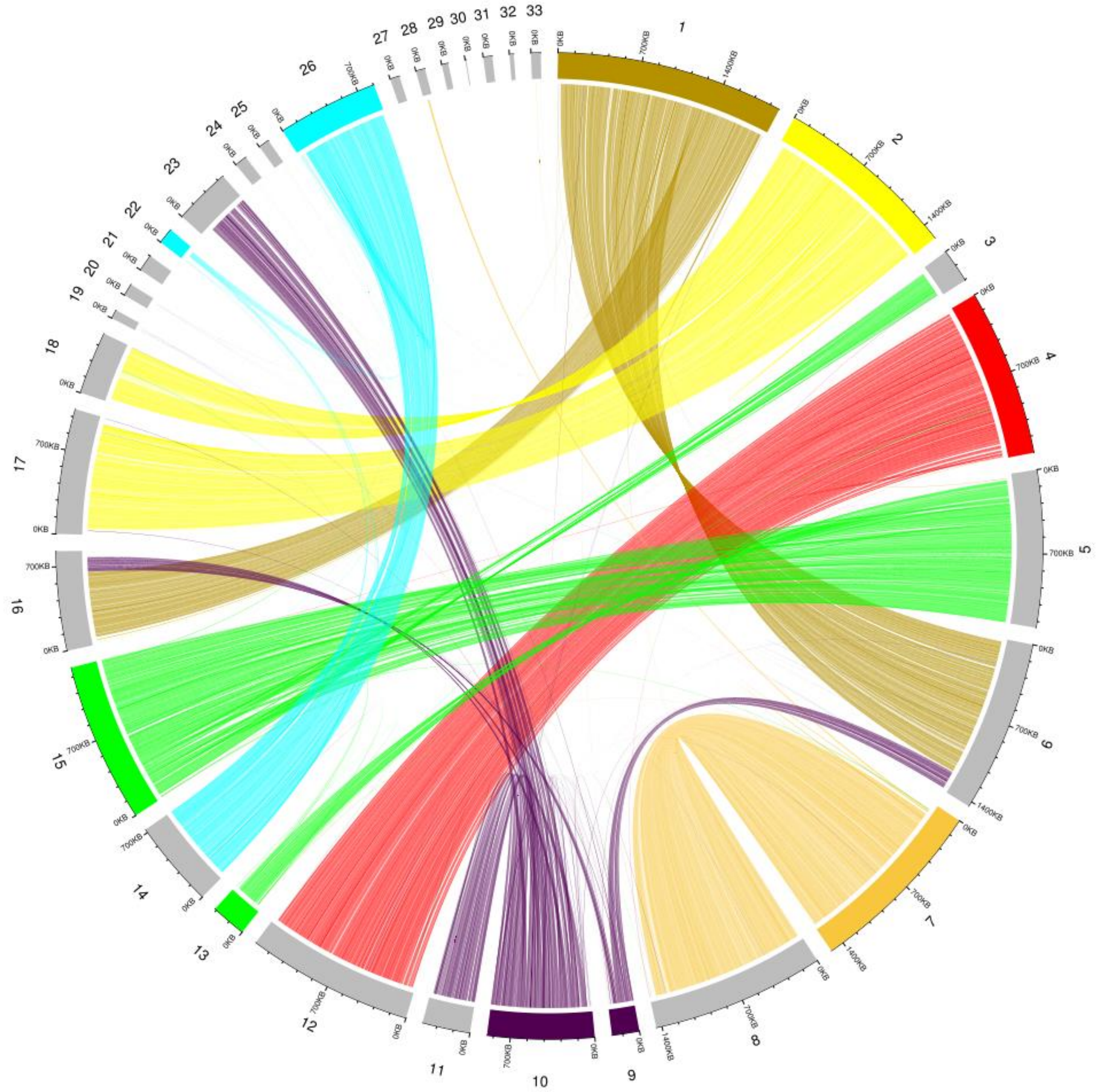


Figure 10. Circos plot of relationships among the *Z. rouxii* ATCC 42981 chromosomes. In the outer arc, coloured segments represent scaffolds attributed to T subgenome from CBS 732^T by SynChro. Arcs in the centre of diagram link homeomologs.



Discussion

For the allodiploid *Z. rouxii* ATCC 42981 genome annotation, we tested two alternative annotation pipelines mainly based on homology and syntheny conservation (YGAP) and *ab initio* and evidence-based prediction (Funannotate). Both methods detected an equivalent number of genes: 11,031 and 11,117, respectively. The protein coding ORFs annotated by YGAP represented the 69.4% of ATCC 42981 whole genome sequence. Protein coding genes provided with introns represented a small fraction (0.95%). On the contrary, Funannotate produced a more fragmented gene structures: genes provided with introns accounted for 14.5% with a genome coverage of 74.0%.

The sensitivity of the two pipelines in detecting the CBS 732^T orthologs by cluster analysis (see below) was similar. YGAP annotated gene products have been included in Funannotate pipeline as evidence-based protein dataset. Therefore, the contribute of *ab initio* gene prediction in Funannotate pipeline decreased the gene prediction quality, this could be due to a suboptimal self and data set training of Genmark-ES and Augustus using Saccharomycetales database.

The analysis of protein identity of homeomologs gene products revealed a limited divergence between T and P parental subgenomes. The *Z. rouxii* genome complement (T subgenome) and its syntheny information are included as training data set in YGAP. The close phylogenetic distance between T and P parental subgenomes probably makes YGAP pipeline, alone, enough accurate in gene prediction, while the introduction of *ab initio* tools trained with a more divergent dataset increased the background noise without any sensitivity improvement.

Cluster analysis of gene products by OrthoVenn showed that these two pipelines shared around 72% of the predicted genes. Among them, about the 44% represented putative homologs in the haploid reference CBS 732^T genome. Since the clustering was performed by setting the highest stringency *E*-value (10^{-5}) and inflation value (1.5), the remaining fraction probably belongs to the non *Z. rouxii* parental complement in agreement with the hybrid nature of ATCC 42981 genome, or, alternatively, represents a small fraction of inaccurate gene annotations. Most of the shared predicted genes received a GO functional annotation. Among the top 20 biological process terms, two of the most represented were the nitrogen compound (GO:0006807) and the heterocycle (GO:0046483) metabolic processes (**Fig. 3**). Both of them are involved in the production of flavor compounds in fermented foods, such as soy sauce and miso, by Valine, Leucine, Isoleucine and Alanine degradation.

The first degradation pathway (KEGG ID:00280) exploites Valine, Leucine and Isoleucine as precursors to generate secondary flavor compounds, such as isoamyl alcohol, isobutyl alcohol, isobutanol, 2-methylpropanol, 2-methylbutanol and 2-phenylethanol (Jansen et al., 2003; Van der Sluis et al., 2001). In *Z. rouxii* CBS 732^T, this metabolic pathway includes multicopy genes: two copies of *ARO8* (ZYRO0C06028g), *ARO9* (ZYRO0C06028g) and *ADH1* (ZYRO0B05940g) genes; three copies of *THI3* (ZYRO0A08426g), *ARO10* (ZYRO0F01606g), *PDC1* (ZYRO0A08426g), *PDC5* (ZYRO0A08426g) and *PDC6* (ZYRO0A08426g). In CBS 732^T, these genes are located on different chromosomes (Chr. A, B, C, E, F and G). We also found single copy genes, such as *BAT1* (ZYRO0G00396g) and *SFA1* (ZYRO0F11704g). YGAP annotation pipeline failed in retrieving *S. cerevisiae* ortholog to *ADH4* aldeide dehydrogenase. Indeed, also in CBS 732^T, as in other pre-WGD species included in YGOB database, the function of this enzyme is shared with *ADH1* and *SFA1* genes.

The second degradation pathway involves Alanaine as precursor to generate 4-Hydroxy-2(or 5)-ethyl-5(or 2)-methyl-3(2/f)-furanone (HEMF), which is a key antioxidant and flavour compound in soy sauce (Sasaki et al., 1991, 1996; Ohata et al., 2007; Uehara et al., 2017). In CBS 732^T, we found only one copy of the gene coding for the secondary metabolic compound HEMF (ZYRO0G05984g), to which no KO was assigned.

Overall, in accordance with the hybrid nature of ATCC 42981 allodiploid genome, we annotated two orthologs for each gene. The protein identity range with CBS 732^T for each ATCC 42981 ortholog pair spanned on average from 82.0 to 100.0%. Accordingly, syntheny mapping with SynChro showed that for each ortholog pairs, T subgenome harboured one copy and P subgenome contained the second one.

Among the top 20 cellular components, the three GO terms associated to the nucleus (GO:0005634), organelle (GO:0043226) and membrane (GO:0016020) are equally distributed, suggesting that the genome annotation was not biased towards a particular cytological district (**Fig. 4**).

Ion binding GO term (GO:0043167) resulted the most represented one in the top 20 molecular functions (**Fig. 5**). This finding is consistent with the genome enrichment in ion membrane transporter coding genes due to the high ATCC 42981 tolerance to solutes, weak organic acids and xenobiotics and to its peculiar adhesive phenotype. In particular, ATCC 42981 multicopy genes found orthologs for ABC transporters, plasma membrane multidrug transporters and flocculation in CBS 732^T (XP_002497583, XP_002496702, XP_002494950, XP_002497211). A similar enrichment in metal cation transporter encoding genes was observed in the extremely halotolerant black yeast *Hortaea werneckii* (Lenassi et al., 2013).

Most of gene products annotated by YGAP and Funannotate pipelines were mapped on KEGG pathways: 63.1% and 62.0%, respectively. KO terms are equally distributed among the main KEGG class pathways (**Fig. 6**). Concerning metabolism class, the most represented pathways were carbohydrate and amino acid metabolisms. A high number of KO terms were attributed to translation among the genetic information class. Considering the remaining pathways, signal transduction, cell growth, death, transport and catabolism were particularly enriched in KO terms.

Another main trait of ATCC 42981 is the inability to produce gametes. Therefore, we focused on KEGG meiosis pathway gene members in order to unravel possible molecular causes of this sterility. KO annotation showed that the meiosis pathway in ATCC 42981 was complete, since all the CBS 732^T orthologs that mapped on this pathway were present. Each KO term was attributed to, at least, two genes: one from T and the other from P subgenome. Nine KO terms were mapped on multicopy ATCC 42981 genes (**Fig. 7**). Four of them were indeed multicopy also in CBS 732^T with a gene copy ratio between orthologs CBS 732^T:ATCC 42981 2:4, 3:6 and 4:9. These findings suggest that probably gene loss is not responsible for ATCC 42981 sterility, since all gene members of meiosis pathway are present, at least, in two copies belonging to T and P subgenomes. It cannot be excluded that other molecular mechanisms, such as gene products or cytogenetic incompatibilities could account for post-zygotic barriers.

Reconstruction of syntheny blocks by RBH using SynChro identified scaffolds that mapped on T subgenome. Both proteome data set from YGAP and Funannotate pipelines gave the same results. Full representation of CBS 732^T complement was achieved: chromosomes B, C, D, E and G were covered by a single ATCC 42981 genome scaffold, while chromosomes A and F mapped mainly on two scaffolds (**Fig. 8**). However, in these cases the detailed analysis of syntheny blocks allowed the relative scaffolds orientation on reference chromosomes and the localization of the breakpoint regions. For chromosome A, syntheny of ATCC 42981 scaffolds 09 and 10 was interrupted by genes without RBH orthologs, which mapped on scaffold 11. For chromosome F, the loss of syntheny between scaffolds 13 and 15 was due to inverted RBH orthologs. The left ATCC 42981 twenty four scaffolds putatively belong to P subgenome or alternatively are too short (<10,000 bp) to provide enough syntheny block length and/or harboured prevalently non protein coding genes.

P subgenome reconstruction was based on the identification of the homeomologs gene pairs through YGAP annotation: two ATCC 42981 gene products were considered homeomologs when mapped on the same CBS 732^T ortholog and were harboured by different scaffolds. Most of ATCC 42981 homeomologs, which mapped on T subgenome, formed collinear pairs with genes on scaffolds not assigned to *Z. rouxii* by reciprocal blast (**Fig. 9**). Moreover, sections of *Z. rouxii*-like scaffolds are collinear with sections of P subgenome scaffolds. Relaying on homeomolog pairs, ATCC 42981 eleven scaffolds were attributed to P complement, while the assignment of the left shorter scaffolds remained uncertain. The presence of false negative homeomologs was checked by manually blasting the gene products harboured by these shorter scaffolds against ATCC 42981 and CBS 732^{T} proteomes. As a result, most of these ORFs displayed partial aa identity and query coverage (on average >74%) with CBS 732^{T} putative orthologs and ATCC 42981 putative homeomologs. However, the subjects' coverage was often less than 50% and a few of these proteins corresponded to a CBS 732^{T} multicopy genes. As a whole, these findings suggested that the homeomologs detection, based on YGAP and cluster analysis, was sub-optimal owing to many fragmentary genes and sequencing artefacts produced by the hybrid assembling approach and/or by the presence of relics of a further divergent N (new) parental subgenome.

Concluding remarks

- YGAP pipeline, based on post-WGD gene homology and synteny conservation, was more accurate than Funannotate pipeline in ATCC 42981 gene prediction and genome annotation.
- Orthologs cluster analysis, syntheny mapping and homeomologs identification confirmed the hybrid origin of ATCC 42981 genome, which includes at least two parental haplotypes.
- The analysis of protein identity among putative orthologs with reference *Z. rouxii* type-strain CBS 732^{T} ranged was about 80-100%, suggesting that one of the two parentals was *Z. rouxii*-like (named T subgenome) and the other one was phylogenetically close related (named P subgenome).
- ATCC 42981 shorter scaffolds were enriched of ORFs coding for proteins with a higher level of divergence than that between T and P subgenome gene products, suggesting the occurrence of recursive hybridization events involving more than two parentals.
- Most of ATCC 42981 genome was duplicated in collinear subgenomes, supporting that the hybridization event occurred recently with limited chromosomal rearrangements or gene loss.

- In accordance to the high ATCC 42981 tolerance to environmental stresses, functional annotation revealed a genome enrichment in genes belonging to ion binding, flocculation, aromatic and heterocycle compounds GO terms.
- Meiosis pathway analysis highlighted that parental subgenomes equally contributed to the sex reproduction gene repertoire, since all pathway members were duplicated and, in a few cases, extra copies were also detected. This finding supports that ATCC 42981 hybrid sterility is not related to gene loss, but might be due to functional network incompatibility.

PART III SUMMARISING DISCUSSION

Summarising Discussion

This thesis deals with the non-conventional yeasts belonging to the *Z. rouxii* species complex. Inhabiting stressful environmental and food-related niches, they represent promising microorganisms for future development in bio-economy and industrial biotechnology and encompass relevant species for fermentative bioprocesses and foodstuff spoilage.

We decided to focus the attention on non-conventional yeasts, since they are attracting increasing attention in basic research and biotechnological applications. Due to their exceptional metabolic pathways, they have been used in various biotechnological processes for producing foods or food additives, drugs or a variety of biochemicals.

Among *Z. rouxii* strains, we chose ATCC 42981 allodiploid strain as a case of study, since it shows interesting industrial features, such as great fermentative performance even at high salt and sugar concentrations. However, despite these properties, ATCC 42981 cannot be used in strain development due to its sterility, which precludes genetic analysis and prevents the application of breeding programs.

Since the breeding system and lifestyle are crucial in shaping the biodiversity of yeasts, we investigated the role of mating-type regions (*MAT*) and the *MAT*-encoded transcriptional factors in governing downstream effectors involved in ATCC 42981 cell-type development, such as mating, meiosis commitment and adhesive phenotype.

To investigate the genetic and molecular basis of hybrid sterility we developed synthetic biology tools and genetically engineered approaches that will be useful for future analysis aimed to restore complex phenotypic traits in hybrid and sterile yeasts, leading to improved industrial phenotypes.

Every chapter of this thesis has its own discussion section. Here I will summarize some broader implications of my research, which can be grouped into three main categories: methodology, genomics and functional study.

- **Methodology**: The methodological results span two information levels. Firstly, synthetic tools and protocols were developed and validated to genetically manipulate *Z. rouxii* allodiploid and prototrophic strains. These advances circumvent the paucity of selectable markers, which, coupled with diploidy and variable transformation efficiency, makes *Z. rouxii* knockout of a single or a few genes a considerable task. These improvements also promise to expand our capabilities in the biotechnological exploitation of this non-conventional species, which displays a variety of interesting but yet poorly explored industrial traits. Secondly, we provided a comprehensive workflow, starting from HMW gDNA extraction to hybrid *de novo* assembly, which circumvents the numerous challenges in sequencing the highly heterozygous genomes. ATCC 42981 sequencing and assembly strategy relied on MinION long reads and MiSeq short reads and allowed the assembling of hybrid genome into a relatively reduced number of scaffolds in a reasonable amount of time and using limited hardware resources.

- **Genomics**: ATCC 42981 genome complements were successfully distinguished in a Z. *rouxii*-like T subgenome and a divergent P subgenome through synteny mapping with Z. *rouxii* reference haploid type-strain CBS 732^{T}. In addition, traces of a third haplotype were recovered, suggesting an evolutionary model based on recursive hybridization. In accordance to the peculiar ATCC 42981 adaptation to salt stresses, functional annotation revealed a genome enrichment in genes belonging to ion binding and flocculation. Genes coding for aromatic and heterocycle compounds were also found enriched in agreement with the ATCC 42981 capability to produce aa-derived secondary metabolites. Comparative genomics further revealed that most of the meiotic genes were duplicated in ATCC 42981, indicating that in this strain post-zygotic barriers that hamper sporulation could be related to transcriptional and/or post translational incompatibilities between T and P subgenomes, rather than a mere gene loss.

- **Functional study**: *MTL* loci turned out to be reasonable candidates for transcriptional and/or post translational incompatibilities, which regulate cell identity and meiosis commitment in the model yeast *S. cerevisiae*. Preliminary analysis of *MTL* loci in two stocks of the reference strain *Z. rouxii* CBS 732^{T}, demonstrated that these loci are recombination hot spots. Cells undergone mating-type switching displayed different *MAT***a**2 copy variants, leading to intra-strain genetic and phenotypic variability. In ATCC 42981 *MTL* repertoire reconstruction was particularly challenging and required a complex strategy that integrated two different genome assemblies with long PCR *in vitro* validation. This strategy succeeded in demonstrating that ATCC 42981 in-house culture retains a *MAT***a**/*MAT*α genotype in contrast to the sister Japanese stock JCM22060, which only displays *MAT*α. These differences point out that a chimeric **a**1-α2 heterodimer accounts for ATCC 42981 hybrid sterility, while a functional α1 activator of αsgs accounts for the JCM22060 ability to mate and regain fertility. In attempt to rescue ATCC 42981 ability to mate and sporulate, we deleted *MAT*α locus in ATCC 42981 and tested whether it behaved as *MAT***a** haploid. However, the resulting *MAT***a**/- hemizygous mutants did not rescue fertility, as incomplete silencing of chimeric *HML*α cassette masked the loss of heterozygosity at the *MAT* locus induced by gene deletion. Overall, these findings highlight the cutting-edge and yet unexplored role of *HMR*/*HML* silencing in establishing or altering allodiploid cell identity.

CONCLUSION

In this Thesis, we provide a comprehensive picture that enables to understand how the interplay of at least two phylogenetically divergent haplotypes in ATCC 42981 genome shapes the *MTL* loci genomic structure, unlocks the *HML* silencing mechanism and, consequently, modifies cell identify circuits, leading to the inability to produce gametes.

Hybridization is an evolutionary force that shapes genome structure, triggers speciation and modifies regulatory circuits, resulting in phenotypic novelties driving adaptation. The findings arose from this work represent novel achievements in the field of genomic and functional aftermaths of allodiploidization in a pre-WGD species, which stands on the crossroad where different and relevant evolutionary events take their way. Our results improve the understanding of post-zygotic barriers as relevant factors in determining sterility and provide tools and fruitful suggestions for future studies on the mechanisms involved in cell identity determination and gene silencing in chimeric genomes, where differently evolved haplotypes have to cooperate.

Overall, this work provides relevant and groundbreaking advances in the genetic investigation of the non-conventional yeasts, a field of study still poorly developed and whose corresponding data are few and dispersed.

The era of the non-conventional yeasts has just begun.

APPENDIX: LIST OF PUBLICATIONS

Appendix: List of publications

1. Cassanelli S., **Bizzarri M**., and Solieri L. (2016). Recent advances in understanding yeast genetics of sex determination*. Editorial Fungal Genomics & Biology*. 6:e122. doi:10.4172/2165-8056.1000e122
2. **Bizzarri, M**., Giudici, P., Cassanelli, S., and Solieri, L. (2016). Chimeric sex-determining chromosomal regions and dysregulation of cell-type identity in a sterile *Zygosaccharomyces* allodiploid yeast. *Plos One.* 11(4), e0152558. doi:10.1371/journal.pone.0152558
3. **Bizzarri, M**., Cassanelli, S., and Solieri, L. (2017). Mating-type switching in CBS 732^T derived subcultures unveils potential genetic and phenotypic novelties in haploid *Zygosaccharomyces rouxii. FEMS Microbiology Letters.* 365(2). doi:10.1093/femsle/fnx263
4. **Bizzarri, M**., Cassanelli, S., Pryszcz, L. P., Gawor, J., Gromadka, R., and Solieri, L. (2018). Draft genome sequences of the highly halotolerant strain *Zygosaccharomyces rouxii* ATCC 42981 and the novel allodiploid strain *Zygosaccharomyces sapae* ATB301^T obtained using the MinION platform. *Microbiology Resource Announcements,* 7(4). doi:10.1128/mra.00874-18
5. **Bizzarri M**., Dušková M., Sychrovà H., Cassanelli S., and Solieri L. (In preparation). Development of plasmids harbouring antibiotic resistance selection markers and Cre recombinase for genetic engineering of non-conventional *Zygosaccharomyces rouxii* yeasts.
6. **Bizzarri M**., Cassanelli S., Bartolini L., Pryszcz L. P., Dušková M., Sychrová H., and Solieri L. (2018)**.** Interplay of chimeric mating-type loci reconstructed by MinION sequencing determines yeast sexual incompetence. Paper accepted for publication in *Frontiers in Genetics*.

SUPPLEMENTARY MATERIALS

Supplementary materials

Chapter 4

Table S1. **Yeast strains used in this work.** Abbreviations: UMCC, Unimore Microbial Culture Collection.

Strain	Stocks	Code	Species	Mating-type/thallism	Spore	Ploidy Ratio	Citation
CBS 732^T	UMCC, Reggio Emilia, Itay	CBS 732_R	*Z. rouxii*	*MAT*α/homothallic	-	1.3	Sacchetti (1932)
	Institute of Physiology, Academy of Sciences of the Czech Republic	CBS 732_P*		*MAT*α/homothallic			
CBS 4837	UMCC, Reggio Emilia, Itay	CBS 4837	Mosaic lineage	*MAT***a**/heterothallic	+	1.96	Mori and Onishi (1967); James et al., (2005)
CBS 4838	UMCC, Reggio Emilia, Itay	CBS 4838	Mosaic lineage	*MAT*α/heterothallic	+	1.90	Mori and Onishi (1967); Solieri et al., (2008)

*a gift from H. Sychrovà

Table S2. **Composition of the media used in this study.**

Media			
Usage	**Abbreviation**	**Composition**	**Growth conditions (temperature; time)**
Routine growth	YPD	1% w/v Yeast Extract, 1% w/v Peptone, 2% w/v Dextrose	28°C; 24-48 h
RNA extraction	YNB5%G	1% w/v Yeast Extract, 5% w/v Dextrose, 6.7 g/l Yeast Nitrogen Base	28°C; 24-48 h
	YNB5%G NaCl	1% w/v Yeast Extract, 5% w/v Dextrose, 6.7 g/l Yeast Nitrogen Base, 116.8 g/l NaCl	28°C; 24-48 h
Morphology and fertility assay	MEA	5% w/v Malt Extract, 1.5% w/v Agar	27°C; 14 d
	6%NaCl-MEA	5% w/v Malt Extract, 1.5% w/v Agar, 60 g/l NaCl	27°C; 14 d
	YM	0,3%w/v Yeast Extract, 0,5% w/v Peptone, 0,3% w/v Malt Extract, 1% w/v Dextrose, 1.5% w/v Agar	27°C; 14 d

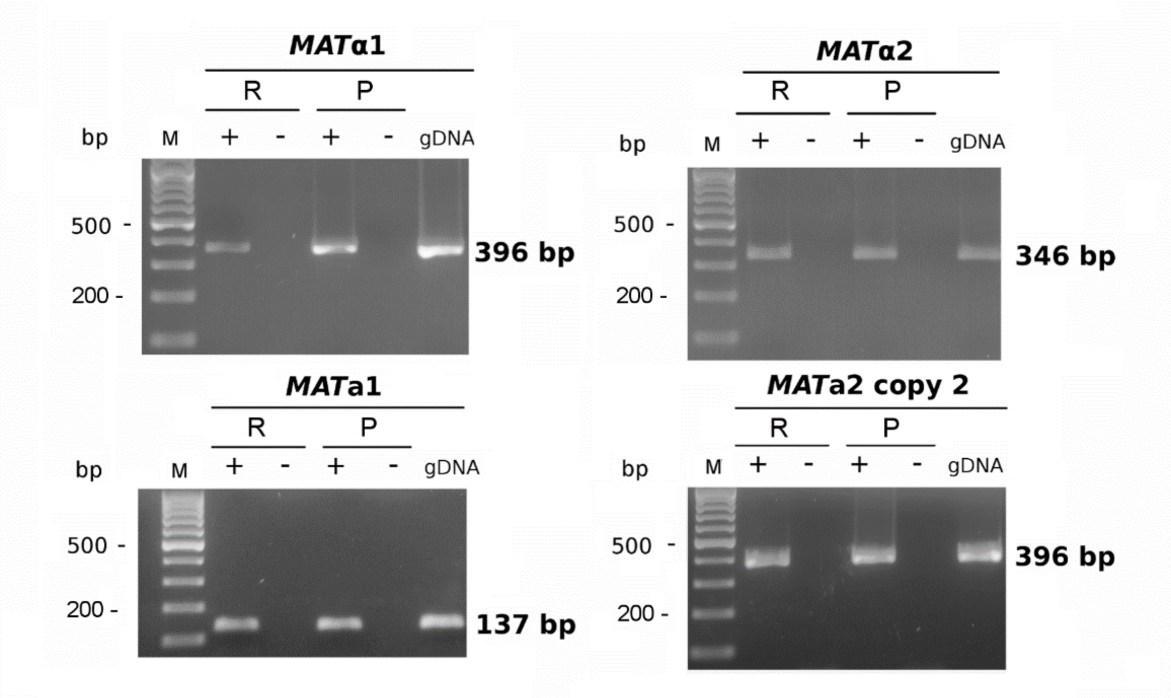
Table S3. List of primers used in this study.

