University of Huddersfield

Doctoral Thesis

Evolutionary Genomics of Transposable Elements in the *Saccharomyces sensu lato* Complex

---

Author:

Cooper Alastair Grace

Supervisors:

Dr. Martin Carr

Dr. Patrick McHugh

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Department of Biological and Geographical Sciences

School of Applied Sciences

University of HUDDERSFIELD
Inspiring tomorrow's professionals

Copyright Statement

Declaration of Authorship

I, Cooper Alastair Grace, declare that this thesis titled, “Evolutionary Genomics of Transposable Elements in the *Saccharomyces sensu lato* Complex” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- I have acknowledged all main sources of help.

Signed:

_____

Date:

_____

"*Believing that I was viewing a basic genetic phenomenon, all attention was given, thereafter, to determining just what it was that one cell had gained that the other cell had lost. These proved to be transposable elements that could regulate gene expression in precise ways. Because of this I called them 'controlling elements'. Their origins and their actions were a focus of my research for many years thereafter."*

Barbara McClintock, 1984

Abstract

Transposable elements (TEs) are almost ubiquitous components of eukaryotic genomes that have long been considered solely deleterious or 'junk DNA'. They are split into two main forms, retrotransposons and DNA transposons, depending on the method of replication employed. Hosts have developed strategies for combating TEs including RNAi, methylation and copy number control. TEs have also evolved ways of persisting in the genome in order to survive, such as target site specificity. Two additional ways which may be utilised by TEs, positive selection and horizontal transfer, were investigated here primarily using the budding yeasts in the *Saccharomyces sensu lato* complex. These species typically contain up to five families of retrotransposons, designated *Ty1-5*, and multiple subfamilies, all of varying transpositional activity.

Discoveries of insertions evolving under positive selection and providing benefits to their hosts have been sporadic and serendipitous findings in a number of organisms. Full genome screenings for such insertions are rarely published, despite the impact TE insertions have upon their hosts. A population genomics approach was performed to address this issue in the genomes of *Saccharomyces cerevisiae* and sister species *S. paradoxus*. Signatures of positive selection acting upon *Ty* insertions were identified using Tajima's statistical *D* test. Neighbouring genes were also analysed to ascertain the true target of selection where hitchhiking linked the two. A subset of LTR-gene pairings were explored using qPCR in order to identify any effects on host gene expression the occupied loci may cause. Two genes displayed significantly increased levels of expression, which may be due to the presence of positive selection candidate LTRs, which in turn may contribute to improving host fitness.

This thesis further documents the systematic screening for *Ty*-like elements of all available genomes of budding yeast and related species. Extensive phylogenetic analyses estimated evolutionary relationships and possible horizontal transfer events of elements between the species. Evidence for in excess of 75 horizontal transfer events was uncovered here, around half of which

were successful in propagating in new genomes. The occurrence of horizontal transfer of TEs in the genomes of budding yeast is therefore far more common than previously documented.

During screening of genomes, a further potential method of avoiding host defences was uncovered. The divergence of the highly active *Ty4* family, which coincided with population isolation of multiple *Saccharomyces* species into subfamilies, was surprising given previous reports of this family being of particularly low activity. Such events are rarely recorded in eukaryotic genomes, and may also illustrate the compulsive spread of a new subfamily via horizontal transfer.

The investigations reported here represent the first genomic screening of *Ty* insertions in *Saccharomyces* for signatures of positive selection, and an updated, comprehensive search for evidence of HT between species of budding yeast. Both may act as methods for TE families to persist in the genomes of their hosts, and represent far more than simply 'junk DNA'.

Acknowledgements

I would like to acknowledge the following:

The entertaining folks in the office and hotdesking area – Marina, Katharina, Alessandro, George, Bobby, Gonzalo, Georgia, Fabiola, Daisy, Adam, Alex, Sophie, Zoe, Asma, Makhosi and Kim. Keep on procrastinatin' Claire. Each of you helped me in various ways, even if it was just random conversations here and there that prevented me from completely losing my mind. I only hope my badgering you about writing worked!

Marisa for giving me a push towards writing my thesis in LaTex rather than watching me fight through another day with Microsoft Word.

The amazing Dr. Catherine Finch, who understood the difficulties in completing a PhD and experienced them alongside me. Your friendship and support in the past year or so have been incredibly important to me.

Dr Patrick McHugh as my secondary supervisor and Cathy McHugh for your help with qPCR, and their students Bethan and Laura with qbase+. Drs Jarek Bryk, Chris Cooper and Dougie Clarke for helpful discussions and always being such nice guys. My study mentor Mary for her help and support over the last few years.

The amazing biology lab technicians Sophie, Felix, Elena and Maggie, without whom I'd probably still be looking for a sterile loop and a tub of sorbitol. They helped me become a better scientist while appreciating my sense of humour, for which I will always be grateful.

The other members of the research team Bernardo (I'll always bring you water), Holly for random lab conversations, Ginés, and of course Jade, who picked me up when my motivation waned and watched me draw crazy mind maps on whiteboards. Andromeda is also rather grateful for the endless supply of almost empty peanut butter jars…

Last but not least, my primary supervisor Dr Martin Carr. If you were hoping to inspire me to pursue a life of researching transposable elements, well done, you succeeded. So long, and thanks for all the tea.

Contents

List of Figures

List of Tables

List of Abbreviations

ARS	autonomously replicating sequence
BI	Bayesian Inference; phylogenetic method
BLAST	Basic Local Alignment Search Tool
cDNA	complementary DNA
DBS	double-strand break
ENA	European Nucleotide Archive
FASTA	FAST-All, or Pearson; nucleotide or protein sequence format
FLE	full-length element
GAG; *gag*	capsid-like domain of transposable elements
GCR	gross chromosomal rearrangement
gDNA	genomic DNA
GIRI	Genetic Information Research Institute
GO	gene ontology
HGT	horizontal gene transfer
HTT	horizontal transfer (of) transposable elements
IN	integrase; enzyme catalysing the integration of DNA into a genome
LBA	long-branch attraction
LINE	long interspersed nuclear elements
LTR	long terminal repeat
MAFFT	Multiple Alignment using Fast Fourier Transform
MCMC	Markov chains Monte Carlo
ML	Maximum Likelihood; phylogenetic inference method
mRNA	messenger RNA
NBRC	NITE Biological Resource Centre
NCBI	National Centre for Biotechnology Information
NEXUS	sequence file format

NGS	Next Generation Sequencing
ORF	open reading frame
PCR	polymerase chain reaction
PGC	Peterhof Genetic Collection
PLE	Penelope-like elements
POL; *pol*	polyprotein domain of transposable elements
Pol II or III	RNA Polymerase II or III
PP	posterior probability
PPT	polypurine tract
PR	protease
PRE	pheromone responding element
qPCR	quantitative polymerase chain reaction
RAxML	Randomised Axelerated Maximum Likelihood; phylogenetic inference program
RH	ribonuclease H
RIP	repeat-induced point mutations
RT	reverse transcriptase; enzyme catalysing the synthesis of cDNA from RNA template
SGD	*Saccharomyces* Genome Database
SGRP	*Saccharomyces* Genome Re-Sequencing Project
SINE	short interspersed nuclear repeat
SNP	single nucleotide polymorphism
TE	transposable element
tRNA	transfer RNA
TSS	transcription start site
UAS	upstream activator sequence
UCSC	University of California, Santa Cruz
UTR	untranslatable region
VLP	virus-like particle
WGD	whole genome duplication

Chapter 1

Introduction

1.1 Transposable Elements

Transposable elements (TEs) are almost ubiquitous components of eukaryotic genomes. They have long been assumed to be detrimental to the host in which they reside due to their replicative autonomy and have an ability to survive on a long-term basis within a genome (Biémont *et al.*, 1997; Kidwell and Lisch, 1997; Montchamp-Moreau, 1990; Arkhipova and Morrison, 2001; Wloch *et al.*, 2001; Dolgin and Charlesworth, 2008; Johnson *et al.*, 2010; Startek *et al.*, 2013). Discovered in maize by Barbara McClintock in the 1940s (McClintock, 1944), it took decades for the existence of widespread mobile genetic elements to be accepted.

There are some striking differences in TE variety, activity, copy number, distribution and genomic fraction across species. Genomic fractions range from as little as 2.7% of the puffer fish genome (Aparicio *et al.*, 2002) to more than 90% in maize (Jiao *et al.*, 2017). TEs are absent in a small number of species, such as the malaria parasite *Plasmodium falciparum* (Gardner *et al.*, 2002).

TEs are divided into two main groups based upon their transposition mechanisms: retrotransposons (Class I) and DNA transposons (Class II; reviewed in Slotkin and Martienssen, 2007). Both classes form families which can be active or inactive, the elements of which can be present in a genome as full-length, mutated and fossil copies (Le Rouzic and Capy, 2006; Pritham, 2009).

Over 45% of the human genome consists of TE insertions (Cordaux and Batzer, 2009) and it is likely that far more is derived from older insertions (Burns and Boeke, 2012). Approximately 3% of the human genome comprises DNA transposons (Cordaux and Batzer, 2009) and none are currently active (Burns and Boeke, 2012). The majority of TEs found in humans are non-LTR retrotransposons such as *L1*, *Alu* and *SVA* (Cordaux and Batzer, 2009; Burns and Boeke,

2012). Human endogenous retrovirus (HERV) LTR retrotransposons are present in the genome, predominantly as degraded copies (Burns and Boeke, 2012).

1.1.1 DNA transposons

DNA transposons primarily move their DNA in a 'cut and paste' fashion via double strand breaks (reviewed in Slotkin and Martienssen, 2007). Many DNA transposons encode a transposase enzyme flanked by inverted terminal repeats (ITRs; reviewed in Kazazian, 2004). Non-autonomous copies, in which the transposase has become defective due to mutation, are not always inactive as they can still utilise the functional enzyme of autonomous elements (Hartl and Clark, 2007). Transposons are able to increase copy number if, after transposition, the double-strand break (DSB) left by the element is repaired using the host's sister chromatid, resulting in a copy of the original element being replicated. Transposition may, however, be non-replicative if the DSB is repaired without a template (Hartl and Clark, 2007).

A subclass of DNA transposons contains the helitron family, which lack ITRs and transposase, instead encoding a DNA helicase and replicator initiator protein which enables them to replicate using a rolling circle mechanism (reviewed by Kapitonov and Jurka, 2007).

As DNA transposons are not widespread in the yeast species studied here, the focus is on retrotransposons.

1.1.2 Retrotransposons

Retrotransposons, particularly when they are the only type of elements inhabiting a genome, are used as model systems as they reflect the evolutionary history of their hosts and also the impact that they alone impose on an individual (Bowen *et al.*, 2003) and on a population (Hosid *et al.*, 2012). Constructed elements containing selectable genes can be useful for gene mutation, target gene activation, chromosomal mapping and in gene cloning (Garfinkel *et al.*, 1988).

Retrotransposons replicate in a way similar to that of retroviruses: via reverse transcription using an RNA intermediate to reintegrate cDNA into the host genome. The first retrotransposons were discovered in the yeast *Saccharomyces cerevisiae* (Cameron *et al.*, 1979) and the fly *Drosophila melanogaster* (Mount and Rubin, 1985). Like DNA transposons, non-autonomous retrotransposons are able to utilise the products of autonomous elements in order to transpose, threatening

the continuation of the family to which they belong (Le Rouzic and Capy, 2006; Le Rouzic *et al.*, 2007).

Retrotransposons are divided into three orders: according to the presence or absence of flanking long terminal repeats (LTRs), and also the small group of variable Penelope-like elements (PLE; Eickbush and Jamburuthugoda, 2008). They are then further divided by superfamily (Figure 1.1). Non-LTR retrotransposons consist of long interspersed elements (LINEs) and short interspersed elements (SINEs) that typically possess a 5' untranslatable region (UTR) and a 3' polyadenylate tail (reviewed in Kazazian, 2004). Retrotransposons typically contain two open reading frames (ORFs; Boeke and Devine, 1998).

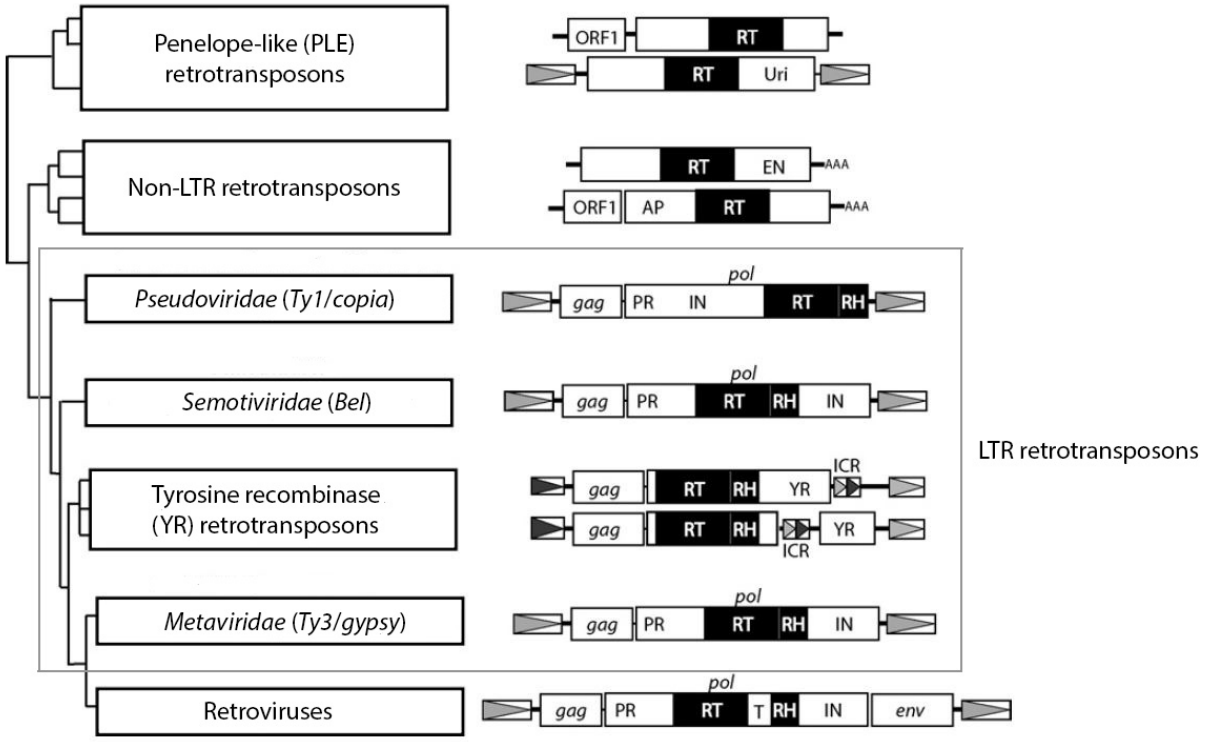

Figure 1.1: **Phylogenetic relationships and structures of the main retrotransposons.** Left: phylogenetic summary of the main lineages using transposase and reverse transcriptase sequences. Right: typical structures of each order of elements. ORF – open reading frame; RT – reverse transcriptase; RH – RNase H; PR – protease; IN – integrase; T – tether domain; APE – apurinic endonuclease; EN – endonuclease; Uri – domain similar to endonuclease; YR – tyrosine recombinase; AAA – polyadenylate tail. Boxed arrows indicate LTRs or internal complimentary repeats (ICRs). Adapted from Eickbush and Jamburuthugoda (2008).

Although the line between retroviruses and LTR retrotransposons is somewhat blurred (reviewed by Sandmeyer and Menees, 1996), retroviruses have most likely descended from retrotransposons (Figure 1.1). Retroelements acquired infectious abilities and the *env* gene, possibly during hybridisation with another virus, in order to evolve into the diverse family of *Retroviridae*.

The presence of *env*, encoding envelope proteins, was once used to distinguish between retroviruses and retrotransposons, but independent work by Laten *et al.* (1998) and Wright and Voytas (1998) has shown that this additional ORF can be present in retrotransposons. Additionally, some elements are endogenous retroviruses that possess these viral characteristics, such as the *gypsy* element in *D. melanogaster*, and act as retroviruses (Kim *et al.*, 1994; Song *et al.*, 1994). However, the overwhelming majority of retrotransposons did not gain an *env*-like gene and so are confined to their hosts.

1.1.2.1 LTR retrotransposons

The order of LTR retrotransposons encompasses BEL (Frame *et al.*, 2001), tyrosine recombinase (YR; which includes DIRS-1, Poulter and Goodwin, 2005), *Pseudoviridae* and *Metaviridae* families (reviewed by Havecker *et al.*, 2004; Figure 1.1). *Pseudoviridae* and *Metaviridae* are the focus of this study, the structures of which are similar to that of retroviruses (Figure 1.1). Their ORFs are homologous to retroviral *gag* and *pol* but lack the retroviral infectious elements and *env* encoding envelope proteins (Wilhelm and Wilhelm, 2001). *Gag* encodes structural proteins to construct the virus-like particles (VLPs) in which reverse transcription occurs (Chapman *et al.*, 1992), whereas *pol* encodes an enzymatic polyprotein with multiple domains. Its catalytic capabilities include protease (PR); a form of reverse transcriptase (RT); integrase (IN) and ribonuclease H (RH), all of which are required for retrotransposition (Lesage and Todeschini, 2005; Wilhelm and Wilhelm, 2001). *Pol* is translated as a single polyprotein which is then processed by the element encoded PR (Garfinkel *et al.*, 1991; Kirchner and Sandmeyer, 1993). LTR retrotransposons fall into two major superfamilies: *Ty1/copia* (*Pseudoviridae*) and *Ty3/gypsy* (*Metaviridae*) depending on the order in which the domains of *pol* are encoded (Wilhelm and Wilhelm, 2001; Figure 1.1). RT is by far the most conserved domain (Xiong and Eickbush, 1988, 1990).

Flanking the ORFs are LTRs, which are functionally similar to those of retroviruses, allowing the element template to correctly be copied and terminated. LTRs are under different evolutionary constraints than internal coding regions. Providing their regulatory regions stay functional (Figure 1.2), LTR sequences appear to be free to diverge beyond that seen in the *gag* and *pol* regions (Benachenhou *et al.*, 2009, 2013).

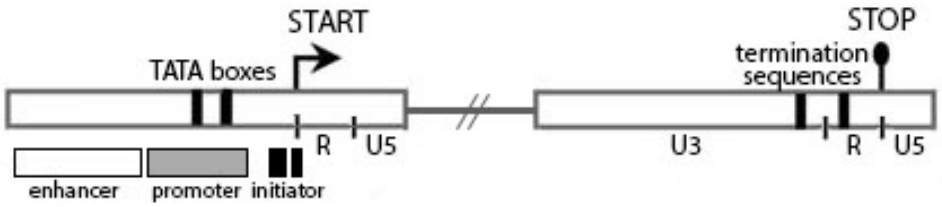

Figure 1.2: **Structure of the LTRs flanking a *Ty1/copia* element.** // indicates the internal coding DNA of the element. Adapted from Cullen *et al.* (1985), Jordan and McDonald (1998) and Curcio *et al.* (2015). The U3 region of the 5' LTR is not labelled due to the visualisation of the enhancer, promoter and initiator regions.

1.1.2.2 LTRs and retrotransposition

LTRs contain three regions: a repeated region R which is located between two unique regions U3 and U5, containing sequences for transcription initiation and the site for polyadenylation (Jordan and McDonald, 1998; Perlman and Boeke, 2004; Varmus, 1988; Figure 1.2). The R and U5 regions are generally more conserved than U3 (Benachenhou *et al.*, 2009). A primer binding site (PBS) is located immediately downstream of the 5' LTR and a polypurine track (PPT) is found immediately upstream of the 3' LTR (Zhang *et al.*, 2014). This basic structure of LTRs is thought to be common across most families of elements.

The mechanism of retrotransposition is believed to be very similar to that of retroviruses, and will only be briefly described here. The process begins when elements are transcribed to mRNA by host RNA polymerase II (Pol II) in their entirety from the 5' R region to the 3' R region, which then function as a template for retrotransposition and translation (Todeschini *et al.*, 2005). During retrotransposition, the mRNA is degraded by RH after being used as a template for reverse transcription of the element into cDNA by RT. In a complex process, the 5'-U3 and 3'-U5 regions are replaced (Elder *et al.*, 1983; Clark *et al.*, 1988; Perlman and Boeke, 2004) before the full-length cDNA copy of the element is incorporated back into the genome by IN (Lauermann and Boeke, 1997). Insertion into the genome causes a target site duplication due to the staggered break made in the host DNA (Gafner and Philippsen, 1980). LTR sequences within the same element are identical upon retrotransposition (Jordan and McDonald, 1998; Bowen and McDonald, 2001) and therefore indicate age relative to their insertion (Dangel *et al.*, 1995; SanMiguel *et al.*, 1996, 1998; Kijima and Innan, 2010).

1.2 Saccharomycetaceae

All yeasts within the *Saccharomyces sensu lato* group are members of the large Ascomycota phylum, Saccharomycotina subphylum, and finally the Saccharomycetaceae family. Other yeasts within the Ascomycota phylum are the Taphrinomycotina, or fission yeasts, such as *Schizosaccharomyces* and Pezizomycotina such as *Neurospora crassa*. Saccharomycotina members as a whole make up virtually two thirds of all known yeast species (reviewed in Dujon and Louis, 2017).

Sensu lato, meaning “in the broad sense”, was part of an older classification system used to differentiate between the members of the *sensu stricto* (meaning “in the strictest sense”) complex and the more divergent species, often still designated *Saccharomyces*. Many were later reclassified by Kurtzman (2003) to new genera after extensive phylogenetic analysis of distinguishing genomic regions. Kurtzman suggested changing the taxonomic status, abandoning the subdivision of *sensu stricto* and *sensu lato* in favour of establishing new genera including *Naumovozyma, Lachancea, Kluyveromyces, Kazachstania*, and *Torulaspora*, among others (James *et al.*, 1997; Muller and McCusker, 2009; Figure 1.3). Despite the newly named genera, the group is still traditionally referred to as *sensu lato*, and consists of the first 12 of the 14 Saccharomycetaceae clades (Figure 1.3; Kurtzman, 2003).

All but one clade (*Zygotorulaspora*, Clade 8) is represented by at least one fully sequenced genome, revealing far more heterogeneity than within *Saccharomyces* species (Muller and McCusker, 2009). However, sequencing quality across the species varies from relatively low coverage Illumina to the single molecule PacBio and Nanopore-based methods (reviewed in Feng *et al.*, 2015; Rhoads and Au, 2015). New species are being discovered continually, even within *Saccharomyces* (Naseeb *et al.*, 2017), causing species phylogenies to frequently become outdated. Species are further complicated by hybridisation and reproductive isolation of individual populations (reviewed in Dujon and Louis, 2017).

First proposed by Wolfe and Shields (1997), it has now been established that the ancestor of clades 1-6 underwent whole genome duplication (WGD; Scannell *et al.*, 2006; Marcet-Houben and Gabaldon, 2015). The *Vanderwaltozyma, Tetrapisispora* and *Naumovozyma* clades for example, diverged soon after the WGD event, accounting for major differences between these species and members of the *Saccharomyces* clade (Scannell *et al.*, 2006, 2007, 2011). More distant still are the pre-WGD species such as *Lachancea waltii*, which contains only around 500 paralogues with the $\sim$6000 genes of *S. cerevisiae* (Replansky *et al.*, 2008). The genes of post-WGD species evolved

new functions, underwent sequence divergence and inactivation due to mutations which then left relics of once functional genes (reviewed by Liti and Louis, 2005).

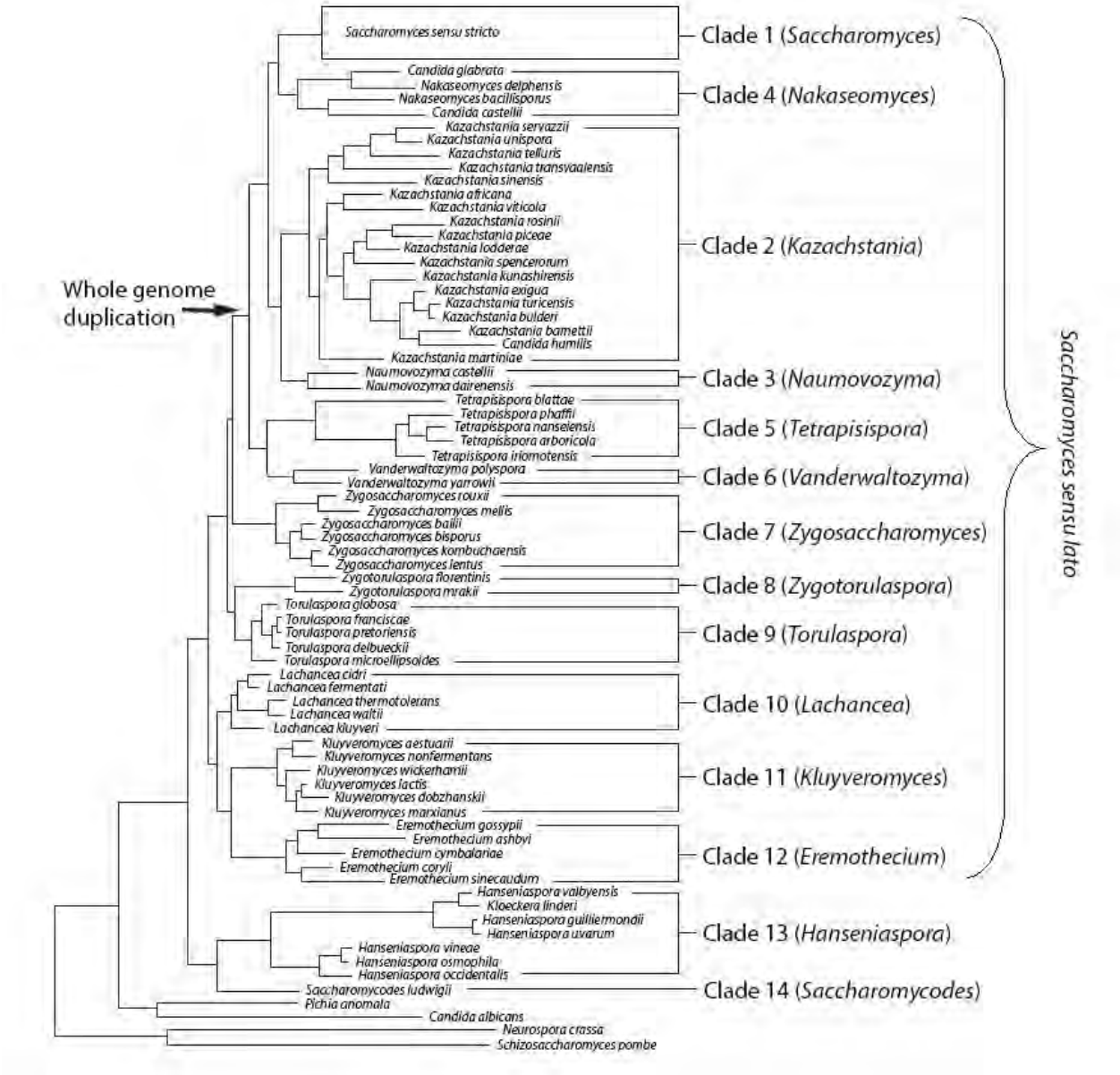

Figure 1.3: **The clades of the *Saccharomyces sensu lato* complex.** This figure illustrates the positions of the 11 clades of the *sensu lato* group. Based on maximum likelihood performed by Hedtke *et al.* (2006), and maximum parsimony analyses in Kurtzman (2003) and Kurtzman and Robnett (2003). Both used concatenated alignments of eight conserved genes. Clade numbers and revised genus names are from Kurtzman (2003). The outgroup consists of sequences from *Neurospora crassa* and *Schizosaccharomyces pombe*. Not all species from each genus are displayed. Adapted from Rozpędowska *et al.* (2011).

1.2.1 The *Saccharomyces sensu stricto* complex

Yeast taxonomy has proven to be a very controversial topic. As most budding yeast are morphologically identical and often physiologically invariable, few species could be distinguished on these characteristics and reproductive studies alone. This classical system of species identification has since been abandoned in favour of the sequencing of species specific regions, such as ribosomal DNA, and whole genome sequencing (WGS). As a result, many species that were once designated *Saccharomyces* have been reclassified as separate genera (Kurtzman, 2003; Kurtzman and Robnett, 2003; Kurtzman *et al.*, 2011).

The *Saccharomyces sensu stricto* complex originally consisted of *S. cerevisiae, S. paradoxus* and *S. bayanus* (Vaughan-Martini and Kurtzman, 1985; Vaughan-Martini, 1989). *S. pastorianus* (syn. *S. carlsbergensis*) was added as a member of the complex until it was realised to be a natural sterile hybrid of *S. cerevisiae* and an unknown *S. bayanus*-like species (Casey and Pedersen, 1988; Hansen and Kielland-Brandt, 1994). Naumov *et al.* (2000) added three new species to the *sensu stricto* complex: *S. mikatae, S. kudriavzevii* and *S. cariocanus*, the latter of which has been labelled the 'South American *S. paradoxus*', and was later discounted as a separate species by Liti *et al.* (2006, 2009). The debate as to whether these are separate species is ongoing (Naumov, G. pers. comm., 2016). The novel species *S. arboricola* was later added to the complex (Wang and Bai, 2008; previously incorrectly designated *S. arboricolus*).

Taxonomic and debates have surrounded *S. bayanus* for many years (Nguyen *et al.*, 2011). *S. uvarum* was long considered a variant of *S. bayanus*, mostly due to identification errors (Perez-Traves *et al.*, 2014). Many research teams argued that *S. uvarum* should be reclassified as its own distinct species (Rainieri *et al.*, 1999; Pulvirenti *et al.*, 2000), and after careful consideration of the available data and arguments, it is treated as such during this thesis. As of the most recent Yeast Taxonomy (Kurtzman *et al.*, 2011), it has not yet been granted taxonomic recognition, but may be rectified in light of so much new information. *S. bayanus* is generally accepted as a hybrid by geneticists, who mostly ignored the taxonomic debate and so adopted a pure *S. uvarum* as the reference strain (reviewed by Hittinger, 2013). *S. bayanus*' hybrid state was finally recognised when the missing parental species of fellow hybrid *S. pastorianus* was revealed as *S. eubayanus* (Libkind *et al.*, 2011). It was thought that much of the *Saccharomyces* mystery had been solved as of 2015, but very recently an additional species was isolated, *S. jurei*, by Naseeb *et al.* (2017). Table 1.1 summarises the species details of *Saccharomyces*.

Species	Genome size (Mb)	Intraspecies variation (%)	Known populations	Divergence from *S. cerevisiae**	Reference(s) for variation and populations
S. arboricola	11.6	unknown	3	Unknown	Naumov *et al.*, 2013; Gayevskiy and Goddard, 2016
S. cerevisiae	12.2	1.4	13	-	Wang *et al.*, 2012
S. eubayanus	11.7	6.02-7.57	5	20 mya	Bing *et al* 2014; Peris *et al.*, 2014
S. jurei	unknown	unknown	1	unknown	Naseeb *et al.*, 2017
S. kudriavzevii	11.3	4.1	3	15-20 mya	Hittinger *et al.*, 2010
S. mikatae	11.4	unknown	1	10-15 mya	Naumov *et al.*, 2000
S. paradoxus	12.0	3.8	5	5-10 mya	Liti *et al.*, 2009; LeDucq *et al.*, 2014
S. uvarum	11.5	4.4	3	20 mya	Almeida *et al.*, 2014

Table 1.1: **Summary of the *Saccharomyces* species.** *data from Replansky *et al.* (2008), achieved with whole genome alignments. mya - million years ago.

The *Saccharomyces* genus (Figure 1.4) is thought to have originated within the last 20 million years (Kellis *et al.*, 2003; Taylor and Berbee, 2006), but the exact location of its origin is unclear. Wang *et al.* (2012) and Bing *et al.* (2014) both suggest a Far Eastern origin, whereas Almeida *et al.* (2014) and Peris *et al.* (2014) instead hypothesised the ancient supercontinent of Gondwana or Patagonia. Both the Far East and Patagonia show diversity in species and abundant populations, so a compromise hypothesis is that the *Saccharomyces* ancestor originated in Gondwana before subsequently moving to the Far East where it diverged further (Almeida *et al.*, 2014). The most recent work strongly supports an out-of-China origin for the genus (Peter *et al.*, 2018).

Although the species are closely related, protein divergence is similar to that of humans and chickens (reviewed by Dujon, 2006). Nucleotide sequence divergence between *S. cerevisiae* and *S. paradoxus* is similar to that of humans and mice; that between *S. eubayanus* and *S. uvarum* is likened to that of humans and macaque, and differing populations of *S. kudriavzevii* or *S. paradoxus* is equated to the differences between humans and chimps (Hittinger, 2013).

Saccharomyces species occupy many habitats and environments (Landry *et al.*, 2006; Goddard and Greig, 2015). Wild *Saccharomyces* strains and species are most often found to be associated with tree bark, soil and leaves (Sniegowski *et al.*, 2002; Glushakova *et al.*, 2007; Libkind *et al.*, 2011; Wang *et al.*, 2012), and have also been isolated from insects such as *Drosophila* (Naumov *et al.*, 2000) and wasps (Stefanini *et al.*, 2012, 2016). However, this could simply be down to sampling bias, as yeasts which are colonised in the laboratory on rich, incubated media may not accurately reflect the true range of species in any given sampled environment. Domesticated strains and species are used in a vast variety of industrial settings, such as in the wine and lager

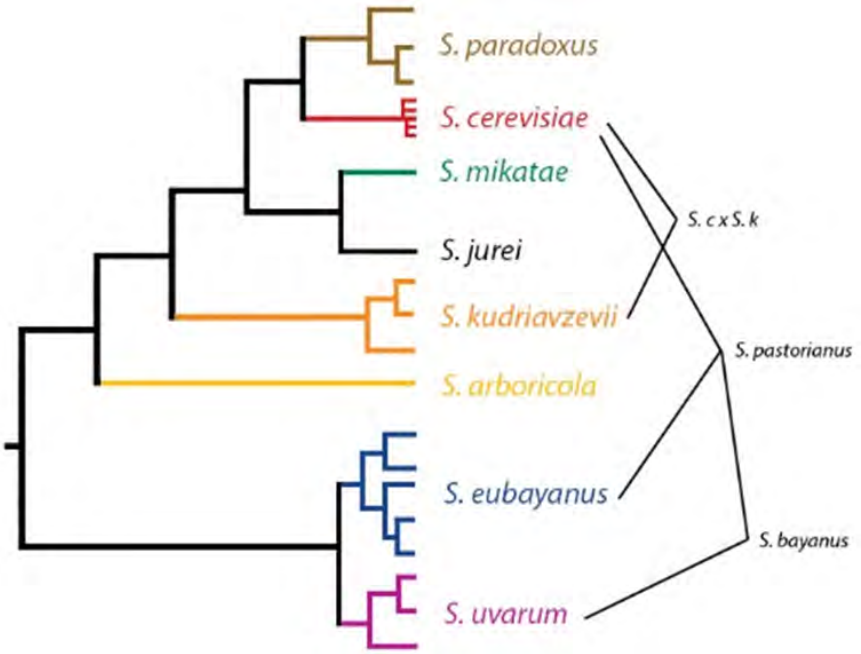

Figure 1.4: **Cladogram of species and populations of the *Saccharomyces sensu stricto* complex.** Topology is adapted from Dujon and Louis (2017). Taxonomic species on the left follow the colour scheme used throughout this thesis. *S. jurei* is yet to be sequenced and so has not been assigned a colour. Major hybrids are on the right; currently only two of which are named taxonomically (*S. pastorianus* and *S. bayanus*). As it was not recognised as a species separate from *S. paradoxus*, *S. cariocanus* was omitted here, but would lie within one of the existing populations of *S. paradoxus*. Additional populations of *S. arboricola* has since been discovered, but their relationship to the Chinese type strain is yet to be elucidated. Due to their abundance, not all *S. cerevisiae* populations are shown.

industries and also for baking (reviewed in Pretorius, 2000; Sicard and Legras, 2011). Hidden fungal diversity studies suggest further environments are yet to be discovered (Bass *et al.*, 2007; Bass and Richards, 2011; Dunthorn *et al.*, 2017).

Speciation in *Saccharomyces* is a complex matter. Although it is most likely to occur through gradual accumulation of sequence divergences, it is clear that gross chromosomal rearrangements (GCRs) encourage the process (Delneri *et al.*, 2003). Due to hybridisation and population divergence, it is becoming difficult to establish boundaries between species. Arguments have been made to split existing species, for example, Taylor *et al.* (2006) suggested splitting *S. cerevisiae* into a minimum of four separate species, allowing subspecies such as the probiotic *S. boulardii* to be classified separately. Rather than the established bifurcating phylogeny as in Figure 1.4, the evolution of the *Saccharomyces* clade appears to be reticulate, with gene flow between species and populations far more common than previously thought (reviewed in Dujon and Louis, 2017). A balanced approach in determining species boundaries, taking into account genome sequence,

ecology, geographical origin and reproductive isolation would allow for gene flow caused by horizontal transfer and introgression (see Section 1.4).

1.2.2 *Sensu stricto* species

1.2.2.1 *S. cerevisiae*

The original species of the *sensu stricto* complex, *S. cerevisiae* has been known by many synonyms and was the first eukaryote to have its genome sequenced (Goffeau *et al.*, 1996). The reference strain S288c was later discovered to be one of a number of strains possessing a mosaic genome (Engel *et al.*, 2014). Conversely, "clean" lineages are free from gene flow from other populations (Liti *et al.*, 2009).

Despite some initial debate, it has now been established that *S. boulardii* is a variant of *S. cerevisiae* (van der Aa Kuhle and Jespersen, 2003; Fietto *et al.*, 2004; MacKenzie *et al.*, 2008; Khatri *et al.*, 2017). Many other variants were described as separate species of *Saccharomyces*, but modern techniques such as genome sequencing have indicated that none are diverse enough to be classed as entirely new species (Kurtzman, 2003).

Isolated on all continents (Liti *et al.*, 2009), *S. cerevisiae* consists of several populations and subpopulations (Figure 1.4), distinct on the basis of geographical origin as well as the extent of association with human activities. As also seen in other yeasts, *S. cerevisiae* populations can become reproductively isolated (reviewed in Greig, 2009), such as the distinct Malaysian population, which has undergone ten GCRs in order to prevent further mating with other populations (Liti *et al.*, 2009; Cubillos *et al.*, 2011). Many authors have conducted population genomics studies on strains of *S. cerevisiae* (reviewed in Liti and Schacherer, 2011; Louis, 2011; Dujon and Louis, 2017; individual reports include Fay and Benavides, 2005; Aa *et al.*, 2006; Liti *et al.*, 2009; Schacherer *et al.*, 2009; Almeida *et al.*, 2015), but it is beyond the scope of this work to discuss them here.

1.2.2.2 *S. paradoxus* and *S. cariocanus*

In comparison to the extensively studied and domesticated *S. cerevisiae*, far less is known about its wild sister species *S. paradoxus*. Of more limited geographical distribution than *S. cerevisiae*, *S. paradoxus* contains a number of reproductively isolated populations and is highly divergent (Sniegowski *et al.*, 2002; Johnson *et al.*, 2003; Tsai *et al.*, 2008; Liti *et al.*, 2009), so much so that

it may be on its way to becoming three separate species, each occupying a different continent (Taylor *et al.*, 2006).

S. cariocanus was the first of three species described by Naumov *et al.* (2000) and for a time considered a distinct species within the *sensu stricto* complex, but classified as a subpopulation of *S. paradoxus* in the most recent Yeast Taxonomy (Kurtzman *et al.*, 2011). It has yet to be isolated outside of Brazil (Morais *et al.*, 1992), and is very close in sequence to the North American population of *S. paradoxus* (Liti *et al.*, 2006, 2009). It is unable to form viable spores with *S. paradoxus*, which has been attributed to four major reciprocal translocations in both existing strains of *S. cariocanus* (Liti *et al.*, 2006).

Whereas some authors believe it is a *S. paradoxus* variant (Liti *et al.*, 2006, 2009), others maintain that *S. cariocanus* is undergoing speciation (Naumov, G. pers. comm., 2016). After careful consideration, it was decided that *S. cariocanus* would be treated as a separate species for the purpose of the work conducted here.

1.2.2.3 *S. mikatae*

Very little documentation exists for *S. mikatae* and it has yet to be isolated outside of Japan (NBRC, 2010). It was first isolated by Yamada *et al.* (1993) and formally described by Naumov *et al.* (2000). The single strain was later sequenced by Kellis *et al.* (2003) and vastly improved by Scannell *et al.* (2011). No natural hybrids involving *S. mikatae* were reported until Bellon *et al.* (2013), but a 4.5kb *S. mikatae* introgression region was previously reported in some strains of *S. cerevisiae* (Dunn *et al.*, 2012). The type strain shares 73% of overall nucleotide identity with that of *S. cerevisiae* (Bellon *et al.*, 2013).

1.2.2.4 *S. kudriavzevii*

Originally isolated by Kaneko and Banno (1991), *S. kudriavzevii* was the final of the three species described by Naumov *et al.* (2000). The type strain was later sequenced by Scannell *et al.* (2011). The number of studies involving *S. kudriavzevii* is on the increase, now that a further population in Europe has been discovered (Sampaio and Gonçalves, 2008; Figure 1.4) outside of the original Japanese population. Hybridisation with *S. cerevisiae* is regularly reported (Belloch *et al.*, 2009; Borneman *et al.*, 2012; Combina *et al.*, 2012; Lopandic *et al.*, 2007).

1.2.2.5 *S. arboricola*

A new species, *S. arboricola*, was discovered by Wang and Bai (2008) and sequenced by Liti *et al.* (2013). The species was thought to be geographically restricted as it had not been isolated outside of Eastern Asia (Naumov *et al.*, 2013) until sampling in New Zealand forests revealed a further population of this species (Gayevskiy and Goddard, 2016).

1.2.2.6 *S. eubayanus*

The existence of *S. eubayanus* was hypothesised when it became clear that *S. bayanus* was not the parent species of hybrid *S. pastorianus*. The fact that *S. eubayanus* was discovered initially in South America (Libkind *et al.*, 2011), puzzled yeast geneticists as the hybridisation event that resulted in *S. pastorianus* predated the discovery of the Americas. Further populations were discovered in China (Bing *et al.*, 2014), North America (Peris *et al.*, 2014) and more recently New Zealand (Gayevskiy and Goddard, 2016), providing an answer to the problem. Once the reference strain had been sequenced, it was apparent that *S. eubayanus* is 99.5% identical to the non-*S. cerevisiae* portion of *S. pastorianus* (Baker *et al.*, 2015) and allowed taxonomists to correctly identify *S. uvarum* and the hybrid species *S. bayanus*. Asian populations have been reported as genetically distinct from the South American strains (Gibson *et al.*, 2015).

1.2.2.7 *S. uvarum*

Originally considered a variant of *S. bayanus*, Nguyen and Gaillardin (2005) proved that *S. uvarum* is in fact a separate species. It is also a parental species of hybrid *S. bayanus* alongside *S. eubayanus* and *S. cerevisiae* (Libkind *et al.*, 2011). *S. uvarum* was sequenced by Kellis *et al.* (2003) and is the sister species of *S. eubayanus*, with a similar relationship to that of *S. cerevisiae* and *S. paradoxus*. Diverse populations are found in South America and across Europe (Libkind *et al.*, 2011; Almeida *et al.*, 2014; Sylvester *et al.*, 2015). This species has now been isolated on every continent except Africa and Antarctica (Hittinger, 2013).

1.2.2.8 A new species: *Saccharomyces jurei*

Isolated in the French Alps by Naseeb *et al.* (2017), *S. jurei* appears to be the sister species to *S. mikatae* but also shows similarity to *S. paradoxus* based on ITS, 26s rRNA D1/D2 regions. It

is reproductively isolated from the other *Saccharomyces*, as only up to 2% of spores were viable depending on the mating species (Naseeb *et al.*, 2017).

1.2.3 The complexity of the *Saccharomyces sensu stricto* species: hybridisation

There is evidence to suggest that, despite being classified as distinct species, *Saccharomyces* are not truly reproductively isolated. Spontaneous hybridisation occurs between most *Saccharomyces* species, commonly in association with industrial fermentation and vineyards (Barros Lopes *et al.*, 2002; Liti *et al.*, 2005; González *et al.*, 2006; Peris *et al.*, 2012; Perez-Torrado *et al.*, 2015; Sipiczki, 2008; Groth *et al.*, 1999) and is thought to be a result of evolutionary adaptation to the harsh industrial environments (Zeyl *et al.*, 1996; Matzke *et al.*, 1999; Tofalo *et al.*, 2013; Marti-Raga *et al.*, 2017). However, it appears that away from industry, stable hybrids are far less common as the interfertility of species is generally low (Naumov, 1996; Marinoni *et al.*, 1999; Muller and McCusker, 2009). Interfertile hybrids are however able to backcross with one of the parental species, with each successive mating cycle becoming easier (Hittinger, 2013). Due to this high rate of variability caused by introgression and hybridisation, species boundaries can be remarkably ambiguous (Muller and McCusker, 2009).

Before the Kurtzman (2003) taxonomic changes, almost 10% of previously classified *sensu stricto* species were actually hybrids (Liti and Louis, 2005). Now, only two have been recognised taxonomically and therefore named: *S. bayanus* (a complex triple hybrid of *S. cerevisiae, S. eubayanus* and *S. uvarum*) and *S. pastorianus* (*S. cerevisiae* x *S. eubayanus*, sometimes referred to as *S. carlsbergensis*) (Kurtzman, 2003). These are both the products of artificial environments and are rarely found in the wild (Tofalo *et al.*, 2013). Away from human influence and industrial settings, it is unclear how often hybridisations naturally occur (Landry *et al.*, 2006). Unnamed natural hybrids include *S. cerevisiae* x *S. kudriavzevii* (Borneman *et al.*, 2012; Peris *et al.*, 2012) and *S. cerevisiae* x *S. paradoxus* (Barbosa *et al.*, 2016), whereas others have to be artificially created in a laboratory environment such as *S. cerevisiae* x *S. uvarum* (Dunn *et al.*, 2013). Besides *S. bayanus*, only two reports of triple hybridisation have been made (Groth *et al.*, 1999; González *et al.*, 2006).

1.3 Transposable elements in *Saccharomyces cerevisiae*: a model organism

S. cerevisiae is one of the most thoroughly studied eukaryotes in biology, and is a model organism in the study of population and evolutionary genetics (reviewed by Zeyl, 2000; Liti and Louis, 2005; Landry *et al.*, 2006; Ruderfer *et al.*, 2006; Replansky *et al.*, 2008). It was also used as an early system for studying TE insertions in eukaryotes (e.g. Ciriacy and Breilmann, 1982; Errede *et al.*, 1980, 1985, 1987; Klein and Petes, 1984; Silverman and Fink, 1984; Adams and Oeller, 1986; Paquin and Williamson, 1986; van Arsdell *et al.*, 1987; Clark *et al.*, 1988; Stucka *et al.*, 1989). With few exceptions, such as the study by Liti *et al.* (2005), the TE content of the other *Saccharomyces* species has not been extensively explored. TEs are also rarely investigated in other yeast species. Neuvéglise *et al.* (2002) used data generated by the Génolevures Consortium (Souciet *et al.*, 2000) to survey the TE content of 13 yeasts. Using random sequence tags (RSTs), the authors were able to construct the sequences of many of the elements in most species. However, the state of TEs over the entirety of the genomes could not be truly elucidated due to low coverage and small fraction of the genomes sequenced ($\sim$20%). It is thought that most hemiascomycetous yeasts have undergone a massive loss of their TEs (Neuvéglise *et al.*, 2002; Dujon, 2006), as other yeasts such as *Candida* and *Yarrowia* also possess DNA transposons and non-LTR retrotransposons (Goodwin *et al.*, 2001; Casarégola *et al.*, 2002; Neuvéglise *et al.*, 2005).

Various strains of *S. cerevisiae* have undergone extensive TE screenings which have shown that the yeast contains only LTR retrotransposons (Figure 1.5), with the exception of two strains, AWRI1631 (Borneman *et al.*, 2008) and M2ONO800-1A (Legras *et al.*, 2018), which contain degenerate copies of the DNA transposon family *Rover* (Sarilar *et al.*, 2015). It is most likely to have arisen by horizontal transfer (Section 1.4) from a closely related member of the Saccharomycetaceae taxonomic group such as *Torulaspora* (Legras *et al.*, 2018), some species of which contain DNA transposons, but these elements are, as a whole, absent from the *Saccharomyces sensu stricto* complex.

Initial full genome surveys of TEs in the reference strain S288c were independently performed by Kim *et al.* (1998) and Hani and Feldmann (1998). These have since been updated by Carr *et al.* (2012), who used less strict constraints when identifying elements, and included partial insertions in the total copy number. 51 insertions in the reference strain are intact and likely functional; the remaining 432 are incomplete due to degradation and age. The TE content of *S. cerevisiae*

is estimated at ∼3% (Carr *et al.*, 2012). However, the TE fraction of a genome will always be an underestimate as older insertions become increasingly difficult to recognise due to sequence divergence. There is also a degree of disruption of existing insertions caused by the repeated transposition of new elements into occupied loci (Carr *et al.*, 2012).

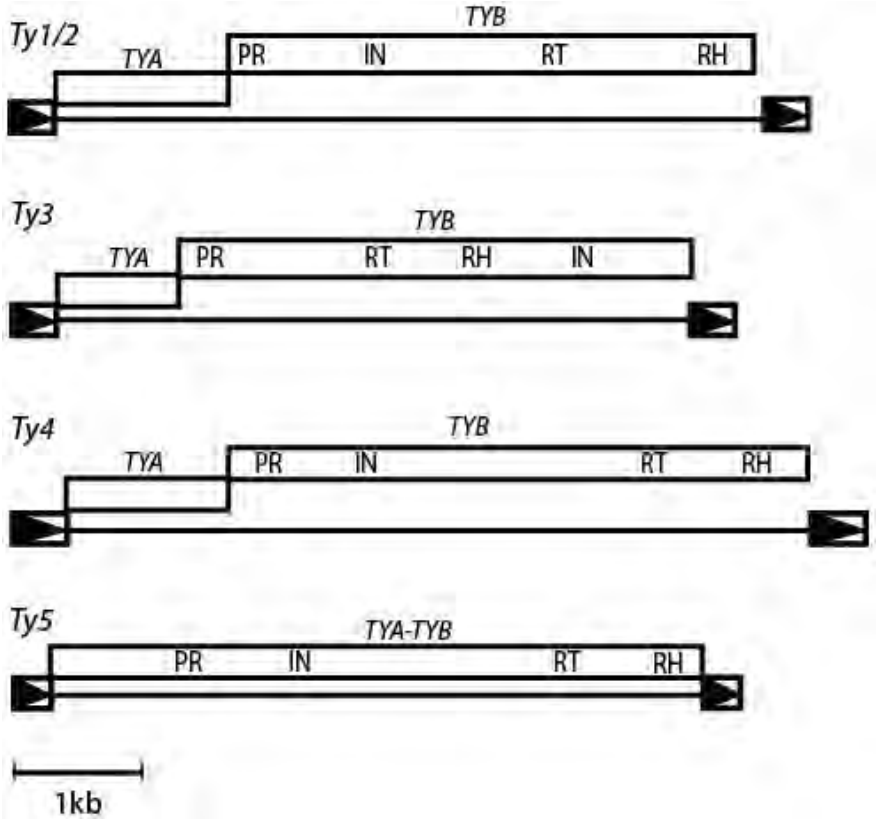

Figure 1.5: **Basic structure of the main families of retrotransposons in *S. cerevisiae*.** Elements within the *Ty1/2* superfamily are grouped together as they share a structure that is virtually identical. Boxed arrows represent LTRs, joined by the mRNA template produced. *Ty1-4* contain a frameshift between *TYA* and *TYB*, whereas *Ty5* elements possess fused ORFs. Adapted from Kim *et al.* (1998), Jordan and McDonald (1999*b*) and Lesage and Todeschini (2005).

The elements in *S. cerevisiae* are representative of retrotransposons in that they contain two open reading frames, *TYA* and *TYB* (homologous to retroviral genes *gag* and *pol*, respectively) and are flanked by 5' and 3' LTRs (Figure 1.5). Based on structure and sequence differences, they form a number of variant families (*Ty1-5*), subfamilies and chimeric families. The main five families were named in the order in which they were discovered (Lesage and Todeschini, 2005).

Numerous investigations into the genomic content of other strains have now been completed (e.g. Wei *et al.*, 2007; Borneman *et al.*, 2008; Argueso *et al.*, 2009; Fritsch *et al.*, 2009; Novo *et al.*, 2009). Bleykasten-Grosshans *et al.* (2013), alongside the reference strain, surveyed a further 40 strains obtained from various sources and countries worldwide. The authors could not find

a correlation between strain source or use and TE content. Small insights into strain content have been made by other authors, including Dunn *et al.* (2005) and Carreto *et al.* (2008), who discovered that wine strains typically contain lower *Ty* content than S288c-related laboratory and pathogenic strains. An inversion of the *Ty1/Ty2* ratio with respect to the reference strain has also been observed in wine strains (Ibeas and Jimenez, 1996; Novo *et al.*, 2009).

1.3.1 Recombination and solo LTRs

Around 85% of TEs in the *S. cerevisiae* reference genome exist as remnant LTRs in a 'solo' state due to recombination events (Lesage and Todeschini, 2005; Bleykasten-Grosshans and Neuvéglise, 2011; Chan and Kolodner, 2011). These recombination events are most common between the two identical LTRs flanking the same element, and the internal element DNA and a single copy of an LTR is excised as circular DNA (illustrated in Figure 1.6; Kim *et al.*, 1998). This extracircular DNA usually degrades, but has been known to be reintegrated into a new site in the genome (Ciriacy and Breilmann, 1982; Garfinkel *et al.*, 2006; Møller *et al.*, 2015, 2016). Genomic rearrangements due to recombination between full-length elements (FLEs) have been extensively documented (Kupiec and Petes, 1988*a*,*b*; Umezu *et al.*, 2002; Lemoine *et al.*, 2005; VanHulle *et al.*, 2007; Argueso *et al.*, 2008; Pennaneach and Kolodner, 2009 and reviewed by Chan and Kolodner, 2011).

Recombination can also occur in the same element if one of the LTRs has degraded, between an element and an existing solo LTR (Downs *et al.*, 1985) or between autonomous and non-autonomous elements (Bleykasten-Grosshans *et al.*, 2011) as long as the sequences have not diverged beyond the level of complementarity required for efficient recombination (Moore *et al.*, 2004). Not all recombination events occur between LTRs, as breakpoints are also observed in coding regions of elements (Jordan and McDonald, 1998, 1999*a*; Garfinkel *et al.*, 2005). Other observed rearrangements between elements have also been attributed to recombination (Liebman *et al.*, 1981; Downs *et al.*, 1985).

Transposition rates of *Ty1* elements are estimated at 1x10^{-8} to 1x10^{-9} per locus per generation, or 1x10^{-4} to 1x10^{-5} per genome (Roeder *et al.*, 1984), whereas the rate of recombination has been estimated at 10^{-5} events per element, which is far higher than transposition frequency (Liebman *et al.*, 1981; Winston *et al.*, 1984) but lower than recombination rates between other repeated sequences (Kupiec and Petes, 1988*a*,*b*).

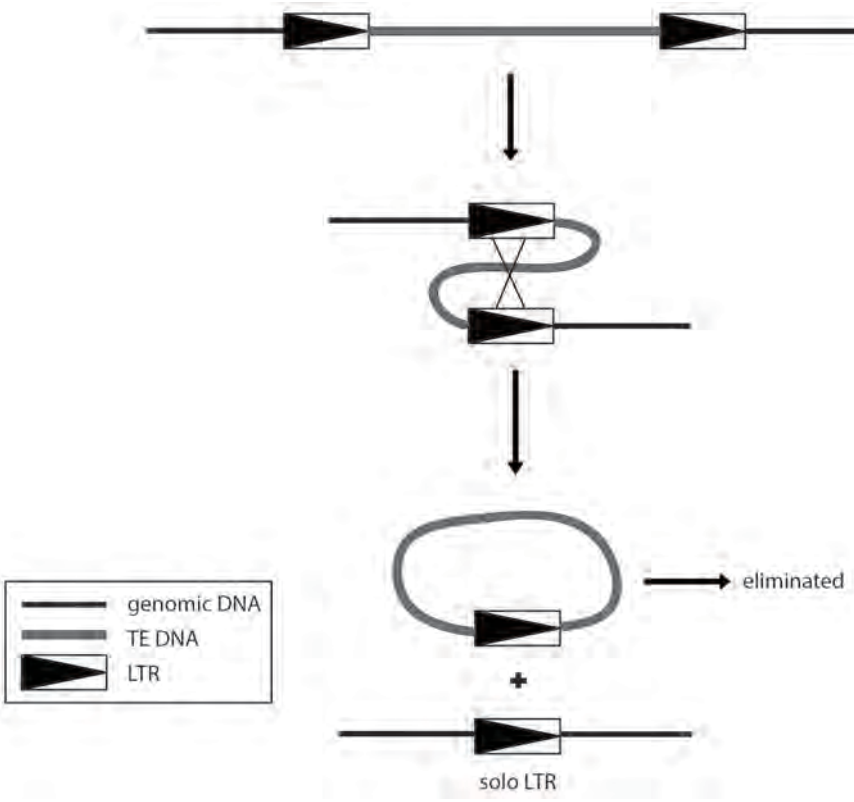

Figure 1.6: **Recombination between the LTRs of an element.** The recombination process results in the excision of circular extrachromosomal DNA containing the coding region of an element and a single LTR. The remnant solo LTR remains in the host's genome whereas the circular element is usually lost.

1.3.2 The *Ty1/2* Superfamily

The *Ty1/2* superfamily comprises the subfamilies of *Ty1*, *Ty2*, recombinational hybrids and *Ty1'*, all of which are *Ty1/copia* elements. It is also the largest family, making up almost 75% of the TE content in the *S. cerevisiae* genome (Hani and Feldmann, 1998; Kim *et al.*, 1998). *Ty1* and *Ty2* are structurally highly similar, but previous attempts at differentiating between the two was complicated by the presence of *Ty1/2* recombinational hybrids (Section 1.3.1; Jordan and McDonald, 1999*a*). *Ty2* has arisen in *S. cerevisiae* via horizontal transfer, having been donated from *S. mikatae* (Liti *et al.*, 2005; Carr *et al.*, 2012). *Ty1'* is characterised by its divergence from *Ty1* in the *gag* ORF (Kim *et al.*, 1998) and like *Ty2*, is believed to have arisen in *S. cerevisiae* by horizontal transfer rather than speciation (Bleykasten-Grosshans *et al.*, 2011).

Ty1 elements were first observed gaining considerable nucleotide variation during the reverse transcription process by Xu and Boeke (1987), which was further investigated and confirmed by Gabriel *et al.* (1996) and Wilhelm *et al.* (1999). The elements also undergo extensive recombination, even with non-autonomous copies, the progeny of which can then be autonomous

(Bleykasten-Grosshans *et al.*, 2011). The high success of non-autonomous *Ty1* elements observed by Bleykasten-Grosshans *et al.* (2011) may become a problem for the family if there is competition between functional and non-functional copies, and reflects a delicate balance between the elements (Le Rouzic *et al.*, 2007).

1.3.3 *Ty3*

Ty3 is the sole member of the *Ty3/gypsy* group of retrotransposons. The reference strain contains two copies of *Ty3* FLEs (Kim *et al.*, 1998) and also *Ty3p*, an extinct family in the form of degenerate LTRs, the true origin of which is still unclear (Fingerman *et al.*, 2003; Carr *et al.*, 2012). This family is still thought to be active, but undergoes effective recombination between elements, resulting in a high frequency of solo LTRs (Carr *et al.*, 2012).

Ty3 activity is typically low, but can be induced by exposing cells to the mating pheromone α factor (van Arsdell *et al.*, 1987; Clark *et al.*, 1988). Bilanchone *et al.* (1993) discovered that the pheromone responding element (PRE) present in *Ty3* LTRs was responsible for this effect.