Usage	Target	Primer	Sequence (5'->3')	Description
***CHA1-MAT-SLA2* locus characterization** 5' PCR-walking				
	*MAT*α2	CHA1-F2	ACTCTTGACTACGTGCAGAAATATG	forward primer specific for *CHA1* gene in CBS 732^{T} genome
		301_MATα2CP1R1	CTTGGTAATACAGGTAAAGAGGGT	*MAT*α2 specific reverse primer
	*MAT***a**2 copy 2	CHA1-F2	ACTCTTGACTACGTGCAGAAATATG	forward primer specific for *CHA1* gene in CBS 732^{T} genome
		301_MATA2R4	GAAACCAACAGAGTTGCAGAAATA	*MAT***a**2 copy 2 specific reverse primer
	MATα1	301MATα1α2-cp1F1	TTCCTTCACCGCCAGAGGTTC	*MAT*α1 specific forward primer
		SLA2-R1	GCATGTATTGTGTA ACCTGGGA	reverse primer for *SLA2* gene in CBS 732^{T} genome
	*MAT***a**1 copy 2	301_MATA1F3	GTAGCTTCCACAAGGTCTTCAAGG	*MAT***a**1 specific forward primer
		SLA2-R1	GCATGTATTGTGTA ACCTGGGA	reverse primer for *SLA2* gene in CBS 732^{T} genome
Colony PCR	*MAT*α2	301_MATA2F5cp2	GTAAACGGTAAAGCTGAAACGC	forward primer designed on the intergenic region between *CHA1* gene and *MAT*α2/**a**2 gene (at the 5' end of the X region)
		301_MATα2CP1R1	CTTGGTAATACAGGTAAAGAGGGT	*MAT*α2 specific reverse primer
	*MAT***a2** copy 2	301_MATA2F5cp2	GTAAACGGTAAAGCTGAAACGC	forward primer designed on the intergenic region between *CHA1* gene and *MAT*α2/**a**2 gene (at the 5' end of the X region)
		301_MATA2R4	GAAACCAACAGAGTTGCAGAAATA	*MAT***a**2 copy 2 specific reverse primer
	*MAT*α1	301_MATα1α2Cp1F1	TTCCTTCACCGCCAGAGGTTC	*MAT*α1 specific forward primer
		301_MATα1R6	ACCCACACTTCGGCTTGACAAA	*MAT*α1 specific reverse primer
Non quantitative RT-PCR	*MAT*α2	301_MATα2Cp1F10	TTGCAGAACTACTTTACTCGTTCA	*MAT*α2 specific forward primer
		301_MATα2CP1R1	CTTGGTAATACAGGTAAAGA GGGT	*MAT*α2 specific reverse primer
	*MAT***a**1 copy 2	301_MATA1F3	GTAGCTTCCACAAGGTCTTCAAGG	*MAT***a**1 copy 2 specific forward primer
		301_MATA1R1	GTCCTCTTTCTCTCAAATACACG	*MAT***a**1 copy 2 specific reverse primer
	*MAT***a**2 copy 2	301_MATA2F3	ATTGGCGATGAGCGAAGAAG	*MAT***a**2 copy 2 specific forward primer
		301_MATA2R4	GAAACCAACAGAGTTGCAGAAATA	*MAT***a**2 copy 2 specific reverse primer
	5' end *HO*	301_5'HOF1	CTACGTCGAGAGATCCATCATAG	*5' HO* specific forward primer
		301_verylikeHOR3	CGCGAATCTACCGGTACTATT	*5' HO* specific reverse primer
	3’ end *HO*	301_verylikeHOF6	ATGGTGAAAAGCAAATTCCATCA	*3' HO* specific forward primer
		301_very_likeHOR5	GGTGTACGACCGGCTAAATG	*3' HO* specific reverse primer

Table S4. **Results of *MAT* genotyping of CBS 732^{T} stock colonies.**

Strain stock	N° of colonies	*MAT* genotyping (%)		
		*MAT*a/α	*MAT*a	*MAT*α
CBS 732_R	36	58.3	39.0	2.8
CBS 732_P	50	36.0	2.0	62.0

Figure. S1 Expression pattern of CBS 732_R and CBS 732_P *MAT* genes. The figure shows amplified cDNAs obtained with *MAT* specific primers. RNA was extracted from unstressed cells harvested at stationary phase for CBS 732_R and CBS 732_P stocks. +/- indicate reverse transcription positive and negative controls. gDNA amplification was used as positive PCR control. Abbreviations: M, molecular weight marker; R, CBS 732_R stock; P, CBS 732_P stock.



Chapter 5

Table S1. List of primers used in the present study.

Usage	Target	Primer	Sequence (5'->3')	Description
***MTL*α loci characterization**				
	*MAT*α copy 1	301MATα2Cp1F10	TTGCAGAACTACTTTACTCGTTCA	copy1-specific forward primer built on *MAT*α2 codon stop (referred to as αCp1_P in Fig. S1)
		301MATα1R6	ACCCACACTTCGGCTTGACAAA	reverse primer built on *MAT*α1 (referred to as αCp1_P in Fig. S1)
	*MAT*α copy 2	301MATα1α2cp2F1	GTTTAGATGCCAGTGCTCTTCA	forward primer built on UTR of *MAT*α2 (referred to as αCp2_P in Fig. S1)
		301MATα1α2cp2R2	TGTCATCCACATTGAAATCATCTC	*MAT*α1 copy2-specific reverse primer (referred to as αCp2_P in Fig. S1)
		301MATα1α2Cp2F3	TTCCTTCACCTCCGGAGAACC	3' end *MAT*α1 copy 2-specific forward primer (referred to as αCp2_P in Fig. S1)
	*MAT*α copy 3	301MATα1rouF1	CCGCCGAAGAATTTACTTAGAG	*MAT*α1 copy 3-specific forward primer (referred to as αCp3_P in Fig. S1)
		301MATα1R6	ACCCACACTTCGGCTTGACAAA	reverse primer built on *MAT*α1 (referred to as αCp3_P in Fig. S1)
		301MATα1α2cp2F1	GTTTAGATGCCAGTGCTCTTCA	forward primer built on UTR of *MAT*α2 (referred to as αCp3_P in Fig. S1)
		301MATα2RouR1	TTAGGAGATAAAGGTAAGAATAGG	*MAT*α2 copy 3-specific forward primer (referred to as αCp3_P in Fig. S1)
***MAT*α1 ORF completion**				
	SLA2	A	CCAGTTAGTGTGTTATCGATAAGTC	reverse primer specific for *SLA2* gene in CBS 732^T genome
	SLA2_D	A_D	TTYGARTTYTAYGCNGAYTG	reverse degenerate primer targeting FEFYADC conserved amino acid sequence of *Z. rouxii* CBS 732^T *SLA2* gene
	ZYRO0C18392	B	TCTATTTCGTCCGTTTATCGTTGGT	reverse primer specific for locus ZYRO0C18392g in CBS 732^T genome

	ZYRO0F18634	C	TCAGTACCAGAAGTGGTCTTTGAAA	reverse primer specific for locus ZYRO0F18634g in CBS 732^T genome
	*MAT*α1 copy 1	301MATα1α2-cp1F1	TTCCTTCACCGCCAGAGGTTC	*MTL*α copy 1-specific forward primer combined with A,B and C reverse primers to complete *MAT*α1 ORF (referred to as for-αCp1 in Fig. S1)
	*MAT*α1 copy 2	301MATα1α2Cp2F3	TTCCTTCACCTCCGGAGAACC	3' end *MAT*α1 copy 2-specific forward primer (referred to as for-αCp2 in Fig. S1)
***MTL*a loci characterization**				
	*MAT***a**1	301_MATa1F3	GTAGCTTCCACAAGGTCTTCAAGG	*ZsMAT***a**1 ORF specific forward primer (referred to as P**a**1 in Fig. S1)
		301_MATa1R3	GTGTCCAATCTACTTGTCAGACCCA	*ZsMAT***a**1 ORF specific reverse primer (referred to as P**a**1 in Fig. S1)
	*MAT***a**2	301_MATa2F2	ACAGGTCTTCGACGTTTAGC	*ZsMAT***a**2 ORF specific forward primer (referred to as P**a**2 in Fig. S1)
		301_MATa2R2	CATGTGTCTGCAATCACTTCAC	*ZsMAT***a**2 ORF specific reverse primer (referred to as P**a**2 in Fig. S1)
***MAT*a2 ORF completion**				
	CHA1	1	GCTACTCCCTCATTAGAACATGAAA	forward primer specific for *CHA1* gene in CBS 732^T genome
	DIC1	2	CGCATGATATGAAACGAAGATGCAA	forward primer specific for *DIC1* gene in CBS 732^T genome
	CHA1$_L$	3	TACTTACTGGATGAATCTTCTGTGA	forward primer specific for *CHA1* paralog (ZYRO0F18524g) near to *HML* silent cassette in CBS 732^T genome
	*MAT***a**2	301_MATa2R2	CATGTGTCTGCAATCACTTCAC	1,2 and 3 reverse primers to complete *MAT***a**2 ORF (referred to as rev-**a**2 in Fig. S1)
System cassette analysis				
5' PCR-walking	*MAT*α2 copy 1	301_MATα2CP1R1	CTTGGTAATACAGGTAAAGAG GGT	*MAT*α2 copy 1-specific reverse primer (referred to rev-αcp1 in S2 Fig)
	*MAT*α2 copy 2	301_MATα1α2Cp2R1	GACACATTGCATTCTGTTAAACGT	*MAT*α2 copy 2-specific reverse primer (referred to rev-αcp2 in S2 Fig)
	*MAT***a**2	301_MATa2R2	CTCTTTCTCTCAAATACACGTTC	*MAT***a**2-specific reverse primer (referred to rev-**a** in S2 Fig)
	CHA1	1	GCTACTCCCTCATTAGAACATGAAA	forward primer specific for *CHA1* gene in CBS 732^T genome
	DIC1	2	CGCATGATATGAAACGAAGATGCAA	forward primer specific for *DIC1* gene in CBS 732^T genome
	CHA1$_L$	3	TACTTACTGGATGAATCTTCTGTGA	forward primer specific for *CHA1* paralog (ZYRO0F18524) located near to the silent *HML*cassette in CBS 732^T genome

3' PCR-walking	*MAT***a1**	301_MATa1F3	GTAGCTTCCACAAGGTCTTCAAGG	*MAT***a**1-specific forward primer (referred to for-**a** in S2 Fig)
	*MAT*α1 copy 1	301_MATα1α2Cp1F1	TTCCTTCACCGCCAGAGGTTC	*MAT*α1 copy 1-specific forward primer (referred to for-αcp1 in S2 Fig)
	*MAT*α1 copy 2	301_MATα1α2Cp2F3	TTCCTTCACCTCCGGAGAACC	*MAT*α1 copy 2-specific forward primer (referred to for-αcp2 in S2 Fig)
	SLA2	A	CCAGTTAGTGTGTTATCGATAAGTC	reverse primer specific for *SLA2* gene in CBS 732^T genome
	ZYRO0C18392	B	TCTATTTCGTCCGTTTATCGTTGGT	reverse primer specific for locus ZYRO0C18392 in CBS 732^T genome
	ZYRO0F18634	C	TCAGTACCAGAAGTGGTCTTTGAAA	reverse primer specific for locus ZYRO0F18634 in CBS 732^T genome
***HO* genes characterization**				
	HO copy 1	UpHOCBS 732F2	ACGAGTGGTGGTGGGATAGACTTA	5' UTR of CBS 732T *HO*-specific forward primer
		301_verylikeHOR4	TCGTGGGCCACTGAACATT	*ZsHO* copy 1-specific reverse primer
		301_verylikeHOF6	ATGGTGAAAAGCAAATTCCATCA	*ZsHO* copy 1-specific forward primer
		DownHOCBS 732R2	ATCTGACGCTTTGGCCTCTTTGGA	3' UTR of CBS 732T *HO*- specific reverse primer
	HO copy 2	301_likeHOR4	CTGATGTGCCACTGAGCACC	*ZsHO* copy 2-specific reverse primer
		301_likeHOF6	GATGGTGAGAAACAAATTCCATTG	*ZsHO* copy 2-specific forward primer
		DownHOCBS 732R1	TCACCAAGGCTATGTCTTCTCGCT	3' UTR of CBS 732T *HO*- specific reverse primer
Probe synthesis in PFGE-Southern blotting				
	*MAT*α1	301_MATa1F2	GTTCGGAGAAGCCACTCAATTC	*MAT*α1 specific digoxigenin labeled probe
		301_MATa1R3	GCTGGCACAAGCTTCTCAACTCTA	
	*MAT***a**1	301_MATA1F3	CGAAGAAGCTGTTCGGAGAAGCCACTCAAT	*MAT***a**1 specific digoxigenin labeled probe
		301_MATA1R3	GTGTCCAATCTACTTGTCAGACCCA	
	HO copies 1 and 2	301_5'HOF4	CGCTGAGGACATCGATGAAA	*HO* digoxigenin labeled probe
		301_3'HOR2	CTTCAAATTCACCACGCAGTTCC	
PCR-based sub-genome assignment of *MTL* and *HO* genes				
	HO copy 1	301_5'HOF1	CTACGTCGAGAGATCCATCATAG	T-subgenome *HO* specific primers
		301_verylikeHOR3	CGCGAATCTACCGGTACTATT	
		301_very_likeHOR5	GGTGTACGACCGGCTAAATG	

		HO_like_F8	GGTAGTTGTGCAAAGGTCACTG	
	HO copy 2	HO_like_R8	TAAATGGGAGTCCTGTCAACGA	P-subgenome *HO* specific primers
		301_likeHOF7	ATGTTGTGGGCGTAACAGTTG	
		HO_like_R9	TCTAATAAAATTCTTTTATCAGAATCAACT	
	*MAT***a**2 copy 1	301_MATA2F7cp1	CAGGTCTTCGACGTTTAGCCATG	*MAT***a**2 copy 1 specific forward primer
	*MAT***a**2 copy 2	301_MATA2F5cp2	GTAAACGGTAAAGCTGAAACGC	*MAT***a**2 copy 2 specific forward primer
	*MAT***a**2 copy 3	301_MATA2F6cp3	TACTTACTGGATGAATCTTCTCTG	*MAT***a**2 copy 3 specific forward primer
	*MAT***a**2	301_MATA2R4	GAAACCAACAGAGTTGCAGAAATA	Reverse primer to amplify all *MAT***a**2 ORFs
Non-quantitative RT-PCR				
	*MAT*α1 copy 1	301MATα1α2-cp1F1	TTCCTTCACCGCCAGAGGTTC	*MAT*α1 copy 1 specific forward primer
		301MATα1R6	ACCCACACTTCGGCTTGACAAA	*MAT*α1 copy 1 specific reverse primer
	*MAT*α1 copy 2	301_MATa1F7	TGGATCTTAGACAGTGGTGTAAGG	*MAT*α1 copy 2 specific forward primer
		301MATα1α2cp2R2	TGTCATCCACATTGAAATCATCTC	*MAT*α1 copy 2 specific reverse primer
	*MAT*α2 copy 1	MATa2_cp1F10	TTGCAGAACTACTTTACTCGTTCA	*MAT*α2 copy 1 specific forward primer
		301_MATα2CP1R1	CTTGGTAATACAGGTAAAGA GGT	*MAT*α2 copy 1 specific reverse primer
	*MAT*α2 copy 2	301MATα1α2cp2F1	GTTTAGATGCCAGTGCTCTTCA	*MAT*α2 copy 2 specific forward primer
		301_MATα1α2Cp2R1	GACACATTGCATTCTGTTAAACGT	*MAT*α2 copy 1 specific reverse primer
	*MAT***a**1 copy 2	301_MATa1F3	GTAGCTTCCACAAGGTCTTCAAGG	*MAT***a**1 copy specific forward primer
		301_MATA1R1	GTCCTCTTTCTCTCAAATACACG	*MAT***a**1 copy specific reverse primer
	*MAT***a**2 copy 2	301_MATA2F5cp2	GTAAACGGTAAAGCTGAAACGC	*MAT***a**2 copy specific forward primer
		301_MATA2R4	GAAACCAACAGAGTTGCAGAAATA	*MAT***a**2 copy specific reverse primer
	HO copy 1	301_5'HOF1	CTACGTCGAGAGATCCATCATAG	*HO* copy 1 specific forward primer
		301_verylikeHOR3	CGCGAATCTACCGGTACTATT	*HO* copy 1 specific reverse primer
	HO copy 2	301_5'HOF4	CGCTGAGGACATCGATGAAA	*HO* copy 2 specific forward primer
		301_likeHOR3	CTACAAACCTACCGGTGTTAGA	*HO* copy 2 specific reverse primer
	IME4	ZrIME4_F2	TGTGAGGAATTCGATTTAG	Primer forward designed to amplify an internal region common to both *IME4* sense mRNA and *IME4* lncRNA
		ZrIME4_R2	GAATGATTTAGCCAATGACC	Primer reverse designed to amplify an internal region common to both *IME4* sense mRNA and *IME4* lncRNA
	AS-*IME4*	ZrIME4_F1	GCGATTGTTCATATTTGGATAC	Primer forward specifically designed for cDNA synthesis (RT) of AS-*IME4*
	S-*IME4*	ZrIME4_R1	TCAGGTTTTCTGCTGGTTTCTCTGGT	Primer forward specifically designed for cDNA synthesis (RT) of S-*IME4*
RT-qPCR				
	ZrACT	301_ACTF1	GGTCGCAGCTTTGGTTATTG	Oligonucleotide RT-qPCR *ACT* specific forward primer

	301_ACTR1	GGCCCATACCAACCATGATA	Oligonucleotide RT-qPCR *ACT* specific reverse primer
*MAT*α1 copy 2	301MATα1-cp2F4	AGAATCGACCCAGACACCAA	Oligonucleotide RT-qPCR *MAT*α1 copy 2 specific forward primer
	301MATα1-cp2R3	TATCAGGTTCTCCGGAGGTG	Oligonucleotide RT-qPCR *MAT*α1 copy 2 specific reverse primer
*MAT*α2 copy 2	301_MATα2cp2F5	TAAACCAAGTTCTAGTGAGTAC	Oligonucleotide RT-qPCR *MAT*α2 copy 2 specific forward primer
	301_MATα2cp2R5	GAAGCTGCACTTGGAAATAAA	Oligonucleotide RT-qPCR *MAT*α2 copy 2 specific reverse primer
*MAT***a**1	301MATA1F4	TCGTCGTCGAAGGAGGTATC	Oligonucleotide RT-qPCR *MAT***a**1 specific forward primer
	301MATA1R4	GCTGCTACAGCTTCCCTTTC	Oligonucleotide RT-qPCR *MAT***a**1 specific reverse primer
HO copy 1	301_very_likeHOF8	GTCGCTAGGTATGCCCGTTA	Oligonucleotide RT-qPCR *HO* copy 1 specific forward primer
	301_very_likeHOR5	GGTGTACGACCGGCTAAATG	Oligonucleotide RT-qPCR *HO* copy 1 specific reverse primer
HO copy 2	301_likeHOF10	GAAACCGCCATCTGAGAAAG	Oligonucleotide RT-qPCR *HO* copy 2 specific forward primer
	301_likeHOR10	AAAATGCTTCACGCACCTGT	Oligonucleotide RT-qPCR *HO* copy 2 specific reverse primer
	301_likeHOF12	CATCGTAGAAACCGCCATC	Oligonucleotide RT-qPCR *HO* copy 2 specific forward primer
	301_likeHOR12	CCGTCTGTATGGAAACCTTG	Oligonucleotide RT-qPCR *HO* copy 2 specific reverse primer

Table S2. PCR-based subgenome assignment of mating-type and *HO* gene copies in *Zygosaccharomyces pseudorouxii* (nom. inval.) NCYC 3042.

Target gene	**NCYC 3042**		
	copy 1	**copy 2**	**copy 3**
***MAT*α1**	**-**	**+**	**nd**
***MAT*α2**	**-**	**+**	**nd**
***MAT*a1**	**-**	**-**	**-**
***MAT*a2**	**-**	**-**	**-**
HO	**-**	**+**	**na**

nd, not determined; na, not applicable; +, positive result; -, negative result

Figure S1. Outlined experimental strategy used for characterizing *MTL*α (A) and *MTL*a loci (B) in the ATCC 42981 genome. Panel A shows *ZsMTL*α variants represented in blue and surrounded in grey (copy 1), green (copy 2) and orange (copy 3), respectively, while panel B represents *MTL***a** loci coloured in red. Primer pairs specific for *ZsMTL*α copies 1 to 3 are arbitrarily referred to as αCp1-P, αCp2-P, and αCp3-P. *ZsMAT***a**1 and *ZsMAT***a**2-specific primer pairs are arbitrarily referred to as P**a**1 and P**a**2, while the primer pair termed P**a**12 spans the complete *MAT***a**1 ORF and a portion of *MAT***a**2 gene. Solid grey arrows indicate generic flanking genes, dotted borders represent uncompleted sequences and small arrows (solid) designate gene-specific primers. Primer sequences are reported in **Table S1**, according to Watanabe et al., 2013 and Solieri et al., 2014b. Abbreviations: Zs, *Zygosaccharomyces sapae*; cp, copy; P, primer; for, forward; rev, reverse.

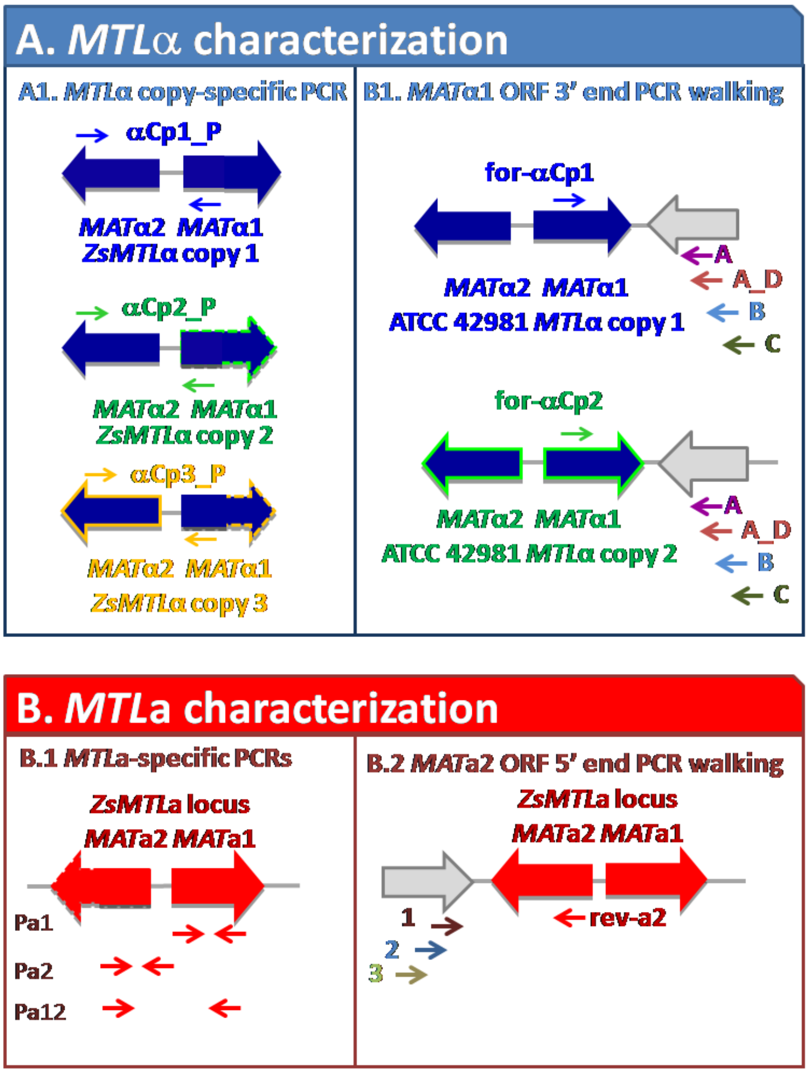


Figure S2. Polymerase chain reaction (PCR)-based strategies used for determining the system of cassette-based arrangement in the ATCC 42981 genome. Forward and reverse *MTL*-specific internal primers were used to screen PCR products obtained using all possible combinations of primers spanning putative *MTL*-flanking genes (semi-nested PCR approach); in cases of negative results, 5' and 3' PCR walking was performed using all possible combinations of *MTL*-specific internal primers and *MTL*-flanking gene primers (direct PCR approach). Small arrows (solid) indicate gene-specific primers and degenerate primers (dotted lines). *CHA1*$_L$ indicates the ZYRO0F18524g locus. Primer sequences are reported in **Table S1**, according to Watanabe et al., 2013 and Solieri et al., 2014b. Abbreviations: cp, copy; for, forward; rev, reverse.