1.3.4 *Ty4*

The next *Ty1/copia* family to be discovered, *Ty4*, was documented by Stucka *et al.* (1989) and Janetzky and Lehle (1992). It is unclear whether *Ty4* is still active as direct transposition has never been observed (Stucka *et al.*, 1992), and Hug and Feldmann (1996) found that most transcripts prematurely terminated at the 3' LTR boundary. Past activity is evident as *S. cerevisiae* contains multiple copies of *Ty4* along with solo LTRs as evidence of recombination events (Kim *et al.*, 1998; Carr *et al.*, 2012). The current global population of *Ty4* in *S. cerevisiae* is polymorphic, consistent with current activity (Carr *et al.*, 2012).

Ty4 is unusual in that its LTRs contain a transcription repression region, unlike *Ty1/2* whose LTRs contain positive regulating regions (Hug and Feldmann, 1996). *Ty4* LTRs also contain a sequence similar to the consensus sequence of the PRE present in *Ty3* LTRs, but is unresponsive to α factor (Stucka *et al.*, 1992). *Ty4* is thought to be the most recently acquired family, and arose in the *Saccharomyces* ancestor before the speciation of the *sensu stricto* species (Neuvéglise *et al.*, 2002).

1.3.5 *Ty5*

The final *Ty1/copia* family, *Ty5* is extinct and non-functional in *S. cerevisiae* (Voytas and Boeke, 1992; Kim *et al.*, 1998; Le Rouzic and Capy, 2006; Carr *et al.*, 2012) but is transpositionally competent in *S. paradoxus* (Zou *et al.*, 1996*b*). Like *Ty3*, *Ty5* is also regulated by the pheromone response pathway and is induced during mating and exposure to α factor (Ke *et al.*, 1997). Unlike other families however, the *gag* and *pol* ORFs have fused (Figure 1.5). *Ty5* displays poor sequence conservation, most likely due to age (Neuvéglise *et al.*, 2002). It may also contribute to the organisation of chromosome ends due to its tendency to integrate into telomeres and form tandem insertions (Figure 1.7; Ke and Voytas, 1997).

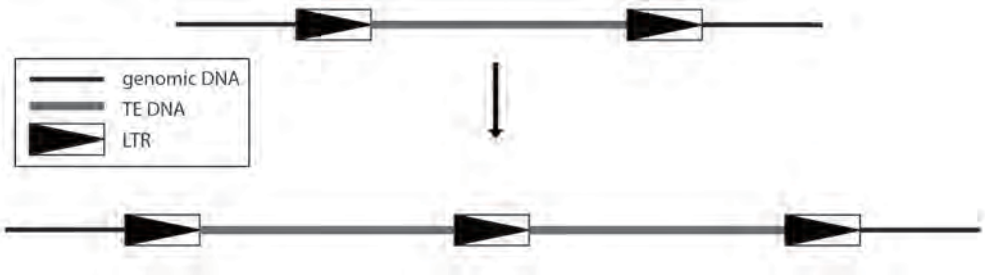

Figure 1.7: **Formation of a tandem element.** Recombination between a new and an existing element can result in the elements sharing an LTR in a tandem formation. In this simple model, the elements are in the same orientation, but recombination can also result in other placements of elements.

1.3.6 Target site specificity

Ty1-4 are transcribed by Pol II (Bolton and Boeke, 2003) and generally integrate into areas 1kb upstream of genes transcribed by Pol III such as tRNA genes and 5S ribosomal genes. *Ty3* has a highly defined integration window of just one to three bases upstream of Pol III genes (Kinsey and Sandmeyer, 1991; Voytas and Boeke, 2002; Sandmeyer, 2003; Lesage and Todeschini, 2005). Some tRNA genes are hotspots – areas of frequent retrotransposition – especially for *de novo Ty1* insertions (Kim *et al.*, 1998) such as chromosome III (Warmington *et al.*, 1986, 1987), and conversely, 33% of tRNAs are not associated with insertions, suggesting that not all tRNAs are ideal targets (Kim *et al.*, 1998). *Ty1* elements, alongside the preference for tRNAs, have been shown to target sites nearby or containing a degenerate element or solo LTR (Roeder *et al.*, 1980; Liebman *et al.*, 1981; Ciriacy and Williamson, 1981). Bachman *et al.* (2004) found that the higher the degeneracy of the *Ty* sequence, the more frequent the rate of a new element insertion, suggesting *Ty1*'s target site selection mechanism may be more complex than initially thought. However, this

may in fact be due to selection acting upon different sites, as a locus containing multiple insertions is more likely to be a 'safe' region of the genome. *Ty5*, when active, favoured insertion sites within telomeres and heterochromatin (Boeke and Devine, 1998; Zou *et al.*, 1996*b*).

Exactly why element target site specificity has evolved in this way is unclear. Boeke and Devine (1998) believe it to be a method of protecting the host's coding regions, which most likely arose due to intergenic regions upstream of tRNAs being larger than other intergenic regions in yeast (Bolton and Boeke, 2003). It could also be a method of maximising the element's chances of survival or, more likely, it simply reflects the conserved mechanism used for integration. Element integration complexes, as a result of interactions with host proteins, become tethered to specific regions upstream of coding regions, causing their transposition to happen there rather than immediately within a gene (Bushman, 2003). *Ty3* IN interacts with two subunits of RNA Pol III preinitiation factors, TFIIC and TFIIB (Yieh *et al.*, 2000; Aye *et al.*, 2001). *Ty5* integrates into silent heterochromatin though an interaction with the *Sir4* protein (Zou and Voytas, 1997; Zhu *et al.*, 1999, 2003; Xie *et al.*, 2001). The interaction of *Ty4* has not yet been elucidated as transposition is yet to be observed (Stucka *et al.*, 1992), and elements from the *Ty1/2* superfamily are still under investigation as the interactions have proved to be far more complex than the other families. Possible sites of interaction for *Ty1* have so far been shown to include *Esp1* (Ho *et al.*, 2015), nucleosomes (Baller *et al.*, 2012) and a subunit of Pol III (Bridier-Nahmias *et al.*, 2015) among others.

Interestingly, the upstream targeting system has evolved in parallel multiple times in *Saccharomyces* (both *Ty1/copia* and *Ty3/gypsy* families), but also in other species. For example, *Dictyostelium discoideum*, a species of slime mould, possesses a family of non-LTR retrotransposons which has a preference site upstream of Pol III genes (Winckler *et al.*, 2002) much like *Ty* insertions in *Saccharomyces*.

1.4 Movement of DNA: horizontal transfer and introgression

The homologous introduction of sequences via horizontal transfer (HT) and introgression is considered rare in yeast (Dujon, 2006; Liti and Louis, 2005), but recently there has been an increase in the reports of foreign DNA in varying species. It is also thought to contribute to eukaryotic evolution (reviewed by Keeling and Palmer, 2008; Oliver and Greene, 2009). When only concerning host genes, the event is known as horizontal gene transfer (HGT) in order to make the distinction

between HT events involving TEs (sometimes referred to as horizontal transfer of transposable elements; HTT).

1.4.1 Mechanisms of HT and introgression

The mechanisms underlying transfer of DNA is well understood in prokaryotes (reviewed in Frost *et al.*, 2005) but is less clear in eukaryotes. It is likely to differ depending on the host species and the genomic regions or TE families involved. In plants, a direct transfer of DNA between species in close contact or hybridisation events can explain some cases of HT and introgression. However, in those plants that do not hybridise or share an environment, it is most likely that a parasitic, viral, bacterial or fungal vector is the reason behind DNA movement. Similar vectors are also likely to be the mechanism for HT events in animals (Piskurek and Okada, 2007; Shen *et al.*, 2003). Baculoviruses are a possible vector for insects (Gilbert *et al.*, 2014) while *Drosophila* parasites and parasitoids such as wasps and mites and also its symbiotic bacteria *Wolbachia*, could all act as potential vectors for transmission (Ortiz *et al.*, 2015). In some species, HT events are not confined to a single mechanism, and could well occur simultaneously (Loreto *et al.*, 2008).

In yeast, genetic transfer is most likely to occur via hybridisation and then backcrossing (Marinoni *et al.*, 1999). The species involved therefore need to share the same environment. As interspecies hybridisation results in very low levels of spore viability, repeated backcrossing with one of the parental species must occur in order for the yeast to avoid reaching an evolutionary impasse (reviewed in Dujon and Louis, 2017). In numerous wine strains, the structure of the DNA inserts indicated a circular intermediate that was gained independently in different genomic locations (Borneman *et al.*, 2011; Galeote *et al.*, 2011).

1.4.2 Horizontal gene transfer and introgression in fungi

HGT is the transfer of genes between species (Andersson, 2009, 2012) whereas introgression is the transfer of DNA as a result of species hybridisation, such as the 23kb subtelomere gained by European strains of *S. paradoxus* from *S. cerevisiae* (Liti *et al.*, 2006).

Reports of HGT and introgression are widespread, particularly in prokaryotes. Cheeseman *et al.* (2014) documented a large genomic region containing 250 genes transferred between *Penicillium* species of fungi. Interestingly, the functions of the 250 genes appear to convey advantages to the recipient species. Reviews of HGT and introgression events in fungi by Richards (2011) and

Fitzpatrick (2012) suggest that the acquisition of some genes is likely to be highly advantageous, allowing fungi to thrive in new environments and even utilise new nutrient sources.

Many examples have also been documented between yeasts, such as the *ASP3* locus in *S. cerevisiae* which originated in *Wickerhamomyces* (League *et al.*, 2012). Marinoni *et al.* (1999) described small scale transfer events, whereas Novo *et al.* (2009) reported three large foreign regions in the strain EC1118 of *S. cerevisiae*, one of which was acquired from *Zygosaccharomyces bailii*. The region from *Z. bailii* has since been found in different strains of *S. cerevisiae* but in varying genomic locations, suggesting that a circular episome is the mechanism behind the transfer (Borneman *et al.*, 2011; Galeote *et al.*, 2011). Another of the regions in strain EC1118 containing *FOT* genes, was later found to be donated by *Torulaspora microellipsoides* (Marsit *et al.*, 2015) after polymerase chain reaction (PCR) amplification. Despite the increasing numbers of transfers being reported, the evolutionary advantage of such events is unclear in many cases. The strains and species involved are usually in close contact with one another, often in industrial or other man-made unusual and stressful environments, suggesting this could play a part in the transfer events. These environments cause a strong selective pressure on the organisms, with the HT events often resulting in favourable adaptations, such as those documented by Marsit *et al.* (2015).

Wei *et al.* (2007) reported introgression on a large scale between species of yeast, something that has now been reported between almost every species of *Saccharomyces* (reviewed by Hittinger, 2013). As each newly discovered population is analysed, further examples of introgression are documented (Almeida *et al.*, 2014). Dunn *et al.* (2012) discovered genomic regions of *S. mikatae* origin had been introgressed into *S. kudriavzevii*. Not surprising for sister species, one of the highest rates of introgression is between *S. cerevisiae* and *S. paradoxus*, with the latter being the donor more often than not (Naumova *et al.*, 2005; Liti *et al.*, 2006, 2009; Wei *et al.*, 2007; Muller and McCusker, 2009; Dunn *et al.*, 2012). Similarly, there are numerous introgressions between *S. eubayanus* and *S. uvarum* (Almeida *et al.*, 2014; Peris *et al.*, 2014). *S. kudriavzevii* has also gained regions from both of these species (Peris *et al.*, 2014).

Away from sister species, *Saccharomyces* isolated from natural sources seem less susceptible to introgression and HT events (Hittinger *et al.*, 2010; Libkind *et al.*, 2011), which would suggest that a human domestication pressure on industrial strains has had an important role in selecting for these rare occurrences (Hittinger, 2013).

Unfortunately, there is a bias towards genes as studies tend to focus on coding areas of the

genome, overlooking TEs and non-coding regions. This is due to the fact that the impact of HGT on a recipient can be more easily evaluated than those caused by non-coding regions and TE insertions (Gilbert and Cordaux, 2013).

1.4.3 HT of TEs

1.4.3.1 HT is widespread in many species

Despite most elements lacking an apparent infectious ability, HT has been proposed as a natural part of the TE life cycle (Le Rouzic and Capy, 2005). A successful HT event requires the transfer of an active TE into the recipient, followed by a burst of transposition in order to propagate throughout the genome, and then further into the population by vertical transmission (Silva *et al.*, 2004; Le Rouzic and Capy, 2005). Depending on the recipient organism, the newly acquired TEs may be regulated or suppressed by the hosts' defences until copy number is reduced to an equilibrium or extinction of the family. Random mutations, excision via recombination leading to stochastic loss and purifying selection all contribute to a reduction in copy number (Le Rouzic and Capy, 2005; Le Rouzic *et al.*, 2007). Should the element escape host defences and selection, HT therefore becomes a way for TEs to avoid extinction.

HT between prokaryotes and eukaryotes has been observed, but fewer incidences still involving TEs. Rolland *et al.* (2009) located six genes present in *Lachancea kluyveri* that were likely to have originated as members of the bacterial TE family IS607, and Gilbert and Cordaux (2013) discovered that this same family had spread into a further 13 species.

HT between eukaryotes has been well documented throughout most kingdoms, for both classes of TEs (reviewed in Schaack *et al.*, 2010). Examples of the transfer of DNA transposons include the *OC1* family in an ancestor of the Tasmanian devil (Gilbert *et al.*, 2013); *SPIN* transposons in mammals (Pace *et al.*, 2008), reptiles (Gilbert *et al.*, 2012) and other tetrapods through a parasitic vector (Gilbert *et al.*, 2010). There are numerous examples in *Drosophila* (Loreto *et al.*, 2008; Sánchez-Gracia *et al.*, 2005) including the *P* element family (Silva and Kidwell, 2000). There are also a number of incidences of HTT in plants (reviewed in Fortune *et al.*, 2008), such as the *MULE* elements (Diao *et al.*, 2006), and *PIF*-like elements in grasses (Markova and Mason-Gamer, 2015). Also reported to have undergone HT is the *mariner* family between many species, such as crustaceans and insects (Hartl *et al.*, 1997; Dupeyron *et al.*, 2014).

For retrotransposons, the majority of documented HT events have been in *Drosophila*, spanning at least 20 families (Sánchez-Gracia *et al.*, 2005; Bartolome *et al.*, 2009; de Setta *et al.*, 2009, 2011). Due to the high level of research interest that this genus has received, at least 100 putative events have been suggested (Loreto *et al.*, 2008; reviewed in Schaack *et al.*, 2010). In other genera besides *Drosophila*, events include the transfer of the *RIRE1* LTR retrotransposon in *Oryza* (Roulin *et al.*, 2008); the *SURL* family in the Echinoidea (González and Lessios, 1999); *Route66* in plants (Roulin *et al.*, 2009) and also the centromeric retrotransposons (CRs) in grasses (Sharma and Presting, 2014).

It is more likely for DNA transposons to undergo HT than retrotransposons, and of the latter, non-LTR retrotransposons are the least likely (Silva and Kidwell, 2000). In *Drosophila*, the number of transferred transposons and LTR retrotransposons is roughly equal, but non-LTR retrotransposons are transferred at a very low frequency (Loreto *et al.*, 2008). Some TE families are also more likely to undergo HT than others (Silva and Kidwell, 2000) and can be pervasive in new species (Sánchez-Gracia *et al.*, 2005; Gilbert *et al.*, 2010; Thomas *et al.*, 2010).

1.4.3.2 HT of elements is less frequent in Saccharomycetaceae

Few HT events of TEs have been recorded in *Saccharomyces*: *Ty2* in *S. cerevisiae* obtained from *S. mikatae* (Liti *et al.*, 2005; Carr *et al.*, 2012), and potentially *Ty3p* in *S. cerevisiae* obtained from *S. paradoxus* or *S. mikatae* (Carr *et al.*, 2012). However, this may reflect the limited number of investigations into the TE content of other *Saccharomyces* species. In addition, many previous studies utilised molecular laboratory techniques to provide presence/absence results (e.g. Liti *et al.*, 2005), rather than in depth phylogenies that were utilised in the work here.

A small number of copies of *Rover*, a *hAT* DNA transposon family, were gained by strains of both *S. cerevisiae* and *Naumovozyma dairenensis*, most likely sourced from *Torulaspora delbrueckii* (Legras *et al.*, 2018) and a *Lachancea* species (Sarilar *et al.*, 2015), respectively. Additionally, the main *Ty1/copia* retrotransposon family, *Tsk1*, in *Lachancea kluyveri*, is thought to have been gained from an unidentified source (Neuvéglise *et al.*, 2002; Payen *et al.*, 2009).

1.4.4 Detection of HT events

The evolutionary history of a TE family is expected to somewhat follow that of its host, therefore deviations from the expected phylogeny can strongly suggest that HT events have occurred (Silva

and Kidwell, 2000). However, HT can be difficult to detect because TEs naturally show differing rates of evolution, and also experience degeneration and stochastic loss - the elimination of all autonomous copies.

There are numerous methods employed for the detection of HT of TEs. For example, sequence similarity searches can highlight differences in codon usage or nucleotide composition, which could expose TEs that share more identity with donor than recipient species (Lerat *et al.*, 2002). This method is mostly used for prokaryotes, but can also be applied to sufficiently divergent eukaryotic species (Rödelsperger and Sommer, 2011). Also, specific to LTR retrotransposons, is to test for divergence between LTRs in two distinct species (Fortune *et al.*, 2008).

The method used here is that of phylogenetic analysis, which has the benefit of allowing HT events to be identified at the same time as generally exploring the evolutionary relationships of TEs between species. There are a number ways in which HT events can be inferred using TE phylogenies. First, highly similar elements present in multiple, often phylogenetically distant species, present on relatively short branch lengths, can indicate that there is less divergence between elements than between non-TE areas of the genome. Second, discrepancies between the topologies of TE phylogenies when compared to those of host species may indicate HT events. Finally, and offering weaker evidence for HT, is the patchy distribution of elements within sister species. However, while indicative of HT, this incongruence can also be the result of paralogous sequences, stochastic loss and the domestication of elements by their hosts. The domestication of TE domains can result in conservation across species, assuming the event occurred prior to speciation (Silva and Kidwell, 2000; Wallau *et al.*, 2012).

The direction of HT and the exact donor and recipient species are not always clear, as events may have occurred in a recent common ancestor rather than the present species. The donor species usually displays more diversity in its TE sequences due to increased age, allowing mutations to accumulate, resulting in relatively long-branched phylogenies, whereas the recipient's element(s) likely underwent transposition upon HT, therefore the copies generated will be far more homogenous. The well-supported nesting of sequences from one species within another also strongly suggests a donor-recipient relationship, as the donor typically possesses copies of higher nucleotide diversity than those of the recipient (Brookfield, 2005; Carr *et al.*, 2012).

1.4.5 Limiting factors

A number of factors may impact the rate of success of a newly transferred family into a naive genome. It is thought that effective population size plays a role in the propagation of elements through HT, as species with higher effective population sizes are more efficient at removing TEs from the population (Charlesworth and Charlesworth, 1983; Brookfield and Badge, 1997; Groth and Blumenstiel, 2016). Selection (Section 1.6; reviewed in Nuzhdin, 1999), as well as the compatibility of host factors with a newly acquired element, may influence the success of the transfer (Silva and Kidwell, 2000). For example, activity of *Ty* elements in *S. cerevisiae* and *Tf* elements in *Sz. pombe* is dependent on numerous host factors (Aye and Sandmeyer, 2003; Aye *et al.*, 2004; Maxwell and Curcio, 2007; Risler *et al.*, 2012; Ahn *et al.*, 2017; Rai *et al.*, 2017). A naive genome lacking such factors may halt the transposition of newly acquired elements. The introduction of a new family, the activity of which is uncontrollable by the host, such as LINEs in *S. cerevisiae* (Dong *et al.*, 2009), may have deleterious effects, preventing propagation of the family further. For LTR retrotransposons specifically, a high likelihood of intra-element recombination could quickly render a FLE as a solo LTR (Section 1.3.1), unable to transpose in a new genome.

1.5 Element-host interactions

The relationship between a host and its TEs is complex. Initially viewed as a conflict between element and host, it now appears that the relationship has co-evolved into a give and take nature (reviewed by Kidwell and Lisch, 2000; Beauregard *et al.*, 2008). Elements are able to harness the transcriptional machinery of a host cell for their own replication, but in turn the host itself has control over its own insertions to an extent, as *Ty* activity is dependent on host cell-type regulation. In diploid a/α cells, studies have shown that *Ty1* transcription (Elder *et al.*, 1983; Errede *et al.*, 1985, 1987; Fulton *et al.*, 1988), transposition (Paquin and Williamson, 1986) and expression are all lowered, as well as the expression of neighbouring genes (Errede *et al.*, 1980) due to mating loci repression (Elder *et al.*, 1983; Herskowitz, 1988). Even if a given transposition event does not result in harm to a host *per se*, the insertion of elements can have significant effects on genomic structure and profound influence over the expression levels of neighbouring genes (Winston *et al.*, 1984; Roeder *et al.*, 1985; Natsoulis *et al.*, 1989; Kinsey and Sandmeyer, 1991; Bolton and Boeke, 2003).

1.5.1 Effects on gene expression

The fact that *Ty* insertions can result in the increased, decreased or constitutive expression of nearby host genes was established in the 1980s (Chaleff and Fink, 1980; Farabaugh and Fink, 1980; Roeder *et al.*, 1980; Roeder and Fink, 1982; Roeder *et al.*, 1984; Williamson, 1983; Silverman and Fink, 1984; Roeder *et al.*, 1985; Coney and Roeder, 1988; Goel and Pearlman, 1988). Observations made on the changes in expression due to differing *Ty* insertions between strains are still being made 30 years later (e.g. Fritsch *et al.*, 2009).

In comparison however, few studies have focused on the effects of solo LTRs, despite 85% of insertions existing in this form in the *S. cerevisiae* reference genome (Lesage and Todeschini, 2005; Bleykasten-Grosshans and Neuvéglise, 2011; Chan and Kolodner, 2011). Solo insertions usually retain their regulatory sequences and transcription factor binding sites required for interaction with the host cell's transcriptional machinery, therefore the excision of the coding region of the element after recombination does not always allow the host to revert back to its original phenotype (Roeder *et al.*, 1980; Paquin and Adams, 1983). Despite receiving less attention, solo LTRs have been known to affect transcription of adjacent gene promoters (Roeder *et al.*, 1980; Winston *et al.*, 1984; reviewed in Thompson *et al.*, 2016). The effects of LTR sequences of three elements were studied by Roeder *et al.* (1985) and found to be homologous with enhancers of the *HIS4* gene. When inserted 5' to the gene, the LTRs could enhance transcription and present a strong phenotype, a weak phenotype or prevent transcription altogether. Roelants *et al.* (1997) later used the ability of a *Ty1* LTR to initiate transcription of the gene *URA2*.

1.5.2 Host defences against TE insertions

Due to their persistent nature, many organisms employ specific genomic defence mechanisms such as RNA interference (RNAi) and DNA methylation in order to minimise the effects of TE insertions (Hartl and Clark, 2007; Slotkin and Martienssen, 2007). Another method of defence is repeat-induced point mutations (RIP), which are only present in certain fungi such as *Microbotryum violaceum* (Johnson *et al.*, 2010) and *Neurospora crassa* (reviewed by Selker, 2002).

Saccharomyces and most related species have lost one or more crucial aspects of the RNAi pathway (Wolfe *et al.*, 2015). Drinnenberg *et al.* (2009) were able to show that the system could be restored once core components of the pathway obtained from *Naumovozyma castellii* were reinserted into *S. cerevisiae*. The results also suggest that the common ancestor of budding yeast

possessed the pathway and that *S. cerevisiae* has subsequently lost RNAi as its current forms of control are enough to keep the TE insertions contained. Methylation is also virtually absent in *S. cerevisiae* (Proffitt *et al.*, 1984) and species of Saccharomycetaceae, but just how many species fail to employ methylation as a defence mechanism is unclear (Benachenhou *et al.*, 2013).

However, *S. cerevisiae* is far from defenceless. In order to counteract the additive effect of transposition, the yeast and its insertions have developed alternative methods of controlling copy number in a number of intuitive and coordinated ways. For example, some families of elements have developed very pronounced target site selection in order to avoid important areas of the genome, thus increasing their chances of survival (Section 1.3.6; Boeke and Corces, 1989; Blanc and Adams, 2003). *Ty* elements also undergo copy number dependent transcriptional and post-transcriptional silencing (cosuppression) to limit further transposition (Jiang, 2008; Garfinkel *et al.*, 2003). Additionally, *Ty* elements require very specific host cofactors (Aye and Sandmeyer, 2003; Aye *et al.*, 2004; Maxwell and Curcio, 2007; Risler *et al.*, 2012) and conditions for integration into the genome such as optimal temperature (Lawler *et al.*, 2002). Furthermore, the 3' end of *Ty1* elements possess a high rate of mutation and addition of non-templated nucleotides, both of which have been shown to reduce the rate of integration into host DNA. This may have evolved as an additional method of keeping *Ty1* integration to a minimum (Gabriel and Mules, 1999).

1.6 Selection

Since the 1980s, TEs have been assumed to be purely deleterious due to their apparent parasitic nature (e.g. Orgel and Crick, 1980; Doolittle and Sapienza, 1980; Hickey, 1982). TE insertions appear to be mainly under the influence of negative (purifying) selection and are quickly and efficiently removed if causing detrimental effects (Charlesworth and Charlesworth, 1983; Fink *et al.*, 1986; Jordan and McDonald, 1999*c*,*b*; Le Rouzic and Deceliere, 2005). Despite this, they have an extraordinary survival ability within a host genome by avoiding selection altogether when causing only mild or neutral effects on fitness, especially in species with a small effective population size (N_e; Wilke and Adams, 1992; Biémont *et al.*, 1997; Charlesworth *et al.*, 1997; Kidwell and Lisch, 2000).

Upon arrival of a new TE into a genome or population, it must undergo immediate and efficient transposition or be lost by genetic (random) drift or purifying selection (Le Rouzic and Capy, 2005); an event that has been observed in *Drosophila* concerning the *P* element family (Anxolabehere

et al., 1988; Biémont, 1994; Biémont *et al.*, 1994; Kimura and Kidwell, 1994). It is thought that after the initial invasion, transposition is balanced by the loss of elements, and a long-term equilibrium is reached (Charlesworth *et al.*, 1994; Le Rouzic and Deceliere, 2005).

Left unchecked, TEs can in theory increase their copy number indefinitely, but host defences are able to control the rate of transposition before becoming fatal. Organisms that need to replicate their DNA quickly or have found recombinational mechanisms to remove insertions seem to be able to avoid the accumulation of elements (Charlesworth *et al.*, 1994). Insertions cause deleterious effects in a genome in a number of ways that would likely cause loss of fitness and consequently the loss of the TE from the population (Eickbush and Jamburuthugoda, 2008). Transposition into vital coding regions or regulatory sequences, chromosomal rearrangements and the physiological impact of protein synthesis during transposition are all deleterious effects highly likely to negatively affect host fitness (Dolgin and Charlesworth, 2008; Le Rouzic and Capy, 2006). All of these events are selected against (Carr *et al.*, 2002), and Pasyukova *et al.* (2004) have shown a clear link between an increase in TE copy number and a decrease in fitness in *D. melanogaster*.

Even with selection working against them, large numbers of highly successful and diverse retrotransposon families persist in the genomes of most organisms (reviewed by Eickbush and Jamburuthugoda, 2008), yet their contributions to host evolution are rarely investigated (Hoban *et al.*, 2016). This is due in part to the complications that arise when sequencing and assembling TE-rich areas of a genome (Treangen and Salzberg, 2011; Hoban *et al.*, 2016), that may change given the increasing use of 3^{rd} generation single-molecule sequencing (Chaisson *et al.*, 2015). Although the current methods of identifying signatures of selection acting upon TEs is limited, they are on the increase (Casacuberta and González, 2013; Villanueva-Cañas *et al.*, 2017).

1.6.1 Genetic hitchhiking

LTRs and neighbouring genes are linked in a fashion comparable to that of alleles, and are therefore often inherited together in an effect known as linkage. Although it does not uncouple the inheritance of the linked sites, the association between the two can be somewhat eroded by recombination (Barton, 2000; Gillespie, 2000). Favourable mutations cause a loss of nucleotide variation within other areas of the genome, and so the effects of positive selection on a favoured locus can also be extended to any other loci with which it is linked. In addition, insertions under purifying selection can cause lower nucleotide variation due to background selection (Maynard Smith

and Haigh, 1974; Kofler *et al.*, 2012). The term hitchhiking refers to a neutral variation, along with its linked advantageous mutation, being driven to fixation in a population due to positive selection (Maynard Smith and Haigh, 1974; Page and Holmes, 1998; Barton, 2000; Kim and Stephan, 2002).

1.6.2 Statistical methods

The effects of selection on loci can be calculated with statistical tests. The neutral theory of molecular evolution was first proposed by Kimura (1968) who suggested many polymorphisms are not under the influence of natural selection but as they confer no advantage on fitness, they are simply neutral. Within this theory, positive selection is considered to be a rare event, and negatively selected mutations are eliminated when affecting functionally constrained areas of the genome.

1.6.2.1 Tajima's *D*

A statistical method for testing the neutrality of nucleotide variants within a population was developed by Tajima (1989) and relies on sequence polymorphism data from only one species. In Tajima's method, the expected amount of genetic variation per nucleotide (denoted θ) is equal to $4N_e\mu$, in which N_e is the effective population size, and μ is equal to the mutation rate per generation. θ is estimated using S, the number of segregating sites, or π (Nei and Li, 1979), the average number of paired nucleotide or pairwise differences in the sample sequences (Page and Holmes, 1998). Under the neutral theory model, using either S or π with a constant population size would give identical values of θ and thus *D*, based upon the difference between those values of θ, would theoretically be zero. Tajima's *D* statistic can therefore indicate the frequency of nucleotide variants under neutrality on a spectrum (Page and Holmes, 1998; Rozas, 2009) and requires a minimum sample size (n=4) in order to function (Tajima, 1989). The test relies on a calculated value of *D* to signify statistical deviation from zero. If the *D* value has significantly deviated from zero, the neutral hypothesis is rejected, and conversely if the deviation is not significant, the neutral hypothesis cannot be rejected. A significantly negative value of *D* may be obtained by population growth generating an excess of rare alleles, or positive selection. Conversely, significantly positive values are a result of population subdivision or balancing selection and is observed as an excess of intermediate frequency alleles (Tajima, 1989).

In this research, Tajima's test was applied to LTRs and their neighbouring genes in order to identify potentially positively selected insertions, and possibilities of hitchhiking between the two.

If it was the LTR that achieved a more statistically significant negative *D* value (i.e. deviated further from zero) in comparison to its neighbouring gene, then it was more likely that positive selection was occurring at the locus of the LTR rather than the gene. Therefore, it is the gene that is said to be hitchhiking. Conversely, if it was the gene that achieved a more statistically significant value of *D*, then the gene is more likely to be under positive selection than the LTR, which is therefore hitchhiking along with the gene at that locus. Tajima's *D* was also selected for use here as, unlike other tests which are not based on a frequency of mutations, it is conservative on regions that have undergone recombination (Ramírez-Soriano *et al.*, 2008).

1.6.2.2 Fu and Li's *D* statistic

Fu and Li (1993) built on the work of Tajima (1989) and proposed a genealogical based method to test their hypothesis that all mutations at any given locus have a neutral effect. Their simulations showed that the expected number of derived mutations present in the population N_e, is equal to θ. Like Tajima's *D*, Fu and Li's *D* statistic fits onto a scale and significantly negative results are indicative of positive selection. Unlike Tajima's *D* however, Fu and Li's test compares the number of derived singleton mutations, rather than pairwise differences. A negative *D* statistic value indicates an excess of singletons (equivalent to rare alleles for Tajima's *D*) whereas a positive value implies a lack of singletons (intermediate alleles in Tajima's test; Fu and Li, 1993; Ramírez-Soriano *et al.*, 2008). It can be considered as a more sensitive method with which to confirm the results of Tajima's *D*.

1.6.3 Impacts on host evolution: the benefits of TEs

Despite often being called "junk" and assumed to confer no advantage to their host (Wong *et al.*, 2000), there has been speculation in the past 20 years that the presence of TEs and some insertions, albeit rarely, can in fact be beneficial to their host. This may be on a genomic scale by increasing genetic diversity, or on a localised scale by affecting gene expression. It was the preferential insertion sites in regulatory regions of genes of *S. cerevisiae* that first inspired the thought that specific *Ty* insertions could cause adaptive changes to their hosts from these critical genomic positions (Roeder and Fink, 1982; Roeder *et al.*, 1984).

In recent years, there has been evidence to implicate TEs in providing general benefits to their hosts and contributing to adaptation and genome evolution in a variety of species (reviewed in

Oliver and Greene, 2009, 2012; Belyayev, 2014; Rey *et al.*, 2016). Genetic diversity generated by TEs allow a host to adapt to new environments, increasing their chances of survival (reviewed by Chenais *et al.*, 2012; Hosid *et al.*, 2012; Casacuberta and González, 2013). Adams and Oeller (1986) were one of the first to document adaptations selected for in a population due to large scale genomic rearrangements caused by *Ty* elements. More recently, Franco-Duarte *et al.* (2015) were able to show that a particular strain of *S. cerevisiae*, VL1, has adapted to vineyard environments at least in part due to its *Ty* insertions. In other species, González *et al.* (2010) documented TE induced adaptations in *Drosophila*, whereas Grandaubert *et al.* (2014) recorded evidence of beneficial changes in the plant pathogen *Leptosphaeria*. Such extreme TE-induced diversity in a host to the point of speciation has also been recognised (Ginzburg *et al.*, 1984; Hurst and Schilthuizen, 1998; de Boer *et al.*, 2007; Zhang and Gao, 2017).

Potentially beneficial insertions are becoming increasingly discovered and their effects are being further investigated to a successful degree in many organisms, such as a solo LTR of the *roo* family in *D. melanogaster* (Merenciano *et al.*, 2016). In *S. cerevisiae*, a small number of potentially beneficial insertions have been discovered and documented (e.g. Brady *et al.*, 2008; Fraser *et al.*, 2010; Servant *et al.*, 2008). However, they have mostly been serendipitous findings and few full genome screenings for beneficial insertions have been completed. The method employed in this work was inspired by that of Kofler *et al.* (2012), using a Portuguese population of *D. melanogaster*. The authors calculated Tajima's *D* values for 500bp windows across the genome, and using this method, identified 13 candidate TE insertions for positive selection, based on their high population frequency and statistically significant negative Tajima's *D* values. However, the authors recognised that the insertions were identified as just candidates for positive selection during the research, and there may be other reasons for the Tajima's *D* values, such as fluctuations in population size. This research by Kofler *et al.* (2012) on *D. melanogaster* is comparable to *S. cerevisiae* as the organisms share a similar effective population size (Tsai *et al.*, 2008; Samani and Bell, 2010).

1.7 Aims and rationale

The work presented in this thesis aimed to investigate the extent to which two possible mechanisms allow *Ty* families to be maintained in their *Saccharomyces* host genomes. This was divided into two main aspects: via positive selection of insertions, and also the HT of elements between species. A

full genome screening for potentially beneficial insertions has not yet been performed on *S. cerevisiae*. Now that many strains of the yeast and its sister species, *S. paradoxus*, are available, they make ideal candidates for comparative studies on the frequency of positively selected insertions. Selection acting upon solo LTRs is only of benefit to the host, as a family's survival depends on positive selection acting upon autonomous FLEs. As deleterious elements are quickly removed by selection or drift, it would be possible that persisting insertions may be under the influence of positive selection, particularly if they provide a beneficial effect on host gene expression.

HT may be an additional method of allowing elements to persist, as they are able to cross species barriers during hybridisation and backcrossing. This work aimed to identify putative HT events through phylogenetic analysis.

The evolutionary relationships of *Ty* elements, particularly within *Saccharomyces* species, have also yet to be explored, having been confined to a very small number of species and strains. Internal coding regions of elements are under differing evolutionary constraints than LTRs, therefore analysing both can provide results that complement one other. Although protein phylogenies retain the signal of HT for a longer period of time, they lack resolution and direction of HT events. In contrast, LTR trees provide resolution and direction, but may miss older events due to the rapidity of LTR evolution. Analyses performed here were designed to elucidate the evolutionary history of elements in all fungal species containing *Ty*-like insertions.

Chapter 2

Materials and Methods

This chapter provides the methods and protocols used in this work to conduct a systematic search of all available Ascomycota yeast genomes containing *Ty*-like TE insertions, dataset construction (Section 2.2), sequence (Section 2.3) and phylogenetic analysis (Section 2.4). In addition, the abundance of population data for species *S. cerevisiae* and *S. paradoxus* provided the ideal basis with which to conduct genome-wide screenings of signatures of selection acting upon *Ty* insertions. This was performed using a method similar to that of Kofler *et al.* (2012) and is described here. Finally, the methods used to identify changes in the expression of genes neighbouring those insertions that may be under positive selection are detailed (Section 2.6).

A list of all software used and website addresses is located in Appendix A. Unless stated, default parameters were used.

2.1 Genome construction and mapping

Where available, previously assembled yeast genomes were downloaded from GenBank. Additional *S. cerevisiae* (Barbosa *et al.*, 2016; Drozdova *et al.*, 2016; Chapter 6), *S. kudriazevii* (Hittinger *et al.*, 2010; Chapter 4), *S. arboricola* (Gayevskiy and Goddard, 2016; Chapter 4) and *S. uvarum* (Sylvester *et al.*, 2015; Chapter 4) genomes analysed here required assembly from raw reads into contigs and/or scaffolds, as full-length sequences, particularly LTRs and RT were required for analysis. Genomes were constructed *de novo* or with the use of a reference, where available. Figure 2.1 displays the workflow used for genome assembly and mapping of raw reads.

2.1.1 Genome assembly

Where genomes were not available on GenBank (Benson *et al.*, 2015) as searchable scaffolds, the raw data were obtained in the form of FASTQ files from the NCBI Sequence Read Archive (SRA)

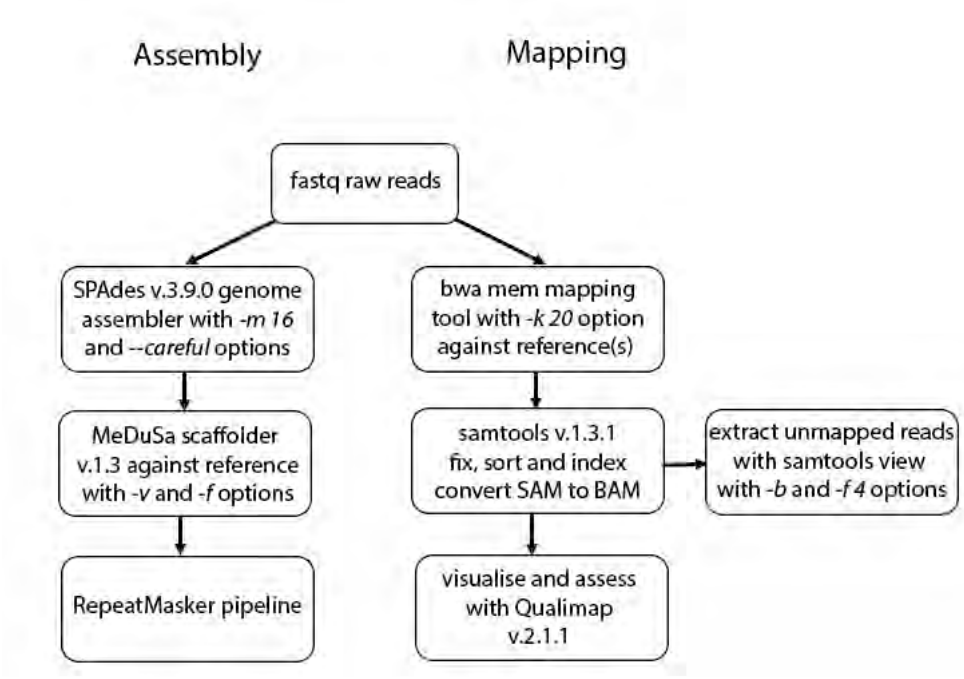

Figure 2.1: **FASTQ workflow for those species whose assembled genomes were not available.**

and the European Nucleotide Archive (ENA). FASTQ reads were assembled with SPAdes v.3.6.2 (Bankevich *et al.*, 2012), with the options *–careful* (lowers rate of mismatching reads) and *-m 16* (limits memory usage to 16Gb). Kmer sizes were reduced if the initial assembly failed. Trimming and removal of poor quality reads were performed as part of the SPAdes program with default parameters. Contigs produced by SPAdes were scaffolded with MeDuSa v.1.3 (Bosi *et al.*, 2015) using a reference genome (where available) to minimise the number of unplaced scaffolds.

2.1.2 Mapping raw reads as a method of identifying introgression

To check for sites of introgression and possible hybridisation primarily in *Saccharomyces* strains and species, FASTQ reads were mapped onto a concatenated reference file consisting of the *Saccharomyces sensu stricto* species using bwa v.0.7.13 (Li and Durbin, 2009). The resulting SAM output was converted to BAM (*samtools view*) and sorted into coordinate order (*samtools sort*) with SAMtools v.1.3.1 (Li *et al.*, 2009a). Unmapped FASTQ reads were exported as a separate file (*samtools* sort with option *-f 4*) and built as in Section 2.1.1. Quality of mapping and visualisation was ascertained with Qualimap v.2.1.1 (Okonechnikov *et al.*, 2016). Areas of interest were visualised and exported as FASTA format using Unipro UGENE v.1.26 (Okonechnikov *et al.*, 2012).

2.2 RT and LTR datasets

Query sequences in the form of full transposable elements, LTRs and translated Reverse Transcriptase (RT) sequences from each *Ty* family (accession numbers of which are listed in Table 2.1) were used to obtain hits. Two main methods were used to acquire hits: BLAST (Basic Local Alignment Search Tool; Altschul *et al.*, 1990) searches and screening genomes with RepeatMasker v.4.0.6 (Smit *et al.*, 2013). Sequences were compiled in order to obtain datasets for each species for phylogenetic analyses (Section 2.4). Table 2.1 displays the accession numbers used as queries for BLAST searches.

Family	GenBank accession number		
	RT	LTR*	FLE
Ty1	NP_058183	1253547873	GU224294
Ty2	NP_058163	1253546814	KT203716
Ty3	NP_012184	1253546229	YSCTY31A
Ty4	PC1253	1253551750	AJ439550
Ty5	AAC02631	1751269299	SPU19263

Table 2.1: **Accession numbers used as queries for BLAST searches.** *LTR accession numbers are from the NCBI trace archive. Full-length elements (FLEs) taken from *S. cerevisiae* (*Ty1-3*), *S. uvarum* (*Ty4/Tsu4*) and *S. paradoxus* (*Ty5)*.

Figure 2.2 displays the workflow employed for both protein and LTR datasets through to phylogenetic analysis (Sections 2.2.1-2.2.5 and 2.4).

2.2.1 Compiling RT datasets

For each *Ty* family, the query sequences used were 300 residues in length, as this ensured that the entire RT domain (approx. 220 residues) would be returned. Primarily BLASTp and tBLASTn were utilised to obtain hits in all species containing *Ty*-like elements, and then expanded to Fungal WU-BLAST on the *Saccharomyces* Genome Database (SGD) server to acquire any species that were not present in GenBank. Searches that returned unclassified/unannotated proteins were checked with reciprocal BLAST (using BLASTp) to ensure hits shared similarity with the appropriate *Ty1/copia* or *Ty3/gypsy* families of retrotransposons. If these searches were inconclusive, the hit was compared with BLAST2 to the query sequences of *S. cerevisiae* to distinguish which family shared the highest similarity with the hit. At this stage any indeterminate sequences were discarded. Sequences were also no longer considered for inclusion in the phylogenetic analyses

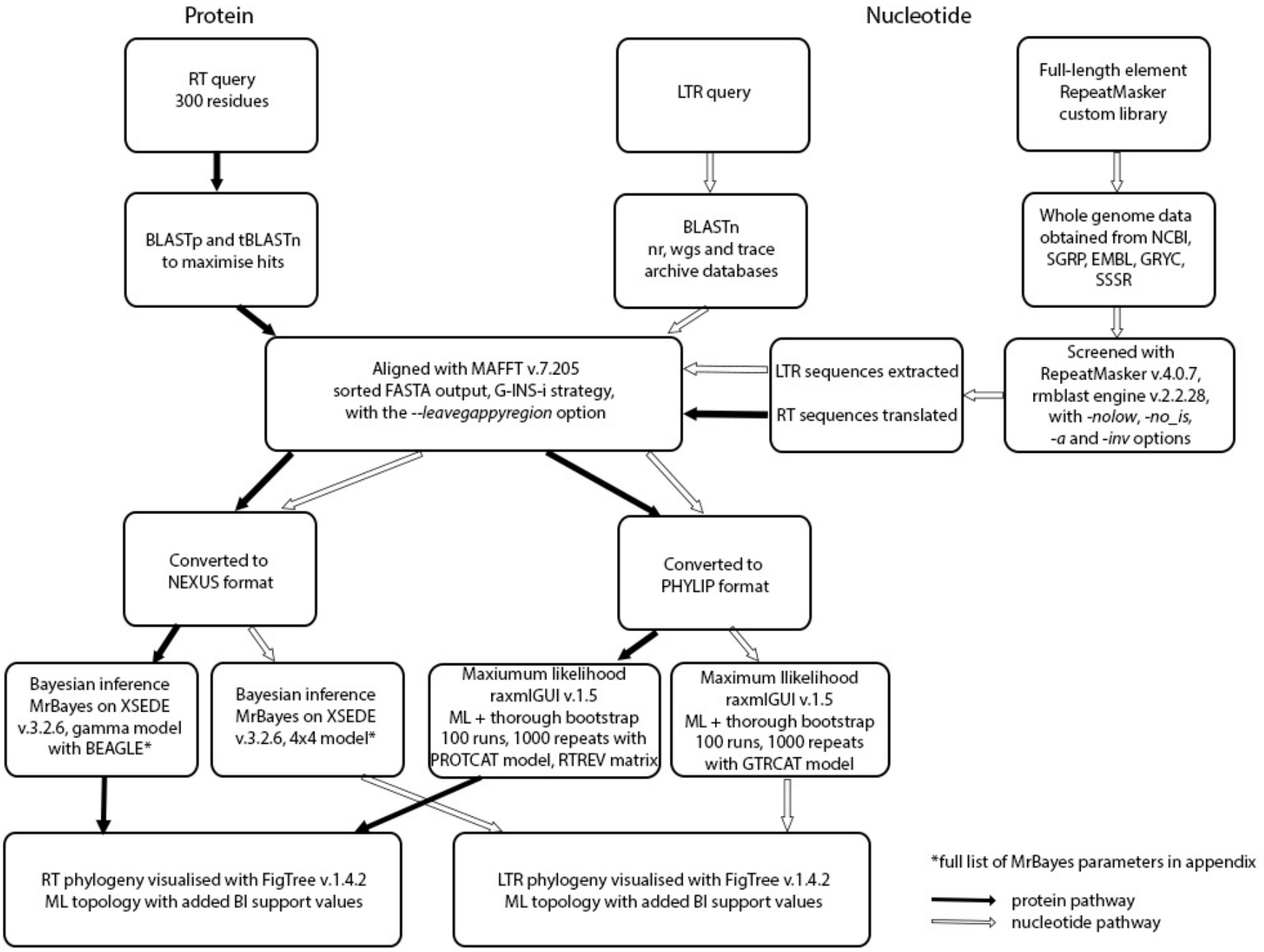

Figure 2.2: **Sequence similarity searches and phylogenetic workflow.** LTR queries ranged from 251bp to 340bp in length for Saccharomyces species. RepeatMasker was initially used with the standard RepBase library.

when a dramatic drop in e value was observed (Chapter 5) or the hit was too short in length to be confidently aligned.

2.2.2 Compiling *Saccharomyces* LTR datasets

For all ascomycete species possessing *Ty*-like elements, datasets were constructed containing all full-length LTRs in all available strains of each species. Partial sequences (i.e. those lacking boundary sequences) were not included in datasets as analyses focussed on potentially active insertions and/or those with intact regulatory regions. This also prevented host genomic DNA from being incorporated into alignments. Identical paralogous copies and those shared across multiple strains (as determined by target site duplications; TSDs) were removed for the purposes of phylogenetic analysis (Section 2.4). Datasets for *S. cerevisiae* and *S. paradoxus* were limited to the *Saccharomyces* Genome Resequencing Project (SGRP) strains (Section 2.3.1). Searches for LTRs utilised the NCBI Trace Archive, Sanger BLAST servers and the SGD WU-BLAST databases. All other species were primarily searched using BLASTn, targeting the nucleotide (nr) and whole-genome sequences (wgs) databases. The initial query sequences (Table 2.1) were not always successful in returning LTR hits from *S. uvarum, S. eubayanus* and *S. kudriavzevii*, therefore searches were repeated with LTR sequences that had been obtained from *S. mikatae*.

2.2.3 Alternative approaches to identifying LTRs in Saccharomycetaceae

Two methods were used to obtain TE sequences in species which had yet to be screened for transposable element insertions, or in those that did not share enough identity with LTRs in *Saccharomyces* species. In the first method, whole genomes were obtained or built from raw reads and screened with RepeatMasker v.4.0.7 using the default RepBase library (*-species fungi*). In the second method, the previously obtained RT sequence was used as a query in tBLASTn in order to acquire a contig large enough to contain a full TE. The RT sequence was identified in the DNA sequence, which allowed for the flanking DNA (5kb up- and downstream) to be compared with BLAST2. The identified nucleotide regions which shared high identity which would in theory be the LTRs flanking the element. These potential LTRs were visually inspected to ensure they possessed the characteristic inverted terminal repeats of 5'-TG-CA-3' (Wicker *et al.*, 2007). Newly obtained LTRs were then used as a search query for the particular species in order to obtain all copies of LTRs across the genome(s).

2.2.4 Constructing a custom RepeatMasker library

Coding regions (where present) and LTRs from each family in each species were collated and correctly formatted in order to construct a custom RepeatMasker library. All strains and species were screened to obtain an accurate TE genome content percentage and to identify any insertions that were not present when using BLAST with a query. This also provided an extensive method to identify any potential horizontal transfers in species that had only initially been screened with *Saccharomyces* BLAST queries. The custom Saccharomycetaceae library is available at `https://github.com/coopergrace/transposable_elements`.

2.2.5 Alignment

As outlined in the workflow (Figure 2.2), familial RT and LTR sequences were aligned with the MAFFT (Multiple Alignment using Fast Fourier Transform) alignment program v.7.205 with an accurate strategy (option 4), but minimising unnecessary gaps (*–leavegappyregion*). Alignments were visually inspected and corrected if required. Regions of poor similarity/identity were excluded from analysis with square brackets in NEXUS format and removed entirely from the corresponding PHYLIP alignment in preparation for phylogenetic analysis (Section 2.4).

2.3 Sequence preparation and analysis

2.3.1 *S. cerevisiae* and *S. paradoxus* population data preparation

Multiple SGRP strains of *S. cerevisiae* (*n*=52) and *S. paradoxus* (*n*=32), sequenced by Liti *et al.* (2009), were ideal for use in population genomics. Only complete LTR sequences from each family were collected, as it is possible that incomplete LTRs do not retain the regulatory (or at least fully functional) sequences which were of interest in this study. All sequences, including those with indels, were added to the datasets providing they retained recognisable 5' and 3' boundaries. Up to two substitutions in the inverted repeats (typically 5'-TGTTG-CAACA-3' or 5'-TGTTG-TACTA-3' for *Ty1/2*) were allowed.

The NCBI Trace Archives were utilised due to the short length of the contigs ($\sim$1kb), and the availability of the SGRP strains. The search was optimised for somewhat similar sequences (blastn) and the maximum number of hits within the search parameters was increased to 20,000. Two versions of each set of hits were downloaded: complete reads which included flanking DNA

and aligned sequences which contained only LTR sequence matching that of the query. The file of purely LTR sequences was used to readily identify duplicate sequences by way of accession numbers using Galaxy, and incomplete LTRs using Galaxy sortbylength (*Ty1-4* minimum length 300; *Ty5* minimum length 200; no maximum). False positive hits were determined by visual inspection. The usable sequences were extracted from the file of complete reads. MAFFT was used to align the sequence files (as in Section 2.2.5), in which the flanking DNA and target site duplications were used again to identify further duplicates across strains. LTRs in the incorrect orientation were reverse complimented using revseq. The resulting dataset consisted of a single copy of each complete LTR from all strains available. For an insertion present in multiple strains, priority was given to the copy present in the reference strain S228c wherever possible due to the higher quality of sequencing/assembly.

2.3.2 Frequency of insertions in *S. cerevisiae* and *S. paradoxus*

Each unique insertion in the data set along with 100bp flanking DNA both boundaries was BLASTed within the trace archives facility of the blastn suite in turn. Flanking DNA was used as part of the search to ensure the correct LTR was identified, due to the high level of similarity between LTRs of the same family. Frequency (polymorphic/fixed) was established by calculating the ratio between the number of insertions (present) vs the flanking DNA (scored). LTRs that were present in a minimum of four strains (henceforth referred to as candidate LTRs) were downloaded as aligned sequences into a separate FASTA file. They were again aligned with MAFFT, the flanking DNA removed and converted to PHYLIP format to be analysed for signatures of positive selection (Section 2.3.4).

2.3.3 Genomic position of candidate LTRs

Each LTR present in ≥4 strains of *S. cerevisiae* and *S. paradoxus* along with 100bp flanking DNA was used as a query sequence in the BLAT feature of the University of California, Santa Cruz (UCSC) genome browsers to determine its genomic co-ordinates. Using the visualisation of the genome browsers, neighbouring genes and/or genomic features were documented, along with distance (bp) between LTR and gene. Those insertions absent from the reference strains were located in their original strain and the regions up- and downstream were searched for host genes.

The respective genome browsers were used to obtain the DNA for each neighbouring gene/feature and downloaded in FASTA format. This was used as a query sequences in a BLASTn search of the *Saccharomyces cerevisiae* taxid (4392) and *Saccharomyces paradoxus* taxid (27291) respectively, to obtain DNA sequences from each of the strains in which the LTR was also present at that particular locus. Flanking DNA was not acquired in order to distinguish between genes unless the gene of interest was present in multiple copies. Genes were also generally well annotated within the genome browsers, therefore in the event that a strain contained more than one copy of a gene, it could be quickly established that the copy is indeed the correct one (e.g. *PAU24* in *S. paradoxus* is present in multiple copies on different chromosomes depending on the strain).

2.3.4 Tajima's *D*

Tajima's *D* (Tajima, 1989) test, part of the DnaSP v.5.10 software package (Rozas and Rozas, 1999) was employed in two main approaches during the work here: to identify candidates for positive selection and to test for recent ancestry of TE copies within families.

2.3.4.1 Positive selection

In Chapter 3, Tajima's *D* test was used on the alignments of single insertions present in ≥4 strains of *S. cerevisiae* and *S. paradoxus* (Section 2.3.1). The test identified signatures of selection acting on any one insertion. The workflow for determination of LTR candidacy for positive selection is displayed in Figure 2.3.

The *D* values were compared to ascertain whether the LTR or neighbouring gene at that locus was more likely to be under positive selection. If it was the LTR that achieved the more negative value (i.e. the further away from zero), then it was more likely that the neighbouring gene was hitchhiking along with the LTR. In this situation, the insertion retained its candidacy for positive selection. Conversely if it was the gene that achieved the more negative *D* value, it therefore the more likely target of positive selection and so the insertion was no longer a candidate. Fu and Li's *D* statistical test (Fu and Li, 1993) was performed on those insertions that showed a significantly negative *D* value in Tajima's test to further strengthen candidacy. The twelve candidate LTRs with the most significantly negative *D* value were selected for expression analysis with qPCR (Section 2.6).

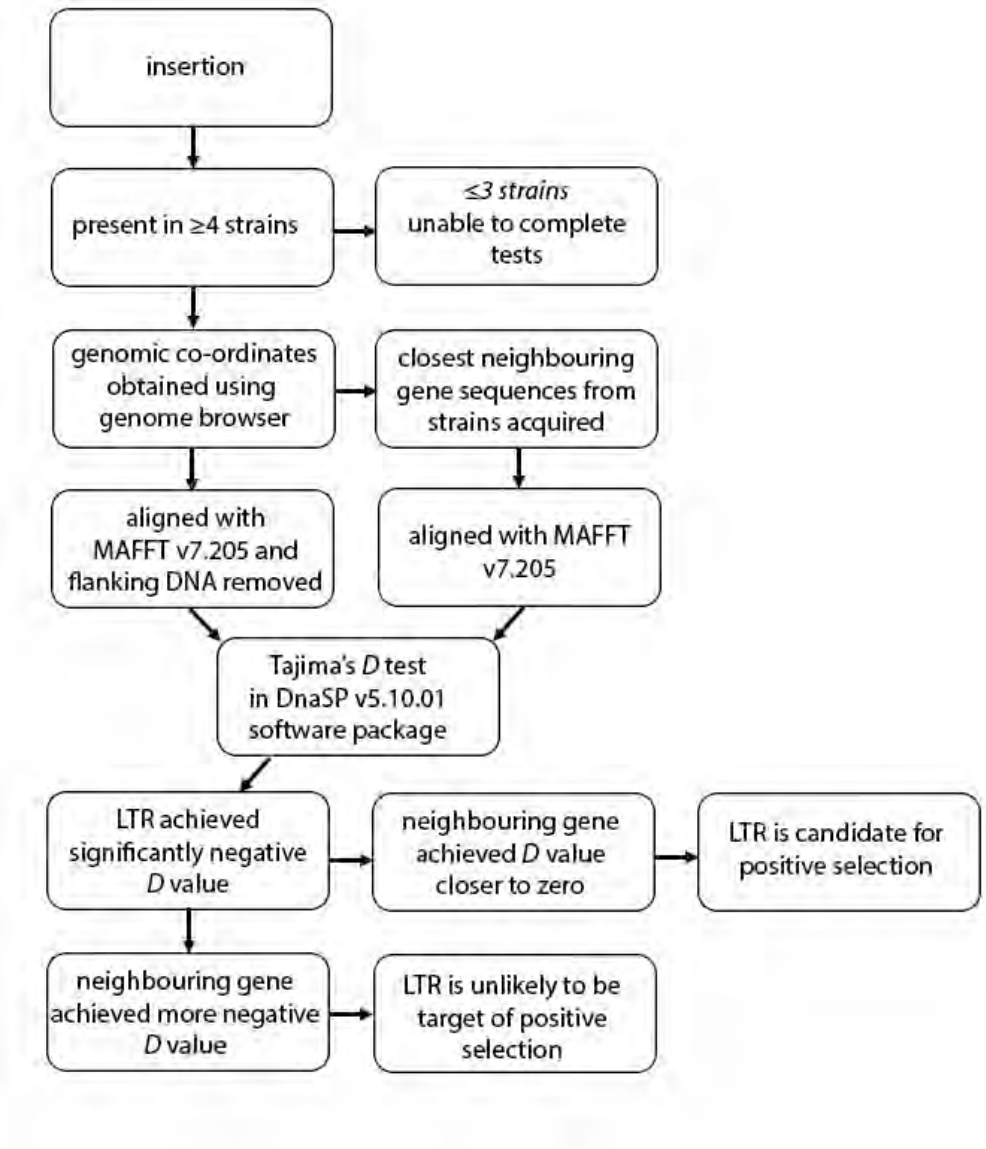

Figure 2.3: **Determining LTR candidacy for positive selection workflow.** Those candidates as determined by Tajima's *D* were retested with Fu and Li's *D* test to confirm the signature of selection.

2.3.4.2 Recent ancestry

In Chapters 4 and 5, Tajima's *D* was used to analyse alignments of entire families of LTR sequences from each species. Families with multiple active lineages in a genome may return a positive value of *D* as there are likely nucleotide variants of intermediate frequency present. Conversely, families with recent ancestry and transposition activity are more likely to receive a significantly negative value of *D*, as nucleotide variants may be at low frequency. Values close to zero are likely the result of neutral evolution or conflicting signals, therefore Tajima's *D* values were evaluated alongside corresponding phylogenetic analyses.