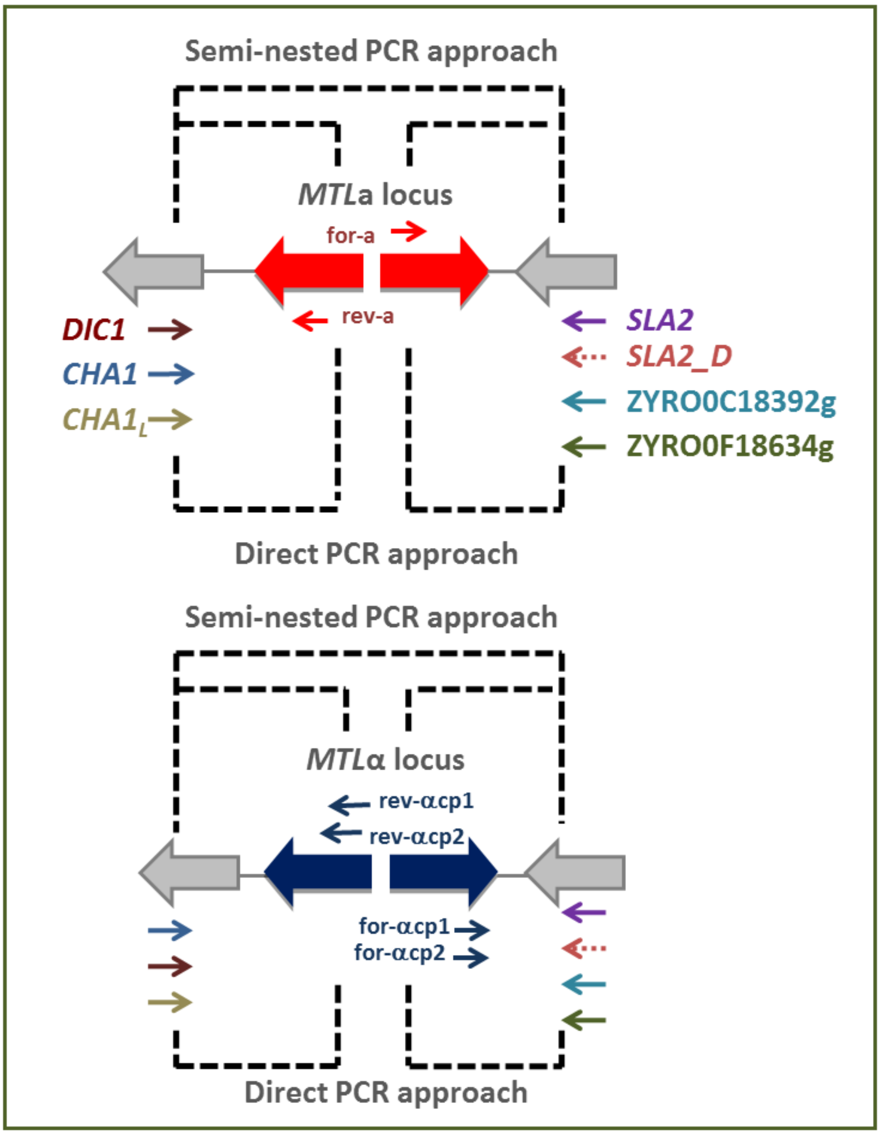


Figure S3. Outlined experimental approach used for HO gene characterization in the ATCC 42981 genome. Dotted lines represent undetermined sequences. Primer sequences are reported in **Table S1**. Abbreviation: ZsHO, *Zygosaccharomyces sapae HO* gene.

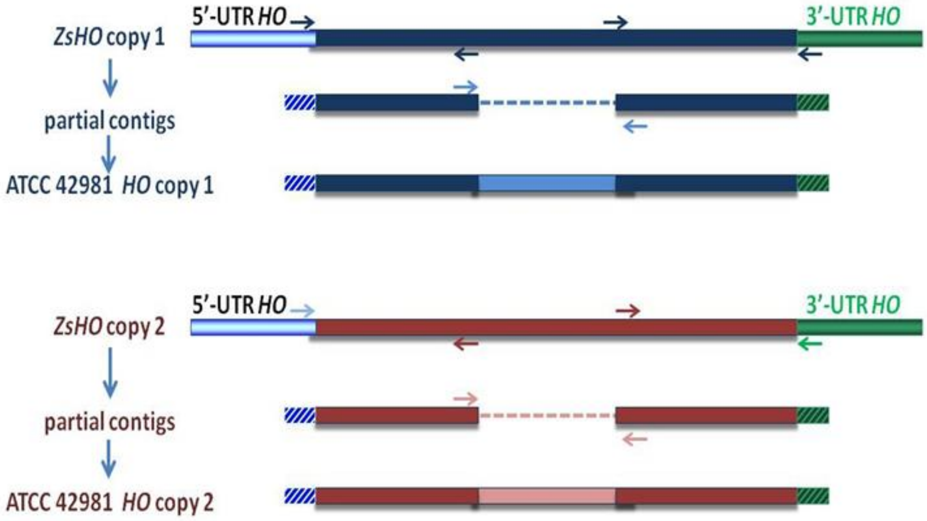


Figure S4. Chromosomal mapping of ATCC 42981 HO genes. Chromosomes were separated by PFGE for ATCC 42981 (1), *Z. rouxii* CBS 732^T (2), and *Z. sapae* ABT301^T (3). Southern blotting analysis was carried out with probe labelled to *HO* genes. M indicates the chromosomal size ladder (*Saccharomyces cerevisiae* S288C, BioRad Laboratories) in megabase pairs. ATCC 42981 chromosomes are indicated in uppercase letters (from A to G).

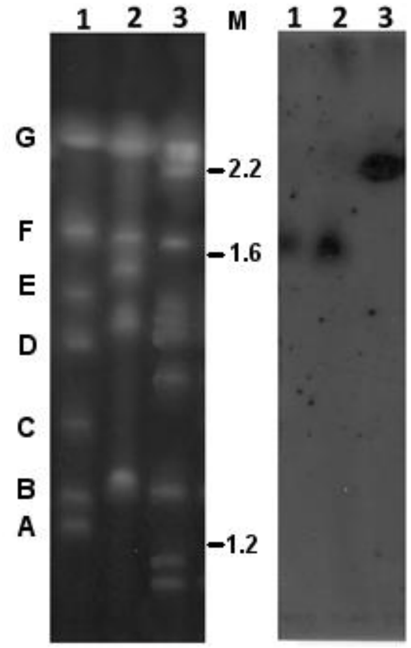
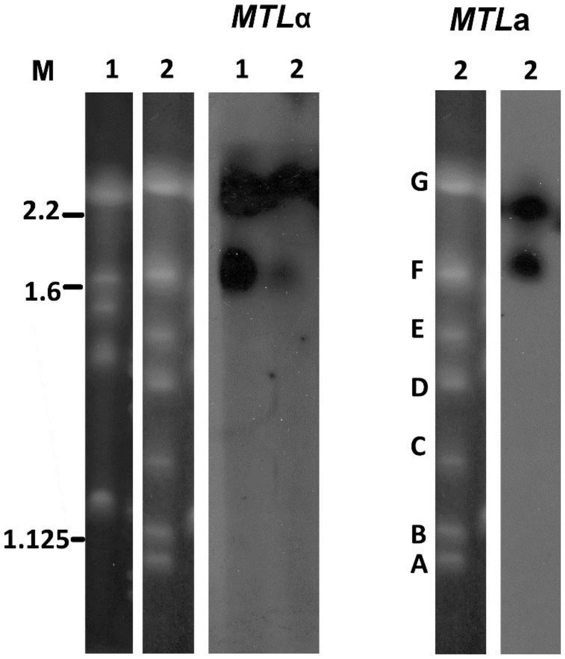


Figure S5. Chromosomal mapping of ATCC 42981 MTLα and *MTL*a loci. Chromosomes were separated by PFGE for *Z. rouxii* CBS 732^T (1) and ATCC 42981 (2). Southern blotting analyses were carried out with probes labelling the α- and **a**-idiomorph loci. The left panel shows signals from *MTL*α loci *Zygosaccharomyces rouxii* CBS 732^T (1) and the ATCC 42981 strain (2), respectively. The right panel reports separated chromosomes and signals from *MTL***a** loci in ATCC 42981 (2). M indicates the chromosomal size ladder (Saccharomyces cerevisiae S288C, BioRad Laboratories) in megabase pairs. ATCC 42981 chromosomes are indicated in uppercase letters (from A to G).



Chapter 6

Table S1. Primers used in this work. This table shows the primers used, their usages, sequences and descriptions. Annealing and melting temperatures are calculated using OligoAnalyzer 3.1 tool (idtdna.com) with oligo, dNTPs and Mg^{++} concentrations of 0.5 µM, 0.2 and 1.5 mM, respectively.

Experiment	Primer	Sequence (5'-> 3')	Amplicon (bp)	PCR conditions	Thermal profiles
Diagnostic PCR for transformation with plasmid YEp352-SAT	SAT1-1	gacctcaagtctcgaacg	206	0.025 U/ml DreamTaq™, 1x DramTaq ™ Green Buffer, 2.0 mM MgCl2, 0.5 µM SAT1-1-, 0.5 µM pGRB2-SAT1-pHl-R, 200 µM dNTPs	95°C x 3'; (95°C x 30''; 53° C x 30''; 72°C x 45'')$_{x35}$; 72° x 10'
	pGRB2-SAT1-pHl-R	cctgcaggaccacctttgattg			
Diagnostic PCR for transformation with plasmid pCg2XpH-N	Down-pHluorin-F	cacaacccacagctaccaccatc	436	0.025 U/ml DreamTaq™, 1x DramTaq ™ Green Buffer, 2.0 mM MgCl2, 0.5 µM Down-pHluorin-F, 0.5 0.5 µM pGRB-R4, 200 µM dNTPs	95°C x 3'; (95°C x 30''; 56.3° C x 30''; 72°C x 1')$_{x35}$; 72° x 10'
	pGRB-R4	gcgcgcgtaatacgactcacta			
Diagnostic PCR to verify the targeted integration of *loxP-kanMX-loxP* cassette	MATα2cp2-893up-F	cggtaacgactgtatagctaag	956	0.025 U/ml DreamTaq™, 1x DramTaq ™ Green Buffer, 2.0 mM MgCl2, 0.5 µM MATα2cp2-893up-F, 0.5 µM KanMX-R1, 200 µM dNTPs	95°C x 5'; (95°C x 30''; 55°C x 30''; 72°C x 1'30'')$_{x35}$; 72° x 10'
	KanMX-R1	catttgatgctcgatga			
	KanMX-F1	ctctggcgcatcgggc	574	0.025 U/ml DreamTaq™, 1x DramTaq ™ Green Buffer, 2.0 mM MgCl2, 0.5 µM KanMX-F1, 0.5 µM MATα1cp2-374down-R, 200 µM dNTPs	95°C x 5'; (95°C x 30''; 55°C x 30''; 72°C x 1')$_{x35}$; 72° x 10'
	MATα1cp2-374down-R	ccaaactttatggatatgagttctagc			

Diagnostic PCR for pZEG construction	pZEU-2F	gctcactgcccgctttccagtcggg	726	0.025 U/ml DreamTaq™, 1x DramTaq™ Green Buffer, 2.0 mM MgCl2, 0.5 µM pZEU-2F, 1 µM KanMX-R1, 200 µM dNTPs	95°C x 5’; (95°C x 30’’; 55°C x 30’’; 72°C x 1’)$_{x35}$; 72° x 10’
	KanMX-R1	catttgatgctcgatga			
	KanMX-F1	ctctggcgcatcgggc	626	0.025 U/ml DreamTaq™, 1x DramTaq™ Green Buffer, 2.0 mM MgCl2, 1 µM KanMX-F1, 0.5 µM pZEU-2R, 200 µM dNTPs	95°C x 5’; (95°C x 30’’; 60°C x 30’’; 72°C x 1’)$_{x35}$; 72° x 10’
	pZEU-2R	cgcaaaccgcctctccccgcgcg			
Diagnostic PCR for pZCAG	pZCA-2F	ccgaggaactcttggtattcttgcc	768	0.025 U/ml DreamTaq™, 1x DramTaq™ Green Buffer, 2.0 mM MgCl2, 0.5 µM pZCA-2F, 1 µM KanMX-R1, 200 µM dNTPs	95°C x 5’; (95°C x 30’’; 55°C x 30’’; 72°C x 1’)$_{x35}$; 72° x 10’
	KanMX-R1	catttgatgctcgatga			
	KanMX-F1	ctctggcgcatcgggc	679	0.025 U/ml DreamTaq™, 1x DramTaq™ Green Buffer, 2.0 mM MgCl2, 1 µM KanMX-F1, 0.5 µM pZCA-2R, 200 µM dNTPs	95°C x 5’; (95°C x 30’’; 55°C x 30’’; 72°C x 1’)$_{x35}$; 72° x 10’
	pZCA-2R	cgaaaagtgccacctgacgtc			
Diagnostic PCR for pZEN	SAT1-1	gacctcaagtctcgaacg	336	0.025 U/ml DreamTaq™, 1x DramTaq™ Green Buffer, 2.0 mM MgCl2, 1 µM SAT1-1, 0.5 µM pZEU-2R, 200 µM dNTPs	95°C x 5’; (95°C x 30’’; 55°C x 30’’; 72°C x 45’’)$_{x35}$; 72° x 10
	pZEU-2R	cgcaaaccgcctctccccgcgcg			
Diagnostic PCR for pZCAN	SAT1-1	gacctcaagtctcgaacg	389	0.025 U/ml DreamTaq™, 1x DramTaq™ Green Buffer, 2.0 mM MgCl2, 1 µM SAT1-1, 0.5 µM pZCA-2R, 200 µM dNTPs	95°C x 5’; (95°C x 30’’; 55°C x 30’’; 72°C x 45’’)$_{x35}$; 72° x 10
	pZCA-2R	cgaaaagtgccacctgacgtc			
Diagnostic PCR for transformation with pGRCRE	pGRB-F4	tttgagtgagctgataccgct	552	0.025 U/ml DreamTaq™, 1x DramTaq ™ Green	95°C x 3’; (95°C x 30’’; 57° C x 30’’; 72°C x 1’)$_{x35}$; 72° x 10’

	GAL-F1	gccaggttactgccaatttttcc		Buffer, 2.0 mM MgCl2, 0.5 µM pGRB-F4, 0.5 GAL-F1, 200 µM dNTPs	
	pGRB-R4	gcgcgcgtaatacgactcacta	723	0.025 U/ml DreamTaq™, 1x DramTaq ™ Green Buffer, 2.0 mM MgCl2, 0.5 µM pGRB-R4, 0.5 µM CRE-R1, 200 µM dNTPs	95°C x 3’; (95°C x 30’’; 57° C x 30’’; 72°C x 1’)$_{x35}$; 72° x 10’
	CRE-R1	ttgccgggtcagaaaaaatg			
Diagnostic PCR for *kanMX-loxP* pop-out	MATα2cp2-893up-F	cggtaacgactgtattagctaag	2889	0.025 U/ml DreamTaq™, 1x DramTaq ™ Green Buffer, 2.0 mM MgCl2, 0.5 µM forward primer, 0.5 µM reverse primer, 200 µM dNTPs	95°C x 3’; (95°C x 30’’; 54° C x 30’’; 72°C x 2’)$_{x35}$; 72° x 10’
	MATα1cp2-374down-R	ccaaactttatggatatgagttctagc			
	KanMXF1	catttgatgctcgatga	600		95°C x 3’; (95°C x 30’’; 55° C x 30’’; 72°C x 1’)$_{x35}$; 72° x 10’
	MATα1cp2-374down-R	ccaaactttatggatatgagttctagc			
***SOD*2 RT-PCR**	SOD_1R	tccttaacaaacaatgctaagt	600	0.025 U/ml rTaq™, 1x Buffer, 2.0 mM MgCl2, 0.3 µM SOD-1F, 0.3 µM SOD-1R, 200 µM dNTPs	95°C x 5’; (95°C x 1’; 55° C x 2’; 72°C x 2’)$_{x35}$; 72° x 10’
	SOD_1F	acgtatgaattcgatgcagat			
***cre* RT-PCR**	RT-CRE_F2	gagtgatgaggttcgcaaga	587	0.1 U/ml DreamTaq™, 2x Buffer green (10x), 1.6 µM dNTPs, 1 µM RT-CRE-F2, 1 µM RT-CRE-R2	95°C x 3’; (95°C x 30’’; 57° C x 30’’; 72°C x 1’)$_{x35}$; 72° x 10’
	RT-CRE_R2	ggctaagtgccttctctacac			

Chapter 8

Table S1. Primer sets used in the present study. Primers named as old are from Watanabe et al. (2013), whereas primers named as new derive from Watanabe et al. (2017). DNA sequences complementary to the *kanMX* gene sequence and plasmid sequences are written in lower-case.

Primer name	**Sequence (5' to 3')**	**Corresponding Figure**
DIC1_P_F1	GGTATTTGATGGGAGCAGCA	Figure 1A
3_old	TACTTACTGGATGAATCTTCTGTGA	Figure 1B
C_old	TCAGTACCAGAAGTGGTCTTTGAAA	Figures 1A, 1B
MATα2cp2-893up-F	CGGTAACGACTGTATAGCTAAG	Figures 1C, S1
SLA2_P_R3	GGACAGTTGGGAGACACTGAA	Figure 1C
3_old	TACTTACTGGATGAATCTTCTGTGA	Figure 1D
SLA2_P_R2	GGGAGACACTGAAGCGTTAGAT	Figure 1D
CHA1L_P_F1	AGACAGCTACAAGGTGTTGTGA	Figure 1E
SLA2Tr_R1	AAAGTCCTATTCACGTGACGAA	Figure 1E
DIC1_P_F2	GAGTGGTATGGTGAAGCTGTG	Figure 1F
SLA2_uni_R1	ATATATCTTATCGAGACAGTGTATTTC	Figure 1F
1_old	GCTACTCCCTCATTAGAACATGAAA	Figure 1G
B_old	TCTATTTCGTCCGTTTATCGTTGGT	Figure 1G
6_new	TGTATTGACCAGCTTCGTTTGA	Figure 1H
F_new	ATGGACTACACGTACCACAA	Figure 1H
301_MATa1F7	TGGATCTTAGACAGTGGTGTAAGG	Figure 3A
301MATα1α2cp2R2	TGTCATCCACATTGAAATCATCTC	Figure 3A
301MATa1a2-cp2F1	GTTTAGATGCCAGTGCTCTTCA	Figure 3B
301MATa1a2-cp2R1	GACACATTGCATTCTGTTAAACGT	Figure 3B
301_MATA1F3	GTAGCTTCCACAAGGTCTTCAAG	Figure 3C
301_MATA1R4	GCTGCTACAGCTTCCCTTTC	Figure 3C
301_MATA2F8	AGCCAAGTGGGCGATTTA	Figure 3D
301_MATA2R2	CATGTGTCTGCAATCACTTCAC	Figure 3D
MATA1_008_F1	ATTCTCCAAATGATCTTCAGA	Figure 3E
MATA1_008_R1	ATACCCATATTCTTACTTGAAGT	Figure 3E
MATα1/2cp2-kanMX-F-80nt	CATGTTTGAACGAGTGTTTTGTTCATTGGTTTGGAATAAACAGGTCTTCGACGTTTAGCCATGTCGAGGATTTAAACGTTTGACAttcgtacgctgcaggtcgac	Figure S1
MATα1/2cp2-kanMX-R-80nt	CAACCGGTAAGTGTTCTTTCAATAAGTCAGTTGTGCAATGAAGTGGCAAGTCAGTTTTTAAGCAACACACCGCACGTACCGgcataggccactagtggatctg	Figure S1
kanMX-R1	CTCTGGCGCATCGGGC	Figure S1
kanMX-F1	CATTTGATGCTCGATGA	Figure S1
MATα1cp2-374down-R	CCAAACTTTATGGATATGAGTTCTAGC	Figure S1
AGA2_backbone1_rouxii_F	CATGTACCACTGTACCCAGTAAG	Data not shown
AGA2_backbone1_rouxii_R	ACCGTAGTAGTCCCGATTGA	
STE2_backbone15_rouxii_F	CCTATTGGCCTCGTCTGTTAAT	
STE2_backbone15_rouxii_R	TAGGCGGACAAGATATGAGGT	
STE2_backbone5_no_rouxii_F	CCTATTGGCCTCGTCTGTTAAT	

STE2_backbone5_no_rouxii_R	TAGGCGGACAAGATATGAGGT
STE6_backbone1_rouxii_F	TAACACTACCAGTGGGTAA
STE6_backbone1_rouxii_R	TCATAAGTGGACGTTTTGAAA
STE6_backbone6_no_rouxii_F	ATATCAAAATCGATGGCATGGA
STE6_backbone6_no_rouxii_R	AGCGGTTATTTTGTTGCCT

Table S2. **Inventory of *MTL* cassettes in ATCC 42981_R draft genome derived from DBG2OLC and MaSuRCA *de novo* assemblies.** DBG2OLC scaffolds (accession numbers: UEMZ01000001.1-UEMZ01000033.1) were derived from ATCC 42981_R BioProject PRJEB26771. Grey shadow indicates *MTL* cassettes found in both the assemblies. JCM66020 *MTL* cassettes were described according to the nomenclature reported by(Watanabe, Uehara, Mogi, & Tsukioka, 2017) Watanabe et al. (2017). Briefly, numbers from 1 to 6 indicate 5' *MTL*-flanking genes *DIC1*^{T}, *CHA1*$_{L}^{T}$, *CHA1*^{T}, *DIC1*^{P}, *CHA1*$_{L}^{P}$, and *CHA1*^{P}, respectively. Capital letters A to F indicate 3' *MTL*-flanking genes *SLA2*^{T}, ZYRO0F18634g^{T}, ZYRO0C18392g^{T}, *SLA2*^{P}, ZYRO0F18634g^{P} and ZYRO0C18392g^{P}, respectively. Abbreviation: r.c., reverse complement.

Cassettes	**Bizzarri *et al*. (2016)**	**JCM22060**	**PCR**	***In silico* analysis**	**NBRC110957 (Accession number)**
Yα^{T}					
DIC1^{P}-*MTL*α^{T}-ZYRO0F18634g^{T}	-	4B	+	DBG2OLC, MaSuRCA	*CHA1*^{P}-*MTL*α^{P}-ZYRO0C18392g^{P} (BDGX01000045)
CHA1$_{L}^{T}$-*MTL*α^{T}-ZYRO0F18634g^{T}	+	-	+	DBG2OLC	*CHA1*$_{L}^{T}$-*MTL*α^{T}-ZYRO0F18634g^{T} (BDGX01000025)
Yα^{P}					
DIC1^{T}-*MTL*α^{P}-*SLA2*^{P}	+	1D	+	DBG2OLC,MaSuRCA	*DIC1*^{T}-*MTL***a**^{P}-*SLA2*^{P} (BDGX01000009)
CHA1$_{L}^{T}$-*MTL*α^{P}-*SLA2*^{P}	+	2D	+	DBG2OLC,MaSuRCA	-
CHA1$_{L}^{P}$-*MTL*α^{P}-ZYRO0F18634g^{P}	-	5E	+	DBG2OLC,MaSuRCA	*CHA1*$_{L}^{P}$-*MTL*α^{P}-ZYRO0F18634g^{P} (BDGX01000013)
DIC1^{T}-*MTL*α^{P}-*SLA2*$^{\mathbf{N}}$	-	-	-	DG2OLC	-
Ya					
DIC1$^{\mathbf{N}}$-*MTL***a**$^{\mathbf{N}}$-*SLA2*^{T}	+ (partially)	-	+	DBG2OLC	*DIC1*^{P}-*MTL***a**^{T}-ZYRO0C18392g^{T} (BDGX01000035)
CHA1^{T}-*MTL***a**^{T}-ZYRO0C18392g^{T}	+	3C	+	DBG2OLC,MaSuRCA	*CHA1*^{T}-*MAT***a**^{P}-*SLA2*^{P} (BDGX01000021)
CHA1^{P}-*MTL***a**^{P}-ZYRO0C18392g^{P}	-	6F	+	MaSuRCA	-

Figure S1. Validation of *MAT*α^{P} deletion in ATCC 42981_R. Panel (**A**) outlines the diagnostic PCR strategy used to verify the integration of *loxP-kanMX-loxP* disruption cassette in the *DIC1*^{T}-*MAT*α^{P}-*SLA*2^{P} locus of ATCC 42981_R genome. Full-length, 5' and 3'-PCRs are shown. Flanking genes from T and P-subgenomes are marked with T and P superscripts, respectively. *MAT*α1 and *MAT*α2 genes from P-subgenome are indicated as α^{P}1 and α^{P}2. In panel (**B**) the *loxP-kanMX-loxP* module was integrated at the *DIC1*^{T}-*MTL*α^{P}-*SLA2*^{P} cassette in four ATCC 42981_R clones, as demonstrated by size difference in full-length PCR products between *kanMX* (2,164 bp) and *MAT*α^{P} (2,889 bp) cassettes. The correct orientation of *loxP-kanMX-loxP* module was confirmed by the 5' and 3' PCRs. Numbers from 1 to 5 indicate *MAT*α^{P}Δ clone_6, _65, _74, and _177, and ATCC 42981_, respectively (**Table 1**), while M represents molecular weight marker GeneRuler 100 bp or 1 Kb Plus DNA Ladders (Thermo Scientific, Waltham, MA). Abbreviations: wt, wild type; HR, homologous recombination; *kanMX*, kanamycin resistance gene; NTC, no template control.

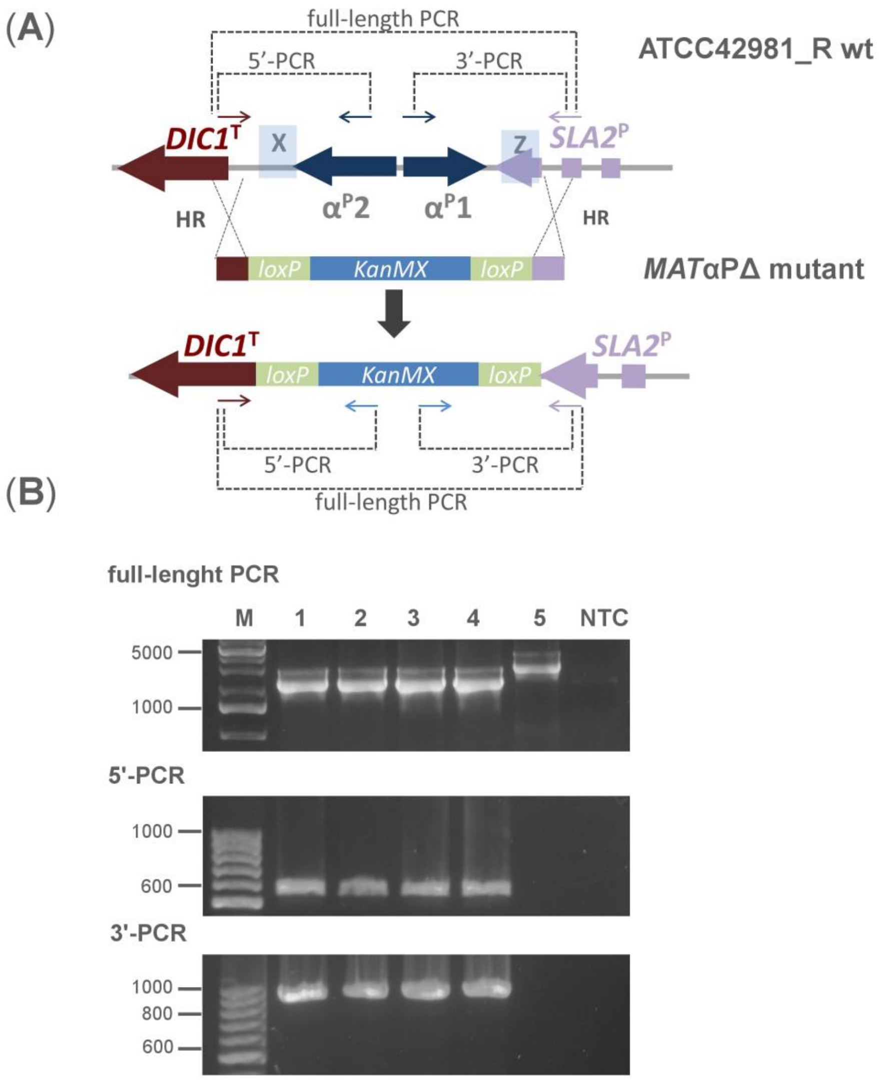
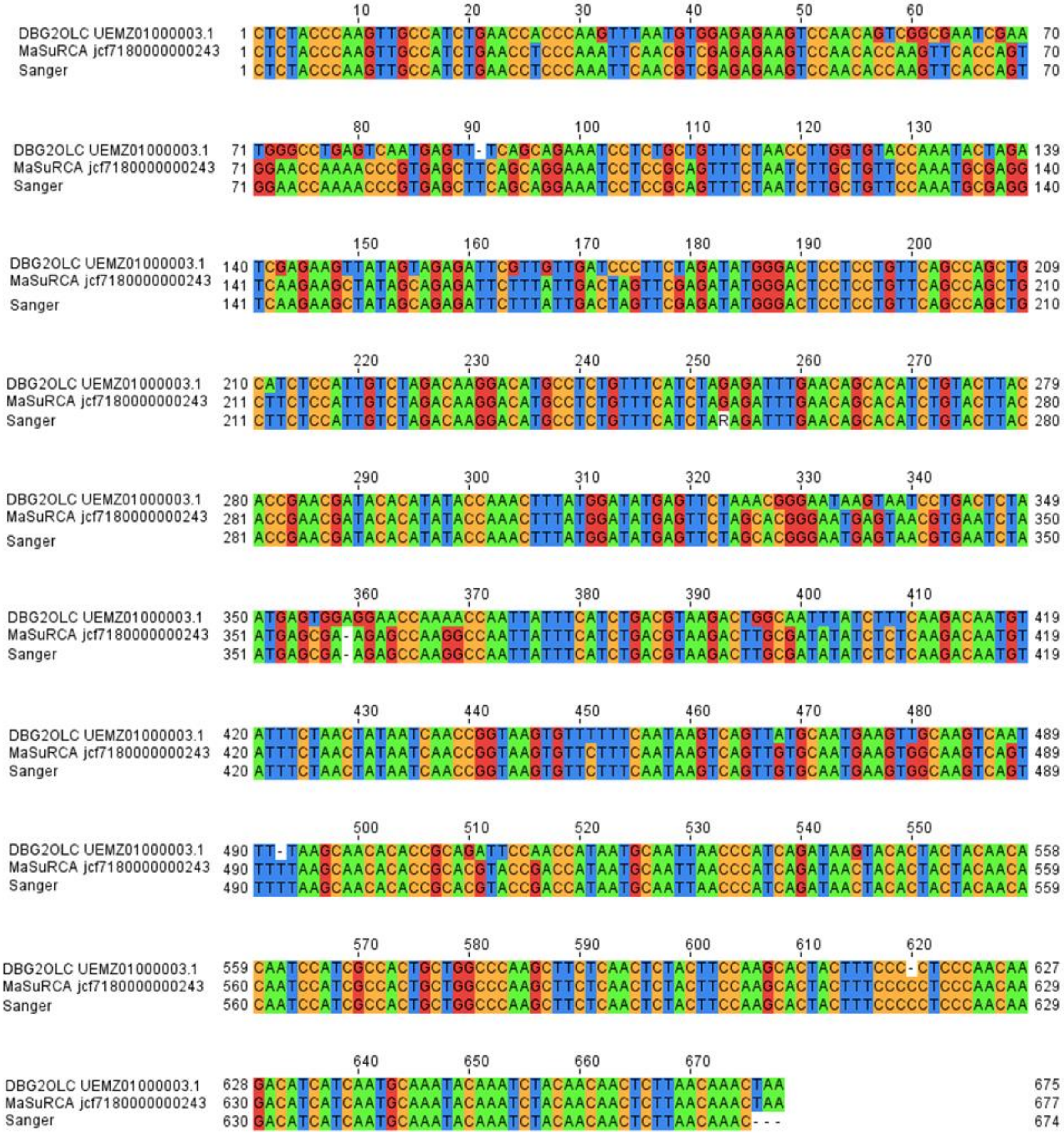


Figure S2. Multiple sequence alignment of *SLA2* genes from DBG2OLC, MaSuRCA assemblies and Sanger sequences. Alignment was carried out using Clustal Omega. Formatting of aligned sequences was done in Jalview alignment viewer. Residues in the alignment follow the default Clustal colour scheme of Jalview.



```
                                       10        20        30        40        50        60
DBG2OLC UEMZ01000003.1        1 CTCTACCCAAGTTGCCATCTGAACCACCCAAGTTTAATGTGGAGAGAAGTCCAACAGTCGGCGAATCGAA  70
MaSuRCA jcf7180000000243      1 CTCTACCCAAGTTGCCATCTGAACCTCCCAAATTCAACGTCGAGAGAAGTCCAACACCAAGTTCACCAGT  70
Sanger                        1 CTCTACCCAAGTTGCCATCTGAACCTCCCAAATTCAACGTCGAGAGAAGTCCAACACCAAGTTCACCAGT  70

                                       80        90       100       110       120       130
DBG2OLC UEMZ01000003.1       71 TGGGCCTGAGTCAATGAGTT-TCAGCAGAAATCCTCTGCTGTTTCTAACCTTGGTGTACCAAATACTAGA 139
MaSuRCA jcf7180000000243     71 GGAACCAAAACCCGTGAGCTTCAGCAGGAAATCCTCCGCAGTTTCTAATCTTGCTGTTCCAAATGCGAGG 140
Sanger                       71 GGAACCAAAACCCGTGAGCTTCAGCAGGAAATCCTCCGCAGTTTCTAATCTTGCTGTTCCAAATGCGAGG 140

                                      150       160       170       180       190       200
DBG2OLC UEMZ01000003.1      140 TCGAGAAGTTATAGTAGAGATTCGTTGTTGATCCCTTCTAGATATGGGACTCCTCCTGTTCAGCCAGCTG 209
MaSuRCA jcf7180000000243    141 TCAAGAAGCTATAGCAGAGATTCTTTATTGACTAGTTCGAGATATGGGACTCCTCCTGTTCAGCCAGCTG 210
Sanger                      141 TCAAGAAGCTATAGCAGAGATTCTTTATTGACTAGTTCGAGATATGGGACTCCTCCTGTTCAGCCAGCTG 210

                                      220       230       240       250       260       270
DBG2OLC UEMZ01000003.1      210 CATCTCCATTGTCTAGACAAGGACATGCCTCTGTTTCATCTAGAGATTTGAACAGCACATCTGTACTTAC 279
MaSuRCA jcf7180000000243    211 CTTCTCCATTGTCTAGACAAGGACATGCCTCTGTTTCATCTAGAGATTTGAACAGCACATCTGTACTTAC 280
Sanger                      211 CTTCTCCATTGTCTAGACAAGGACATGCCTCTGTTTCATCTARAGATTTGAACAGCACATCTGTACTTAC 280

                                      290       300       310       320       330       340
DBG2OLC UEMZ01000003.1      280 ACCGAACGATACACATATACCAAACTTTATGGATATGAGTTCTAAACGGGAATAAGTAATCCTGACTCTA 349
MaSuRCA jcf7180000000243    281 ACCGAACGATACACATATACCAAACTTTATGGATATGAGTTCTAGCACGGGAATGAGTAACGTGAATCTA 350
Sanger                      281 ACCGAACGATACACATATACCAAACTTTATGGATATGAGTTCTAGCACGGGAATGAGTAACGTGAATCTA 350

                                      360       370       380       390       400       410
DBG2OLC UEMZ01000003.1      350 ATGAGTGGAGGAACCAAAACCAATTATTTCATCTGACGTAAGACTGGCAATTTATCTTTCAAGACAATGT 419
MaSuRCA jcf7180000000243    351 ATGAGCGA-AGAGCCAAGGCCAATTATTTCATCTGACGTAAGACTTGCGATATATCTCTCAAGACAATGT 419
Sanger                      351 ATGAGCGA-AGAGCCAAGGCCAATTATTTCATCTGACGTAAGACTTGCGATATATCTCTCAAGACAATGT 419

                                      430       440       450       460       470       480
DBG2OLC UEMZ01000003.1      420 ATTTCTAACTATAATCAACCGGTAAGTGTTTTTTCAATAAGTCAGTTATGCAATGAAGTTGCAAGTCAAT 489
MaSuRCA jcf7180000000243    420 ATTTCTAACTATAATCAACCGGTAAGTGTTCTTTCAATAAGTCAGTTGTGCAATGAAGTGGCAAGTCAGT 489
Sanger                      420 ATTTCTAACTATAATCAACCGGTAAGTGTTCTTTCAATAAGTCAGTTGTGCAATGAAGTGGCAAGTCAGT 489

                                      500       510       520       530       540       550
DBG2OLC UEMZ01000003.1      490 TT-TAAGCAACACACCGCAGATTCCAACCATAATGCAATTAACCCATCAGATAAGTACACTACTACAACA 558
MaSuRCA jcf7180000000243    490 TTTTAAGCAACACACCGCACGTACCGACCATAATGCAATTAACCCATCAGATAACTACACTACTACAACA 559
Sanger                      490 TTTTAAGCAACACACCGCACGTACCGACCATAATGCAATTAACCCATCAGATAACTACACTACTACAACA 559

                                      570       580       590       600       610       620
DBG2OLC UEMZ01000003.1      559 CAATCCATCGCCACTGCTGGCCCAAGCTTCTCAACTCTACTTCCAAGCACTACTTTCCC-CTCCCAACAA 627
MaSuRCA jcf7180000000243    560 CAATCCATCGCCACTGCTGGCCCAAGCTTCTCAACTCTACTTCCAAGCACTACTTTCCCCCTCCCAACAA 629
Sanger                      560 CAATCCATCGCCACTGCTGGCCCAAGCTTCTCAACTCTACTTCCAAGCACTACTTTCCCCCTCCCAACAA 629

                                      640       650       660       670
DBG2OLC UEMZ01000003.1      628 GACATCATCAATGCAAATACAAATCTACAACAACTCTTAACAAACTAA                       675
MaSuRCA jcf7180000000243    630 GACATCATCAATGCAAATACAAATCTACAACAACTCTTAACAAACTAA                       677
Sanger                      630 GACATCATCAATGCAAATACAAATCTACAACAACTCTTAACAAAC---                       674
```

Figure S3. Sequence alignment highlighting the 27 bp indel in X regions downstream the *DIC1*^{N} and *DIC1*^{T} gene variants. X region from CBS 732^T was used as reference. Sequences were retrieved from Sanger sequencing and aligned using Clustal Omega. Formatting of aligned sequences was done in Jalview alignment viewer. Residues in the alignment follow the default Clustal colour scheme of Jalview.

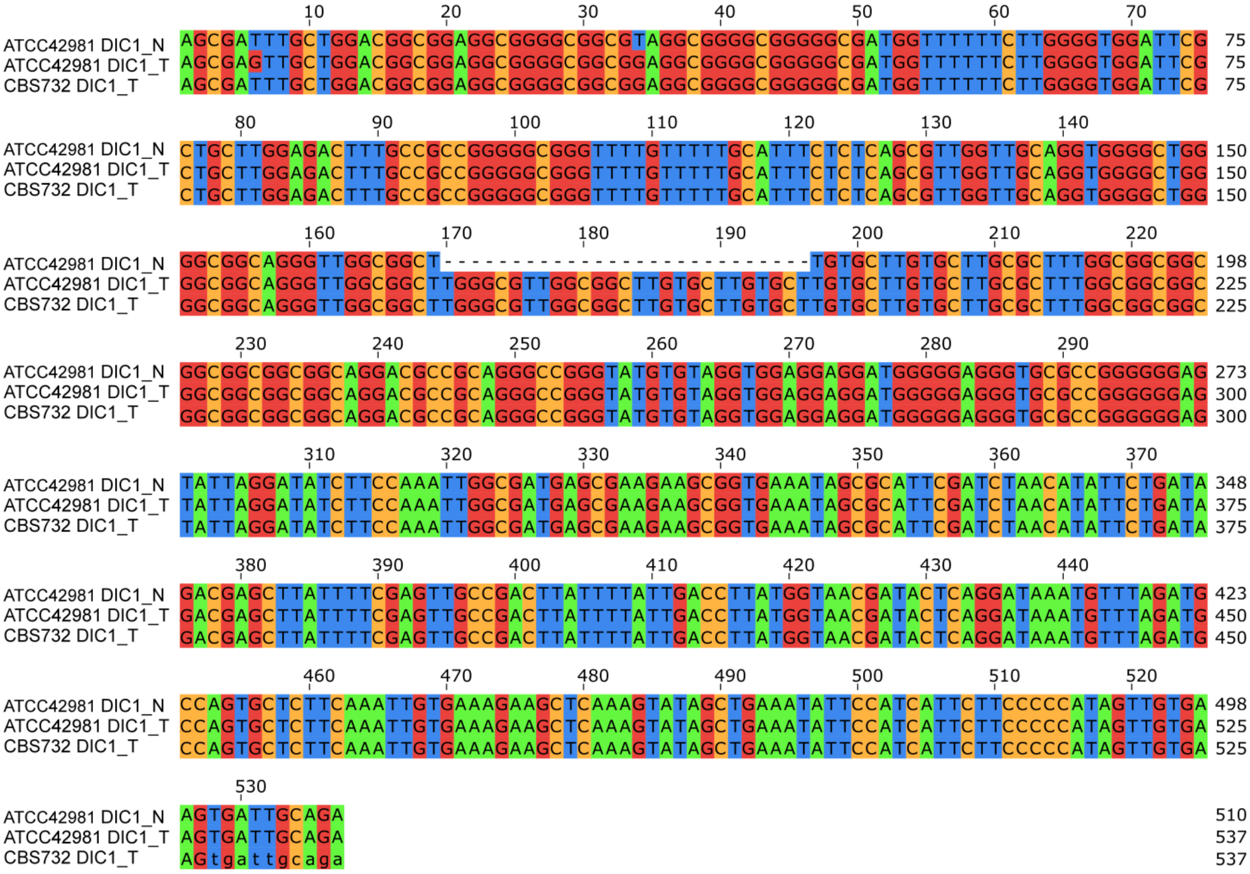


Figure S4. Cartoon illustration of main differences between MaSuRCA and DBG2OLC assemblies in *MTLα* loci–containing scaffolds. In panel (**A**) black horizontal bar represents size scale with tick marks for every 100 Kb. MaSuRCA and DBG2OLC scaffolds are represented as green and orange rectangles, respectively. Synteny representation around *MTLα* loci omits X and Z regions for brevity. T and P variants are depicted as filled, dot arrows, and marked with T and P superscripts, respectively. Panel (**B**) details collinear sets of homologous regions between the jcf180000000243 (r.c) MaSuRCA scaffold and either UEMZ01000013.1 or UEMZ01000003.1 (r.c) DBG2OLC scaffolds. Gene size and distance are not in scale. The yellow rectangle represents the starting point of scaffold synteny. Abbreviation: r.c., reverse complement.

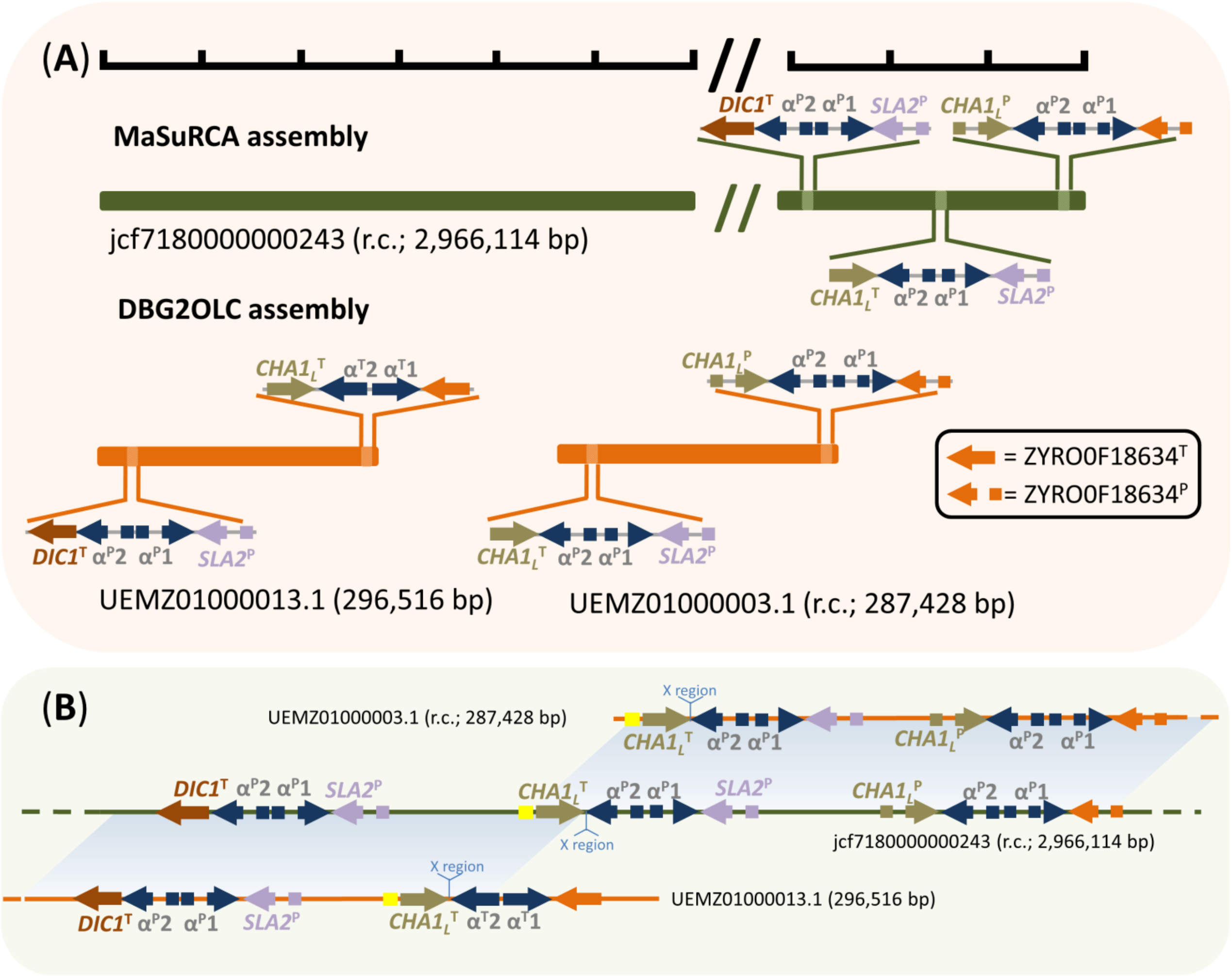
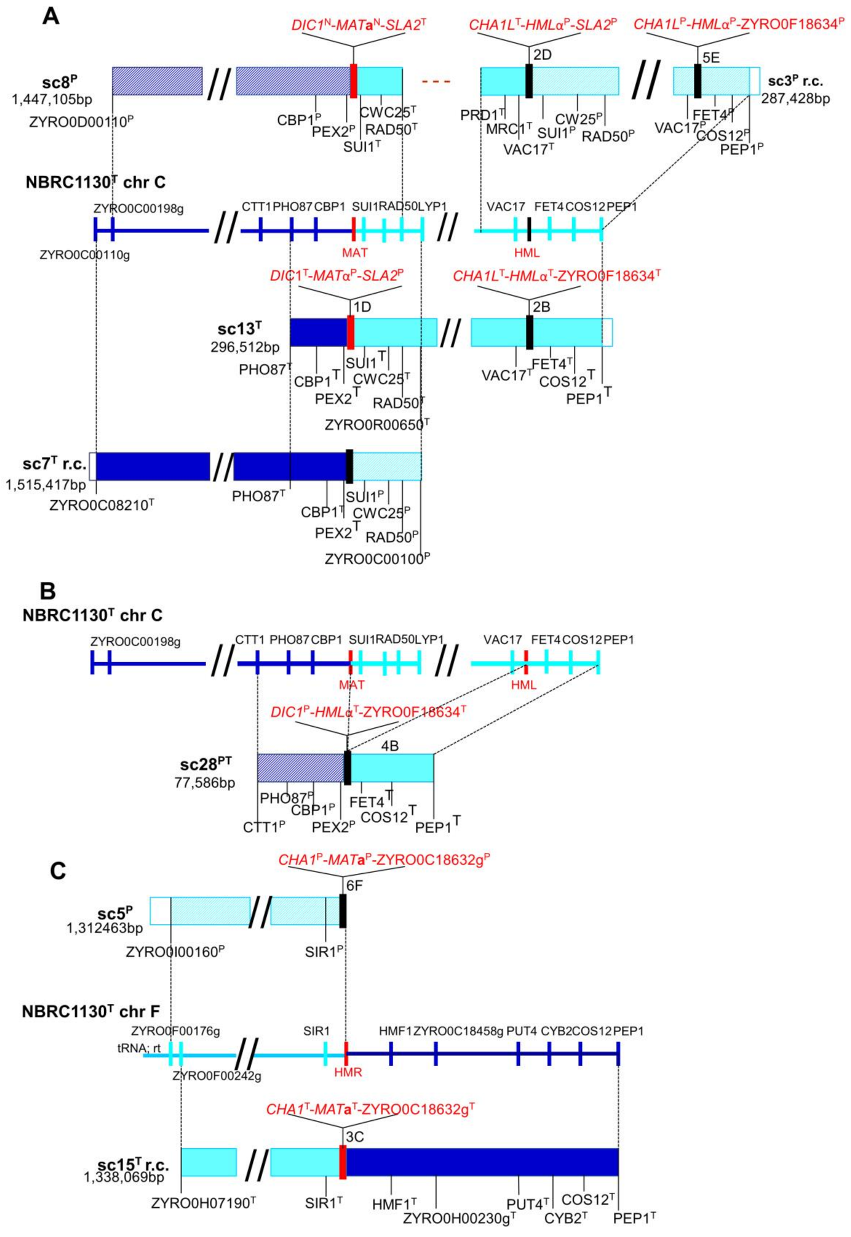

(A)
MaSuRCA assembly
$DIC1^{T}$ $\alpha^{P}2$ $\alpha^{P}1$ $SLA2^{P}$ $CHA1_{L}^{P}$ $\alpha^{P}2$ $\alpha^{P}1$
jcf7180000000243 (r.c.; 2,966,114 bp)
$CHA1_{L}^{T}$ $\alpha^{P}2$ $\alpha^{P}1$ $SLA2^{P}$
DBG2OLC assembly
$CHA1_{L}^{T}$ $\alpha^{T}2$ $\alpha^{T}1$
$CHA1_{L}^{P}$ $\alpha^{P}2$ $\alpha^{P}1$
= ZYRO0F18634^{T}
= ZYRO0F18634^{P}
$DIC1^{T}$ $\alpha^{P}2$ $\alpha^{P}1$ $SLA2^{P}$
$CHA1_{L}^{T}$ $\alpha^{P}2$ $\alpha^{P}1$ $SLA2^{P}$
UEMZ01000013.1 (296,516 bp)
UEMZ01000003.1 (r.c.; 287,428 bp)
(B)
UEMZ01000003.1 (r.c.; 287,428 bp)
X region
$CHA1_{L}^{T}$ $\alpha^{P}2$ $\alpha^{P}1$ $CHA1_{L}^{P}$ $\alpha^{P}2$ $\alpha^{P}1$
$DIC1^{T}$ $\alpha^{P}2$ $\alpha^{P}1$ $SLA2^{P}$ $CHA1_{L}^{T}$ $\alpha^{P}2$ $\alpha^{P}1$ $SLA2^{P}$ $CHA1_{L}^{P}$ $\alpha^{P}2$ $\alpha^{P}1$
X region
jcf7180000000243 (r.c.; 2,966,114 bp)
X region
UEMZ01000013.1 (296,516 bp)
$DIC1^{T}$ $\alpha^{P}2$ $\alpha^{P}1$ $SLA2^{P}$ $CHA1_{L}^{T}$ $\alpha^{T}2$ $\alpha^{T}1$

Figure S5. Cartoon detailing gene order and synteny blocks around *MTL* loci in ATCC42981_R. Rectangles represent regions in ATCC42981_R scaffolds that are collinear with a region of the reference strain NBRC1130^T ancestral genome. Blue and light blue colours correspond to regions that in CBS732^T genome are on chromosomes C and F, respectively. Solid rectangles and rectangles with diagonals represent T- and P-sequences, respectively. Scaffold (sc) numbers referred to the DBG2OLC genome assembly deposited in European Nucleotide Archive under accession number PRJEB26771 (Bizzarri et al., 2018); for simplicity the last number of ENA code marked each scaffold (*i.e*. UEMZ01000028.1 in short sc28). Genes from T and P-subgenomes are marked with T and P superscripts, respectively, while *DIC1* and *MAT*a2 new variants with N superscript. Scaffolds are not in scale.