2.3.5 Nucleotide diversity (π)

Nucleotide diversity across familial LTR alignments was calculated using the Nucleotide Diversity test in the DnaSP software package (Rozas and Rozas, 1999). In Chapter 3, the sliding window option was used, with a step size of 5 and a window length of 30bp.

2.3.6 Recombination

Any potential recombination events were identified with the TOPALi v.2 recombination (Milne *et al.*, 2009) and SplitsTree4 v.4.14.4 Phi tests (Huson and Bryant, 2006) on familial alignments. Alignments were also visually inspected to locate potential breakpoints using ClustalX v.2.1 (Larkin *et al.*, 2007). Annotations to the figures generated with ClustalX were added with Adobe Photoshop CS6 v.13.

2.3.7 Chromosomal rearrangements and synteny

To determine the translocation and inversion breakpoints of *S. cariocanus* in relation to parental species *S. paradoxus*, the genomes of *S. cariocanus* strains were aligned against the *S. paradoxus* reference genomes of each population using Mauve v2.4.0 (Darling *et al.*, 2010). Seed weight was increased to 18 due to the low level of divergence between subspecies.

To visualise synteny between *Saccharomyces* genomes, assemblies at contig, scaffold or chromosome level were aligned along with genomic annotation files (.gff) acquired from GenBank using Symap v4.2 (Soderlund *et al.*, 2011). Annotations were added with Adobe Photoshop CS6 v.13.

2.4 Phylogenetic analysis

Two methods of phylogenetic analysis were used on the protein and nucleotide datasets: Bayesian Inference (BI; Yang and Rannala, 1997) and Maximum Likelihood (ML; Huelsenbeck and Crandall, 1997). As outlined in Figure 2.2, prior to phylogenetic analysis alignments were converted to both NEXUS and PHYLIP formats for use with each program respectively.

2.4.1 Bayesian Inference

BI was performed with the MrBayes program v.3.2.6 (Ronquist and Huelsenbeck, 2003) on XSEDE available on the CIPRES (**C**yber **I**nfrastructure for **P**hylogenetic **RES**earch) Science Gateway v.3.3

(Miller *et al.*, 2013). Full parameters to generate both nucleotide and protein phylogenies are located in Appendix B.

2.4.2 Maximum Likelihood

Preliminary ML trees were built with the inference program RAxML v.8 on XSEDE (Stamatakis, 2014) on the CIPRES Science Gateway. Excessively long-branched sequences and those causing disruption to rooting were removed from the LTR alignments to minimise incorrect phylogenetic signals. Corrected alignments were loaded into RAxMLGUI v.1.5 beta (Stamatakis, 2014) for analysis. Full parameters to generate both nucleotide and protein phylogenies are located in Appendix B.

2.4.3 Visualisation of phylogenetic trees

The main output files generated by RAxMLGUI and Mr Bayes were visualised using the FigTree v.1.4.2 Tree Figure Drawing Tool. Each tree was rooted with an appropriate outgroup sequence(s) (or mid-point rooted in the case of *Ty3/gypsy* RT). In LTR trees, distant species were used as the root, usually species which served as an outgroup or assumed the basal position within the corresponding RT phylogeny. ML topologies were used as final figures with added support values from BI when branch topologies were consistent between the two methods. The colour scheme followed that of Carr *et al.* (2012) where possible and additional species and genera were assigned new colours as required. Species names and additional details, grouping names etc. and indicators of significant bootstrap values were added with Adobe Photoshop CS6 v.13.

2.5 Yeast husbandry

The SGRP Strain Set 1 of *S. cerevisiae* yeast was purchased from the National Collection of Yeast Cultures (NCYC). Strains were provided frozen in a single use glycerol stock plate. Upon receipt, the strains were stored at -80°C for 48 hours to allow excess CO2 to dissipate. The *S. cerevisiae* strains were cultured in liquid YPD media (Fisher Scientific) as described by Amberg *et al.* (2005). 1ml per litre of media ampicillin (100mg; Invitrogen) was added to media before use. 20ml of liquid media was transferred to 50ml tissue culture flasks (Starstedt) before inoculation with the strains. Upon opening each supplied strain tube, 200µl aliquots were transfered to new cryopreservation

tubes (Fisher Scientific) containing 1ml of 15% glycerol (Sigma) YPD media. The remaining stock (∼800µl) of each strain was added to the culture flasks and incubated at 30°C for 72 hours before use or until turbidity was observed.

2.6 Expression studies

2.6.1 RNA extraction of *S. cerevisiae* SGRP strains

The protocol of RNASwift, developed by Nwokeoji *et al.* (2016), was used with minor adjustments to extract RNA from *S. cerevisiae* cultures. 2ml of inoculated media was transferred into a 2ml microcentrifuge tube (Starstedt) and centrifuged at 4500rpm (Sigma 1-14 microfuge) for ten minutes to form a pellet. The supernatant was removed and 100µl of lysis buffer, consisting of 0.5M NaCl (Fisher Scientific), 4% SDS (Fisher Scientific), pH 7.5, added to each tube. The cells were homogenised by pipetting and incubated (Labret Accublock) at 90˚C for four minutes. The lysate was centrifuged at 13,000rpm for one minute and the supernatant transferred to a new 2ml microcentrifuge tube, taking care to avoid also transferring the proteins, gDNA and other cellular remnants present in the pellet. 540µl of purification reagent, consisting of 40µl 5.0M NaCl, 250µl 1.0M guanidine HCl (Sigma) and 250µl isopropanol (Fisher Scientific), was added to each sample tube and gently mixed by pipetting. This was transferred onto a silica gel extraction column (Qiagen) and centrifuged at 13,000rpm for one minute. The flow-through was discarded and 700µl of wash buffer (15mM Tris HCl, 85% ethanol, pH 7.4; both Fisher Scientific) added onto each column. The columns were again centrifuged at 13,000rpm for one minute and the flow-through discarded. The dry columns were once again centrifuged at 13,000rpm for one minute before eluting the RNA with 100µl of RNase free water (Fisher Scientific) into fresh microcentrifuge tubes. RNA was stored at -80˚C until required.

A 1µl aliquot of each sample was removed for a quality check and determination of yield concentration with the NanoDrop system (Fisher Laboratories). A further aliquot, depending on calculated yield, was added to 1µl of GelPilot 5x loading dye (Qiagen). Total RNA was electrophoresed in a 1% agarose gel (Cleaver), with 1x DEPC-treated TBE (Fisher Scientific) buffer for 40 minutes at 80V and visualised under UV light to check for quality and DNA contamination.

2.6.2 *S. cerevisiae* cDNA synthesis

cDNA synthesis was completed using the Tetro cDNA synthesis kit (Bioline). The protocol was adjusted to a final volume of 10µl instead of 20µl. All steps were performed on ice to avoid nonspecific amplification, as the polymerase enzyme is active at room temperature.

RNA concentration was previously determined using NanoDrop, and the volume was adjusted to use 500ng of RNA (Appendix C), without exceeding 4µl, and added to a 0.2ml PCR tube (Thermo Fisher). If the volume was less than 4µl, the difference was made up with DEPC-treated RNase free water (Fisher). Genomic DNA was removed from the samples by adding 0.5µl of DNase buffer and 0.5µl of DNase enzyme (both New England BioLabs), mixed gently and then incubated at 37˚C for 60 minutes in a Nexus SX1 Mastercycler (Eppendorf). The reaction was stopped by adding 0.5µl of 0.5M EDTA and further incubation in the Mastercycler at 75˚C for ten minutes. The total volume of each sample was therefore 5.5µl.

To synthesise cDNA, a bulk mix was created from the kit solutions up to a total volume of 10µl. For each sample, the bulk mix contained: 2µl of 5x RT buffer, 1µl of dNTP mix, 0.375µl of Random Hexamer Primer Mix, 0.125µl of Oligo (dT)15 Primer Mix and 0.5µl of RiboSafe RNase Inhibitor. The components were gently spun to mix and a 4µl aliquot of the mix added to each sample PCR tube, and also to the RT- control. Finally, 0.5µl of Tetro RT enzyme was added to each of the RT+ tubes. 0.5µl of RNase free water was added to the RT- control which was used to highlight background signals from contaminating gDNA. The final volume of each sample was therefore 10µl. In the Nexus SX1 Mastercycler (Eppendorf), the samples were then incubated at 25˚C for ten minutes, followed by 45˚C for 30 minutes when using random hexamers. The reaction was stopped by incubating at 85˚C for five minutes, and then the samples were returned to ice. Samples were stored at -20˚C long term or used immediately for qPCR. cDNA was diluted to 1:5 with low concentrate TE (Fisher Scientific) and stored at -20°C. This was then diluted to a working solution of 1:25 with DEPC-treated RNase free water (Fisher Scientific) and stored at 4°C for use in qPCR.

2.6.3 qPCR for determination of expression levels of neighbouring genes

2.6.3.1 Primer design

Primers were designed using sequences from the reference strain of S. cerevisiae S288c in the online primer design software Primer3 v4.0.0 (Untergasser *et al.*, 2012). The size of the PCR

product required was set to 80-130bp; number of options to return increased to 20; primer Tm was set to between 59°C and 61°C with an optimum of 60°C; max Tm difference to 1°C; algorithm of thermodynamic parameters used by Breslauer *et al.* (1986); GC% content maximum set to 70%; max self-complementarity set to 3 and max 3' self-complementarity set to 1 as ideal values, but these were increased if the software failed to return any potential primers. All other parameters were kept as default. Primers generated are located in Appendix D. The genes neighbouring ten most significantly negative candidates for positive selection were selected, plus three housekeeping/reference genes determined in previous study by Teste *et al.* (2009): *ALG9*; *TAF10*; *UCB6*. Primers were ordered from Eurofins MWG Operon. Upon receipt, primers were diluted with low concentrate TE (Fisher Scientific) to a stock solution of 50μM and stored at -20˚C. Primers were then diluted to a working solution (10μM) with ultrapure water and also stored at -20˚C.

2.6.3.2 qPCR protocol

For each gene of interest and housekeeping gene, a bulk mix was prepared on ice. For each strain, the bulk mix contained: 3μl of iTAQ Mastermix (Bio-Rad), 0.3μl of forward primer, 0.3μl of reverse primer and 1.4μl of UV-treated ultrapure water. The bulk mix was gently centrifuged and mixed by pipetting. Despite the iTAQ being a hot start polymerase, plate preparation was also performed on ice as a necessary precaution to minimise background amplification. 5μl of the bulk mix was added to each plate well, along with 5μl of cDNA from each strain. All reactions were carried out in triplicate, with three negative controls for each gene. A 0.3x allowance for pipetting error was used. 96 well plates (Bio-Rad; layout is located in Appendix E) were used, as well as the corresponding plate seals (Bio-Rad). After ensuring the plates were well sealed, they were very gently spun in a plate spinner (Labnet International) for 20-30 seconds. The qPCR protocol used was as follows: initial denaturation stage of 30 seconds at 95°C, then an annealing/extension stage at 60°C for 30 seconds followed by denaturation at 95°C for five seconds. The annealing-denaturation stages were repeated a total of 40 times. The melt curve analysis was then performed by increasing the temperature in 0.5°C increments every five seconds from 65-95°C, during which the plate would be read at each step by the CFX96 Touch Thermal Cycler (Bio-Rad).

2.6.3.3 qPCR data analysis

The data were exported from the CFX machine (Bio-Rad) and imported into CFX Manager v. 2.1 (Bio-Rad) to add the plate details (sample name, target etc.) to the data. This was then imported into qbase+ software for analysis. The two housekeeping genes were identified as reference genes to the software, which then normalised the rest of the data. This was then exported without logarithmic scale as a Microsoft Excel file. Expression data were separated into two columns for each gene, depending on whether the insertion was present or absent at the adjacent locus. These columns were then imported into GraphPad Prism v4.0 for analysis with unpaired two-tailed t-tests with 95% confidence intervals. Box-plots were generated from the data for each adjacent gene.

Chapter 3

Identifying positive selection acting upon *Ty* insertions in *Saccharomyces*

TE insertions are typically evolving under purifying selection, particularly as they can transpose into important areas of the genome, causing deleterious effects such as chromosomal rearrangements (Charlesworth *et al.*, 1993, 1994). TEs are also known to cause changes in gene expression in a variety of host organisms (reviewed in Elbarbary *et al.*, 2016). Examples include plants (Wang *et al.*, 2013; Le *et al.*, 2015; Hirsch and Springer, 2017), mammals (Pereira *et al.*, 2009; Nakanishi *et al.*, 2010; Warnefors *et al.*, 2010; Trizzino *et al.*, 2017) and insects (Cridland *et al.*, 2015). In yeast, this was first documented in the 1980s with the expression control of *HIS4* by an adjacent *Ty1* element (Roeder *et al.*, 1985, 1986; Coney and Roeder, 1988). These effects have also more recently been explored in other fungi, such as Basidiomycetes (Castanera *et al.*, 2016) and *Coccidioides sp.* (Kirkland *et al.*, 2018).

The discovery of potentially beneficial insertions has predominantly been the result of serendipitous findings, such as insecticide and viral resistance in *Drosophila* (Daborn *et al.*, 2002; Aminetzach *et al.*, 2005; Magwire *et al.*, 2011; Guio *et al.*, 2014; Mateo *et al.*, 2014) and mosquitoes (Paris and Despres, 2012), breeding compatibility in *Arabidopsis* sp. (Shimizu *et al.*, 2008), and endogenous retroviruses which may contribute to the immune response in humans (Chuong *et al.*, 2016, 2017). Conferring benefits such as these may allow insertions to persist in the genomes of their hosts. Speculation into the possibility of insertions conferring benefits beyond genomic diversity led to genome-wide screenings for selection acting upon TE insertions. Using methods based upon haplotype structure to detect signatures of selection, screenings have been completed on the genomes of human (Kuhn *et al.*, 2014), *Drosophila melanogaster* (Macpherson *et al.*, 2008; Ullastres *et al.*, 2015; Merenciano *et al.*, 2016) and periwinkle (Wood *et al.*, 2008). Site frequency

spectrum based methods, such as that of Tajima's *D* (Tajima, 1989), have identified selection signatures on TEs in the genomes of other organisms, including *D. melanogaster* (Macpherson *et al.*, 2008; Kofler *et al.*, 2012; Ullastres *et al.*, 2015; Merenciano *et al.*, 2016), *D. simulans* (Schlenke and Begun, 2004), *Arabidopsis* (Hazzouri *et al.*, 2008), *Caenorhabditis elegans* (Laricchia *et al.*, 2017) and the silk moth *Bombyx mori* (Sun, Shen, Han, Cao and Zhang, 2014). In this chapter, population genomics data of two *Saccharomyces* species, *S. cerevisiae* and *S. paradoxus*, enabled the examination of insertions present in multiple strains for signatures of selection. Tajima's *D* test uses the relationship between pairwise nucleotide diversity and the frequency of segregating sites to generate a value representative of selection that may be acting upon that site. Significantly positive values are consistent with balancing selection due to intermediate frequencies of polymorphisms (Kreitman, 2000; Barreiro and Quintana-Murci, 2010), whereas significantly negative values are consistent with positive selection due to an excess of rare polymorphisms (Tajima, 1989; Rozas, 2009).

While previous studies focussed primarily on the variability of full-length elements (FLEs; e.g. Bleykasten-Grosshans *et al.*, 2013; Sasaki *et al.*, 2013), few investigations into solo LTRs have been performed, therefore the work presented here is one of the first in *S. cerevisiae* and *S. paradoxus*.

3.1 Data collection

The population genomics data of sister species *S. cerevisiae* and *S. paradoxus* (Liti *et al.*, 2009) were utilised in this investigation into insertions potentially under positive selection. This high quality and availability of data is currently unavailable for strains of other *Saccharomyces* species.

The *Saccharomyces* Genome Resequencing Project (SGRP) *S. cerevisiae* (*n*=52) and *S. paradoxus* (*n*=32) strains were screened for TE insertions via the NCBI trace archive. LTR sequences were extracted from reads along with 100bp flanking DNA to aid identification. Insertions present in <4 strains at any given locus (i.e. possessing the same flanking DNA) were discarded as the statistical tests required sequences from ≥4 strains. Those insertions that are present in ≥4 strains were aligned with MAFFT, flanking DNA removed and tested for the signature of positive selection with Tajima's *D* test, part of the DnaSP software package (Rozas and Rozas, 1999). If the insertion possessed a significantly negative value of *D*, the LTR's position in the reference genome was

ascertained using the UCSC's *S. cerevisiae* and *S. paradoxus* genome browsers, respectively. Regions up- and downstream from the insertion were explored for neighbouring genes, which were also subjected to the same statistical tests. If, of the pair, the LTR possessed the more significantly negative value of *D*, it retained its candidacy for positive selection. If the adjacent gene possessed the more negative value of *D* however, the LTR was no longer considered the target of positive selection.

The total numbers of sequences screened, unique copies, those present in ≥4 strains and final numbers of candidates are displayed in Tables 3.1 and 3.2 for each species, respectively.

Family	Sequences screened	Unique copies	Present in ≥4 strains	significant -ve *D* value	Candidate insertions
Ty1/2	38,000	656	115	21	19
Ty3	7,080	346	81	18	15
Ty4	4,900	212	26	4	3
Ty5	1,500	48	19	3	3
Total	51,480	1,262	241	46	40

Table 3.1: **Summary of sequences screened for positive selection in *S. cerevisiae*.** Numbers of initial sequences screened (approx.), unique copies, those present in ≥4 strains and those that received significantly negative values of Tajima's *D*. True candidate insertions for the SGRP strains of *S. cerevisiae* were determined from those with more significant *D* values than neighbouring genes within effective distance (Section 3.2).

Family	Sequences screened	Unique copies	Present in ≥4 strains	Significant -ve *d* value	Candidate insertions
Ty1/2	32,600	315	108	15	17*
Ty3p	4,500	131	22	1	1
Ty4	5,400	194	13	2	2
Ty5	1,000	53	11	1	1
Total	43,500	693	154	19	21

Table 3.2: **Summary of sequences screened for positive selection in *S. paradoxus*.** Numbers of initial sequences screened (approx.), unique copies, those present in ≥4 strains and those that received significantly negative values of Tajima's *D*. As in *S. cerevisiae*, true candidate insertions for the SGRP strains of *S. paradoxus* were determined from those with more significant *D* values than neighbouring genes within effective distance (Section 3.2). *Two candidates in *Ty1/2* are added as although they did not gain a value of *D* due to lack of polymorphisms, they are conserved in ≥10 strains (detailed in Section 3.2.1).

Although the initial numbers of screened sequences in each species does not greatly differ, the numbers of unique copies, those present in ≥4 strains and the candidate insertions are all approximately double in *S. cerevisiae* than those in *S. paradoxus*. The differences between copy numbers in each species are not significant however (*P*=0.15; paired two-tailed t-test).

While the number of candidate LTRs in the *Ty1/2* superfamily is similar in both species, *S. cerevisiae* possesses a greater number of candidates in the other families, particularly those of *Ty3*. In both species, candidates represent 3% of unique LTRs across all strains. In *S. paradoxus*, 5% of unique *Ty1/2* LTRs are candidates, whereas this was reduced to 2% in *S. cerevisiae*.

Details of candidates for positive selection are listed in Section 3.2, and interesting observations are discussed in Sections 3.1.1 to 3.1.4.

3.1.1 YDRWdelta11: a *Ty1* relic in *S. cerevisiae* may be under positive selection

YDRWdelta11 received a significantly negative value of Tajima's *D* (Section 3.2), and upon examination of this insertion in the reference genome browser, it was discovered to be the 5' LTR of a *Ty1* relic element (YDR170W-A) on chromosome IV. The element has retained the most 5' ∼1.3kb containing *TYA/gag* and frameshift region associated with the LTR (Figure 3.1 A).

Testing of the insertion's neighbouring features, *SEC7* and autonomously replicating sequence ARS425, indicated that the signature of positive selection is centred on the LTR at this locus and therefore remains a candidate. As this relic is present in multiple strains (*n*=22), the alignment was expanded to test the coding region for signatures of selection with a sliding window of Tajima's *D*. Figure 3.1 (B) shows that the two regions possessing the most significantly negative *D* values - and therefore the likely regions upon which selection is strongest - are within the LTR and *TYA*.

3.1.2 Candidates within regions of nested LTRs possess signatures of positive selection

Four candidate LTRs are nested within other insertions and elements in the genomes of *S. cerevisiae* SGRP strains. Those features around the candidates were also tested with Tajima's *D*, the majority of results of which were not significant. Figure 3.2 displays genome browser views of regions where candidates are nested or otherwise closely associated with other insertions.

All neighbouring features were also tested with Tajima's *D* and as each received a less significant value of *D*, LTRs all retained their candidacy. However, YGRWdelta23 (Figure 3.2 D) is

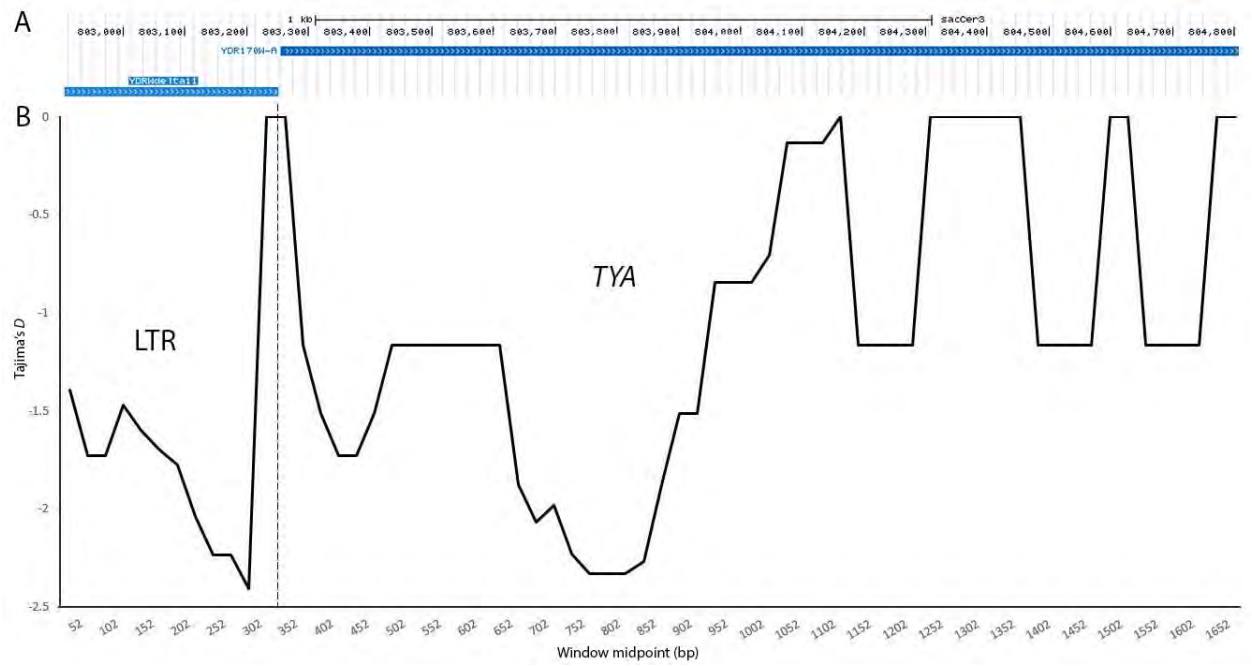

Figure 3.1: **Sliding window of Tajima's *D* values over a *S. cerevisiae Ty1* relic.** The genomic region containing the relic element visualised in the UCSC genome browser (A); resulting *D* values calculated across 25bp sliding windows of the 5' LTR and remaining coding region of the *Ty1* relic on chromosome IV in 22 strains (B).

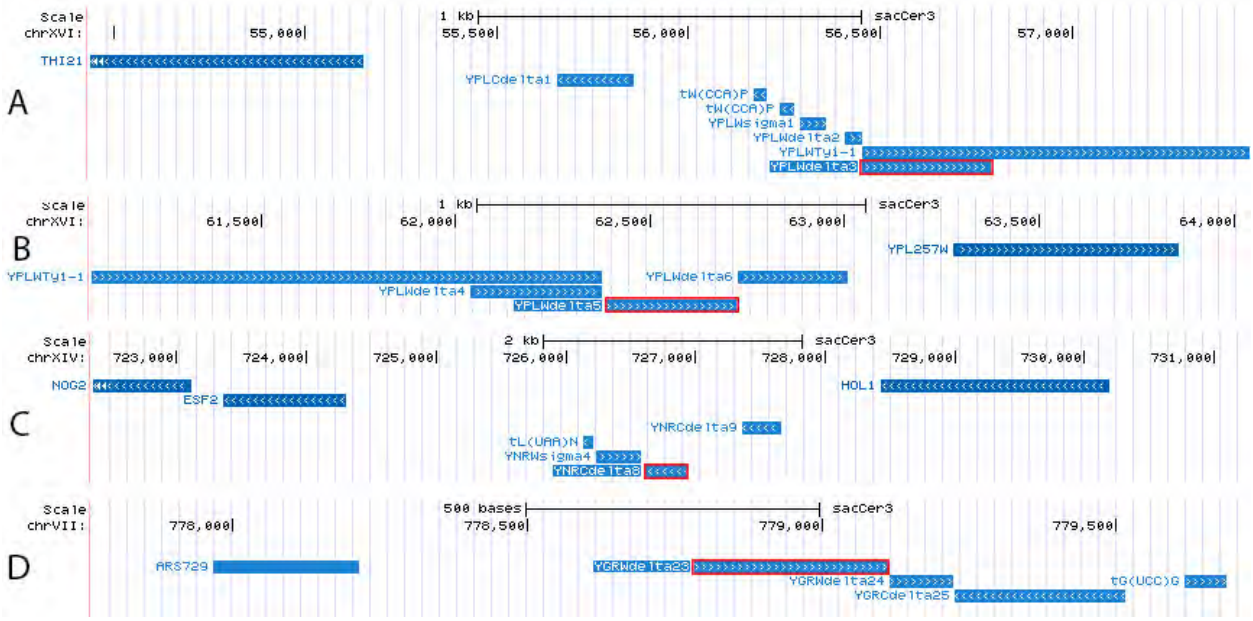

Figure 3.2: **Genome browser views of regions of nested LTRs containing candidate insertions.** A: YPLWdelta3 is present in the reference strain as the 5' LTR of YPLW*Ty1*-1 but as a solo LTR nested within other insertions and tRNAs in all other strains (*n*=10; 3.1.3). B: YPLWdelta5 is located immediately downstream of the FLE visualised in A. C: YNRCdelta8 is located between YNRWsigma4 (solo *Ty3*) and YNRCdelta9 (truncated *Ty1*). D: YGRWdelta23 is the most recent insertion in this region of chromosome VII, having disrupted two previous solo LTRs. Candidates are outlined in red. Scale bars are positioned at the top of each window. Arrows within features indicate orientation.

located upstream of a solo *Ty1* LTR, YGRCdelta25, which is truncated in multiple strains (281bp, *n*=15; also full-length in four strains; Section 3.2.1), having been later disrupted by partial LTR YGRWdelta24, which in turn is the insertion site of the candidate. YGRCdelta25 also possesses a strongly negative value of *D* (-2.156, *P*<0.01), with no significant difference to that of the candidate (*D*=-2.248, *P*<0.01; two-tailed paired t-test, *P*=0.5). Therefore, the exact target of positive selection at this locus may be spread over the two LTRs, or may have been acting upon the older insertion (YGRCdelta25) before the more recent LTRs caused the disruption of this sequence.

3.1.3 Deviations from the reference genome of *S. cerevisiae* illustrate the variability of SGRP strains

The *S. cerevisiae* reference genome is based predominately on S288c, but was actually constructed using multiple derivative strains (Goffeau *et al.*, 1996; Engel *et al.*, 2014). Other authors have reported updates to the TE content of S288c having found that insertions differ to those presented in the UCSC genome browser (Wheelan *et al.*, 2006; Shibata *et al.*, 2009; Bleykasten-Grosshans *et al.*, 2011; Carr *et al.*, 2012; Istace *et al.*, 2017). One such region observed here is that of an insertional hotspot around *LEU2* on chromosome III (Figure 3.3).

Figure 3.3: **Genome browser view of the variable region surrounding *LEU2*.** Candidate III-92884 (far right of the figure) is unannotated in the reference genome. Poor quality sequencing of this area in the other SGRP strains prompted investigation into this region of chromosome III. Multiple partial *Ty1-3* LTR sequences were discovered in this region of most strains, in addition to the *Ty1* FLE (black arrow; Shibata *et al.*, 2009) upstream of YCLWTy2-1 (compressed in this view, also note the change in co-ordinates at the top of the figure).

A search for good quality sequence data of solo *Ty3* insertion III-92884 - absent from the reference genome - provided results in relatively few strains (*n*=11). Trace archive reads and SGRP assemblies for the region of *LEU2* and surrounding areas are poor quality and/or contained unassigned bases (N). *LEU2* is necessary for the leucine biosynthesis pathway (Brisco and Kohlhaw, 1990), therefore its apparent absence in more than half of *S. cerevisiae* strains suggests this may be a difficult area of the genome to sequence. Visual inspection of the genome browser (Figure 3.3) shows that this region is populated by multiple *Ty* insertions (Warmington *et al.*, 1986;

Wheelan *et al.*, 2006; Shibata *et al.*, 2009; Istace *et al.*, 2017). Highly repetitive regions are notoriously difficult to sequence with short read methods (Treangen and Salzberg, 2011; Hoban *et al.*, 2016). In addition, expression data from 22 SGRP strains of *S. cerevisiae* (Skelly *et al.*, 2013) were analysed using the same method as in Section 3.7. *LEU2* was found to be consistently expressed, regardless of insertion presence (P=0.14; two-tailed t-test). Further deviations from the *S. cerevisiae* reference genome were also observed in the other SGRP strains, such as that displayed in Figure 3.4. While previous studies focussed primarily on the variability of FLEs (e.g. Bleykasten-Grosshans *et al.*, 2013; Sasaki *et al.*, 2013), few investigations into solo LTRs have been performed. Determining candidacy was complicated by the polymorphic state of insertions across strains (also Section 3.2.1); for example, FLEs in the reference genome are not necessarily present as such in other strains, but rather as solo LTRs or absent altogether, and vice versa.