REFERENCES

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., et al. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252, 1651–6. doi:10.1126/SCIENCE.2047873.

Alani, E., Cao, L., and Kleckner, N. (1987). A method for gene disruption that allows repeated use of *URA3* selection in the construction of multiply disrupted yeast strains. *Genetics* 116, 541–545. doi:10.1534/genetics.112.541.test.

Albertin, W., and Marullo, P. (2012). Polyploidy in fungi: evolution after whole-genome duplication. *Proc. R. Soc. B Biol. Sci.* 279, 2497–2509. doi:10.1098/rspb.2012.0434.

Albertin, W., Marullo, P., Aigle, M., Bourgais, A., Bely, M., Dillmann, C., et al. (2009). Evidence for autotetraploidy associated with reproductive isolation in *Saccharomyces cerevisiae*: towards a new domesticated species. *J. Evol. Biol.* 22, 2157–2170. doi:10.1111/j.1420-9101.2009.01828.x.

Alby, K., Schaefer, D., and Bennett, R. J. (2009). Homothallic and heterothallic mating in the opportunistic pathogen *Candida albicans*. *Nature* 460, 890–893. doi:10.1038/nature08252.

Aleksey V. Zimin, Daniela Puiu, Ming-Cheng Luo, Tingting Zhu, Sergey Koren, James A.Yorke, Jan Dvorak, S. L. S. (2016). Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a grogenitor of bread wheat, with the mega-reads algorithm. *BioRxiv*, 1–12. doi:10.1101/066100.

Alipaz, J. A., Karr, T. L., and Wu, C. I. (2005). Evolution of sexual isolation in laboratory populations: fitness differences between mating types and the associated hybrid incompatibilities. *Am. Nat.* 165, 429–438. doi:10.1086/428407.

Alper, H., Fischer, C., Nevoigt, E., Stephanopoulos, G., Demasi, J., Huh, K., et al. (2006). Tuning genetic control through promoter engineering. *Pnas*. 103, 3006–3007. doi:10.1073/pnas.0507062103.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI- BLAST: a new generation of protein database search programs. *Nucleic acids Res* 25, 3389–3402. doi:10.1093/nar/25.17.3389.

Ampe, F., ben Omar, N., Moizan, C., Wacher, C., and Guyot, J. P. (1999). Polyphasic study of the spatial distribution of microorganisms in Mexican pozol, a fermented maize dough, demonstrates the need for cultivation-independent methods to investigate traditional fermentations. *Appl. Environ. Microbiol.* 65, 5464–73. doi:10.1128/aem.00451-09.

Anderson, J. B., Sirjusingh, C., and Ricker, N. (2004). Haploidy, diploidy and evolution of antifungal drug resistance in *Saccharomyces cerevisiae*. *Genetics* 168, 1915–1923. doi:10.1534/genetics.104.033266.

Anderson, J. S., Forman, M. D., Modleski, S., Dahlquist, F. W., and Baxter, S. M. (2000). Cooperative ordering in homeodomain-dna recognition: solution structure and dynamics of the *MAT***a**1 homeodomain. *Biochemistry* 39, 10045–10054.

Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. *N. Biotechnol.* 25, 195–203. doi:10.1016/J.NBT.2008.12.009.

Apweiler, R., Martin, M. J., O'Donovan, C., Magrane, M., Alam-Faruque, Y., Antunes, R., et al. (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 40, 71–75. doi:10.1093/nar/gkr981.

Araki, H., Jearnpipatkul, A., Tatsumi, H., Sakurai, T., Ushio, K., Muta, T., et al. (1985). Molecular and functional organization of yeast plasmid pSR1. *J. Mol. Biol.* 182, 191–203. doi:10.1016/0022-2836(85)90338-9.

Asensio, L., González, I., García, T., and Martín, R. (2008). Determination of food authenticity by enzyme-linked immunosorbent assay (ELISA). *Food Control* 19, 1–8. doi:10.1016/J.FOODCONT.2007.02.010.

Aylon, Y., and Kupiec, M. (2004). DSB repair: The yeast paradigm. *DNA Repair (Amst).* 3, 797–815. doi:10.1016/j.dnarep.2004.04.013.

Baker Brachmann, C., Davies, A., Cost, G. J., Caputo, E., Li, J., Hieter, P., et al. (1998). Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: A useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* 14, 115–132. doi:10.1002/(sici)1097-0061(19980130)14:2<115::aid-yea204>3.0.co;2-2.

Baker, C. R., Booth, L. N., Sorrells, T. R., and Johnson, A. D. (2012). Protein modularity, cooperative binding, and hybrid regulatory states underlie transcriptional network diversification. Cell, 151(1), 80-95. doi: 10.1016/j.cell.2012.08.018

Baker, C. R., Tuch, B. B., and Johnson, A. D. (2011). Extensive DNA-binding specificity divergence of a conserved transcription regulator. *Proc. Natl. Acad. Sci.* 108, 7493–7498. doi:10.1073/pnas.1019177108.

Bakhrat, A., Jurica, M. S., Stoddard, B. L., and Raveh, D. (2004). Homology modeling and mutational analysis of ho endonuclease of yeast. *Genetics* 166, 721–728. doi:10.1534/genetics.166.2.721.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi:10.1089/cmb.2012.0021.

Barsoum, E., Rajaei, N., and Åstro, S. U. (2011). RAS/Cyclic AMP and transcription factor Msn2 regulate mating and mating-type switching in the yeast *Kluyveromyces lactis*. *Eukaryotic Cell*. 10, 1545–1552. doi:10.1128/EC.05158-11.

Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F., and Cullin, C. (1993). A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae*. *Nucl. Acids Res.* 21, 3329–3330. doi:10.1093/nar/21.14.3329.

Bayes, J. J., and Malik, H. S. (2009). Altered heterochromatin binding by a hybrid sterility protein in *Drosophila* sibling species. *Science.* 326, 1538–1541.

Belfort, M., and Roberts, R. J. (1997). Homing endonucleases: keeping the house in order. *Nucleic Acids Res.* 25, 3379–3388. doi:10.1093/nar/25.17.3379.

Bentley, D. R. (2006). Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.* 16, 545–552. doi:10.1016/J.GDE.2006.10.009.

Bianchi, M. M., Brambilla, L., Protani, F., Liu, C. L., Lievense, J., and Porro, D. (2001). Efficient homolactic fermentation by *Kluyveromyces lactis* strains defective in pyruvate utilization and transformed with the heterologous *LDH* gene. *Appl. Environ. Microbiol.* 67, 5621–5. doi:10.1128/AEM.67.12.5621-5625.2001.

Billiard, S., López-Villavicencio, M., Hood, M. E., And Giraud, T. (2012). Sex, outcrossing and mating types: unsolved questions in fungi and beyond. *J. Evol. Biol.* 25, 1020–1038. doi:10.1111/j.1420-9101.2012.02495.x.

Birky, C. W. (1996). Heterozygosity, heteromorphy, and phylogenetic trees in asexual eukaryotes. *Genetics* 144, 427–437.

Bizzarri, M., Cassanelli, S., and Solieri, L. (2018). Mating-type switching in CBS 732^T derived subcultures unveils potential genetic and phenotypic novelties in haploid *Zygosaccharomyces rouxii*. *FEMS Microbiol. Lett.* 365, 1–8. doi:10.1093/femsle/fnx263.

Bizzarri, M., Cassanelli, S., Pryszcz, L. P., Gawor, J., Gromadka, R., and Solieri, L. (2018). Draft genome sequences of the highly halotolerant strain *Zygosaccharomyces rouxii* ATCC 42981 and the novel allodiploid strain *Zygosaccharomyces sapae* ATB301^T obtained using the MinION Platform. *Microbiol. Res. Announc*. 7(4), e00874-18. doi: 10.1128/MRA.00874-18

Bizzarri, M., Giudici, P., Cassanelli, S., and Solieri, L. (2016). Chimeric sex-determining chromosomal regions and dysregulation of cell-type identity in a sterile *Zygosaccharomyces* allodiploid yeast. *Plos One*. 1–23. doi:10.1371/journal.pone.0152558.

Björkqvist, S., Ansell, R., Adler, L., Lidén, G., Stahl, U., and Stephanopoulos, G. (1997). Physiological response to anaerobicity of glycerol-3-phosphate dehydrogenase mutants of *Saccharomyces cerevisiae*. *Appl. Environ. Microbiol.* 63, 128–32. doi:10.1128/aem.00530-06.

Blazeck, J., Hill, A., Liu, L., Knight, R., Miller, J., Pan, A., et al. (2014). Harnessing *Yarrowia lipolytica* lipogenesis to create a platform for lipid and biofuel production. *Nat. Commun.* 5, 3131. doi:10.1038/ncomms4131.

Blazeck, J., Liu, L., Redden, H., and Alper, H. (2011). Tuning gene expression in *Yarrowia lipolytica* by a hybrid promoter approach. *Appl. Environ. Microbiol.* 77, 7905–7914. doi:10.1128/AEM.05763-11.

Bligh, H. F. J. (2000). Detection of adulteration of Basmati rice with non-premium long-grain rice. *Int. J. Food Sci. Technol.* 35, 257–265. doi:10.1046/j.1365-2621.2000.00390.x.

Boeke, J. D., La Croute, F., and Fink, G. R. (1984). A positive selection for mutants lacking orotidine-5′-phosphate decarboxylase activity in yeast: 5-fluoro-orotic acid resistance. *MGG Mol. Gen. Genet.* 197, 345–346. doi:10.1007/BF00330984.

Boeke, J.D., Garfinkel, D.J., Styles, C.A., and Fink, G.R. (1985). Ty elements transpose through an RNA intermediate. *Cell* 40:491–500. https://doi.org/10.1016/0092-8674(85)90197-7.

Boisnard, S., Li, Y. Z., Arnaise, S., and Sequeira, G. (2015). Efficient mating-type switching in *Candida glabrata* induces cell death. 1–18. doi:10.1371/journal.pone.0140990.

Bond, U., Neal, C., Donnelly, D., and James, T. C. (2004). Aneuploidy and copy number breakpoints in the genome of lager yeasts mapped by microarray hybridisation. *Curr. Genet.* 45, 360–370. doi:10.1007/s00294-004-0504-x.

Booth, L. N., Tuch, B. B., and Johnson, A. D. (2010). Intercalation of a new tier of transcription regulation into an ancient circuit. *Nature* 468, 959–963. doi:10.1038/nature09560.

Borneman, A. R., Chambers, P. J., and Pretorius, I. S. (2007). Yeast systems biology: modelling the winemaker's art. *Trends Biotechnol.* 25, 349–355. doi:10.1016/J.TIBTECH.2007.05.006.

Boynton, P. J., Janzen, T., and Greig, D. (2018). Modeling the contributions of chromosome segregation errors and aneuploidy to *Saccharomyces* hybrid sterility. *Yeast*. 35, 85–98. doi: 10.1002/yea.3282

Braslavsky, I., Hebert, B., Kartalov, E., and Quake, S. R. (2003). Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci. U. S. A.* 100, 3960–4. doi:10.1073/pnas.0230489100.

Braun, B. R., Van het Hoog, M., d'Enfert, C., Martchenko, M., Dungan, J., Kuo, A., et al. (2005). A human-curated annotation of the *Candida albicans* genome. *PLoS Genet.* 1, e1. doi:10.1371/journal.pgen.0010001.

Braun-Galleani, S., Ortiz-Merino, R. A., Wu, Q., Xu, Y., and Wolfe, K. H. (2018). *Zygosaccharomyces pseudobailii*, another yeast interspecies hybrid that regained fertility by damaging one of its *MAT* loci. *FEMS Yeast Res.*, 1–9. doi:10.1093/femsyr/foy079.

Brideau, N. J., Flores, H. a, Wang, J., Maheshwari, S., Wang, X., and Barbash, D. a (2006). Two Dobzhansky-Muller genes interact to cause hybrid lethality in *Drosophila*. *Science*. 314, 1292–1295.

Brierley, R. A. (1998). Secretion of recombinant human Insulin-Like Growth Factor I (IGF-I). *Pichia Protocols*. 149–178. doi:10.1385/0-89603-421-6:149.

Bryan, G. J., Dixon, A., Gale, M. D., and Wiseman, G. (1998). A PCR-based method for the detection of hexaploid bread wheat adulteration of durum wheat and pasta. *J. Cereal Sci.* 28, 135–145. doi:10.1006/JCRS.1998.0182.

Butler, G., Kenny, C., Fagan, A., Kurischko, C., Gaillardin, C., and Wolfe, K. H. (2004). Evolution of the *MAT* locus and its Ho endonuclease in yeast species. *Proc. Natl. Acad. Sci. U. S. A.* 101, 1632–1637. doi:10.1073/pnas.0304170101.

Byrne, K. P., and Wolfe, K. H. (2006). Visualizing syntenic relationships among the hemiascomycetes with the Yeast Gene Order Browser. *Nucleic acids res*. 34, D452-D455. doi:10.1093/nar/gkj041.

Campbell, M. A., Ganley, A. R., Gabaldon, T., and Cox, M. P. (2016). The case of the missing ancient fungal polyploids. *Am. Nat*. 188, 602-614. doi: 10.1086/688763

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., et al. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 18, 188–96. doi:10.1101/gr.6743907.

Cao, X., Hou, L., Lu, M., and Wang, C. (2009). Improvement of soy-sauce flavour by genome shuffling in *Candida versatilis* to improve salt stress resistance. *Int. J. Food Sci. Technol.* 45, 17–22. doi:10.1111/j.1365-2621.2009.02085.x.

Cassanelli, S., Bizzarri, M., and Solieri, L. (2016). Recent advances in understanding yeast genetics of sex determination. *Fungal Genomics & Biology*, *6*(1), 1-3. doi: 10.4172/2165-8056.1000e122.

Chattoo, B. B., Sherman, F., Azubalis, D. A., Fjellstedt, T. A., Mehnert, D., and Ogur, M. (1979). Selection of lys2 mutants of the yeast *Sccharomyces cerevisiae* by the utilization of α-aminoadipate. *Genetics.* 93.1: 51-65

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). Fastp: an ultra-fast all-in-one FASTQ preprocessor. *BioRxiv*, 274100.

Chen, Y., and Gartenberg, M. R. (2015). Epigenetics: Silencing sounds off. *eLife*, 2015, 4: e06717. doi:10.7554/eLife.05007.

Chen, Y., and Nielsen, J. (2013). Advances in metabolic pathway and strain engineering paving the way for sustainable production of chemical building blocks. *Curr. Opin. Biotechnol.* 24, 965–972. doi:10.1016/j.copbio.2013.03.008.

Chien, C. Ting, Buck, S., Sternglanz, R., and Shore, D. (1993). Targeting of SIR1 protein establishes transcriptional silencing at HM loci and telomeres in yeast. *Cell* 75, 531–541. doi:10.1016/0092-8674(93)90387-6.

Chopra, R., Sharma, V. M., and Ganesan, K. (1999). Elevated growth of *Saccharomyces cerevisiae* ATH1 null mutants on glucose is an artifact of nonmatching auxotrophies of mutant and reference strains. *Appl. Environ. Microbiol.* 65, 2267–8. doi:10.1128/aem.68.5.2095-2100.2002.

Cicardi, M., Levy, R. J., McNeil, D. L., Li, H. H., Sheffer, A. L., Campion, M., et al. (2010). Ecallantide for the treatment of acute attacks in hereditary angioedema. *N. Engl. J. Med.* 363, 523–531. doi:10.1056/NEJMoa0905079.

Coenye, T., Gevers, D., de Peer, Y. Van, Vandamme, P., and Swings, J. (2005). Towards a prokaryotic genomic taxonomy. *FEMS Microbiol. Rev.* 29, 147–167. doi:10.1016/j.fmrre.2004.11.004.

Cole, C., Barber, J. D., and Barton, G. J. (2008). The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.* 36, W197–W201. doi:10.1093/nar/gkn238.

Comitini, F., Capece, A., Ciani, M., and Romano, P. (2017). New insights on the use of wine yeasts. *Curr. Opin. Food Sci.* 13, 44–49. doi:10.1016/j.cofs.2017.02.005.

Consortium, I. H. G. S. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi:10.1038/35057062.

Coppin, E., Debuchy, R., Arnaise, S., and Picard, M. (1997). Mating types and sexual development in filamentous ascomycetes. *Microbiol. Mol. Biol. Rev.* 61, 411–428.

Costello, C. A., Payson, R. A., Menke, M. A., Larson, J. L., Brown, K. A., Tanner, J. E., et al. (2000). Purification, characterization, cDNA cloning and expression of a novel ketoreductase from *Zygosaccharomyces rouxii*. *Eur. J. Biochem.* 267, 5493–5501. doi:10.1046/j.1432-1327.2000.01608.x.

Covitz, P. a., Herskowitz, I., and Mitchell, A. P. (1991). The yeast *RME1* gene encodes a putative zinc finger protein that is directly repressed by **a**1-α2. *Genes Dev.* 5, 1982–1989. doi:10.1101/gad.5.11.1982.

Cox, H., Mead, D., Sudbery, P., Eland, R. M., Mannazzu, I., and Evans, L. (2000). Constitutive expression of recombinant proteins in the methylotrophic yeast *Hansenula polymorpha* using the P*MAI* promoter. *Yeast* 16, 1191–1203. doi:10.1002/1097-0061(20000930)16:13<1191::AID-YEA589>3.0.CO;2-2.

Coyne, J. A., and Orr, H. A. (2004). Speciation. Sunderland, MA.

Cullen, P. J. (2012). The regulation of filamentous growth in yeast. 190, 23–49. doi:10.1534/genetics.111.127456.

Dakal, T. C., Giudici, P., and Solieri, L. (2016). Contrasting patterns of rDNA homogenization within the *Zygosaccharomyces rouxii* species complex. *PloS One*, *11*(8), e0160744. doi: 10.1371/journal.pone.0160744.

Dakal, T. C., Solieri, L., and Giudici, P. (2014). Adaptive response and tolerance to sugar and salt stress in the food yeast *Zygosaccharomyces rouxii*. *Int. J. Food Microbiol.* 185, 140–157. doi:10.1016/J.IJFOODMICRO.2014.05.015.

Daniels, K. J., Lockhart, S. R., Staab, J. F., Sundstrom, P., and Soll, D. R. (2003). The adhesin Hwp1 and the first daughter cell localize to the **a**/**a** portion of the conjugation bridge during *Candida albicans* mating. *Mol. Biol. Cell* 14, 4920–4930. doi:10.1091/mbc.E03-04-0264.

Dave, M. N., and Chattoo, B. B. (1997). A counter-selectable marker for genetic transformation of the yeast *Schwanniomyces alluvius*. *Appl. Microbiol. Biotechnol.* 48, 204–207. doi:10.1007/s002530051039.

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*. doi:10.1093/bioinformatics/bty149.

De Lannoy, C., de Ridder, D., and Risse, J. (2017). The long reads ahead: de novo genome assembly using the MinION. *F1000Research* 6, 1083. doi:10.12688/f1000research.12012.2.

De Montigny, J., Straub, M., Potier, S., Tekaia, F., Dujon, B., Wincker, P., et al. (2000). Genomic exploration of the hemiascomycetous yeasts: 8. *Zygosaccharomyces rouxii*. *FEBS Lett.* 487, 52–55. doi:10.1016/S0014-5793(00)02279-1.

Del Angel, D. V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Pettersson, O. V., et al. (2018). Ten steps to get started in Genome Assembly and Annotation. F1000Research. 7: ELIXIR-148. doi: 10.12688/f1000research.13598.1

DeLisi, C. (2008). Santa Fe 1986 : Human genome baby-steps. *Nature*. 455:876–7.

Delneri, D., Tomlin, G. C., Wixon, J. L., Hutter, A., Sefton, M., Louis, E. J., et al. (2000). Exploring redundancy in the yeast genome: an improved strategy for use of the *Cre-loxP* system. *Gene* 252, 127–135. doi:10.1016/S0378-1119(00)00217-1.

Dennison, P. M. J., Ramsdale, M., Manson, C. L., and Brown, A. J. P. (2005). Gene disruption in *Candida albicans* using a synthetic, codon-optimised Cre-*loxP* system. *Fungal Genet. Biol.* 42, 737–748. doi:10.1016/j.fgb.2005.05.006.

Dettman, J. R., Sirjusingh, C., Kohn, L. M., and Anderson, J. B. (2007). Incipient speciation by divergent adaptation and antagonistic epistasis in yeast. *Nature* 447, 585–588.

Di Rienzi, S. C., Lindstrom, K. C., Lancaster, R., Rolczynski, L., Raghuraman, M. K., and Brewer, B. J. (2011). Genetic, genomic, and molecular tools for studying the protoploid yeast, *L. waltii*. *Yeast* 28, 137–151. doi:10.1002/yea.1826.

Dobzhansky, T. (1933). On the sterility of the interracial hybrids in *Drosophila pseudoobscura*. *Proc. Natl. Acad. Sci. U. S. A.* 19, 397–403. doi:10.1073/PNAS.19.4.397.

Domergue, R., Castaño, I., De Las Peñas, A., Zupancic, M., Lockatell, V., Hebel, J. R., et al. (2005). Nicotinic acid limitation regulates silencing of *Candida* adhesins during UTI. *Science* 308, 866–70. doi:10.1126/science.1108640.

Dominguez Del Angel, V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Vinnere Pettersson, O., et al. (2018). Ten steps to get started in Genome Assembly and Annotation. *F1000Research* 7, 148. doi:10.12688/f1000research.13598.1.

Drillon, G., Carbone, A., and Fischer, G. (2014). SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. *PLoS One* 9, e92621. doi:10.1371/journal.pone.0092621.

Drocourt, D., Calmels, T., Reynes, J. P., Baron, M., and Tiraby, G. (1990). Cassettes of the *Streptoalloteichus hindustanus ble* gene for transformation of lower and higher eukaryotes to phleomycin resistance. *Nucleic Acids Res.* 18, 4009. doi:10.1093/nar/18.13.4009.

Dujon, B. (2006). Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends Genet.* 22, 375–387. doi:10.1016/j.tig.2006.05.007.

Dujon, B. (2010). Yeast evolutionary genomics. *Nat. Rev. Genet.* 11, 512–524. doi:10.1038/nrg2811.

Dujon, B. A., and Louis, E. J. (2017). Genome diversity and evolution in the budding yeasts (Saccharomycotina). *Genetics* 206, 717–750. doi:10.1534/genetics.116.199216.

Dunn, B., Paulish, T., Stanbery, A., Piotrowski, J., Koniges, G., Kroll, E., et al. (2013). Recurrent rearrangement during adaptive evolution in an interspecific yeast hybrid suggests a model for rapid introgression. *PLoS Genet* 9, e1003366.

Dutreux, F., Da Silva, C., d'Agata, L., Couloux, A., Gay, E. J., Istace, B., et al. (2018). *De novo* assembly and annotation of three *Leptosphaeria* genomes using Oxford Nanopore MinION sequencing. *Sci. Data* 5, 180235. doi:10.1038/sdata.2018.235.

Elena, S. F., and Lenski, R. E. (2003). Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat. Rev. Genet.* 4, 457–469. doi:10.1038/nrg1088.

Ellahi, A., and Rine, J. (2016). Evolution and functional trajectory of Sir1 in gene silencing. *Mol. Cell. Biol.* 823, MCB.01013-15. doi:10.1128/MCB.01013-15.

Erdman, S., and Snyder, M. (2001). A filamentous growth response mediated by the yeast mating pathway. *Genetics*. 159.3: 919-928.

Fabre, E., Muller, H., Therizols, P., Lafontaine, I., Dujon, B., Fairhead, C., et al. (2004). Evolution of sexual isolation in laboratory populations: fitness differences between mating types and the associated hybrid incompatibilities. *Nucleic Acids Res.* 13, 1726–1733. doi:10.1038/nature05099.

Fabre, E., Muller, H., Therizols, P., Lafontaine, I., Dujon, B., and Fairhead, C. (2005). Comparative genomics in hemiascomycete yeasts: evolution of sex, silencing, and subtelomeres. *Mol. Biol. Evol.* 22, 856–873. doi:10.1093/molbev/msi070.

Farris, M. H., and Olson, J. B. (2007). Detection of *Actinobacteria* cultivated from environmental samples reveals bias in universal primers. *Lett. Appl. Microbiol.* 45, 376–381. doi:10.1111/j.1472-765X.2007.02198.x.

Fickers, P., Le Dall, M. T., Gaillardin, C., Thonart, P., and Nicaud, J. M. (2003). New disruption cassettes for rapid gene disruption and marker rescue in the yeast *Yarrowia lipolytica*. *J. Microbiol. Methods* 55, 727–737. doi:10.1016/j.mimet.2003.07.003.

Finn, R. D., Bateman, a., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families' database. *Nucleic Acids Res.* 42, D222–D230. doi:10.1093/nar/gkt1223.

Fischer, G., Neuvéglise, C., Durrens, P., Gaillardin, C., and Dujon, B. (2001). Evolution of gene order in the genomes of two related yeast species. *Genome Res.* 11, 2009–19. doi:10.1101/gr.212701.

Flagfeldt, D. B., Siewers, V., Huang, L., and Nielsen, J. (2009). Characterization of chromosomal integration sites for heterologous gene expression in Saccharomyces cerevisiae. *Yeast* 26, 545–551. doi:10.1002/yea.

Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenza*. *Science.* 269, 496–512. doi:10.1126/science.7542800.

Foissac, S., Gouzy, J., Rombauts, S., Mathé, C., Amselem, J., Sterck, L., et al. (2008). Genome annotation in plants and fungi: EuGene as a model platform. *Current Bioinformatics*, *3*(2), 87-97. doi:10.2174/157489308784340702.

Förster, A., Aurich, A., Mauersberger, S., and Barth, G. (2007). Citric acid production from sucrose using a recombinant strain of the yeast *Yarrowia lipolytica*. *Appl. Microbiol. Biotechnol.* 75, 1409–1417. doi:10.1007/s00253-007-0958-0.

Fraatz, M. A., Rühl, M., and Zorn, H. (2013). Food and Feed Enzymes. Springer, Berlin, Heidelberg. 229–256. doi:10.1007/10_2013_235.

Fraser, J. A, Giles, S. S., Wenink, E. C., Geunes-Boyer, S. G., Wright, J. R., Diezmann, S., et al. (2005). Same-sex mating and the origin of the Vancouver Island *Cryptococcus gattii* outbreak. *Nature* 437, 1360–1364. doi:10.1038/nature04220.

Friedrich, A., Jung, P., Reisser, C., Fischer, G., and Schacherer, J. (2015). Population genomics reveals chromosome-scale heterogeneous evolution in a protoploid yeast. *Mol. Biol. Evol.* 32, 184–192. doi:10.1093/molbev/msu295.

Gabaldón, T., Martin, T., Marcet-houben, M., Durrens, P., Bolotin-Fukuhara, M., Lespinet, O., et al. (2013). Comparative genomics of emerging pathogens in the *Candida glabrata* clade. *BMC genomics*. 14.1: 623. doi:10.1186/1471-2164-14-623.

Galagan, J. E., Henn, M. R., Ma, L.-J., Cuomo, C. A., and Birren, B. (2005). Genomics of the fungal kingdom: insights into eukaryotic biology. *Genome Res.* 15, 1620–31. doi:10.1101/gr.3767105.

Galgoczy, D. J., Cassidy-Stone, A., Llinas, M., O'Rourke, S. M., Herskowitz, I., DeRisi, J. L., et al. (2004). Genomic dissection of the cell-type-specification circuit in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci.* 101, 18069–18074. doi:10.1073/pnas.0407611102.

Gallagher J. E. G., Babiarz J. E and Teytelman L., Wolfe K. H., and Rine J. (2009). Elaboration, diversification and regulation of the Sir1 family of silencing proteins in *Saccharomyces*. *Genetics*. 181, 1477–1491. doi: 10.1534/genetics.108.099663

Galperin, M. Y., and Koonin, E. V. (1998). *In silico biology.* 1.1: 55-67.

Garraway, L. A., Tosi, L. R. ., Wang, Y., Moore, J. B., Dobson, D. E., and Beverley, S. M. (1997). Insertional mutagenesis by a modified in vitro Ty1 transposition system. *Gene* 198, 27–35. doi:10.1016/S0378-1119(97)00288-6.

Gatignol, A., Baron, M., and Tiraby, G. (1987). Phleomycin resistance encoded by the ble gene from transposon Tn5 as a dominant selectable marker in *Saccharomyces cerevisiae*. *MGG Mol. Gen. Genet.* 207, 342–348. doi:10.1007/BF00331599.

Geisberg, J. V., Moqtaderi, Z., Fan, X., Ozsolak, F., and Struhl, K. (2014). Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast. *Cell* 156, 812–824. doi:10.1016/j.cell.2013.12.026.

Gelfand, B., Mead, J., Bruning, A., Apostolopoulos, N., Tadigotla, V., Nagaraj, V., et al. (2011). Regulated antisense transcription controls expression of cell-type-specific genes in yeast. *Mol. Cell. Biol.* 31, 1701–1709.

Gibson, D. G., Glass, J. I., Lartigue, C., Noskov, V. N., Chuang, R.-Y., Algire, M. a, et al. (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 329, 52–56. doi:10.1126/science.1190719.

Gietz, R. D., and Schiestl, R. H. (2007). Quick and easy yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* 2, 35–37. doi:10.1038/nprot.2007.14.

Giordano, F., Aigrain, L., Quail, M. A., Coupland, P., Bonfield, J. K., Davies, R. M., et al. (2017). *De novo* yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Sci. Rep.* 7, 1–10. doi:10.1038/s41598-017-03996-z.

Giudici, P., Solieri, L., Pulvirenti, A. M., and Cassanelli, S. (2005). Strategies and perspectives for genetic improvement of wine yeasts. *Appl Microbiol Biotechnol.* 66(6), 622-628. doi:10.1007/s00253-004-1784-2.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U. S. A.* 108, 1513–8. doi:10.1073/pnas.1017351108.

Goossens, K. V. Y., Ielasi, F. S., Nookaew, I., Stals, I., Alonso-sarduy, L., Daenen, L., et al. (2015). Molecular mechanism of flocculation self-recognition in yeast and its role in mating and survival. *mBio*. 6, 1–16. doi:10.1128/mBio.00427-15.

Gopal, C., Broad, D., and Lloyd, D. (1989). Bioenergetic consequences of protein overexpression in *Saccharomyces cerevisiae*. *Appl. Microbiol. Biotechnol.* 30, 160–165. doi:10.1007/BF00264005.

Gordon J. L., and Wolfe, K. H. (2008). Recent allopolyploid origin of *Zygosaccharomyces rouxii* strain ATCC 42981*. Yeast*. 25, 449-456. doi:10.1002/yea.1598

Gordon, J. L., Armisen, D., Proux-Wera, E., OhEigeartaigh, S. S., Byrne, K. P., and Wolfe, K. H. (2011). Evolutionary erosion of yeast sex chromosomes by mating-type switching accidents. *Proc. Natl. Acad. Sci.* 108, 20024–20029. doi:10.1073/pnas.1112808108.

Goutte, C., and Johnson, A. D. (1988). **a**1 protein alters the DNA binding of α2 repressor specificity. *Cell*. 52, 875–882.

Granek, J. a., Kayıkçı, Ö., and Magwene, P. M. (2011). Pleiotropic signaling pathways orchestrate yeast development. *Curr. Opin. Microbiol.* 14, 676–681. doi:10.1016/j.mib.2011.09.004.

Greig, D., Borts, R. H., Louis, E. J., and Travisano, M. (2002). Epistasis and hybrid sterility in *Saccharomyces. Proc. R. Soc. B Biol. Sci.* 269, 1167–1171. doi:10.1098/rspb.2002.1989.

Gritz, L., and Davies, J. (1983). Plasmid-encoded hygromycin B resistance: the sequence of hygromycin B phosphotransferase gene and its expression in *Escherichia coli* and *Saccharomyces cerevisiae*. *Gene* 25, 179–188. doi:10.1016/0378-1119(83)90223-8.

Güldener, U., Heck S., Fiedler T., Beinhauer J., and Hegemann, J. H. (1996). A new efficient gene disruption cassette for repeated use in budding yeast. Nucleic Acids Res. 24, 2519-2524. doi: 10.1093/nar/24.13.2519

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075. doi:10.1093/bioinformatics/btt086.

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith Jr, R. K., Hannick, L. I., et al. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654–5666. doi:10.1093/nar/gkg770.

Haber J. E. and Wolfe K. H. (2005). Function and evolution of *HO* and *VDE* endonucleases in fungi. In: *Homing Endonucleases and Inteins*. Springer, Berlin, Heidelberg. p. 161-175.

Haber, J. E. (1998). Mating-type gene switching in *Saccharomyces cerevisiae*. *Annu. Rev. Genet.* 32, 561–599. doi:10.1146/annurev.genet.32.1.561.

Haber, J. E. (2012). Mating-type genes and *MAT* switching in *Saccharomyces cerevisiae*. *Genetics* 191, 33–64. doi:10.1534/genetics.111.134577.

Haber, J. E., and George, J. P. (1979). A mutation that permits the expression of normally silent copies of mating-type information in *Saccharomyces cerevisiae*. *Genetics* 93, 13–35.

Hadfield, C., Cashmore, A. M., and Meacock, P. A. (1986). An efficient chloramphenicol-resistance marker for *Saccharomyces cerevisiae* and *Escherichia coli*. *Gene* 45, 149–158. doi:10.1016/0378-1119(86)90249-0.

Hafez, M., Hausner, G., and Bonen, L. (2012). Homing endonucleases: DNA scissors on a mission. *Genome* 55, 553–569. doi:10.1139/g2012-049.

Hamilton, D. L., and Abremski, K. (1984). Site-specific recombination by the bacteriophage P1 lox-Cre system: Cre-mediated synapsis of two lox sites. *J. Mol. Biol.* 178, 481–486. doi:10.1016/0022-2836(84)90154-2.

Hanson, S. J., and Wolfe, K. H. (2017). An evolutionary perspective on yeast mating-type switching. *Genetics* 206, 9–32. doi:10.1534/genetics.117.202036.

Harashima, S., Miller, a M., Tanaka, K., Kusumoto, K., Mukai, Y., Nasmyth, K., et al. (1989). Mating-type control in *Saccharomyces cerevisiae*: isolation and characterization of mutants defective in repression by **a**1-α2. *Mol Cell Biol* 9, 4523–4530.

Hauck, T., Hübner, Y., Brühlmann, F., and Schwab, W. (2003). Alternative pathway for the formation of 4,5-dihydroxy-2,3-pentanedione, the proposed precursor of 4-hydroxy-5-methyl-3(2H)-furanone as well as autoinducer-2, and its detection as natural constituent of tomato fruit. *Biochim. Biophys. Acta - Gen. Subj.* 1623, 109–119. doi:10.1016/J.BBAGEN.2003.08.002.

Hedtke, S. M., Townsend, T. M., and Hillis, D. M. (2006). Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst. Biol. 55*(3), 522-529. doi:10.1080/10635150600697358.

Heim, U., Tietze, E., Weschke, W., Tschäpe, H., and Wobus, U. (1989). Nucleotide sequence of a plasmid born streptothricin-acetyl-transferase gene (sat-1). *Nucleic Acids Res. 17*(17), 7103.

Heitman, J. (2006). Sexual reproduction and the evolution of microbial pathogens. *Curr. Biol.* 16, R711–R725. doi:10.1016/J.CUB.2006.07.064.

Heitman, J. (2009). Microbial genetics: Love the one you're with. *Nature* 460, 807–808.

Henderson, R. C. A., Cox, B. S., and Tubb, R. (1985). The transformation of brewing yeasts with a plasmid containing the gene for copper resistance. *Curr. Genet.* 9, 133–138. doi:10.1007/BF00436961.

Hennequin, C., Gallaud, J., Dujon, B., Muller, H., Hennequin, C., Gallaud, J., et al. (2008). The asexual yeast *Candida glabrata* maintains distinct a and α haploid mating types. *Eukaryot. Cell* 7, 848–858. doi:10.1128/EC.00456-07.

Henson, J., Tischler, G., and Ning, Z. (2012). Next-generation sequencing and large genome assemblies. *Pharmacogenomics* 13, 901–915. doi:10.2217/pgs.12.72.

Hentges, P., Van Driessche, B., Tafforeau, L., Vandenhaute, J., and Carr, A. M. (2005). Three novel antibiotic marker cassettes for gene disruption and marker switching in *Schizosaccharomyces pombe*. *Yeast* 22, 1013–1019. doi:10.1002/yea.1291.

Heo, J., Hong, W., Cho, E., Kim, M., Kim, J., Kim, C., et al. (2003). Properties of the -derived constitutive promoter, assessed using an HSA reporter gene. *FEMS Yeast Res.* 4, 175–184. doi:10.1016/S1567-1356(03)00150-8.

Herskowitz, I. R. a, and Herskowitz, I. R. a (1988). Life cycle of the budding yeast *Saccharomyces cerevisiae*. *Microbiol. Rev.* 52, 536–553.

Hert, D. G., Fredlake, C. P., and Barron, A. E. (2008). Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis* 29, 4618–4626. doi:10.1002/elps.200800456.

Hewitt, S. K., Donaldson, I. J., Lovell, S. C., and Delneri, D. (2014). Sequencing and characterisation of rearrangements in three *S. pastorianus* strains reveals the presence of chimeric genes and gives evidence of breakpoint reuse. *PLoS One* 9. doi:10.1371/journal.pone.0092203.

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. *Eukaryot. Cell* 10, 1183–1192. doi:10.1128/EC.05123-11.

Hicks, W. M., Kim, M., and Haber, J. E. (2010). Increased mutagenesis and unique mutation signature associated with mitotic gene conversion. *Science* 329, 82–85. doi:10.1126/science.1191125.

Hicks, W. M., Yamaguchi, M., and Haber, J. E. (2011). Real-time analysis of double-strand DNA break repair by homologous recombination. *Proc. Natl. Acad. Sci. U. S. A.* 108, 3108–15. doi:10.1073/pnas.1019660108.

Hittinger, C. T., Rokas, A., and Carroll, S. B. (2004). Parallel inactivation of multiple *GAL* pathway genes and ecological diversification in yeasts. *Proc. Natl. Acad. Sci.* 101, 14144–14149. doi:10.1073/PNAS.0404319101.

Hoess, R. H., and Abremski, K. (1985). Mechanism of strand cleavage and exchange in the Cre-lox site-specific recombination system. *J. Mol. Bio.*, *181*(3), 351-362. doi:10.1016/0022-2836(85)90224-4.

Hoffman, C. S., and Winston, F. (1987). A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformaion of *Escherichia coli*. *Gene* 57, 267–272.

Hold, G. L., Russell, V. J., Pryde, S. E., Rehbein, H., Quinteiro, J., Rey-Mendez, M., et al. (2001). Validation of a PCR-RFLP based method for the identification of salmon species in food products. *Eur. Food Res. Technol.* 212, 385–389. doi:10.1007/s002170000237.

Hongay, C. F., Grisafi, P. L., Galitski, T., and Fink, G. R. (2006). Antisense transcription controls cell fate in *Saccharomyces cerevisiae*. *Cell* 127, 735–745. doi:10.1016/j.cell.2006.09.038.

Horwitz, A. A., Walter, J. M., Schubert, M. G., Kung, S. H., Hawkins, K., Platt, D. M., et al. (2015). Efficient multiplexed integration of synergistic alleles and metabolic pathways in yeasts via CRISPR-Cas. *Cell Syst.* 1, 88–96. doi:10.1016/J.CELS.2015.02.001.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., et al. (2016). EGGNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* 44, D286–D293. doi:10.1093/nar/gkv1248.

Hughes, T. R., Roberts, C. J., Dai, H., Jones, A. R., Meyer, M. R., Slade, D., et al. (2000). Widespread aneuploidy revealed by DNA microarray expression profiling. *Nat. Genet.* 25, 333–337. doi:10.1038/77116.

Hunter, N., Chambers, S. R., Louis, E. J., and Borts, R. H. (1996). The mismatch repair system contributes to meiotic sterility in an interspecific yeast hybrid. *EMBO J.* 15, 1726–1733. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC450085/.

Ira, G., Satory, D., and Haber, J. E. (2006). Conservative inheritance of newly synthesized DNA in double-strand break-induced gene conversion. *Mol. Cell. Biol.* 26, 9424–9. doi:10.1128/MCB.01654-06.

Istace, B., Friedrich, A., d'Agata, L., Faye, S., Payen, E., Beluche, O., et al. (2017). *De novo* assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *Gigascience* 6, 1–13. doi:10.1093/gigascience/giw018.

Iwaki, T., and Takegawa, K. (2004). A set of *loxP* marker cassettes for Cre-mediated multiple gene disruption in *Schizosaccharomyces pombe*. *Biosci. Biotechnol. Biochem.* 68, 545–550. doi:10.1271/bbb.68.545.

Iwaki, T., Higashida, Y., Tsuji, H., Tamai, Y., and Watanabe, Y. (1998). Characterization of a second gene (*ZSOD22*) of Na+/H+ antiporter from salt-tolerant yeast *Zygosaccharomyces rouxii* and functional expression of *ZSOD2* and *ZSOD22* in *Saccharomyces cerevisiae*. *Yeast* 14, 1167–1174. doi:10.1002/(SICI)1097-0061(19980930)14:13<1167::AID-YEA318>3.0.CO;2-5.

Iwaki, T., Tamai, Y., and Watanabe, Y. (1999). Two putative *MAP* kinase genes, *ZrHOG1* and *ZrHOG2*, cloned from the salt-tolerant yeast *Zygosaccharomyces rouxii* are functionally homologous to the *Saccharomyces cerevisiae HOG1* gene. *Microbiology* 145, 241–248. doi:10.1099/13500872-145-1-241.

Jain, M., Fiddes, I. T., Miga, K. H., Olsen, H. E., Paten, B., and Akeson, M. (2015). Improved data analysis for the MinION nanopore sequencer. *Nat. Methods* 12, 351–356. doi:10.1038/nmeth.3290.

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36, 338–345. doi:10.1038/nbt.4060.

Jain, M., Olsen, H. E., Paten, B., and Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* 17, 1–11. doi:10.1186/s13059-016-1122-x.

James, S. A., and Stratford, M. (2010). *Zygosaccharomyces*Barker (1901). *The Yeasts (Fifth Edition)* (pp. 937-947).

James, S. a., Bond, C. J., Stratford, M., and Roberts, I. N. (2005). Molecular evidence for the existence of natural hybrids in the genus *Zygosaccharomyces*. *FEMS Yeast Res.* 5, 747–755. doi:10.1016/j.femsyr.2005.02.004.

Janowicz, Z. A., Melber, K., Merckelbach, A., Jacobs, E., Harford, N., Comberbach, M., et al. (1991). Simultaneous expression of the S and L surface antigens of hepatitis B, and formation of mixed particles in the methylotrophic yeast *Hansenula polymorpha*. *Yeast* 7, 431–443. doi:10.1002/yea.320070502.

Jansen, H. J., Liem, M., Jong-Raadsen, S. A., Dufour, S., Weltzien, F. A., Swinkels, W., et al. (2017). Rapid *de novo* assembly of the European eel genome from nanopore sequencing reads. *Sci. Rep.* 7, 1–13. doi:10.1038/s41598-017-07650-6.

Jansen, M., Veurink, J. H., Euverink, G. J. W., and Dijkhuizen, L. (2003). Growth of the salt-tolerant yeast *Zygosaccharomyces rouxii* in microtiter plates: effects of NaCl, pH and temperature on growth and fusel alcohol production from branched-chain amino acids. *FEMS Yeast Res.* 3, 313–318. doi:10.1016/S1567-1356(02)00162-9.

Jasin, M., and Rothstein, R. (2013). Repair of strand breaks by homologous recombination. *Cold Spring Harb. Perspect. Biol.* 5, a012740. doi:10.1101/cshperspect.a012740.

Jearnpipatkul, A., Hutacharoen, R., Araki, H., and Oshima, Y. (1987). A cis-acting locus for the stable propagation of yeast plasmid pSR1. *MGG Mol. Gen. Genet.* 207, 355–360. doi:10.1007/BF00331601.

Jensen, R., Sprague, G. F., and Herskowitz, I. (1983). Regulation of yeast mating-type interconversion: feedback control of *HO* gene expression by the mating-type locus. *Proc. Natl. Acad. Sci. U. S. A.* 80, 3035–3039. doi:10.1073/pnas.80.10.3035.

Johnson, A. D. (1995). Molecular mechanisms of cell-type determination in budding yeast. *Curr. Opin. Genet. Dev.* 5, 552–558. doi:10.1016/0959-437X(95)80022-0.

Johnson, E. A., and Echavarri-Erasun, C. (2011). Yeast biotechnology. *The Yeasts*. 21-44. doi:10.1016/B978-0-444-52149-1.00003-3.

Johnson, N. A. (2010). Hybrid incompatibility genes: remnants of a genomic battlefield? *Trends Genet.* 26, 317–325. doi:10.1016/j.tig.2010.04.005.

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi:10.1093/bioinformatics/btu031.

Jones, T., Federspiel, N. A., Chibana, H., Dungan, J., Kalman, S., Magee, B. B., et al. (2004). The diploid genome sequence of *Candida albicans*. *Proc. Natl. Acad. Sci. U. S. A.* 101, 7329–34. doi:10.1073/pnas.0401648101.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., et al. (2014). Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* 24, 1384–95. doi:10.1101/gr.170720.113.

Käll, L., Krogh, A., and Sonnhammer, E. L. L. (2005). An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics* 21, 251–257. doi:10.1093/bioinformatics/bti1014.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44, D457–D462. doi:10.1093/nar/gkv1070.

Kannan, V., Narayanaswamy, P., Gadamsetty, D., Hazra, P., Khedkar, A., and Iyer, H. (2009). A tandem mass spectrometric approach to the identification of O-glycosylated glargine glycoforms in active pharmaceutical ingredient expressed in *Pichia pastoris*. *Rapid Commun. Mass Spectrom.* 23, 1035–1042. doi:10.1002/rcm.3965.

Karimi, M., Inzé, D., Van Lijsebettens, M., and Hilson, P. (2013). Gateway vectors for transformation of cereals. *Trends Plant Sci.* 18, 1–4. doi:10.1016/J.TPLANTS.2012.10.001.

Kashyap, P., Sabu, A., Pandey, A., Szakacs, G., and Soccol, C. R. (2002). Extra-cellular L-glutaminase production by *Zygosaccharomyces rouxii* under solid-state fermentation. *Process Biochem.* 38, 307–312. doi:10.1016/S0032-9592(02)00060-2.

Kawai, S., Hashimoto, W., and Murata, K. (2010). Transformation of *Saccharomyces cerevisiae* and other fungi. *Bioeng. Bugs* 1, 395–403. doi:10.4161/bbug.1.6.13257.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423, 241–254. doi:10.1038/nature01644.

Kim, H. S., Huh, J., Riles, L., Reyes, A., and Fay, J. C. (2012). A noncomplementation screen for quantitative trait alleles in *Saccharomyces cerevisiae*. *G3.* 2, 753–760. doi:10.1534/g3.112.002550.

Kinclová, O., Potier, S., and Sychrová, H. (2001). The *Zygosaccharomyces rouxii* strain CBS 732^T contains only one copy of the *HOG1* and the *SOD2* genes. *J. Biotechnol.* 88, 151–158.

Kinclova-Zimmermannova, O., Zavrel, M., and Zavrel, H. (2006). Importance of the seryl and threonyl residues of the fifth transmembrane domain to the substrate specificity of yeast plasma membrane Na $^+$ /H $^+$ antiporters. *Mol. Membr. Biol.* 23, 349–361. doi:10.1080/09687860600738908.

Knop, M. (2006). Evolution of the hemiascomycete yeasts: on life styles and the importance of inbreeding. *BioEssays* 28, 696–708. doi:10.1002/bies.20435.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res* 27:722–736. https://doi.org/10.1101/gr.215087.116.

Koufopanou, V., and Burt, A. (2003). Degeneration and domestication of a selfish gene in yeast: molecular evolution versus site-directed mutagenesis. doi:10.1093/molbev/msi149.

Krappmann, S., Pries, R., Gellissen, G., Hiller, M., and Braus, G. H. (2000). *HARO7* encodes chorismate mutase of the methylotrophic yeast *Hansenula polymorpha* and is derepressed upon methanol utilization. *J. Bacteriol.* 182, 4188–97. doi:10.1128/JB.182.15.4188-4197.2000.

Krasovska, O. S., Stasyk, O. G., Nahorny, V. O., Stasyk, O. V., Granovski, N., Kordium, V. A., et al. (2007). Glucose-induced production of recombinant proteins in *Hansenula polymorpha* mutants deficient in catabolite repression. *Biotechnol. Bioeng.* 97, 858–870. doi:10.1002/bit.21284.

Krauke, Y., and Sychrova, H. (2011). Cnh1 Na+/H+ antiporter and Ena1 Na+-ATPase play different roles in cation homeostasis and cell physiology of *Candida glabrata*. *FEMS Yeast Res.* 11, 29–41. doi:10.1111/j.1567-1364.2010.00686.x.

Kristensen, D. M., Wolf, Y. I., Mushegian, A. R., and Koonin, E. V. (2011). Computational methods for Gene Orthology inference. *Brief. Bioinform.* 12, 379–391. doi:10.1093/bib/bbr030.

Krügel, H., Fiedler, G., Smith, C., and Baumberg, S. (1993). Sequence and transcriptional analysis of the nourseothricin acetyltransferase-encoding gene nat1 from Streptomyces noursei. *Gene* 127, 127–131. doi:10.1016/0378-1119(93)90627-F.

Kües, U., and Casselton, L. a. (1992). Fungal mating type genes - regulators of sexual development. *Mycol. Res.* 96, 993–1006. doi:10.1016/S0953-7562(09)80107-X.

Kurtzman, C. P., and Robnett, C. J. (2003). Phylogenetic relationships among yeasts of the 'Saccharomyces complex' determined from multigene sequence analyses. *FEMS Yeast Res*. *3*(4), 417-432. doi:10.1016/S1567-1356(03)00012-6.

Kurtzman, C. P., and Suzuki M. (2010). Phylogenetic analysis of ascomycete yeasts that form coenzyme Q-9 and the proposal of the new genera *Babjeviella*, *Meyerozyma*, *Millerozyma*, *Priceomyces*, and *Scheffersomyces*. *Mycoscience*. 21: 2-14. doi:10.1007/S10267-009-0011-5Get.

Kurtzman, C., Fell, J. W., and Boekhout, T. (2011). The yeasts: a taxonomic study. Elsevier.

Langkjær, R. B., Nielsen, M. L., Daugaard, P. R., Liu, W., and Piškur, J. (2000). Yeast chromosomes have been significantly reshaped during their evolutionary history. *J. Mol. Biol.* 304, 271–288. doi:10.1006/jmbi.2000.4209.