Figure 3.4 displays a deviation from the reference genome, in which a candidate insertion was lost due to the more recent transposition of a *Ty1* FLE, whereas this high frequency solo LTR is full-length in multiple SGRP strains (n=19). Discrepancies between the reference strain and others such as that of YGRCdelta12 were also observed for insertions YPLWdelta5, YMLWdelta4, YPLWdelta3 and YILWsigma2. These are all present in the reference genome as associated with FLEs, whereas they are present as solo LTRs in the other SGRP strains. Furthermore, candidates were identified here that are absent from the reference genome altogether (n=7).

3.1.4 LTR VI-183598 possesses a significantly positive Tajima's *D* value in *S. paradoxus*

Insertion VI-183598, present in 22 strains of *S. paradoxus*, was investigated further due to its significantly positive Tajima's D value (2.508, P<0.01; Table 3.4). This locus is also occupied in *S. cerevisiae* by a partial *Ty1* insertion, YFRWdelta7, that was not included in the candidacy search due to the loss of the first $\sim$40bp (Figure 3.5). Partial insertions such as this may be lacking the regulatory regions necessary for alteration of host gene expression and therefore not provide an identifiable benefit to the host, despite persisting in the genome.

When testing alignments of individual insertions, a significantly positive Tajima's D value is consistent with balancing selection. It may also be the result of multiple differing *Ty* insertions mistakenly being included in the same alignment, therefore it was visually inspected for evidence of multiple insertions. As the flanking DNA of each insertion is identical, it was therefore concluded

Figure 3.4: **Genome browser view of the region containing candidate insertion YGRCdelta12.** The candidate LTR (outlined in red) is disrupted in the reference strain by a more recent *Ty1* FLE (A), whereas this is present as a full-length LTR in multiple SGRP strains (B, *n*=19). In the reference strain, the disrupted candidate LTR was not able to be recovered from the 3' end of the FLE. Scale bar at the top of the figure is accurate for all strains whereas the co-ordinates apply to the reference strain only (A).

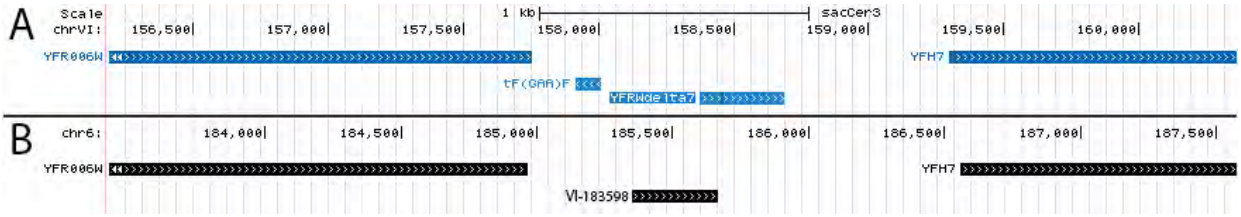

Figure 3.5: **Genome browser view of the region containing VI-183598/YFRWdelta7.** The UCSC genome browser for *S. paradoxus* (A) does not contain annotated *Ty* insertions and other features such as tRNAs, therefore these were manually annotated for the purposes of generating this and similar figures. The insertion site is shared with *S. cerevisiae* (B), suggesting this may be a 'safe' region. In both species, the insertion is located between two ORFs, YFR006W and *YFH7*. YFR006W encodes an aminopeptidase, while the gene product of *YFH7* is a P-loop kinase (SGD).

that the insertion site is consistent in all strains, while the insertion itself varied. Polymorphic sites (n=17) allow the sequences to be split into two groups, each of which possess insignificant D values when tested alone (data not shown). This split is not based upon geographical origin, as the insertion is only present in European strains, indicating that the positive value of D may be due to subdivision of sequences at this particular locus.

As a result of this observation in *S. paradoxus*, the same genomic region was investigated in *S. cerevisiae*. The flanking DNA at this locus is not shared between species; in fact, the *S. paradoxus* flanking DNA is not present anywhere in the *S. cerevisiae* genome. 32 strains of *S. cerevisiae* contain a fixed partial insertion (YFRWdelta7) within the same region (Figure 3.5). The insertion possesses a significantly negative D value (-2.081, P<0.05), consistent with positive selection. However, as this study focusses upon insertions with intact boundaries, it was not considered for expression analysis (Section 3.7).

3.1.5 YJRWdelta18 in *S. cerevisiae* possesses a significantly positive Tajima's *D*

During statistical testing of insertions, a single *Ty1* solo LTR in *S. cerevisiae*, YJRWdelta18, produced a significantly positive value of D (2.60, P<0.01). This is located on chromosome X, alongside a further *Ty1* solo LTR and tRNA, between ARS1018 and an ORF of unknown function, YJR056C (Figure 3.6). None of these other genomic features returned significant D values when tested, but all were slightly negative.

Upon visual inspection, the alignment was easily split into two groupings based on polymorphic sites (n=6), and like the insertion in *S. paradoxus* with a significantly positive D value (Section 3.1.4), this is not due to geographic origin. Flanking DNA and TSDs again indicate that this is

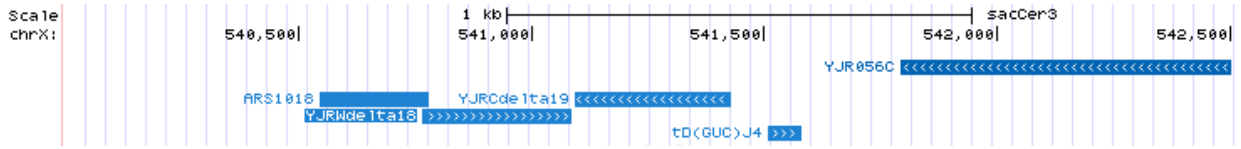

Figure 3.6: **Genome browser view of the region containing YJRWdelta18.** The LTR insertion YJRWdelta18 was shared the TSD with that of the next insertion, YJRCdelta19, which did not receive a significant value of *D*.

the same locus occupied in each strain, yet the positive *D* value is consistent with subdivision of sequences.

Furthermore, this locus is also occupied by a *Ty1p* solo LTR in *S. paradoxus*, but as its *D* value is insignificant, this may be another 'safe' region of the genome in which insertions can reside.

3.2 Signatures of selection acting upon LTRs were identified with Tajima's *D*

All insertions present in ≥4 strains were aligned and analysed with the Tajima's *D* test available within the DnaSP software package (Rozas and Rozas, 1999). Tables 3.3 and 3.4 display the significant results of Tajima's tests on LTRs and their adjacent genes in each species, respectively. The results of tests on all insignificant insertions in *S. cerevisiae* and *S. paradoxus* are available in Appendices G and H, respectively. Tests on candidates were repeated with Fu and Li's *D* statistic (Section 3.3) to strengthen candidacy, and the most significantly negative results were chosen for expression studies in *S. cerevisiae* (Section 3.7).

Only solo LTRs were included in alignments (with the exception of YDRWdelta11, 3.1.1), as those strains that possess FLEs at corresponding loci were excluded as the strongest signature of selection may not be in the associated LTRs. Although excluded from statistical analysis, the strains containing FLEs at otherwise solo loci were included in qPCR analysis as 'present' (Section 3.7). LTRs were also excluded if the distance to an adjacent gene was larger than the 2.2kb effective distance for an enhancer (Elion and Warner, 1984; Johnson and Warner, 1989). Distance between insertions and adjacent genes was checked in each case due to the variability of strains (Section 3.1.3).

When performing multiple statistical significance tests, it is standard to employ a method of correction to minimise false discovery rate (Goeman and Solari, 2014). However, as *P* values are dependent on the strength of selection, the decision was made not to use a multiple test correction.

Weak selection, even positive, would result in a less significant *P* value, therefore a method of correcting for false positives may eradicate these results. In addition, DnaSP does not provide exact *P* values. This was the same decision made by Kofler *et al.* (2012) for similar reasons.

3.2.1 Comparison of candidate LTR and adjacent gene Tajima's *D* values

Tables 3.3 and 3.4 on the following four pages display the *D*, π and θ values calculated for candidate LTRs and adjacent genes in both *S. cerevisiae* and *S. paradoxus*, respectively.

Those LTRs that possess a more negative value of *D* than adjacent genes retained their candidacy for positive selection, whereas those genes that possess a more negative *D* value were considered as the more likely target for positive selection. In these latter cases, the LTR was no longer considered a candidate from this point onwards.

In *S. cerevisiae*, nine genes adjacent to candidates also possess significantly negative values of *D* (Table 3.3). In around half of these (*n*=4), the gene gained a more significantly negative value and therefore the more likely region on which positive selection is acting. In these cases, the adjacent LTR lost its candidacy for positive selection. Due to the compact genome of *Saccharomyces*, the majority of insertions investigated here are within effective distance of adjacent genes. Potential linkage with four genomic features, ARS425, *ORM1*, *ESF2* and *MTC3* was discounted in *S. cerevisiae* due to too great a distance (>2.2kb). In *S. paradoxus*, only *OCA4* was considered beyond the effective distance for a candidate to affect expression. Unlike in *S. cerevisiae*, only one adjacent gene in *S. paradoxus* also achieved a significant value of *D* (Table 3.4). Additionally, all candidate LTRs in *S. paradoxus* retained their candidacy after also testing the adjacent genes, i.e. LTRs always possess the more negative of the two values. Those genes adjacent to LTRs that lost their candidacy due to distance or *D* values were excluded from analysis from this point onwards.

Second only to centromeres, LTR sequences are the most rapidly diverging components of genomes (Bensasson *et al.*, 2008; Bensasson, 2011), therefore, very high conservation of insertions across multiple strains was unexpected. Insertions in both species that failed to produce *D* values due to lack of polymorphisms (*S. cerevisiae n*=8; *S. paradoxus n*=7), were therefore carried forward as candidates if present at relatively high frequency (arbitrary cut off point of ≥10 strains). Two insertions in *S. paradoxus* were included as potential candidates for positive selection (Table 3.4), while *S. cerevisiae* did not gain any additional candidates.

		Candidate LTR					Adjacent gene/feature					Distance (bp)
Family	No.	SGD I.D.	π	Θ	*D*	*P* sig.[c]	SGD I.D.	π	Θ	*D*	*P* sig.[c]	
Ty1/2	1	YERWdelta22	0.004	0.006	**-1.796**	*	*COG3*	0.003	0.004	-0.721		642
	2	YMLCdelta2	0.014	0.020	**-1.665**	*	YML053C	0.005	0.006	-0.285		435
	3	YERCdelta16	0.008	0.014	**-2.060**	**	YER134C	0.001	0.002	-1.448		990
	4	YORWdelta17	0.006	0.013	**-2.192**	**	*MPC54*	0.001	0.002	-1.728	*	640
	5	YHRCdelta10	0.014	0.022	**-1.849**	*	*DCD1*	0.004	0.005	-0.955		452
	6	YPLWdelta5	0.009	0.018	**-1.958**	*	YPL257W	0.001	0.01	-1.129		559
	7	YELWdelta6	0.012	0.027	**-2.280**	**	*GCN4*	0.012	0.021	-1.814	*	365
	8	YDRWdelta11	0.012	0.030	**-2.222**	**	*SEC7*	0.001	0.001	-1.435		684
							ARS425	n/a	n/a	n/a		2818
	9	YCRCdelta6	0.009	0.020	**-2.235**	**	YCR007C	0.001	0.002	-1.712		1546
	10	YHRCdelta3	0.004	0.010	-2.235	**	*TIM10*	0.002	0.003	**-2.346**	*	521
							ARS806	n/a	n/a	n/a		1
	11	YGRWdelta32	0.006	0.012	**-1.998**	*	*CRM1*	0.001	0.001	-0.092		522
	12	YLRCdelta9	0.019	0.038	**-2.130**	**	*ADY4*	0.002	0.003	-1.052		664
	13	YNRCdelta8	0.008	0.016	**-2.073**	**	*HOL1*	0.002	0.002	-0.739		1481
							ESF2	n/a	n/a	n/a		2309
	14	YMLWdelta4	0.007	0.013	**-2.046**	**	*RRN11*	0.002	0.003	-1.413		161
							CAT2	0.002	0.003	-1.236		1705
	15	YPLWdelta3	0.003	0.005	**-1.831**	*	*THI21*	0.001	0.001	0.196		1300
	16	YGRWdelta23	0.004	0.010	**-2.235**	**	ARS729	0.003	0.006	-1.692		571
	17	YLRCdelta21	0.006	0.012	**- 2.120**	**	*GAS2*	0.007	0.008	-0.663		313
							RPL26A	0.006	0.007	-0.890		909
	18	VII-77305	0.001	0.003	**-1.861**	*	*MTC3*	0.003	0.004	-0.703		3594
							VRG4	0.001	0.002	-1.520		269
	19	YJRWdelta18	0.010	0.006	**2.602**[a]	**	YJR056C	0.003	0.003	1.032		705
	20	YOLCdelta3	0.005	0.013	**-2.122**	*	ARS1509	0.009	0.008	0.497		220
							YOL107W	0.001	0.001	-0.567		165
	21	YGRCdelta12	0.023	0.050	**-2.144**	**	*MSP1*	0.002	0.003	-1.050		214[b]
							RPS25A	0.002	0.002	0.303		1278[b]
Ty3	1	YGRWsigma4	0.002	0.004	-1.800	*	*CLB6*	0.003	0.006	**-2.182**	**	695
	2	YDRCsigma3	0.010	0.022	**-2.140**	**	*AMD2*	0.002	0.003	-1.662		512
	3	XIII-760251	0.011	0.019	**-2.010**	*	*RKR1*	0.001	0.001	-1.245		1081[b]

Ty3 cont.	4	XII-681209	0.019	0.036	**-2.155**	**	*YCS4*	0.001	0.001	-0.665		754[b]
	5	YORWsigma3	0.011	0.020	**-1.950**	*	YOR289W	0.002	0.003	-1.767	*	164
	6	VII-569102	0.004	0.009	**-1.914**	*	*ORM1*	0.001	0.001	-1.160		4015[b]
	7	V-487854	0.024	0.040	**-2.008**	**	YER158C	0.003	0.004	-1.123		1861[b]
	8	YGLCsigma1	0.017	0.027	**-1.863**	*	*USE1*	0.001	0.002	-1.448		1345
	9	YILWsigma3	0.027	0.039	**-1.631**	*	*BAR1*	0.001	0.001	-0.676		287
	10	YHRCsigma2	0.013	0.025	**-1.930**	*	YHR020W	0.003	0.003	-0.820		270
	11	VI-212340	0.024	0.043	**-1.980**	*	*MET10*	0.001	0.001	-1.359		1921[b]
	12	YLRWsigma2	0.006	0.014	**-2.261**	**	*AVL9*	0.001	0.002	-2.152	**	606
							ARS1213	0.003	0.008	-1.983	*	667
	13	III-168387	0.021	0.038	**-2.020**	**	*RHB1*	0.001	0.001	-0.778		388
	14	III-92884	0.018	0.027	-1.730	*	*LEU2*	0.051	0.081	**-1.812**	*	466
	15	YBLWsigma1	0.017	0.026	**-1.720**	*	ARS231	0.006	0.007	-0.876		418
	16	YNLWsigma3	0.014	0.020	**-1.670**	*	YNL042W-B	n/a	n/a	n/a		36
							BOP3	0.001	0.002	-1.337		1023
	17	YERCsigma3	0.002	0.005	**-1.831**	*	*GLC7*	0.001	0.001	1.444		674
	18	YILWsigma2	0.010	0.015	**-1.633**	*	*AIR1*	0.003	0.004	-0.389		276
Ty4	1	YERCtau2	0.016	0.028	**-2.010**	**	*UBP9*	0.001	0.001	-1.595		736
	2	YMRCtau3	0.007	0.014	**-2.132**	**	*SCS7*	0.001	0.001	-1.610		717
	3	YNRCtau3	0.014	0.026	-1.920	*	*ATO2*	0.003	0.006	**-2.084**	**	519
	4	YMRWtau2	0.007	0.012	**-1.961**	*	ARS1319	0.005	0.007	-1.478		1
Ty5	1	YGLWomega1	0.022	0.034	**-1.780**	*	TEL07L	0.097	0.103	0.005		78
	2	YCRWomega3	0.002	0.005	**-2.000**	*	ARS317	0.005	0.008	-1.471		471
							HMR	0.005	0.007	-1.592		222
	3	YHLComega1	0.009	0.013	**-1.674**	*	*ARN2*	0.006	0.006	0.449		126
							ARS802	0.031	0.027	1.728		212

Table 3.3: **Significant Tajima's *D* values for insertions and neighbouring genes in *S. cerevisiae*.** (Table continued from previous page). π, θ and *D* values were calculated for LTRs and neighbouring genes to determine the most likely site of selection. *D* values in bold are those more negative of the pairs. LTRs that achieved a less negative *D* value than adjacent genes were discounted as candidates from this point onwards (*n*=4). Four tests on genes could not be completed, due to lack of polymorphisms and so were not carried forward. [a]As the LTR showed a signature of strong balancing and not positive selection, its investigation was not taken further. [b]Distance in respective strain. All others were present in the reference genome and distances were calculated using the UCSC Genome Browser. ARS425, *ORM1*, *ESF2* and *MTC3* were discounted on the basis of distance. [c]*P* value significance for Tajima's values: * *P*<0.05; ** *P*<0.01. All others are not significant.

		Candidate LTR					Adjacent gene/feature(s)					Distance (bp)[b]
Family	No.	Designation[a]	π	Θ (per site)	*D* value	*P* sig.[c]	SGD I.D.	π	Θ (per site)	*D* value	*P* sig.[c]	
Ty1/2	1	IV-364275	0.022	0.043	**-2.102**	*	*MCH1*	0.004	0.004	-0.619		401
	2	XI-472807	0.021	0.040	**-1.965**	*	*FOX2*	0.004	0.003	0.936		900
							TOF2	0.005	0.004	1.288		770
	3	X-513745	0.028	0.040	**-1.713**	*	*KCH1*[d]	0.002	0.001	1.912		270
							HIT1	0.002	0.002	-0.275		305
	4	IV-445330	0.009	0.013	**-1.757**	*	YDL007C-A	n/a	n/a	n/a		435
							RPT2	0.000	0.000	1.444		416
	5	VII-326627	0.006	0.011	**-1.82**	*	*TOS8*	0.001	0.001	-0.936		1701
							VPS45	0.001	0.001	-1.088		536
	6	XVI-856470	0.021	0.034	**-1.837**	*	*KRE6*	0.002	0.003	-1.353		821
							CUR1	0.005	0.006	-0.664		340
	7	II-236002	0.011	0.018	**-1.920**	**	*IPP1*	0.001	0.002	-1.610		700
							YBR012C	0.001	0.001	1.444		217
	8	VIII-112510	0.011	0.020	**-1.910**	*	*SPO13*	0.001	0.001	-0.382		937
							MIP6	0.002	0.002	-0.404		813
	9	XV-911530	0.015	0.026	**-1.926**	*	*TYE7*	0.001	0.002	-0.521		1140
	10	II-786219	0.004	0.008	**-2.124**	**	*PAU24*	0.003	0.003	-0.178		1817
							ARR1	0.004	0.004	-0.022		1416
	11	XV-100955	0.007	0.011	**-1.879**	*	*ZEO1*	0.004	0.003	1.459		818
							INO4	n/a	n/a	n/a		500
	12	XV-7897	0.005	0.009	**-1.930**	*	*ENB1*	0.002	0.003	-1.844	*	566
							CSS3[d]	0.001	0.001	-0.462		2117
	13	XIV-704237	0.026	0.037	**-1.710**	*	*HOL1*	0.001	0.001	-1.237		583
	14	VII-318823	0.011	0.020	**-1.980**	*	*USE1*	n/a	n/a	n/a		867
							SRM1	0.001	0.001	0.104		1433
	15	XV-610717	0.021	0.040	**-1.920**	*	*HEM15*	0.003	0.003	-0.703		844
							MPC54	0.004	0.005	-0.711		1240
	16[e]	XV-285166	n/a	n/a	n/a		YOL014W	n/a	n/a	n/a		433
							HRD1					788
	17[e]	VIII-442749	n/a	n/a	n/a		*OYE3*	0.001	0.001	-1.112		458
	18	VI-183598	0.022	0.014	2.508[f]	**	YFR006W	0.004	0.003	0.368		241

Ty3	1	VIII-445249	0.006	0.009	**-1.701**	*	*PFS1*	n/a	n/a	n/a	454
							KOG1	0.000	0.000	0.069	201
Ty4	1	VII-642530	0.010	0.015	**-1.791**	*	*NNF2*	0.001	0.001	1.851	1887
							UTP22	0.001	0.001	1.266	832
	2	III-276699	0.006	0.011	**-2.087**	**	*OCA4*	0.003	0.005	-1.530	2583
							HMLα2	n/a	n/a	n/a	1240
Ty5	1	XVI-24300	0.017	0.027	**-1.802**	*	YHL042W	0.005	0.005	0.254	463
							RMD6	0.002	0.002	-0.525	542

Table 3.4: **Significant Tajima's *D* values for insertions and neighbouring genes in *S. paradoxus*.** (Table continued from previous page). LTR numbers refer to those used in the gene table as shorter designations, rather than accession numbers, which are kept consistent throughout this thesis. Six Tajima's *D* tests on genes could not be completed due to lack of polymorphisms in the sequences. [a]Insertions are not annotated in the reference genome of *S. paradoxus*, therefore assigned names based upon chromosomal co-ordinates. [b]Distance in respective strain, unless calculated using the UCSC genome browser. The association between insertion III-276699 and *OCA4* was discounted from this point onwards based on distance. [c]*P* value significance: * $P<0.05$; ** $P<0.01$. All others are not significant. [d]Where genes have not been named or since been named independently of the UCSC genome browser, the corresponding name in *S. cerevisiae* has been used. The current *S. paradoxus* ORF names are: YJR054W (*KCH1*); YOL159C (*CSS3*). [e]Athough these insertions did not gain a *D* value, they were included as candidates as they were conserved in ≥10 strains without polymorphisms. [f]As the LTR showed a signature of strong balancing and not positive selection, its investigation was not taken further.

3.3 Fu and Li's *D* statistic provides further evidence for positive selection acting upon LTRs

Alignments of candidate insertions were also tested with Fu and Li's *D* statistical test (Fu and Li, 1993), which works similarly to Tajima's *D*. A negative value of Fu and Li's *D* statistic indicates an excess of singletons (equivalent to rare alleles for Tajima's *D*) whereas a positive value implies a lack of singletons (intermediate alleles in Tajima's test; Fu and Li, 1993; Ramírez-Soriano *et al.*, 2008). It was used here as a more sensitive method to confirm the results of Tajima's *D*. Tables 3.5 and 3.6 display the results of Fu and Li's *D* statistical test for *S. cerevisiae* and *S. paradoxus*, respectively. The resulting values confirmed candidacy as determined by Tajima's *D* in all but two cases: YBLWsigma3 in *S. cerevisiae* and XV-610717 in *S. paradoxus*, which received insignificantly negative values of *D* statistic.

3.4 Candidates in the genomes of *S. cerevisiae* and *S. paradoxus* are present at varying frequencies

Significantly negative values of Tajima's *D* indicate an excess of rare mutations (Rozas, 2009), which is a possible signature of a genetic sweep due to positive selection (Braverman *et al.*, 1995). It was therefore hypothesised that those candidates that are fixed or at high frequencies within the population may be due to positive selection. Tables 3.7 and 3.8 display the frequencies of candidates in *S. cerevisiae* and *S. paradoxus*, respectively. The presence of each locus in any given strain was ascertained by BLASTing the DNA flanking candidate LTRs ('scored' in Tables 3.7 and 3.8), while the presence of the LTR counted towards the total in those strains where the locus was occupied consistently by the same insertion, that is, possessing the same insertion point and typically identical TSDs across strains. It is interesting to note that in both species, *Ty1/2* TSDs in particular rarely match, which is indicative of LTR-LTR recombination occurring between multiple elements and/or the accumulation of mutations within TSD sequences.

Each species possesses only one fixed candidate - YELWdelta6 in *S. cerevisiae* and VIII-442749 in *S. paradoxus*. Although the *S. cerevisiae* population contains approximately double the number of candidates compared to that of *S. paradoxus*, there is no significant difference between the frequencies of all insertions in each species (P=0.5, unpaired two-tailed t test).

Family	No.	SGD I.D.	*D* statistic	*P* value significance
	1	YERWdelta22	-2.081	*
	2	YMLCdelta2	-1.765	**
	3	YERCdelta16	-2.497	**
	4	YORWdelta17	-2.888	**
	5	YHRCdelta10	-2.036	*
	6	YPLWdelta5	-2.411	*
	7	YELWdelta6	-3.572	**
	8	YDRWdelta11	-3.456	**
	9	YCRCdelta6	-2.894	**
Ty1/2	11	YGRWdelta32	-2.450	*
	12	YLRCdelta9	-2.698	**
	13	YNRCdelta8	-2.515	**
	14	YMLWdelta4	-2.477	**
	15	YPLWdelta3	-2.229	*
	16	YGRWdelta23	-3.138	**
	17	YLRCdelta21	-2.598	**
	18	VII-77305	-2.578	*
	20	YOLCdelta3	-2.995	**
	21	YGRCdelta12	-2.788	*
	2	YDRCsigma3	-3.031	**
	3	XIII-760251	-2.353	**
	4	XII-681209	-2.605	**
	5	YORWsigma3	-2.374	**
	7	V-487854	-2.386	**
	8	YGLCsigma1	-2.138	*
	9	YILWsigma3	-1.713	*
Ty3	10	YHRCsigma2	-2.593	**
	11	VI-212340	-2.458	**
	12	YLRWsigma2	-3.067	**
	13	III-168387	-2.290	**
	15	YBLWsigma1	-1.924	NS
	16	YNLWsigma3	-1.771	**
	17	YERCsigma3	-2.229	*
	18	YILWsigma2	-1.730	*
	1	YERCtau2	-2.323	**
Ty4	2	YMRCtau3	-2.667	**
	4	YMRWtau2	-2.299	**
	1	YGLWomega1	-2.072	**
Ty5	2	YCRWomega3	-3.050	*
	3	YHLComega1	-1.827	*

Table 3.5: ***S. cerevisiae D* statistic values.** Insertions that are absent from the reference genome and are therefore unannotated are given a designation based on chromosomal coordinates. *P<0.05; **P<0.02; NS – not significant.

3.5 Genomic regions containing candidate insertions are shared between species

Candidate insertions are located close to host genes *USE1*, *HOL1* and *MPC54* in both species (Figure 3.7). In *S. paradoxus*, these three candidates all belong to the *Ty1* family, whereas this differs in *S. cerevisiae*, as multiple insertions are located downstream of *USE1*, and the candidate

Family	No.	Designation	*D* statistic	*P* value significance
Ty1/2	1	IV-364275	-2.568	*
	2	XI-472807	-2.622	**
	3	X-513745	-1.820	**
	4	IV-445330	-1.925	*
	5	VII-326627	-2.061	*
	6	XVI-856470	-2.121	*
	7	II-236002	-2.184	**
	8	VIII-112510	-2.280	*
	9	XV-911530	-2.269	**
	10	II-786219	-2.832	**
	11	XV-100955	-2.087	**
	12	XV-7897	-2.269	*
	13	XIV-704237	-1.815	**
	14	VII-318823	-2.672	**
	15	XV-610717	-2.254	NS
	16	XV-285166	n/a	-
	17	VIII-442749	n/a	-
Ty3	1	VIII-445249	-1.859	*
Ty4	1	VII-642530	-1.966	**
	2	III-276699	-2.598	**
Ty5	1	XVI-24300	-2.190	*

Table 3.6: ***S. paradoxus D* statistic values.** Insertions have not been annotated in the reference genome and are therefore given a designation based on chromosomal coordinates. Tests on alignments of two insertions could not be completed due to lack of polymorphisms, as with Tajima's *D*. *P<0.05; **P<0.02; NS – not significant.

is of the *Ty3* family. Although these insertion sites are shared by candidate LTRs in both species, the distances between the genes and the precise insertion points differ between species. While individual insertions are highly unlikely to survive the divergence of host species, 'safe' regions for insertions have evidently been conserved between *S. cerevisiae* and *S. paradoxus*. These particular host genes may also be ideal targets for positively selected insertions to affect their expression.

As insertions in *S. paradoxus* are unannotated in the reference genome browser, regions surrounding the candidates and adjacent genes were searched for evidence of additional insertions in order to add representations of the insertions in the genome browser figures.