Längle-Rouault, F., and Jacobs, E. (1995). A method for performing precise alterations in the yeast genome using a recycable selectable marker. *Nucleic Acids Res.* 23, 3079–81. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7659536 [Accessed November 8, 2018].

Lannoy, C. de, Risse, J., and Ridder, D. de (2018). PoreTally: run and publish *de novo* Nanopore assembler benchmarks. *bioRxiv*, 424184. doi:10.1101/424184.

Larkin, M. a., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. a., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi:10.1093/bioinformatics/btm404.

Leandro, M. J., Sychrová, H., Prista, C., and Loureiro-Dias, M. C. (2013). ZrFsy1, a high-affinity fructose/h+ symporter from fructophilic yeast *Zygosaccharomyces rouxii*. *PLoS One* 8, e68165. doi:10.1371/journal.pone.0068165.

Lee, C.-S., and Haber, J. E. (2015). Mating-type gene switching in *Saccharomyces cerevisiae*. *Mobile DNA III* (American Society of Microbiology), 491–514. doi:10.1128/microbiolspec.MDNA3-0013-2014.

Lee, H.-Y., Chou, J.-Y., Cheong, L., Chang, N.-H., Yang, S.-Y., and Leu, J.-Y. (2008). Incompatibility of nuclear and mitochondrial genomes causes hybrid sterility between two yeast species. *Cell* 135, 1065–1073. doi:10.1016/j.cell.2008.10.047.

Leggett, R. M., and Clark, M. D. (2017). A world of opportunities with nanopore sequencing. *J. Exp. Bot.* 68, 5419–5429. doi:10.1093/jxb/erx289.

Leh-Louis, V., Wirth, B., Potier, S., Souciet, J.-L., and Despons, L. (2004). Expansion and contraction of the *DUP240* multigene family in *Saccharomyces cerevisiae* populations. *Genetics* 167, 1611–9. doi:10.1534/genetics.104.028076.

Lenassi M, Gostinčar C, Jackman S, Turk M, Sadowski I, Nislow C, et al. (2013). Whole genome duplication and enrichment of metal cation transporters revealed by *de novo* genome sequencing of extremely halotolerant black yeast *Hortaea werneckii*. *PLoS One.* 8(8): e71328. https://doi.org/10.1371/journal.pone.0071328.

Liang, J., Ning, J. C., and Zhao, H. (2013). Coordinated induction of multi-gene pathways in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 41, e54–e54. doi:10.1093/nar/gks1293.

Liti, G., and Louis, E. J. (2005). Yeast evolution and comparative genomics. *Annu. Rev. Microbiol.* 59, 135–153. doi:10.1146/annurev.micro.59.030804.121400.

Liti, G., Barton, D. B. H., and Louis, E. J. (2006). Sequence diversity, reproductive isolation and species concepts in *Saccharomyces*. *Genetics* 174, 839–850.

Liu, L., Redden, H., and Alper, H. S. (2013). Frontiers of yeast metabolic engineering: diversifying beyond ethanol and *Saccharomyces*. *Curr. Opin. Biotechnol.* 24, 1023–1030. doi:10.1016/j.copbio.2013.03.005.

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C}{}_{T}$ method. *Methods* 25, 402–408. doi:10.1006/meth.2001.1262.

Llopis-Torregrosa, V., Hušeková, B., and Sychrová, H. (2016). Potassium uptake mediated by Trk1 is crucial for *Candida glabrata* growth and fitness. *PLoS One* 11, 1–18. doi:10.1371/journal.pone.0153374.

Löbs, A.-K. K., Schwartz, C., and Wheeldon, I. (2017). Genome and metabolic engineering in non-conventional yeasts: Current advances and applications. *Synth. Syst. Biotechnol.* 2, 198–207. doi:10.1016/j.synbio.2017.08.002.

Lockhart, S. R., Daniels, K. J., Zhao, R., Wessels, D., and Soll, D. R. (2003). Cell biology of mating in *Candida albicans*. *Eukaryot. Cell* 2, 49–61. doi:10.1128/EC.2.1.49-61.2003.

Lõoke, M., Kristjuhan, K., Kristjuhan, A. (2011). Extraction of genomic dna from yeasts for pcr-based applications. BioTechniques. 50, 325–328. doi: 10.2144/000113672

Losberger, C., and Ernst, J. F. (1989). Sequence and transcript analysis of the *C. albicans URA3* gene encoding orotidine-5-phosphate decarboxylase. *Curr. Genet.* 16, 153–157. doi:10.1007/BF00391471.

Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research*, *25*(5), 955.

Lu, H., Giordano, F., and Ning, Z. (2016). Oxford Nanopore MinION sequencing and genome assembly. *Genomics, Proteomics Bioinforma.* 14, 265–279. doi:10.1016/j.gpb.2016.05.004.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* 1, 18. doi:10.1186/2047-217X-1-18.

Lynch, M. (2000). The evolutionary fate and consequences of duplicate genes. *Science.* 290, 1151–1155. doi:10.1126/science.290.5494.1151.

Mable, B. K., and Otto, S. P. (2001). Masking and purging mutations following EMS treatment in haploid, diploid and tetraploid yeast (*Saccharomyces cerevisiae*). *Genet. Res.* 77, 9–26. doi:10.1017/S0016672300004821.

Madhani, H. D. (2000). Interplay of intrinsic and extrinsic signals in yeast differentiation. *Proc. Natl. Acad. Sci. U. S. A.* 97, 13461–3. doi:10.1073/pnas.011511198.

Magwene, P. M., Kayıkçı, Ö., Granek, J. a, Reininga, J. M., Scholl, Z., and Murray, D. (2011). Outcrossing, mitotic recombination, and life-history trade-offs shape genome evolution in

Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U. S. A.* 108, 1987–1992. doi:10.1073/pnas.1012544108.

Mallona, I., Weiss, J., and Marcos, E.-C. (2011). pcrEfficiency: a web tool for PCR amplification efficiency prediction. *BMC Bioinformatics* 12, 404. doi:10.1186/1471-2105-12-404.

Malone, R. E., Esposito, R. E., Lin, F., and Åström, S. U. (1980). The *RAD52* gene is required for homothallic interconversion of mating types and spontaneous mitotic recombination in yeast. *Proc. Natl. Acad. Sci. U. S. A.* 77, 503–7. doi:10.1073/pnas.77.1.503.

Manigandan, S., Gunasekar, P., Devipriya, J., and Nithya, S. (2016). Determination of heat flux on dual bell nozzle by Monte Carlo method. *J. Chem. Pharm. Sci.* 9, 3251–3253. doi:10.1002/bit.

Marcet-Houben, M., and Gabaldón, T. (2015). Beyond the Whole-Genome Duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLOS Biol.* 13, e1002220. doi:10.1371/journal.pbio.1002220.

Mardis, E. R. (2007). ChIP-seq: welcome to the new frontier. *Nat. Methods* 4, 613–614. doi:10.1038/nmeth0807-613.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380. doi:10.1038/nature03959.

Markland, W., Ley, A. C., Lee, S. W., and Ladner, R. C. (1996). Iterative optimization of high-affinity protease inhibitors using phage display. 1. Plasmin. *Biochemistry*, *35*(24), 8045-8057.

Martin, T., Lu, S. W., van Tilbeurgh, H., Ripoll, D. R., Dixelius, C., Turgeon, B. G., et al. (2010). Tracing the origin of the fungal α1 domain places its ancestor in the HMG-box superfamily: Implication for fungal mating-type evolution. *PLoS One* 5. doi:10.1371/journal.pone.0015199.

Masneuf, I., Hansen, J., Groth, C., Piskur, J., and Dubourdieu, D. (1998). New hybrids between *Saccharomyces sensu stricto* yeast species found among wine and cider production strains. *Appl Environ Microbiol*. *64*(10), 3887-3892.

Mathias, J. R., Hanlon, S. E., O’Flanagan, R. a., Sengupta, A. M., and Vershon, A. K. (2004). Repression of the yeast *HO* gene by the *MATα2* and *MAT***a***1* homeodomain proteins. *Nucleic Acids Res.* 32, 6469–6478. doi:10.1093/nar/gkh985.

Matthäus, F., Ketelhot, M., Gatter, M., and Barth, G. (2014). Production of lycopene in the non-carotenoid-producing yeast *Yarrowia lipolytica*. *Appl. Environ. Microbiol.* 80, 1660–9. doi:10.1128/AEM.03167-13.

Mayer, A. F., Hellmuth, K., Schlieker, H., Lopez-Ulibarri, R., Oertel, S., Dahlems, U., et al. (1999). An expression system matures: a highly efficient and cost-effective process for phytase production by recombinant strains ofHansenula polymorpha. *Biotechnol. Bioeng.* 63, 373–381. doi:10.1002/(SICI)1097-0290(19990505)63:3<373::AID-BIT14>3.0.CO;2-T.

Merico, A., Sulo, P., Piškur, J., and Compagno, C. (2007). Fermentative lifestyle in yeasts belonging to the *Saccharomyces* complex. *FEBS J.* 274, 976-989. doi:10.1111/j.1742-4658.2007.05645.x.

Metzenberg, R. L., and Glass, N. L. (1990). Mating type and mating strategies in *Neurospora*. *BioEssays* 12, 53–59. doi:10.1002/bies.950120202.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11, 31–46. doi:10.1038/nrg2626.

Miller, M. G., Johnson, A. D., and Francisco, S. (2002). White-Opaque Switching in Candida albicans Is Controlled by Mating-Type Locus Homeodomain Proteins and Allows Efficient Mating. 110, 293–302.

Mirończuk, A. M., Rakicka, M., Biegalska, A., Rymowicz, W., and Dobrowolski, A. (2015). A two-stage fermentation process of erythritol production by yeast *Y. lipolytica* from molasses and glycerol. *Bioresour. Technol.* 198, 445–455. doi:10.1016/J.BIORTECH.2015.09.008.

Mischo, H. E., and Proudfoot, N. J. (2013). Disengaging polymerase: Terminating RNA polymerase II transcription in budding yeast. *Biochim. Biophys. Acta - Gene Regul. Mech.* 1829, 174–185. doi:10.1016/j.bbagrm.2012.10.003.

Mitchell, A. P., and Herskowitz, I. (1986). Activation of meiosis and sporulation by repression of the *RME1* product in yeast. *Nature* 319, 738–742. doi:10.1038/319738a0.

Monerawela, C., and Bond, U. (2017). Recombination sites on hybrid chromosomes in Saccharomyces pastorianus share common sequence motifs and define a complex evolutionary relationship between group I and II lager yeasts. *FEMS Yeast Res.* 17, 1–12. doi:10.1093/femsyr/fox047.

Morales, L., and Dujon, B. (2012). Evolutionary role of interspecies hybridization and genetic exchanges in yeasts. *Microbiol. Mol. Biol. Rev.* 76, 721–739. doi:10.1128/MMBR.00022-12.

Morett, E., Korbel, J. O., Rajan, E., Saab-Rincon, G., Olvera, L., Olvera, M., et al. (2003). Systematic discovery of analogous enzymes in thiamin biosynthesis. *Nat. Biotechnol.* 21, 790–795. doi:10.1038/nbt834.

Mori, H. (1973). Life cycle in a heterothallic haploid yeast, *Saccharomyces rouxii*. *J Ferment Technol*.

Mori, H., and Onishi, H. (1967). Diploid hybridization in a heterothallic haploid yeast, *Saccharomyces rouxii. Appl. Microbiol.* 15, 928–34.

Morozova, O., and Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics*. doi:10.1016/j.ygeno.2008.07.001.

Mortimer, R. K. (2000). Evolution and variation of the yeast (Saccharomyces) genome. *Genome Res.* 10, 403–409. doi:10.1101/gr.10.4.403.

Mortimer, R. K., Romano, P., Suzzi, G., and Polsinelli, M. (1994). Genome renewal: A new phenomenon revealed from a genetic study of 43 strains of *Saccharomyces cerevisiae* derived from natural fermentation of grape musts. *Yeast* 10, 1543–1552. doi:10.1002/yea.320101203.

Muller, H., Hennequin, C., Gallaud, J., Dujon, B., and Fairhead, C. (2008). The asexual yeast Candida glabrata maintains distinct a and α haploid mating-types. Eukaryot. Cell. 7, 848-858. doi: 10.1128/EC.00456-07

Mumberg, D., Müller, R., and Funk, M. (1995). Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds. *Gene* 156, 119–122. doi:10.1016/0378-1119(95)00037-7.

Nacken, V., Achstetter, T., and Degryse, E. (1996). Probing the limits of expression levels by varying promoter strength and plasmid copy number in *Saccharomyces cerevisiae*. *Gene* 175, 253–260. doi:10.1016/0378-1119(96)00171-0.

Nagy, A. (2000). Cre recombinase: the universal reagent for genome tailoring. *Genesis* 26, 99–109. doi:10.1002/(SICI)1526-968X(200002)26:2<99::AID-GENE1>3.0.CO;2-B.

Nakayashiki, T., Ebihara, K., Bannai, H., and Nakamura, Y. (2001). Yeast [PSI+] “Prions” that are crosstransmissible and susceptible beyond a species barrier through a quasi-prion state. *Mol. Cell* 7, 1121–1130. doi:10.1016/S1097-2765(01)00259-3.

Nasmyth, K. (1987). The determination of mother cell-specific mating type switching in yeast by a specific regulator of *HO* transcription. *EMBO J.* 6, 243–248.

Nasmyth, K. (1993). Regulating the *HO* endonuclease in yeast. *Curr. Opin. Genet. Dev.* 3, 286–294. doi:http://dx.doi.org/10.1016/0959-437X(93)90036-O.

Ner, S. S. (1989). Role of intron splicing in the function of the *MAT***a**1 gene of *Saccharomyces cerevisiae*. *Molec and Cell Biol*.9, 4613–4620.

Nolte, A. W., and Tautz, D. (2010). Understanding the onset of hybrid speciation. *Trends Genet.* 26, 54–58. doi:10.1016/j.tig.2009.12.001.

O’Neill, M., McPartlin, J., Arthure, K., Riedel, S., and McMillan, N. (2011). Comparison of the TLDA with the Nanodrop and the reference Qubit system. *J. Phys. Conf. Ser.* 307, 012047. doi:10.1088/1742-6596/307/1/012047.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, *27*(1), 29-34. doi:10.1093/nar/27.1.29.

Oguntoyinbo, F. A., Tourlomousis, P., Gasson, M. J., and Narbad, A. (2011). Analysis of bacterial communities of traditional fermented West African cereal foods using culture independent methods. *Int. J. Food Microbiol.* 145, 205–210. doi:10.1016/J.IJFOODMICRO.2010.12.025.

Ohata, M., Kohama, K., Morimitsu, Y., Kubota, K., and Sugawara, E. (2007). The formation mechanism by yeast of 4-hydroxy-2(or 5)-ethyl-5(or 2)-methyl-3(2 *H*)-furanone in miso. *Biosci. Biotechnol. Biochem.* 71, 407–413. doi:10.1271/bbb.60466.

Ohno, S. (1970). Evolution by gene duplication. Berlin, New York: Springer-Verlag. xv, 160 p.

Oldenburg, K., Vo, K. T., Michaelis, S., and Paddon, C. (1997). Recombination-mediated PCR-directed plasmid construction in vivo in yeast. *Nucleic Acids Res.* 25, 451–452. doi:10.1093/nar/25.2.451.

Olson, M. V (1999). When less is more: gene loss as an engine of evolutionary change. *Am. J. Hum. Genet.* 64, 18–23. doi:10.1086/302219.

Orr, H. A., and Turelli, M. (2001). The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities. *Evolution (N. Y).* 55, 1085–1094.

Orr-Weaver, T. L., Szostak, J. W., and Rothstein, R. J. (1981). Yeast transformation: a model system for the study of recombination. *Proc. Natl. Acad. Sci.* 78, 6354–6358. doi:10.1073/pnas.78.10.6354.

Ortiz-Merino, R. A., Kuanyshev, N., Braun-Galleani, S., Byrne, K. P., Porro, D., Branduardi, P., et al. (2017). Evolutionary restoration of fertility in an interspecies hybrid yeast, by whole-genome duplication after a failed mating-type switch. *PLoS Biol.* 15, 1–25. doi:10.1371/journal.pbio.2002128.

Otto, S. P., and Gerstein, A. C. (2008). The evolution of haploidy and diploidy. *Curr. Biol.* 18, 1121–1124. doi:10.1016/j.cub.2008.09.039.

Paliwal, S., Iglesias, P. A., Campbell, K., Hilioti, Z., Groisman, A., and Levchenko, A. (2007). MAPK-mediated bimodal gene expression and adaptive gradient sensing in yeast. *Nature* 446, 46–51. doi:10.1038/nature05561.

Palková, Z., and Váchová, L. (2006). Life within a community: benefit to yeast long-term survival. *FEMS Microbiol Rev*, 30(5), 806–824. doi:10.1111/j.1574-6976.2006.00034.x

Pâques, F., and Haber, J. E. (1999). Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.* 63, 349–404.

Peter, J., and Schacherer, J. (2016). Population genomics of yeasts: towards a comprehensive view across a broad evolutionary scale. *Yeast*. 33(3), 73-81. doi: 10.1002/yea.3142

Pfliegler, W. P., Antunovics, Z., and Sipiczki, M. (2012). Double sterility barrier between *Saccharomyces* species and its breakdown in allopolyploid hybrids by chromosome loss. *FEMS Yeast Res.* 12, 703–718. doi:10.1111/j.1567-1364.2012.00820.x.

Piatkowska, E. M., Naseeb, S., Knight, D., and Delneri, D. (2013). Chimeric protein complexes in hybrid species generate novel phenotypes. *PLoS Genet.* 9, e1003836. doi:10.1371/journal.pgen.1003836.

Poinar, H. N., Hofreiter, M., Spaulding, W. G., Martin, P. S., Stankiewicz, B. A., Bland, H., et al. (1998). Molecular coproscopy: dung and diet of the extinct ground sloth *Nothrotheriops shastensis*. *Science* 281, 402–6. doi:10.1126/science.281.5375.402.

Porro, D., and Branduardi, P. (2009). Yeast cell factory: fishing for the best one or engineering it? *Microb. Cell Fact.* 8, 51. doi:10.1186/1475-2859-8-51.

Pribylova, L., and Sychrovà, H. (2003). Efficient transformation of the osmotolerant yeast *Zygosaccharomyces rouxii* by electroporation. *J. Microbiol. Methods* 55, 481–484. doi:10.1016/S0167-7012(03)00197-0.

Pribylova, L., De Montigny, J., and Sychrovà, H. (2007a). Tools for the genetic manipulation of *Zygosaccharomyces rouxii*. *FEMS Yeast Res.* 7, 1285–1294. doi:10.1111/j.1567-1364.2007.00308.x.

Pribylova, L., Papouskova, K., and Sychrovà, H. (2008). The salt tolerant yeast Zygosaccharomyces rouxii possesses two plasma-membrane Na+/H+-antiporters (*ZrNha1p* and *ZrSod2-22p*) playing different roles in cation homeostasis and cell physiology. *Fungal Genet. Biol.* 45, 1439–1447. doi:10.1016/j.fgb.2008.08.001.

Pribylova, L., Straub, M. L., Sychrovà, H., and De Montigny, J. (2007b). Characterisation of *Zygosaccharomyces rouxii* centromeres and construction of first *Z. rouxii* centromeric vectors. *Chromosome Res*, *15*(4), 439. doi: 10.1007/s10577-007-1136-z

Priyam, A., Woodcroft, B. J., Rai, V., Munagala, A., Moghul, I., Ter, F., et al. (2015). Sequenceserver: a modern graphical user interface for custom BLAST databases. *BioRxiv*, 033142. doi:10.1101/033142.

Proux-Wéra, E., Armisén, D., Byrne, K. P., and Wolfe, K. H. (2012). A pipeline for automated annotation of yeast genome sequences by a conserved-synteny approach. *BMC Bioinformatics* 13. doi:10.1186/1471-2105-13-237.

Pryszcz, L. P., and Gabaldón, T. (2016). Redundans: An assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* 44, e113. doi:10.1093/nar/gkw294.

Quail, M., Smith, M. E., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., et al. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13, 341. doi:10.1186/1471-2164-13-341.

Rajaei, N., Chiruvella, K. K., Lin, F., and Astrom, S. U. (2014). Domesticated transposase Kat1 and its fossil imprints induce sexual differentiation in yeast. *Proc. Natl. Acad. Sci.* 111, 15491–15496. doi:10.1073/pnas.1406027111.

Rajeh, A., Lv, J., and Lin, Z. (2018). Heterogeneous rates of genome rearrangement contributed to the disparity of species richness in Ascomycota. *BMC Genomics* 19, 1–13. doi:10.1186/s12864-018-4683-0.

Read, T. D., Petit, R. A., Joseph, S. J., Alam, M. T., Weil, M. R., Ahmad, M., et al. (2017). Draft sequencing and assembly of the genome of the world's largest fish, the whale shark: Rhincodon typus Smith 1828. *BMC Genomics* 18, 1–10. doi:10.1186/s12864-017-3926-9.

Reid, J. G., Carroll, A., Veeraraghavan, N., Dahdouli, M., Sundquist, A., English, A., et al. (2014). Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. *BMC Bioinformatics* 15, 30. doi:10.1186/1471-2105-15-30.

Reuß, O., Vik, Å.˚, Kolter, R., and Morschhäuser, J. (2004). The *SAT1* flipper, an optimized tool for gene disruption in *Candida albicans*. *Gene* 341, 119–127. doi:10.1016/j.gene.2004.06.021.

Reuter, J. A., Spacek, D. V., and Snyder, M. P. (2015). High-throughput sequencing technologies. *Mol. Cell* 58, 586–597. doi:10.1016/J.MOLCEL.2015.05.004.

Rieseberg, L. H., Raymond, O., Rosenthal, D. M., Lai, Z., Livingstone, K., Nakazato, T., et al. (2003). Major ecological transitions in wild sunflowers facilitated by hybridization. *Science.* 301, 1211–1216.

Riesenfeld, C. S., Schloss, P. D., and Handelsman, J. (2004). Metagenomics: Genomic Analysis of microbial communities. *Annu. Rev. Genet.* 38, 525–552. doi:10.1146/annurev.genet.38.072902.091216.

Riley, R., Haridas, S., Wolfe, K. H., Lopes, M. R., Hittinger, C. T., Göker, M., et al. (2016). Comparative genomics of biotechnologically important yeasts. *Proc. Natl. Acad. Sci.* 113, 9882–9887. doi:10.1073/pnas.1603941113.

Roberts, C., and der Walt, J. P. (1959). The life cycle of *Kluyveromyces polysporus*. *C. R. Trav. Lab. Carlsberg. Chim.* 31, 129.

Roh, S. W., Abell, G. C. J., Kim, K.-H., Nam, Y.-D., and Bae, J.-W. (2010). Comparing microarrays and next-generation sequencing technologies for microbial ecology research. *Trends Biotechnol.* 28, 291–299. doi:10.1016/J.TIBTECH.2010.03.001.

Ronaghi, M. (2001). Pyrosequencing sheds light on DNA sequencing. *Genome Res.* 11, 3–11. doi:10.1101/GR.150601.

Rooney, A. P., and Ward, T. J. (2005). Evolution of a large ribosomal RNA multigene family in filamentous fungi: birth and death of a concerted evolution paradigm. *Proc. Natl. Acad. Sci. U. S. A.* 102, 5084–5089. doi:10.1073/pnas.0409689102.

Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., Hariharan, I. K., et al. (2000). Comparative genomics of the eukaryotes. *Science.* 287, 2204–2215. doi:10.1126/science.287.5461.2204.

Ruderfer, D. M., Pratt, S. C., Seidel, H. S., and Kruglyak, L. (2006). Population genomic analysis of outcrossing and recombination in yeast. *Nat. Genet.* 38, 1077–1081. doi:10.1038/ng1859.

Ruijter, J. M., Ramakers, C., Hoogaars, W. M. H., Karlen, Y., Bakker, O., Van Den Hoff, M. J. B., et al. (2009). Amplification efficiency: Linking baseline and bias in the analysis of quantitative PCR data. *Nucleic Acids Res.* doi:10.1093/nar/gkp045.

Rusche, L. N., and Rine, J. (2010). Switching the mechanism of mating type switching : a domesticated transposase supplants a domesticated homing endonuclease. *Genes & development*. 10–14. doi:10.1101/gad.1886310.10.

Ryu, S. L., Murooka, Y., and Kaneko, Y. (1996). Genomic reorganization between two sibling yeast species, *Saccharomyces bayanus* and *Saccharomyces cerevisiae*. *Yeast* 12, 757–764. doi:10.1002/(SICI)1097-0061(19960630)12:8<757::AID-YEA970>3.0.CO;2-H.

Salonen, A., Nikkilä, J., Jalanka-Tuovinen, J., Immonen, O., Rajilić-Stojanović, M., Kekkonen, R. A., et al. (2010). Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: Effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *J. Microbiol. Methods* 81, 127–134. doi:10.1016/J.MIMET.2010.02.007.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., et al. (1977b). Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature* 265, 687–695. doi:10.1038/265687a0.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977a). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463–7. doi:10.1073/PNAS.74.12.5463.

Saraya, R., Krikken, A. M., Kiel, J. A. K. W., Baerends, R. J. S., Veenhuis, M., and van der Klei, I. J. (2012). Novel genetic tools for *Hansenula polymorpha*. *FEMS Yeast Res.* 12, 271–278. doi:10.1111/j.1567-1364.2011.00772.x.

Sasaki, M., Matsudo, T., and Nunomura, N. (1991). Biosynthesis of 4-hydroxy-2(or 5)-ethyl-5(or 2)-methyl-3(2h)-furanone by yeasts. *J. Agric. Food Chem.* 39, 934–938. doi:10.1021/jf00005a027.

Sasaki, M., Matsudo, T., and Nunomura, N. (1991). Biosynthesis of 4-hydroxy-2(or 5)-ethyl-5(or 2)-methyl-3(2h)-furanone by yeasts. *J. Agric. Food Chem.* 39, 934-938. doi:10.1021/jf00005a027.