3.6 Host genes adjacent to candidate LTRs display varied functions

The gene ontology (GO) functional categories of genes adjacent to candidate LTRs in both species are listed in Appendix I. A minority of candidates are associated with one of the paralogous pairs of genes that arose during WGD (17% in *S. cerevisiae*; 29% in *S. paradoxus*). Regions surrounding

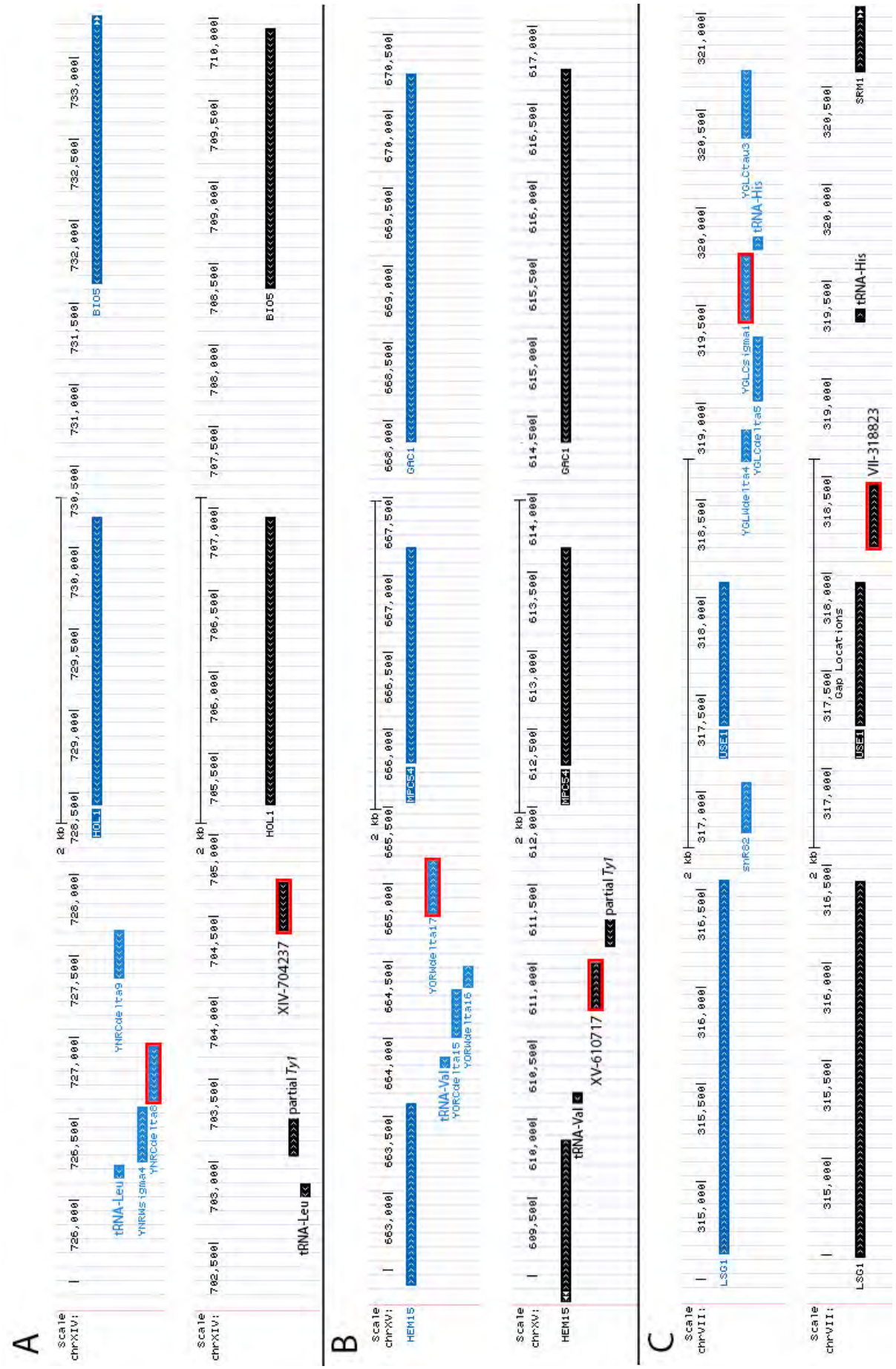

Figure 3.7: **Regions containing candidate insertions shared by both species.** Three regions containing candidates were shared by both *S. cerevisiae* (blue genome browser, top rows) and *S. paradoxus* (black genome browser, bottom rows). A - candidates adjacent to *HOL1*; B - candidates between *HEM15* and *MPC54*; C - candidates adjacent to *USE1*. Candidates are outlined in red. As insertions in the *S. paradoxus* genome browser are unannotated, these were manually added for the purposes of this figure.

Family	No.	I.D.	TSDs 5'	TSDs 3'	Type in reference	Scored	Present	Frequency	
	1	YERWdelta22	AATAG	CAAAT	solo	23	11	0.48	polymorphic
	2	YMLCdelta2	ATAAG	ATAAG	solo	27	7	0.26	polymorphic
	3	YERCdelta16	TTATG	CATTA	solo	17	11	0.65	polymorphic
	4	YORWdelta17	ATAGT	ATAGT	solo	19	14	0.74	polymorphic
	5	YHRCdelta10	AAAAG	AAAAG	solo	21	10	0.48	polymorphic
	6	YPLWdelta5	ATATG	CATAC	3'	17	15	0.88	polymorphic
	7	YELWdelta6	CAAAG	ACTAG	solo	24	24	1.00	fixed
	8	YDRWdelta11	CATAG	TGGTA	solo	20	19	0.95	polymorphic
	9	YCRCdelta6	GAATG	GAATG	solo	26	13	0.50	polymorphic
Ty1/2	11	YGRWdelta32	TCAAT	TGTTG	solo	27	12	0.44	polymorphic
	12	YLRCdelta9	TTAAT	TTAAT	solo	19	14	0.74	polymorphic
	13	YNRCdelta8	AATAA	AATAA	solo	14	11	0.79	polymorphic
	14	YMLWdelta4	AACAA	ATTTT	3'	12	11	0.92	polymorphic
	15	YPLWdelta3	CCAAT	ATTTT	5'	14	11	0.79	polymorphic
	16	YGRWdelta23	AAATT	ATTAT	solo	32	17	0.53	polymorphic
	17	YLRCdelta21	ATAAC	CGAAT	solo	20	11	0.55	polymorphic
	18	VII-77305	TTAAT	TTAAT	n/a	29	19	0.66	polymorphic
	20	YOLCdelta3	GTAGA	GTAGA	solo	24	21	0.88	polymorphic
	21	YGRCdelta12	AATTG	ATGGA	solo	22	19	0.86	polymorphic
	2	YDRCsigma3	AGTAA	AGTAA	solo	31	18	0.58	polymorphic
	3	chrXIII_760251	AATTA	ATTAA	n/a	35	11	0.31	polymorphic
	4	chrXII_681209	GTTTA	TTTAA	n/a	30	11	0.37	polymorphic
	5	YORWsigma3	GTTTT	GTTTT	solo	29	11	0.38	polymorphic
	7	chrV_487854	CACTG	CACTG	n/a	23	10	0.43	polymorphic
	8	YGLCsigma1	AACTT	AACTT	solo	28	10	0.36	polymorphic
	9	YILWsigma3	GTAGT	GTAGC	solo	8	7	0.88	polymorphic
Ty3	10	YHRCsigma2	CATTC	CATTC	solo	25	14	0.56	polymorphic
	11	chrVI_212340	GATAT	GAAAT	n/a	27	12	0.44	polymorphic
	12	YLRWsigma2	AAAAC	AAAAC	solo	31	16	0.52	polymorphic
	13	chrIII_168387	AATAT	AATAT	n/a	28	12	0.43	polymorphic
	15	YBLWsigma1	TTCTC	TTCTC	solo	32	9	0.28	polymorphic
	16	YNLWsigma3	GTAGT	GTATC	solo	8	7	0.88	polymorphic
	17	YERCsigma3	GTTCT	GTCTC	solo	25	12	0.48	polymorphic
	18	YILWsigma2	CACCA	GACCA	3'	30	7	0.23	polymorphic
	1	YERCtau2	GAATC	GAATC	solo	33	10	0.30	polymorphic
Ty4	2	YMRCtau3	ATTTA	ATTTA	solo	31	12	0.39	polymorphic
	4	YMRWtau2	TAAAC	TAAAC	solo	28	10	0.36	polymorphic
	1	YGLWomega1	TTTCA	TCCAA	solo	18	13	0.72	polymorphic
Ty5	2	YCRWomega3	TTTTA	AGGAA	solo	27	25	0.93	polymorphic
	3	YHLComega1	CTTTT	GATAA	solo	20	8	0.40	polymorphic

Table 3.7: **Candidate frequencies in *S. cerevisiae*.** Short read data from the NCBI Trace Archive were used to calculate frequencies of candidates. Insertions in *S. cerevisiae* were named in the UCSC reference genome browser. Those not present in the reference genome were named according to chromosome and co-ordinates. Scored - strains containing the locus; Present - strains containing the occupied locus.

paralogues of adjacent genes were also examined for insertions (Appendix M). Collectively, the associations between paralogue pairs and *Ty* insertions suggest that the 'safe' regions may have been present prior to the WGD event in order for both paralogues to be targeted for adjacent

Family	No.	I.D.	TSDs 5'	TSDs 3'	Type in reference	Scored	Present	Frequency	
	1	IV-364275	ATAAT	ATAAT	solo	36	15	0.42	polymorphic
	2	XI-472807	TTATT	ATCTT	solo	28	15	0.54	polymorphic
	3	X-513745	AATTA	AATTA	solo	13	7	0.54	polymorphic
	4	IV-445330	ATATA	CGGAG	solo	25	8	0.32	polymorphic
	5	VII-326627	AATAT	ATATT	solo	25	9	0.36	polymorphic
	6	XVI-856470	TAATT	ATTCC	solo	28	10	0.36	polymorphic
	7	II-236002	AACTT	TATAC	solo	23	9	0.39	polymorphic
	8	VIII-112510	AAATT	GAATT	solo	29	12	0.41	polymorphic
Ty1/2	9	XV-911530	GAATT	TGGTA	5'	17	10	0.59	polymorphic
	10	II-786219	CATGT	AAAAA	solo	16	15	0.94	polymorphic
	11	XV-100955	AAGAT	AAGAT	solo	10	9	0.90	polymorphic
	12	XV-7897	TTATA	CACAC	solo	15	14	0.93	polymorphic
	13	XIV-704237	AAAAC	AAAAC	solo	32	7	0.22	polymorphic
	14	VII-318823	AATAC	ATACT	solo	34	13	0.38	polymorphic
	15	XV-610717	AAAGA	CGAGG	solo	29	17	0.59	polymorphic
	16	XV-285166	TTTTG	GTAAT	solo	11	10	0.91	polymorphic
	17	VIII-442749	CCGTC	TATTT	solo	11	11	1.00	fixed
Ty3	1	VIII-445249	TAGGA	TAGGA	solo	34	8	0.24	polymorphic
Ty4	1	VII-642530	TAATT	TATTT	solo	17	8	0.47	polymorphic
	2	III-276699	CATCA	CATCA	solo	29	12	0.41	polymorphic
Ty5	1	XVI-24300	CTTTG	TGTTG	solo	25	10	0.40	polymorphic

Table 3.8: **Candidate frequencies in *S. paradoxus*.** Short read data from the NCBI Trace Archive were used to calculate frequencies of candidates. Insertions in *S. paradoxus* have not been designated names as in *S. cerevisiae*, therefore insertions were termed according to chromosome and 5' co-ordinate. Scored - strains containing the locus; Present - strains containing the occupied locus.

insertion.

Around 89% of all *Ty1-4* insertions in the reference genome of *S. cerevisiae* are associated with tRNAs (Hani and Feldmann, 1998; Kim *et al.*, 1998). This close relationship between tRNAs and *Ty* insertions is also displayed by the majority of candidates in *S. cerevisiae* (78%; Appendix J). Fewer candidates are however associated with tRNAs in *S. paradoxus* (55%; Appendix J), suggesting the relationship in this species differs to that in *S. cerevisiae*.

3.6.1 Adjacent genes in *S. cerevisiae*

Figure 3.8 displays the breakdown of adjacent gene categories in *S. cerevisiae* as determined by the Panther gene classification system (Mi *et al.*, 2017). Details for adjacent genes collected are located Appendix K.

There is no significant enrichment for any gene ontology category (P=0.1-1.0; two-tailed t-tests). Candidates are most commonly associated with genes that encode enzymes (n=15), and

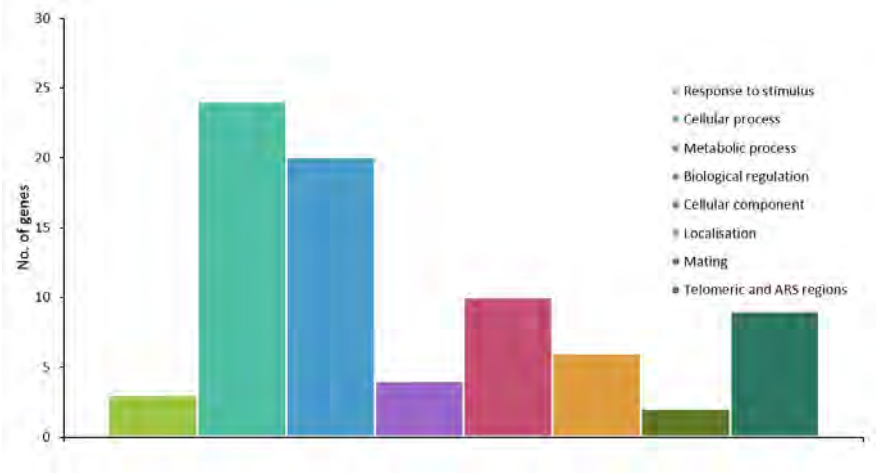

Figure 3.8: **Bar chart of GO major categories of genes adjacent to candidate LTRs in *S. cerevisiae* classified by Panther.** Gene products with multiple functions are counted multiple times by the software. No significant difference from random expectation was observed.

related to transportation involving the Golgi body or functional components of the complex (*n*=5). Additionally, two adjacent genes encode components of the meiotic outer plaque. An adjacent gene of particular interest is the transcription factor *GCN4*, located downstream of the insertion YELWdelta6. Unfortunately, due to this insertion being fixed in the SGRP strains (Table 3.7), any potential effects this may have on gene expression cannot be ascertained with qPCR (Section 3.7). An expanded search of strains indicates that this insertion and flanking DNA is fixed in a subset of currently sequenced *S. cerevisiae* genomes (*n*=151), yet the remaining strains lack this particular locus adjacent to *GCN4*. The Gcn4 gene product has multiple binding sites within *Ty1* LTRs (Servant *et al.*, 2008). The optimal binding site for Gcn4 is typically seven bases in length but minor variability will still result in transcription and expression of the target gene (Arndt and Fink, 1986; Hill *et al.*, 1986; Oliphant *et al.*, 1989; Morillon *et al.*, 2002). A survey of all unique *Ty1* LTRs showed that none possess all five Gcn4 binding sites, and very few possess four (*n*=9; 0.03%), one of which being the candidate YELWdelta6. The majority of insertions (*n*=172; 56%) possess one or no binding sites.

Figure 3.9 displays the locations of candidates and adjacent genes in the genome of *S. cerevisiae*. The average distance between an insertion and adjacent gene is 656bp (Appendix K).

3.6.2 Adjacent genes in *S. paradoxus*

Figure 3.10 displays the breakdown of adjacent gene categories in *S. paradoxus* as determined by the Panther classification system (Mi *et al.*, 2017). Details for adjacent genes collected are in

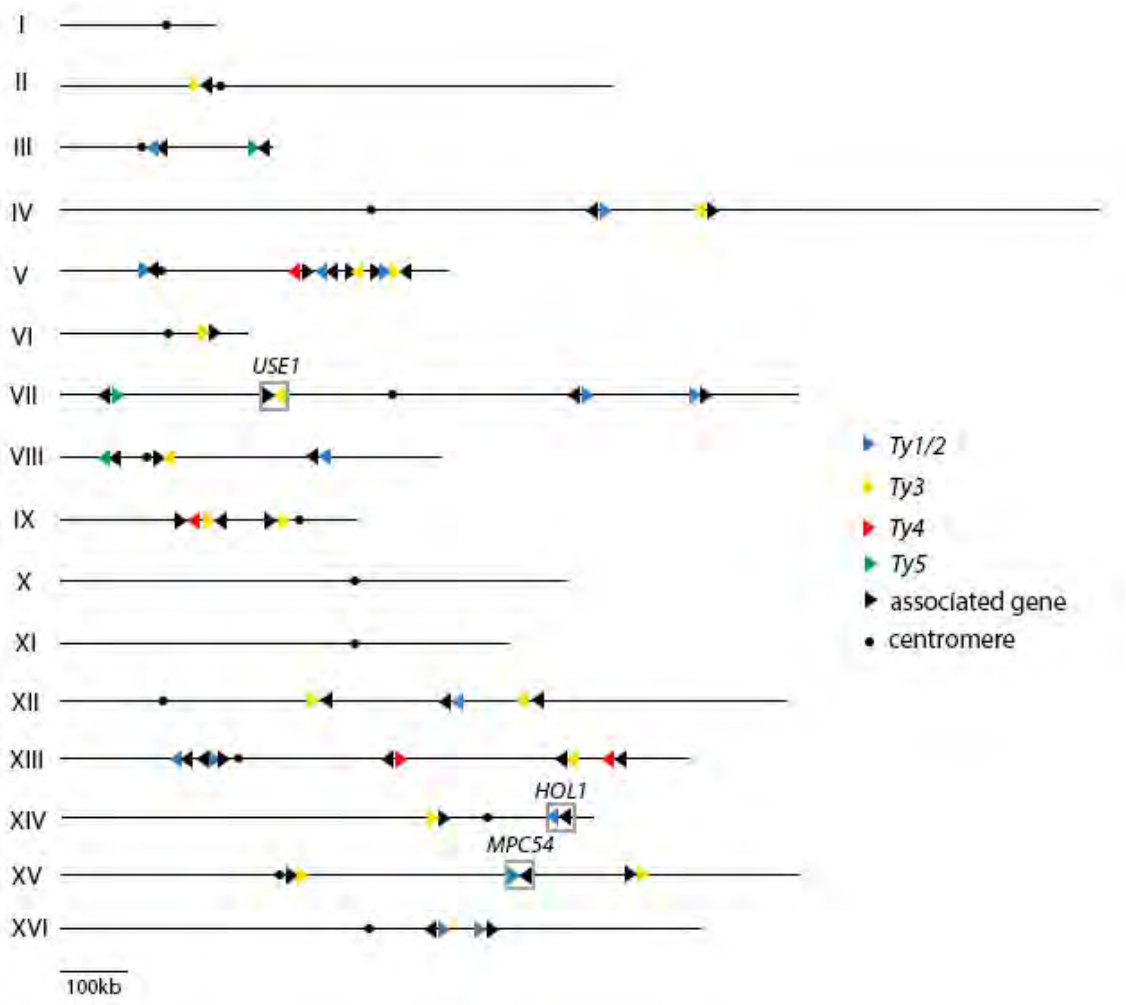

Figure 3.9: **Mapping of candidate insertions and adjacent genes of *S. cerevisiae*.** Chromosome length and centromere data were obtained from SGD. Those regions containing candidates shared with *S. paradoxus* are boxed and named. Candidate families are colour coded, with associated genes in black.

Appendix L.

Adjacent gene functions are highly diverse, with no significant enrichment for any gene ontology category (P=0.2-1.0; two-tailed t-test). In *S. paradoxus*, five associated genes encode enzymes, and as in *S. cerevisiae*, candidates are located close to both Golgi body (n=1) and meiotic outer plaque related genes (n=1). In addition, *CSS3*, a gene currently of unknown function, may suppress *Ty1* transposition, as activity is greatly increased when this gene is deleted (Yofe *et al.*, 2016). Further adjacent genes of interest include transcription factors *ARR1*, *INO4*, *TOS8* and *TYE7*. Downstream of *TYE7*, the polymorphic candidate LTR is associated with a partial *Ty1* element, truncated within the *gag* coding region. Similarly to Gcn4 in *S. cerevisiae*, Tye7 binds to sites within *Ty1 gag*, and is involved in *Ty1* transcription (Servant *et al.*, 2012; see Discussion).

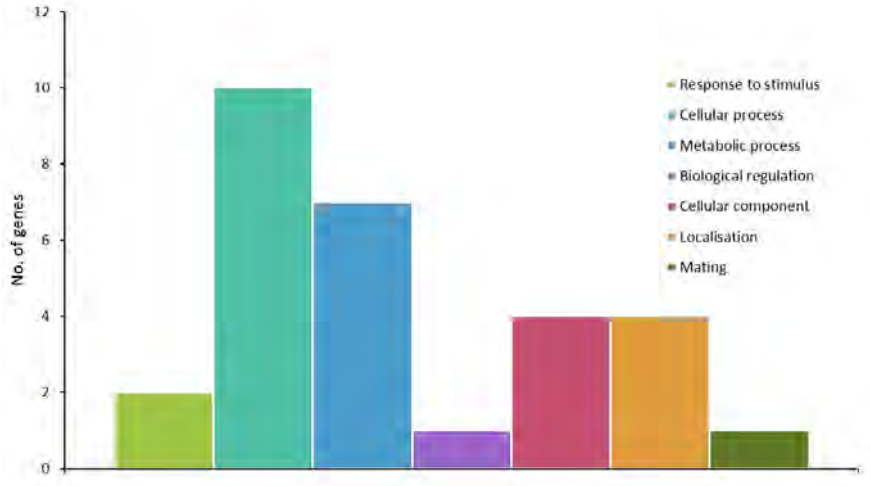

Figure 3.10: **Bar chart of GO major categories of genes adjacent to candidate LTRs in *S. paradoxus* classified by Panther.** Gene products with multiple functions are counted multiple times by the software. Telomeric regions and ARSs are unannotated in the *S. paradoxus* genome browser and therefore any associations with such features were unable to be determined. No significant difference from random expectation was observed.

Figure 3.11 displays the locations of candidates and adjacent genes in the genome of *S. paradoxus*. The average distance between insertions and host genes is 816bp (Appendix L).

Due to the lack of annotations of genomic features such as *Ty* insertions and ARSs in the *S. paradoxus* UCSC genome browser, associations between candidates and features other than host genes and tRNAs could not be determined. ARSs in *S. cerevisiae* are commonly associated with candidates (n=8), a link which cannot be determined at this stage in *S. paradoxus* due to the lack of annotations. The typical consensus region present in ARSs is just 17bp, while the remainder of the sequences can be highly variable (van Houten and Newlon, 1990). Searches of *S. paradoxus* genomes with *S. cerevisiae* query sequences results in potential ARSs sharing <84% identity. Due to this uncertainty, the decision was made only to consider annotated genes as potentially hitchhiking along with candidate LTRs.

3.7 Assessing expression of neighbouring genes with qPCR

S. cerevisiae SGRP strains (n=36) were purchased as a culture plate from NCYC. Strain DBVPG6765 failed to thrive and was therefore discounted from the expression studies. Candidate LTR presence was determined in each of these strains during data collection (Appendix G). No source- or origin-specific pattern of candidate presence was observed. Candidate-gene pairings were chosen from those LTRs with the most significantly negative Tajima's *D* values for expression

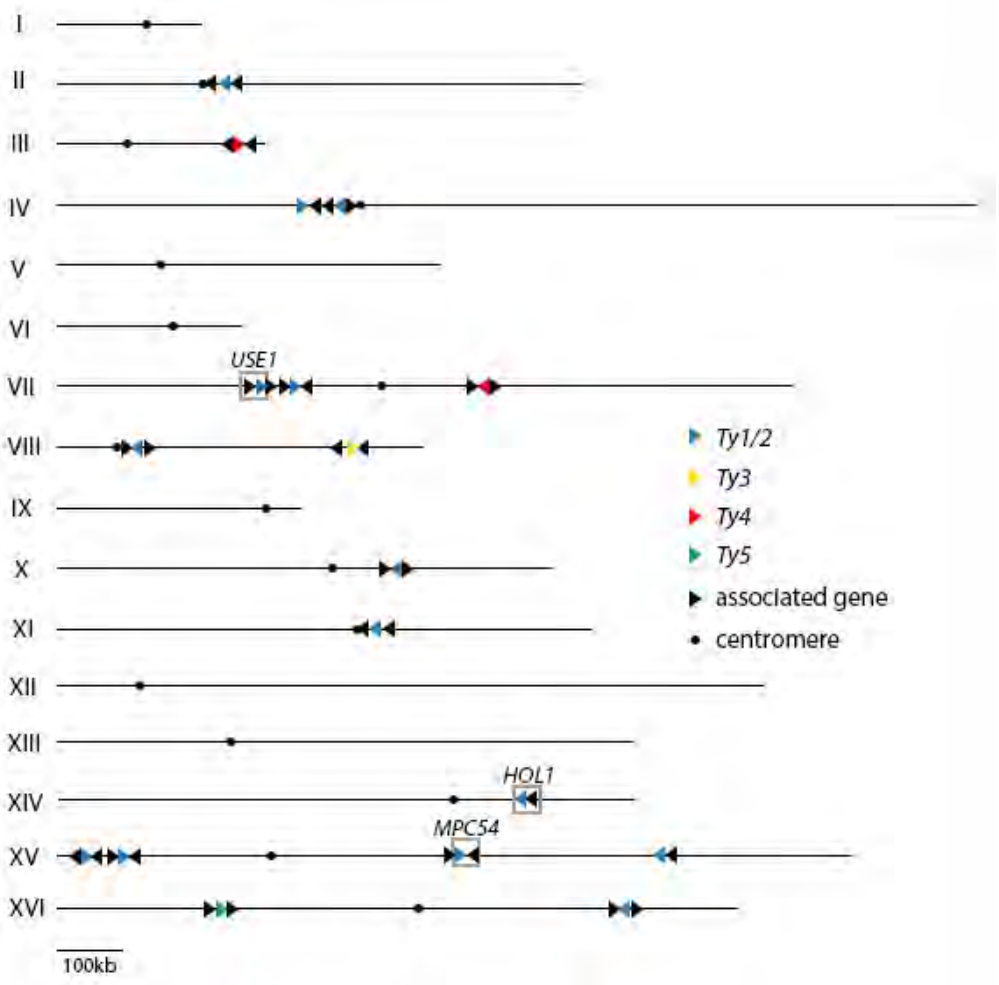

Figure 3.11: **Mapping of candidate insertions and adjacent genes of *S. paradoxus*.** Those regions containing candidate insertions shared with *S. cerevisiae* are boxed and named. Chromosome length data from Liti *et al.* (2009). Centromeric data from Varoquaux *et al.* (2015). Candidate families are colour coded, with associated genes in black.

analysis. Pairings range across four of the five families, with *Ty5* being unrepresented. Excluded for qPCR were ARS and telomeric sequences, or mating loci whose expression would be difficult to determine. Total RNA was extracted from each strain (Appendix D) and cDNA synthesised. cDNA was diluted to a working solution of 1:25 from an initial stock of 500ng/µl. Primers were designed for adjacent genes, ensuring that paralogues or other sequences would not be amplified (Appendix E). Two housekeeping genes, *TAF10* and *UCB6*, were selected due to their consistent expression levels (Teste *et al.*, 2009).

Twelve adjacent genes of interest were amplified and expression levels determined in 35 SGRP strains using qPCR. This method was selected as it was accessible and reliable on a small scale study such as this. After each qPCR reaction, PCR products were electrophoresed to check for the presence of a single band indicating a single amplified PCR product. Expression was determined

by amplification of cDNA which directly correlates with the initial quantity of mRNA extracted from the strains. Incorporating fluorescent-tagged nucleotides into the amplification process allowed the quantity of PCR products to be measured in relative fluorescence units (RFU). BioRad software was used to exclude any regions of primer dimer in the expression graph. Expression values were normalised to the two selected housekeeping genes (M/stability value = <1.0) in QBase (Biogazelle), and then across each triplicated sample. The mean and standard deviation were calculated for each strain and then separated into two groups: expression in strains containing the insertion adjacent to the gene, and those strains lacking the insertion. Box and whisker plots were generated using GraphPad (Figure 3.12). Unpaired, two-tailed t-tests were then performed on the two series for each of the adjacent genes in GraphPad (Table 3.9). Expression for one strain (YJM975; plate position B5) was inconsistent across triplicates despite repeated qPCR reactions and so was entirely excluded from analysis here. This was not caused by positional effect as the strains were split between two PCR plates (Appendix F) and inconsistency was not observed on the second plate.

Gene	No. of strains			Mean expression (RFU)		Difference between means	*P* value
	Present	Absent	Excluded	Present	Absent		
ADY4	10	22	3	0.768±0.103	1.023±0.236	-0.255±0.360	0.483
AMD2	12	22	1	1.097±0.218	0.944±0.099	0.153±0.209	0.470
AVL9	8	26	1	1.762±0.284	0.963±0.177	0.800±0.356	0.032*
CAT2	5	29	1	1.634±0.582	1.519±0.232	0.114±0.616	0.854
HOL1	4	30	1	1.214±0.226	1.037±0.172	0.177±0.318	0.581
MPC54	7	27	1	1.000±0.215	1.108±0.139	-0.108±0.295	0.716
RRN11	5	29	1	1.120±0.220	1.046±0.126	0.074±0.319	0.819
SCS7	6	28	1	2.030±0.402	0.953±0.147	1.078±0.362	0.005**
SEC7	9	25	1	0.930±0.194	1.230±0.213	-0.296±0.369	0.429
UBP9	4	30	1	1.238±0.242	1.745±0.293	-0.507±0.426	0.243
YCS4	7	27	1	1.084±0.231	1.169±0.172	-0.085±0.360	0.816
YER134C	6	28	1	1.172±0.286	1.199±0.213	-0.027±0.483	0.956

Table 3.9: **Comparison of expression levels of genes of interest with and without loci occupied by candidate LTRs.** Expression was measured in RFU and averaged over three qPCR reactions in each strain, ±SE. Results from strain YJM975 were excluded from all analyses due to inconsistent expression across triplicates. Strains UWOPS05.227-2 and 273614N failed to provide results for expression of *ADY4* despite multiple repetitions of qPCR.