Sato, A., Matsushima, K., Oshima, K., Hattori, M., and Koyama, Y. (2017). Draft genome sequencing of the highly halotolerant and allopolyploid yeast *Zygosaccharomyces rouxii* NBRC 1876. *Genome Announc*. doi: 10.1128/genomeA.01610-16

Sauer, B. (1987). Functional expression of the cre-lox site-specific recombination system in the yeast *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* 7, 2087-2096. doi:10.1128/MCB.7.6.2087.Updated.

Scannell, D. R., Frank, A. C., Conant, G. C., Byrne, K. P., Woolfit, M., and Wolfe, K. H. (2007). Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc. Natl. Acad. Sci.* 104, 8397–8402. doi:10.1073/pnas.0608218104.

Schacherer, J., De Montigny, J., Welcker, A., Souciet, J.-L., and Potier, S. (2005). Duplication processes in *Saccharomyces cerevisiae* haploid strains. *Nucleic Acids Res.* 33, 6319–6326. doi:10.1093/nar/gki941.

Schacherer, J., Tourrette, Y., Souciet, J.-L., Potier, S., and De Montigny, J. (2004). Recovery of a function involving gene duplication by retroposition in *Saccharomyces cerevisiae*. *Genome Res.* 14, 1291–7. doi:10.1101/gr.2363004.

Schadt, E. E., Turner, S., and Kasarskis, A. (2010). A window into third-generation sequencing. *Hum. Mol. Genet.* 19, R227–R240. doi:10.1093/hmg/ddq416.

Schalamun, M., Kainer, D., Beavan, E., Nagar, R., Eccles, D., Rathjen, J. P., et al. (2018). A comprehensive toolkit to enable MinION long-read sequencing in any laboratory. *BioRxiv*, 289579. doi:10.1101/289579.

Shah, J. C., and Clancy, M. J. (1992). *IME4*, a gene that mediates *MAT* and nutritional control of meiosis in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* 12, 1078–1086. doi:10.1128/MCB.12.3.1078.Updated.

Shen, W., Xu, G., Luo, M., and Jiang, Z. (2016). Genetic diversity of *Sporisorium scitamineum* in mainland China assessed by SCoT analysis. *Trop. Plant Pathol.* 41, 288–296. doi:10.1007/s40858-016-0099-z.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.* 26, 1135–1145. doi:10.1038/nbt1486.

Shuster, J. R., Moyer, D., and Irvine, B. (1987). Sequence of the *Kluyveromyces lactis URA3* gene. *Nucleic Acids Res.* 15, 8573. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3671096 [Accessed November 8, 2018].

Sievers, F., and Higgins, D. G. (2014). Clustal Omega, accurate alignment of very large numbers of sequences. *Multiple sequence alignment methods* (Springer), 105–116. doi: 10.1007/978-1-62703-646-7

Siliciano, P. G., and Tatchell, K. (1984). Transcription and regulatory signals at the mating type locus in yeast. *Cell* 37, 969–978. doi:10.1016/0092-8674(84)90431-8.

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. doi:10.1093/bioinformatics/btv351.

Simbolo, M., Gottardi, M., Corbo, V., Fassan, M., Mafficini, A., Malpeli, G., et al. (2013). DNA qualification workflow for next generation sequencing of histopathological samples. *PLoS One* 8, e62692. doi:10.1371/journal.pone.0062692.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res.* 19, 1117–23. doi:10.1101/gr.089532.108.

Singh, A., And Sherman, F. (1974). Association of methionine requirement with methyl mercury resistant mutants of yeast. *Nature* 247, 227–229. doi:10.1038/247227a0.

Sipiczki, M. (2008). Interspecies hybridization and recombination in *Saccharomyces* wine yeasts. *FEMS Yeast Res.* 8, 996–1007. doi:10.1111/j.1567-1364.2008.00369.x.

Slater, G. S. C., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 6.1: 31. doi:10.1186/1471-2105-6-31.

Smith, H., Tomb, J., Dougherty, B., Fleischmann, R., Venter, J., Kerlavage, A., et al. (1995). Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae* Rd genome. *Science.* 269, 538–540. doi:10.1126/science.7542802.

Solieri, L. (2010). Mitochondrial inheritance in budding yeasts: towards an integrated understanding. *Trends Microbiol.* 18, 521–530. doi:10.1016/J.TIM.2010.08.001.

Solieri, L., Cassanelli, S., Croce, M. A., and Giudici, P. (2008). Genome size and ploidy level: New insights for elucidating relationships in *Zygosaccharomyces* species. *Fungal Genet. Biol.* 45, 1582–1590. doi:10.1016/j.fgb.2008.10.001.

Solieri, L., Chand Dakal, T., Croce, M. A., and Giudici, P. (2013a). Unravelling genomic diversity of *Zygosaccharomyces rouxii* complex with a link to its life cycle. *FEMS Yeast Res.* 13, 245–258. doi:10.1111/1567-1364.12027.

Solieri, L., Dakal, T. C., and Bicciato, S. (2014a). Quantitative phenotypic analysis of multistress response in *Zygosaccharomyces rouxii* complex. *FEMS Yeast Res.* 14, 586–600. doi:10.1111/1567-1364.12146.

Solieri, L., Dakal, T. C., and Giudici, P. (2013b). *Zygosaccharomyces sapae* sp. nov., isolated from Italian traditional balsamic vinegar. *Int. J. Syst. Evol. Microbiol.* 63, 364–371.

Solieri, L., Dakal, T. C., Giudici, P., and Cassanelli, S. (2014b). Sex-determination system in the diploid yeast *Zygosaccharomyces sapae. G3 (Bethesda).* 4, 1011–25. doi:10.1534/g3.114.010405.

Solieri, L., Gullo, M., and Giudici, P. (2012). Traditional balsamic vinegar: a microbiological overview. *Handbook of Plant-Based Fermented Food and Beverage Technology* (pp. 628-649). *CRC Press.*

Solieri, L., Landi, S., De Vero, L., and Giudici, P. (2006). Molecular assessment of indigenous yeast population from traditional balsamic vinegar. *J. Appl. Microbiol.* 101, 63–71. doi:10.1111/j.1365-2672.2006.02906.x.

Solieri, L., Vezzani, V., Cassanelli, S., Dakal, T. C., Pazzini, J., and Giudici, P. (2016). Differential hypersaline stress response in *Zygosaccharomyces rouxii* complex yeasts : a physiological and transcriptional study. *FEMS Yeast Res.* 16, 1–11. doi:10.1093/femsyr/fow063.

Solis-Escalante, D., Kuijpers, N. G. A., Bongaerts, N., Bolat, I., Bosman, L., Pronk, J. T., et al. (2013). amdSYM, A new dominant recyclable marker cassette for *Saccharomyces cerevisiae*. *FEMS Yeast Res.* 13, 126–139. doi:10.1111/1567-1364.12024.

Sonnhammer, E. L. L., Gabaldon, T., Sousa da Silva, A. W., Martin, M., Robinson-Rechavi, M., Boeckmann, B., et al. (2014). Big data and other challenges in the quest for orthologs. *Bioinformatics* 30, 2993–2998. doi:10.1093/bioinformatics/btu492.

Souciet, J. L., Dujon, B., Gaillardin, C., Johnston, M., Baret, P. V., Cliften, P., et al. (2009). Comparative genomics of protoploid Saccharomycetaceae. *Genome Res.* 19, 1696–1709. doi:10.1101/gr.091546.109.

Sreekrishna K. (1993). Strategies for optimizing protein expression and secretion in the methylotrophic yeast Pichia pastoris. *Ind. Microorg. Basic Appl. Mol. Genet.*

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res.* 34, 435–439. doi:10.1093/nar/gkl200.

Steensels, J., Snoek, T., Meersman, E., Nicolino, M. P., Voordeckers, K., and Verstrepen, K. J. (2014). Improving industrial yeast strains: exploiting natural and artificial diversity. *FEMS Microbiol. Rev.* 38, 947–995. doi:10.1111/1574-6976.12073.

Steensma, H. Y., and Ter Linde, J. J. M. (2001). Plasmids with the Cre-recombinase and the dominant nat marker, suitable for use in prototrophic strains of *Saccharomyces cerevisiae* and *Kluyveromyces lactis*. *Yeast* 18, 469–472. doi:10.1002/yea.696.

Stefanini, I., and Dapporto, L. (2012). Role of social wasps in *Saccharomyces cerevisiae* ecology and evolution. *PNAS*. 109, 13398–13403. https://doi.org/10.1073/pnas.1208362109.

Sternberg, P. W., Stern, M. J., Clark, I., and Herskowitz, I. (1987). Activation of the yeast *HO* gene by release from multiple negative controls. *Cell* 48, 567–577. doi:10.1016/0092-8674(87)90235-2.

Stoddard, B. L. (2005). Homing endonuclease structure and function. *Q. Rev. Biophys.* 38, 49–95. doi:10.1017/S0033583505004063.

Stratford, M. (2006). Food and Beverage Spoilage Yeasts. *Yeasts in Food and Beverages* (Berlin, Heidelberg: Springer Berlin Heidelberg), 335–379. doi:10.1007/978-3-540-28398-0_11.

Strathern, J. N., and Herskowitz, I. (1979). Asymmetry and directionality in production of new cell types during clonal growth: the switching pattern of homothallic yeast. *Cell* 17, 371–381. doi:10.1016/0092-8674(79)90163-6.

Strathern, J. N., Klar, A. J. S., Hicks, J. B., Abraham, J. A., Ivy, J. M., Nasmyth, K. A., et al. (1982). Homothallic switching of yeast mating type cassettes is initiated by a double-stranded cut in the *MAT* locus. *Cell* 31, 183–192. doi:10.1016/0092-8674(82)90418-4.

Strathern, J., Shafer, B., Hicks, J., and McGill, C. (1988). **a**/α-specific repression by *MAT*α2. *Genetics* 120, 75–81.

Tai, M., and Stephanopoulos, G. (2013). Engineering the push and pull of lipid biosynthesis in oleaginous yeast *Yarrowia lipolytica* for biofuel production. *Metab. Eng.* 15, 1–9. doi:10.1016/J.YMBEN.2012.08.007.

Taing, O., and Taing, K. (2007). Production of malic and succinic acids by sugar-tolerant yeast *Zygosaccharomyces rouxii*. *Eur. Food Res. Technol.* 224, 343-347. doi:10.1007/s00217-006-0323-z.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi:10.1093/molbev/mst197.

Teo, W. S., and Chang, M. W. (2014). Development and characterization of AND-gate dynamic controllers with a modular synthetic *GAL1* core promoter in *Saccharomyces cerevisiae*. *Biotechnol. Bioeng.* 111, 144–151. doi:10.1002/bit.25001.

Toyn, J. H., Gunyuzlu, P. L., Hunter White, W., Thompson, L. A., and Hollis, G. F. (2000). A counterselection for the tryptophan pathway in yeast: 5-fluoroanthranilic acid resistance. *Yeast* 16, 553–560. doi:10.1002/(SICI)1097-0061(200004)16:6<553::AID-YEA554>3.0.CO;2-7.

Treangen, T. J., and Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* 13, 36–46. doi:10.1038/nrg3117.

Tsai, I. J., Bensasson, D., Burt, A., and Koufopanou, V. (2008). Population genomics of the wild yeast *Saccharomyces paradoxus*: quantifying the life cycle. *Proc. Natl. Acad. Sci. U. S. A.* 105, 4957–4962. doi:10.1073/pnas.0707314105.

Tschopp, J. F., Sverlow, G., Kosson, R., Craig, W., and Grinna, L. (1987). High-level secretion of glycosylated invertase in the methylotrophic yeast, *Pichia pastoris*. *Nat. Biotechnol.* 5, 1305–1308. doi:10.1038/nbt1287-1305.

Tsong, A. E., Miller, M. G., Raisner, R. M., and Johnson, A. D. (2003). Evolution of a combinatorial transcriptional circuit: a case study in yeasts. *Cell*. 115, 389-399. doi: 10.1016/S0092-8674(03)00885-7.

Tsong, A. E., Tuch, B. B., Li, H., and Johnson, A. D. (2006). Evolution of alternative transcriptional circuits with identical logic. *Nature* 443, 415–420. doi:10.1038/nature05099.

Tuch, B. B., Galgoczy, D. J., Hernday, A. D., Li, H., and Johnson, A. D. (2008). The evolution of combinatorial gene regulation in fungi. *PLoS Biol.* 6, e38. doi:10.1371/journal.pbio.0060038.

Uehara, K., Watanabe, J., Mogi, Y., and Tsukioka, Y. (2017). Identification and characterization of an enzyme involved in the biosynthesis of the 4-hydroxy-2(or 5)-ethyl-5(or 2)-methyl-3(2H)-furanone in yeast. *J. Biosci. Bioeng.* 123, 333–341. doi:10.1016/j.jbiosc.2016.10.004.

Ullah, A., Lopes, M. I., Brul, S., and Smits, G. J. (2013). Intracellular pH homeostasis in *Candida glabrata* in infection-associated conditions. *Microbiol.* 159, 803–813. doi:10.1099/mic.0.063610-0.

Ushio, K., Tatsumi, H., Araki, H., Toh-e, A., and Oshima, Y. (1988). Construction of a host-vector system in the osmophilic haploid yeast *Zygosaccharomyces rouxii*. *J. Ferment. Technol.* 66, 481–488. doi:10.1016/0385-6380(88)90079-9.

Van den Berg, J. A., van der Laken, K. J., van Ooyen, A. J. J., Renniers, T. C. H. M., Rietveld, K., Schaap, A., et al. (1990). *Kluyveromyces* as a host for heterologous gene expression: expression and secretion of prochymosin. *Nat. Biotechnol.* 8, 135–139. doi:10.1038/nbt0290-135.

Van Der Sluis, C., Tramper, J., and Wijffels, R. H. (2001). Enhancing and accelerating flavour formation by salt-tolerant yeasts in Japanese soy-sauce processes. *Trends Food Sci. Technol.* 12, 322–327. doi:10.1016/S0924-2244(01)00094-2.

Van Dyk, D., Hansson, G., Pretorius, I. S., and Bauer, F. F. (2003). Cellular differentiation in response to nutrient availability: the repressor of meiosis, rme1p, positively regulates invasive growth in *Saccharomyces cerevisiae*. *Genetics* 165, 1045–1058.

Van Ooyen, A. J. J., Dekker, P., Huang, M., Olsthoorn, M. M. A., Jacobs, D. I., Colussi, P. A., et al. (2006). Heterologous protein production in the yeast *Kluyveromyces lactis*. *FEMS Yeast Res.* 6, 381–392. doi:10.1111/j.1567-1364.2006.00049.x.

Vaser, R., Sovic, I., Nagarajan, N., and Sikic, M. (2016). Fast and accurate *de novo* genome assembly from long uncorrected reads. *BioRxiv*, 068122. doi:10.1101/068122.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. *Science* 291, 1304–51. doi:10.1126/science.1058040.

Vicenzi, J. T., Zmijewski, M. J., Reinhard, M. R., Landen, B. E., Muth, W. L., and Marler, P. G. (1997). Large-scale stereoselective enzymatic ketone reduction with in situ product removal via polymeric adsorbent resins. *Enzyme Microb. Technol.* 20, 494–499. doi:10.1016/S0141-0229(96)00177-9.

Wach, A., Brachat, A., Alberti-Segui, C., Rebischung, C., and Philippsen, P. (1997). Heterologous HIS3 marker and GFP reporter modules for PCR-targeting in *Saccharomyces cerevisiae*. *Yeast* 13, 1065–1075. doi:10.1002/(SICI)1097-0061(19970915)13:11<1065::AID-YEA159>3.0.CO;2-K.

Wach, A., Brachat, A., Pöhlmann, R., and Philippsen, P. (1994). New heterologous modules for classical or PCR-based gene disruptions in *Saccharomyces cerevisiae*. *Yeast* 10, 1793–1808. doi:10.1002/yea.320101310.

Wagner, J. M., and Alper, H. S. (2016). Synthetic biology and molecular genetics in non-conventional yeasts: current tools and future advances. *Fungal Genet. Biol.* 89, 126–136. doi:10.1016/j.fgb.2015.12.001.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9. doi:10.1371/journal.pone.0112963.

Walther, A., Hesselbart, A., and Wendland, J. (2014). Genome sequence of *Saccharomyces carlsbergensis*, the world's first pure culture lager yeast. *G3: Genes, Genomes, Genetics*. g3-113. doi:10.1534/g3.113.010090.

Wang, M., Beck, C. R., English, A. C., Meng, Q., Buhay, C., Han, Y., et al. (2015a). PacBio-LITS: A large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations. *BMC Genomics* 16, 1–12. doi:10.1186/s12864-015-1370-2.

Wang, Y., Coleman-Derr, D., Chen, G., and Gu, Y. Q. (2015b). OrthoVenn: A web server for genome wide comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* 43, W78–W84. doi:10.1093/nar/gkv487.

Wang, Z., Ye, S., Li, J., Zheng, B., Bao, M., and Ning, G. (2011). Fusion primer and nested integrated PCR (FPNI-PCR): a new high-efficiency strategy for rapid chromosome walking or flanking sequence cloning. *BMC Biotechnol.* 11, 109. doi:10.1186/1472-6750-11-109.

Watanabe, J., Uehara, K., and Mogi, Y. (2013). Diversity of mating-type chromosome structures in the yeast *Zygosaccharomyces rouxii* caused by ectopic exchanges between *MAT*-like loci. *PLoS One* 8, e62121. doi:10.1371/journal.pone.0062121.

Watanabe, J., Uehara, K., Mogi, Y., and Tsukioka, Y. (2017). Mechanism for restoration of fertility in hybrid *Zygosaccharomyces rouxii* generated by interspecies hybridization. *Appl. Environ. Microbiol.* 83, 4–26. doi:10.1128/AEM.01187-17.

Watanabe, J., Uehara, K., Mogi, Y., Suzuki, K., Watanabe, T., and Yamazaki, T. (2010). Improved transformation of the halo-tolerant yeast *Zygosaccharomyces rouxii* by electroporation. *Biosci. Biotechnol. Biochem.* 74, 1092–1094. doi:10.1271/bbb.90865.

Watanabe, Y., Miwa, S., and Tamai, Y. (1995). Characterization of Na+/H+-antiporter gene closely related to the salt-tolerance of yeast *Zygosaccharomyces rouxii*. *Yeast* 11, 829–838. doi:10.1002/yea.320110905.

Watanabe, Y., Tsuchimoto, S., and Tamai, Y. (2004). Heterologous expression of *Zygosaccharomyces rouxii* glycerol 3-phosphate dehydrogenase gene (*ZrGPD1*) and glycerol dehydrogenase gene (*ZrGCY1*) in *Saccharomyces cerevisiae*. *FEMS Yeast Res.* 4, 505–510. doi:10.1016/S1567-1356(03)00210-1.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191. doi:10.1093/bioinformatics/btp033.

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., et al. (2015). AntiSMASH 3.0-A comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* 43, W237–W243. doi:10.1093/nar/gkv437.

Weydemann, U., Keup, P., Piontek, M., Strasser, A. W. M., Schweden, J., Gellissen, G., et al. (1995). High-level secretion of hirudin by *Hansenula polymorpha* - authentic processing of three different preprohirudins. *Appl. Microbiol. Biotechnol.* 44, 377–385. doi:10.1007/BF00169932.

Wilbrandt, J., Misof, B., and Niehuis, O. (2017). COGNATE: Comparative gene annotation characterizer. *BMC Genomics* 18, 1–10. doi:10.1186/s12864-017-3870-8.

Wildt, S., and Gerngross, T. U. (2005). The humanization of N-glycosylation pathways in yeast. *Nat. Rev. Microbiol.* 3, 119–128. doi:10.1038/nrmicro1087.

Wilson, I. G. (1997). Inhibition and facilitation of nucleic acid amplification. *Appl. Environ. Microbiol.* 63, 3741–51.

Wolberger, C., Vershon, A. K., Johnson, A. D., and Pabo, C. (1991). Crystal Structure of a *MAT***a** homeodomain-operator model complex suggests a general for homeodomain-DNA interactions. *Cell* 67, 517–526. doi:10.1016/0092-8674(91)90526-5.

Wolfe, K. H. (2001). Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* 2, 333–341. doi:10.1038/35072009.

Wolfe, K. H. (2004). Evolutionary Genomics: Yeasts Accelerate beyond BLAST. *Curr. Biol.* 14, R392–R394. doi:10.1016/J.CUB.2004.05.015.

Wolfe, K. H. (2015). Origin of the Yeast Whole-Genome Duplication. *PLOS Biol.* 13, e1002221. doi:10.1371/journal.pbio.1002221.

Wolfe, K. H., and Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387, 708–713. doi:10.1038/42711.

Wolfe, K. H., Armiśen, D., Proux-Wera, E., ÓhÉigeartaigh, S. S., Azam, H., Gordon, J. L., et al. (2015). Clade- and species-specific features of genome evolution in the Saccharomycetaceae. *FEMS Yeast Res.* 15, 1–12. doi:10.1093/femsyr/fov035.

Woolfe, M., and Primrose, S. (2004). Food forensics: using DNA technology to combat misdescription and fraud. *Trends Biotechnol.* 22, 222–226. doi:10.1016/J.TIBTECH.2004.03.010.

Xu, Y., and Pan, S. (2013). Effects of various factors of ultrasonic treatment on the extraction yield of all-trans-lycopene from red grapefruit (Citrus paradise Macf.). *Ultrason. Sonochem.* 20, 1026–1032. doi:10.1016/J.ULTSONCH.2013.01.006.

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* 13, 329–342. doi:10.1038/nrg3174.

Ye, C., Hill, C. M., Wu, S., Ruan, J., and Ma, Z. (2016). DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci. Rep.* 6, 1–9. doi:10.1038/srep31900.

You, M., Yue, Z., He, W., Yang, X., Yang, G., Xie, M., et al. (2013). A heterozygous moth genome provides insights into herbivory and detoxification. *Nat. Genet.* 45, 220–225. doi:10.1038/ng.2524.

Yovkova, V., Otto, C., Aurich, A., Mauersberger, S., and Barth, G. (2014). Engineering the α-ketoglutarate overproduction from raw glycerol by overexpression of the genes encoding NADP+-dependent isocitrate dehydrogenase and pyruvate carboxylase in *Yarrowia lipolytica. Appl. Microbiol. Biotechnol.* 98, 2003–2013. doi:10.1007/s00253-013-5369-9.

Yu, Y., Ouyang, Y., and Yao, W. (2018). shinyCircos: an R/Shiny application for interactive creation of Circos plot. *Bioinformatics* 34, 1229–1231. doi:10.1093/bioinformatics/btx763.

Yue, J. X., Li, J., Aigrain, L., Hallin, J., Persson, K., Oliver, K., Bergström, A., Coupland, P., Warringer, J., Lagomarsino, M. C., Fischer, G., Durbin, R., and Liti, G. (2017). Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nat. Genet.* 49, 913-924. doi: 10.1093/bioinformatics/bty614.

Zerbino, D. R., McEwen, G. K., Margulies, E. H., and Birney, E. (2009). Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read *de novo* assembler. *PLoS One* 4, e8407. doi:10.1371/journal.pone.0008407.

Zhang, Y.-P., Chen, X.-J., Li, Y.-Y., and Fukuhara, H. (1992). Yeast sequencing reports. *LEU2* gene homolog in *Kluyveromyces lactis*. *Yeast* 8, 801–804. doi:10.1002/yea.320080914.

Zhao, S., and Fernald, R. D. (2005). Comprehensive algorithm for quantitative real-time polymerase chain reaction. *J. Comput. Biol.* 12, 1047–1064. doi:10.1089/cmb.2005.12.1047.

Zheng, W., Huang, L., Huang, J., Wang, X., Chen, X., Zhao, J., et al. (2013). High genome heterozygosity and endemic genetic recombination in the wheat stripe rust fungus. *Nat. Commun.* 4, 2673. doi:10.1038/ncomms3673.

Zill, O. A., Scannell, D. R., Kuei, J., Sadhu, M., and Rine, J. (2012). Evolutionary analysis of heterochromatin protein compatibility by interspecies complementation in *Saccharomyces*. *Genetics* 192, 1001–1014. doi:10.1534/genetics.112.141549.

Zill, O. A., Scannell, D., Teytelman, L., and Rine, J. (2010). Co-evolution of transcriptional silencing proteins and the DNA elements specifying their assembly. *PLoS Biol.* 8. doi:10.1371/journal.pbio.1000550.

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics* 29, 2669–2677. doi:10.1093/bioinformatics/btt476.

Zimin, A. V., Puiu D., Luo M. C., Zhu T., Koren S., Marcais G., Yorke J. A., Dvorak J., and Salzberg S. L. (2017). Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res*. doi: 10.1101/gr.213405.116.

Zimmermannova, O., Salazar, A., Sychrovà, H., and Ramos, J. (2015). *Zygosaccharomyces rouxii* Trk1 is an efficient potassium transporter providing yeast cells with high lithium tolerance. *FEMS Yeast Res.* 15, 1–11. doi:10.1093/femsyr/fov029.

Zordan, R. E., Ren, Y., Pan, S.-J., Rotondo, G., De Las Peñas, A., Iluore, J., et al. (2013). Expression plasmids for use in *Candida glabrata*. *G3 (Bethesda).* 3, 1675–86. doi:10.1534/g3.113.006908.

Zörgö, E., Chwialkowska, K., Gjuvsland, A. B., Garré, E., Sunnerhagen, P., Liti, G., et al. (2013). Ancient evolutionary trade-offs between yeast ploidy states. *PLoS Genet.* 9, e1003388. doi:10.1371/journal.pgen.1003388.

Zörgö, E., Gjuvsland, A., Cubillos, F. a., Louis, E. J., Liti, G., Blomberg, A., et al. (2012). Life history shapes trait heredity by accumulation of loss-of-function alleles in yeast. *Mol. Biol. Evol.* 29, 1781–1789. doi:10.1093/molbev/mss019.