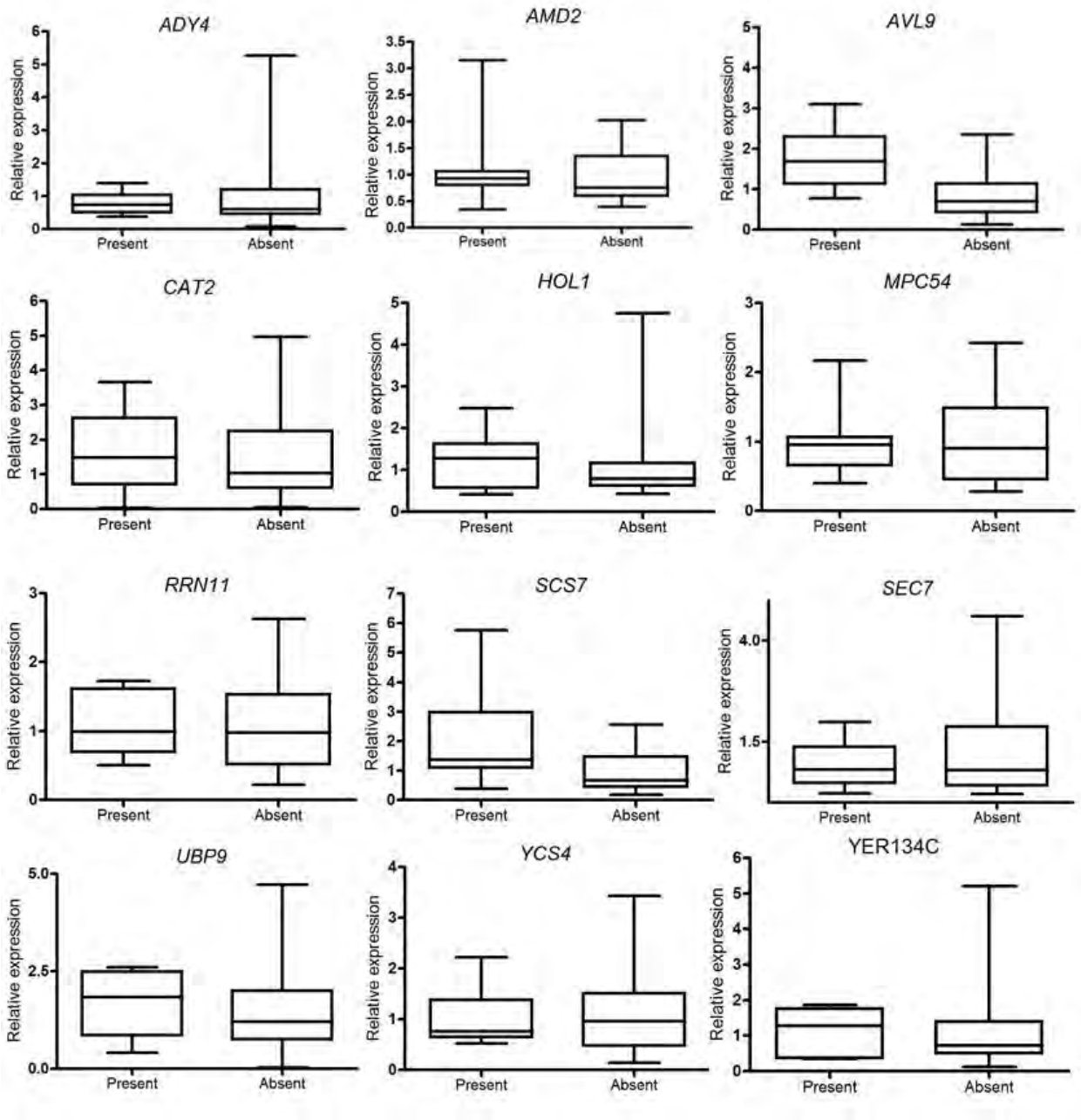

Figure 3.12: **Box and whisker plots for expression of genes adjacent to candidate LTRs in *S. cerevisiae*.** Expression is relative to the housekeeping genes *TAF10* and *UCB6*.

3.7.1 Expression of *AVL9* and *SCS7* are significantly higher in strains possessing adjacent candidate LTRs

Expression of two host genes, *AVL9* and *SCS7*, is significantly higher in strains possessing candidates YLRWsigma2, a *Ty3* solo LTR, and YMRCtau3, a solo *Ty4* LTR, respectively (Table 3.9). *AVL9*, whose average expression increases by >70% in strains with the occupied locus, encodes an exocytic transport protein in the Golgi body (Harsay and Schekman, 2007). Although a non-essential gene, null mutants experience heat sensitivity, decreased fitness and a shortened lifespan (Tkach *et al.*, 2012). Average *SCS7* expression increases by 100% in those strains possessing the candidate. This gene encodes a hydroxylase enzyme which functions in the alpha-hydroxylation of fatty acids and sterols in the production of sphingolipids (Dunn *et al.*, 1998). The predominate sphingolipid in yeast, inositol-phosphorylceramide, may be implicated in the signalling pathways of heat stress response (Jenkins *et al.*, 1997; Chung *et al.*, 2000; Ferguson-Yankey *et al.*, 2002).

The relative orientation and positioning of candidates and adjacent genes is displayed in the genome browser view of Figure 3.13.

Figure 3.13: **Genomic regions containing candidates which may significantly increase expression of adjacent genes.** A - Region of chromosome XII containing candidate YLRWsigma2 and adjacent gene *AVL9*; B - Region of chromosome XIII containing candidate YMRCtau3 and adjacent gene *SCS7*.

3.8 Discussion

In the work here, the SGRP genomes of *S. cerevisiae* and *S. paradoxus* were screened for *Ty* insertions that may be evolving under positive selection. Conferring a benefit to their host may be a way in which insertions persist in the genome. As insertions are known to be able to affect host gene expression (reviewed in Elbarbary *et al.*, 2016), the most significant of these candidates

in *S. cerevisiae* were investigated in the laboratory. Expression levels of adjacent genes were determined in strains both possessing and lacking the insertions.

Recombination events can break the linkage between LTRs and host genes (Barton, 2000). Although the difference in numbers of candidates in *S. cerevisiae* and *S. paradoxus* is not significant, it was hypothesised that fewer genes would be found hitchhiking along with LTRs in *S. cerevisiae* strains due to it possessing one of the highest genomic recombination rates (Esberg *et al.*, 2011). *S. paradoxus* possesses a far lower rate of recombination (Tsai *et al.*, 2008), yet possesses fewer candidates and hitchhiking host genes. However, this observation may simply be the result of the higher *Ty* copy number in *S. cerevisiae* strains in comparison to those of *S. paradoxus*.

The number of candidate LTRs identified here is likely to be an underestimate for a number of reasons. Firstly, many insertions were excluded: only LTRs with intact boundaries were considered during the screening process, as this would avoid the estimation of truncated termini and therefore reduce 'noise' in Tajima's *D* alignments. Full-length LTRs may be more likely to possess the regulatory regions necessary for interaction with host proteins and transcription factors, but it is possible that those truncated LTRs excluded from tests have remained in the genome due to their possessing those regulatory regions. Nested insertions were generally excluded due to variability across strains and the tendency for the signature of selection to be unclear (but see Section 3.1.3). Variable LTRs (i.e. those whose state differed between FLE and solo in strains) were also not considered as candidates, as the presence of a FLE as opposed to a solo LTR can greatly disrupt the sequences around it (e.g. Knight *et al.*, 1996). Secondly, due to the confines of the *D* tests, such as requiring a minimum of four sequences, candidates could be present in three strains or less, or even be confined to providing a benefit for a single strain and have yet to spread to the population. The complexity of the relationship between an insertion and its host means that many additional scenarios in which an insertion provides a benefit or contributes to its host's fitness are possible. Thirdly, quality of the *S. cerevisiae* SGRP/trace archive strains is highly varied, which likely impacted data collection. Around 38 are relatively reliable, whereas reads from the remaining strains are intermittently present in the NCBI Trace Archive. Finally, the SGRP assemblies of both *S. cerevisiae* and *S. paradoxus* are unreliable, filled with gaps and undetermined bases, often in the same region throughout strains of each species, so the screening performed here for candidate insertions is very much a conservative estimate.

Tajima's *D*

Tajima's *D* test is being increasingly used to identify selection signatures of TEs in a variety of genera such as *Drosophila* (Schlenke and Begun, 2004; Macpherson *et al.*, 2008; Kofler *et al.*, 2012; Ullastres *et al.*, 2015; Merenciano *et al.*, 2016), mosquito (Subramanian *et al.*, 2007) and arboreal lizard *Anolis carolinensis* (Ruggiero *et al.*, 2017). The work performed here examined the signatures of selection acting upon LTRs and their adjacent genes in order to find potential candidates for positive selection in two species of *Saccharomyces*. Previous investigations into the LTRs within 11 intergenic regions in the *S. paradoxus* genome provided no significant results (Tsai *et al.*, 2008), in contrast to the results obtained here. In both *S. cerevisiae* and *S. paradoxus*, 3% of unique insertions display evidence of positive selection, which is far higher than the 13 candidates of >10,000 unique insertions (0.1%) identified by Kofler *et al.* (2012) in *Drosophila*. Kofler *et al.* (2012) used sliding windows of Tajima's *D* across the genomes of a *Drosophila* population, rather than the method employed here which identified genes that may be hitchhiking along with positively selected LTRs. Kofler *et al.* (2012) classified insertions as candidates if the strong signature of selection, identified by a significantly negative value of *D*, was within 500bp up- or downstream of the insertion, as the strongest signal of selection may not always be directly at the site of positive selection (Kim and Stephan, 2002). However, the poor quality of the SGRP assemblies and short reads available in the NCBI Trace Archive prevented similar analysis of surrounding regions in the genomes of *S. cerevisiae* and *S. paradoxus*.

Among the candidates for positive selection, Kofler *et al.* (2012) confirmed two previously identified insertions involved in adaptation. In the work here on *Saccharomyces*, two genes were previously reported as being under positive selection (Scannell *et al.*, 2011) are adjacent to candidate insertions in this study: *AMD2* and YJR056C. The former possesses a negative value of *D*, consistent with positive selection, while the latter returned a positive value of *D*. Datasets of positively selected genes generated by Li *et al.* (2009b), Zhou *et al.* (2010), Sawyer and Malik (2006), Zhang *et al.* (2009), Artieri and Fraser (2014) and Fay *et al.* (2004) did not overlap with this study. Each investigation did however, use varying techniques to identify positive selection, rather than Tajima's *D*, and were not looking for positive selection in relation to *Ty* insertions which may account for discrepancies between studies.

Unsurprisingly *Ty5* in *S. cerevisiae* only proved to have three candidates, and with *D* values of relatively low significance. As this family has not been active recently, it was unlikely to be a

target for recent positive selection. It cannot be discounted that this family has in the past been evolving under positive selection, however, as the tests used here lose power when analysing older insertions. The low number of *Ty5* candidates contrasts with the higher numbers of candidates recorded in the active *Ty1/2* superfamily and *Ty3* family. However, signatures of selection may also be difficult to detect if the candidate became selected for very recently (Garud *et al.*, 2015).

Functions of genes adjacent to candidate LTRs

No patterns in function or significant enrichment of GO categories were observed in the adjacent genes of candidate LTRs. There are, however, high numbers of enzymes, likely due to the fact that this class of gene is relatively common. Other than genes of unknown function, *S. paradoxus* possesses more adjacent genes with a function involved in response to stimuli than *S. cerevisiae*, which would be expected in a wild organism with varying environments. Both species share a similar number of adjacent genes involved with mating, metabolic processes and biological regulation and also encoding cellular components. *S. paradoxus* also possesses more genes with functions involved in cellular processes and localisation/transport, also indicative of its adaptation to varying wild environments.

Although no GO categories are significantly enriched, a number of similarities regarding neighbouring genes were observed between the species, such as those involved in chemical resistance. In *S. cerevisiae*, candidate YNLWsigma3 on chrXIV is upstream of *BOP3*, overexpression of which confers organic mercury resistance (Hwang *et al.*, 2005). In *S. paradoxus*, a candidate insertion is found adjacent to *ARR1*, a transcription factor for genes involved in arsenic resistance. If candidate LTRs increased expression of these genes, this may increase the fitness of the host when exposed to arsenic/mercury and therefore increase the likelihood that the insertion would become fixed in the population. However, this would depend on the environment, as an absence of the metals may cause a loss of fitness due to the metabolic cost of constant upregulation of these genes without benefit. Further connections between the adjacent genes in both species were observed: multiple genes related to the Golgi body are also closely situated to candidate LTRs in both *S. cerevisiae* (n=5) and *S. paradoxus* (n=2). Each species also possesses a *Ty5* candidate within 500bp of ARSs present in telomeric regions. >80% of all *Ty5* insertions are present in these regions (Zou *et al.*, 1996*a*), and increase the stability of telomeres in both *S. cerevisiae* and *S. paradoxus* (Zou *et al.*, 1995, 1996*a*). Therefore, the presence of insertions in these areas are

likely to benefit the host while also providing a 'safe' region for *Ty* insertions. Furthermore, candidates may not confer observable benefits to the host, but simply transpose safely into certain regions. For example, candidates in *S. cerevisiae* (*n*=8) are within close proximity to ARSs, which are essential in chromosome maintenance (Nieduszynski *et al.*, 2006). As replication is initiated at multiple chromosomal sites at intervals of 40-100kb (Nieduszynski *et al.*, 2006; Siow *et al.*, 2012), insertion close to these sequences may increase expression of the ARS and the chances of the insertion being replicated in the process. However, assessing the effects of insertions on ARSs is difficult, and unfortunately not something achievable with qPCR.

Candidate *Ty* insertions and transcription factors

Multiple candidates are adjacent to genes encoding transcription factors: *GCN4* and *RRN11* in *S. cerevisiae*, and *TOS8*, *TYE7*, *ARR1* (discussed above) and *INO4* in *S. paradoxus*. *RRN11* encodes a component of the core factor rDNA transcription factor complex required for ribosomal gene transcription (Lalo *et al.*, 1996); *TOS8* is expressed during meiosis and under cell-damaging conditions where it binds to chromatin (Jelinsky *et al.*, 2000; Rabitsch *et al.*, 2001); *INO4* is involved in the activation of phospholipid, sterol and fatty acid biosynthetic enzymes (Bachhawat *et al.*, 1995; Santiago and Mamoun, 2003). Altered expression of these transcription factors may increase host fitness, particularly when under stress in the case of *TOS8*. *GCN4* and *TYE7* are of particular interest as their gene products bind to sites within *Ty1* elements (Figure 3.14).

The product of *GCN4*, adjacent to candidate YELWdelta6, regulates the activation of multiple biosynthesis pathway genes during amino acid, purine and glucose starvation. It is part of the basic leucine zipper (bZIP) family and is tightly regulated, being quickly degraded unless under starvation conditions (Kornitzer *et al.*, 1994; Natarajan *et al.*, 2001). Gcn4 is one of multiple transcription factors that also bind to promoter regions within *Ty1* LTRs and form complexes in order to initiate transcription (Figure 3.14; Servant *et al.*, 2008). Up to five Gcn4 binding sites are present which cause variation in transcription levels; elements which are weakly expressed typically possess fewer than five sites (Morillon *et al.*, 2002; Servant *et al.*, 2008). The majority of *Ty1* LTRs in *S. cerevisiae* may therefore be expressed at a lower level (Section 3.6.1). The insertion at this particular locus has since been driven to fixation in more than 150 strains, as the remaining sequenced genomes lack the 200bp flanking DNA as well as the insertion itself. The insertion may benefit both the *Ty* family and host if it results in the upregulation of *GCN4* expression or positive

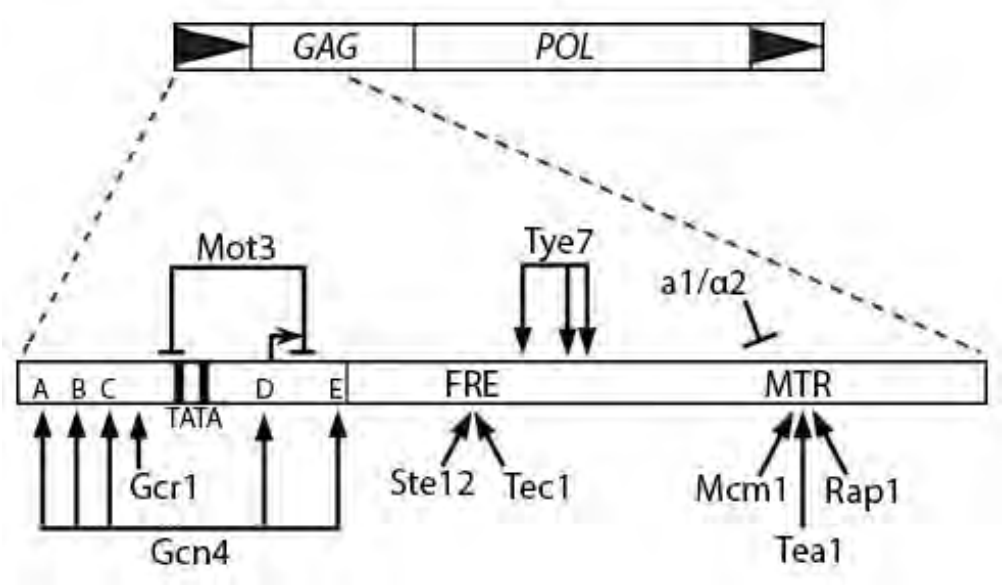

Figure 3.14: **Binding sites within the LTR and *gag* region of *Ty1* elements.** A *Ty1* LTR contains up to five Gcn4 binding sites (denoted A-E) depending on how strongly the element is expressed (Morillon *et al.*, 2002). FRE - filamentous response element, consisting of Ste12 and Tec1 binding sites. Transcription repression sites include Mot3 and a1/α2. Adapted from Servant *et al.* (2008, 2012) and Curcio *et al.* (2015).

feedback loop, perhaps increasing the host's chance of survival during starvation. Consequently, other *Ty1* elements may in turn also experience an increase in expression/transcription.

Furthermore, *TYE7* encodes a transcription factor which is activated during adenine depletion and may aid in the control of *Ty1* antisense RNA synthesis, thus increasing *Ty1* activity (Servant *et al.*, 2012). In *S. paradoxus*, *Ty1* candidate XV-911530 is present downstream of this gene and possesses the most 5' binding site of three within its truncated *gag* (Figure 3.14; Servant *et al.*, 2012). As with Gcn4 in *S. cerevisiae*, if the candidate insertion causes an increase in *TYE7* expression, this in turn may allow further *Ty1* activity due a lower rate of antisense RNA synthesis.

LTRs displaying signatures of positive selection may influence adjacent gene expression in *S. cerevisiae*

Previous work by Feng *et al.* (2013) on *Schizosaccharomyces pombe* determined relative expression in strains with and without loci occupied by insertions adjacent to heat shock response genes. Around 40% of insertions in *Sz. pombe* may increase expression of adjacent genes by providing increased enhancer activity (Sehgal *et al.*, 2007; Leem *et al.*, 2008; Feng *et al.*, 2013). In the work performed here, significant increases in expression levels of two host genes, *AVL9* and *SCS7*, were observed in strains possessing solo LTRs at adjacent loci (Section 3.7.1). *AVL9* encodes a Golgi transport protein (Harsay and Schekman, 2007); the enzyme encoded by *SCS7* may be involved

in the cellular heat shock response (Jenkins *et al.*, 1997; Chung *et al.*, 2000; Ferguson-Yankey *et al.*, 2002). While expression of both *AVL9* and *SCS7* was increased in strains containing the candidates, overexpression of these genes results in decreased cellular growth (Yoshikawa *et al.*, 2011) and chemical resistance (Parsons *et al.*, 2004), respectively. The increased level of expression that may be due to the presence of the candidate insertions is therefore unlikely to be at these problematic levels. Unlike in the study by Feng *et al.* (2013), due to the positioning of candidates relative to adjacent genes (Figure 3.13), the exact method by which candidates promote gene expression is unclear (discussed in the next section).

In the remaining ten qPCR LTR-gene pairings, results failed to show a significant difference in the expression of genes regardless of the presence or absence of an adjacent LTR insertion. There are a number of reasons as to why candidate LTRs had no observable effects on the expression of adjacent genes in the experimental conditions. Insertions may instead act over larger distances on other areas of the genome, rather than immediately adjacent. Effective distance for endogenous enhancers is thought to be a maximum of 2.2kb for RNA pol I genes (Elion and Warner, 1984; Johnson and Warner, 1989) and 1.2kb for RNA pol II genes (Brand *et al.*, 1987; Barberis *et al.*, 1995; Escher *et al.*, 2000). It is unclear as to how far the enhancer region within an LTR can function. Servant *et al.* (2008) reported a maximum distance of 300bp, whereas Petrascheck *et al.* (2005) were able to demonstrate an effective distance of 3kb via DNA looping. The looping mechanism has been detected in yeast between the promoters and terminators of long genes (e.g. O'Sullivan *et al.*, 2004; Ansari and Hampsey, 2005) which does not preclude the possibility that LTRs are acting on other regions, rather than immediately up- or downstream of their insertion point. Additionally, insertions may be part a cumulative effect, e.g. only provides a benefit in conjunction with other loci. Insertions may also be acting on non-coding RNAs rather than coding genes. Furthermore, effects of an insertion may not be directly observable on expression, particularly in laboratory conditions. There have been many reports of stress as an activator of TE insertions in a variety of species (e.g. Kashkush *et al.*, 2003; Grandbastien *et al.*, 2005; Feng *et al.*, 2013; Cavrak *et al.*, 2014; Finatto *et al.*, 2015; Matsunaga *et al.*, 2015; Merenciano *et al.*, 2016 and reviewed in Miousse *et al.*, 2015). Various stress conditions have also been reported to induce *Ty* activity in *Saccharomyces*: ethanol (Stanley *et al.*, 2010), adenine starvation (Todeschini *et al.*, 2005; Servant *et al.*, 2008), cryopreservation (Stamenova *et al.*, 2008) and suboptimal temperatures (Paquin and Williamson, 1986), heat shock and nutrient deprivation (Dai *et al.*, 2007), oxygen (Stoycheva

et al., 2010; VanHoute and Maxwell, 2014), DNA damage (Bradshaw and McEntee, 1989), nitrogen starvation (Ribeiro-dos Santos *et al.*, 1997; Morillon *et al.*, 2000), UV exposure (Rolfe *et al.*, 1986), radiation (Sacerdot *et al.*, 2005), mutagens and carcinogens (Pesheva *et al.*, 2005, 2008; Stoycheva *et al.*, 2007). Therefore, in a laboratory where cultures are provided with consistent temperatures and a steady supply of nutrients required for growth, activity of *Ty* elements may be lower than levels observed in stressful environments. It must also be noted however, that due to the inability to use a form of statistical correction as exact *P* values are not provided by the DnaSP software and to avoid the eradication of LTRs under relatively weak selection, the possibility of falsely positive candidates cannot be discounted.

Expression of adjacent genes may be dependent on relative LTR position and orientation

Results of early studies suggested that the only function of *Ty* LTRs was to provide an efficient insertion site during transposition (Yu and Elder, 1989), and that all transcription regulatory regions resided in *TYA* (Fulton *et al.*, 1988; Yu and Elder, 1989). Since this time, effects on transcription of host genes by solo LTRs have been observed (e.g. Long *et al.*, 1998; Dudley *et al.*, 1999; Medstrand *et al.*, 2001; Landry *et al.*, 2002; Sehgal *et al.*, 2007; Leem *et al.*, 2008; Servant *et al.*, 2008), proving that these sequences do contain regulatory regions comparable to those of host genes (Figure 3.15). TEs can cause a variety of changes upon transposition into, or close to, regulatory regions of host genes (Williamson, 1983; Figure 3.16).

Knight *et al.* (1996) found that in separating a gene from its promoter, a *Ty2* element eliminated expression of a copper transport gene, *CTR3*, altogether (Figure 3.16 B). Similarly, solo LTRs failed to significantly affect expression while in a 5' position tandem to adjacent genes in a number of studies (Williamson *et al.*, 1983; Liao *et al.*, 1987; Fulton *et al.*, 1988; Yu and Elder, 1989). However, Roelants *et al.* (1997) previously showed that *Ty1* LTRs may entirely replace *URA2*'s promoter region, helping to trigger transcription. LTRs have also been adopted as promoters and untranslatable regions in other species, particularly mammals (Jordan *et al.*, 2003; Romanish *et al.*, 2007; Göke and Ng, 2016). Company and Errede (1987) noted further positional flexibility of LTRs in their ability to affect expression of host genes, with greatest effects seen when the *Ty* LTR and gene of interest assumed the divergent position (Figure 3.16 C). This may account for the insignificant differences in expression of host genes where those candidate LTRs were not present

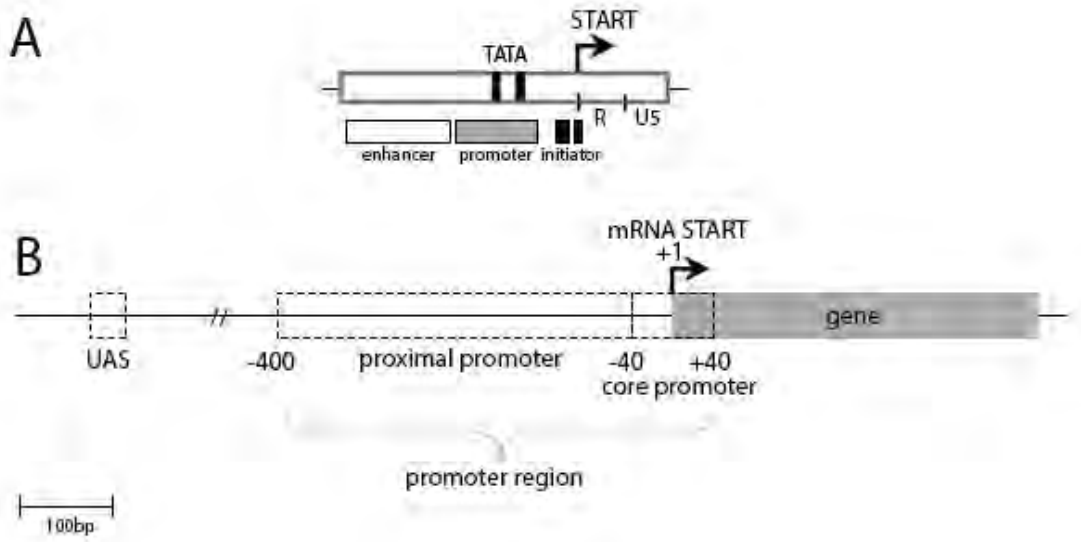

Figure 3.15: **Simplified overview of the regulatory regions required for transcription.** *Ty1* LTRs (A) and typical yeast genes (B) contain similarities in transcription regulatory regions. LTR U3 regions contain the promoter, initiator and enhancer/upstream activator sequence (UAS) (Curcio *et al.*, 2015). Host genes typically possess core and proximal promoters, which contain multiple binding sites for transcription factors, TATA box(es) and initiation regions all within ∼400bp upstream of the transcription start site (TSS; +1; Goffeau *et al.*, 1996; Pelechano *et al.*, 2006). The UAS region is typically within 1.5kb upstream of the TSS, but can be up to 3kb. Both parts of the figure are drawn to scale. UAS - upstream activator sequence.

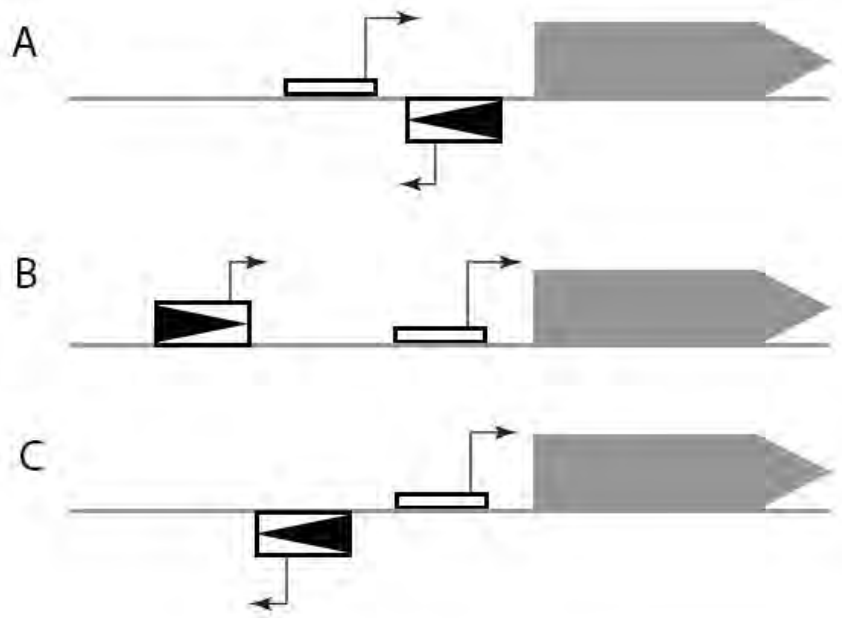

Figure 3.16: **Solo LTR insertion points relative to gene promoters.** The three main positions for an insertion to be present within close proximity to a gene, in relation to its promoter region, in order to affect expression. Additionally, insertions may replace regulatory regions altogether. A - convergent; B - tandem; C - divergent. Grey boxes – genes; boxed arrows – LTRs; black boxes – promoter regions. Small arrows represent initiation codons and the subsequent direction of transcription.

in the divergent position relative to the gene. Furthermore, most candidates are unlikely to have disrupted or replaced a host gene's regulatory region, however, as they are not positioned within

the promoter region, or are positioned downstream of their closest host gene (Appendices K and L). The flexibility in relative insertion sites allows for the significantly increased levels of expression of *AVL9* and *SCS7* observed in the work here, as the candidate LTRs, acting as enhancer sequences, can function in both orientations (Khoury and Gruss, 1983; Serfling *et al.*, 1985; Blackwood and Kadonaga, 1998; Schaffner, 2015).

Insertions may alter chromosome architecture

In addition to the alteration of host gene expression, some TEs may offer other functions to the genome. For example, LINEs and SINEs help to regulate chromatin structure as well as function as enhancers and promoters to alter expression of host genes (reviewed by Elbarbary *et al.*, 2016). Retrotransposon families in maize spread heterchromatin, in turn lowering expression of nearby genes (Eichten *et al.*, 2012). In fission yeast *Sz. pombe*, behaviour of *Ty3/gypsy* family *Tf2* alters depending on the stage of cell cycle (Mizuguchi *et al.*, 2015). Copies of *Tf2* spread across the genome are collected into formations of *Tf* bodies by centromere B proteins (CENP-B), which are themselves derived from *pogo* DNA transposons (Casola *et al.*, 2008). *Tf* bodies interact with histones which then modulate condensin, particularly around centromeres (Tanaka *et al.*, 2012; Mizuguchi *et al.*, 2015). However, the possibility of *Ty* insertions altering chromosomes - for example by opening chromatin in order for transcription factors to bind - has yet to be investigated in *Saccharomyces*. As effects upon chromatin structure by TE insertions were quickly established in other model organisms such as *Drosophila* (Chen and Corces, 2001) however, it is likely that a similar effect in *S. cerevisiae* would have been observed.

3.9 Summary and conclusions

Screening the genomes of *Saccharomyces* for signatures of positive selection acting upon TE insertions identified potential candidates that may persist in populations due to benefits conferred to their host. A minority of these candidates may influence expression of adjacent host genes, while any effects the remaining candidates cause were not able to be elucidated using the qPCR technique employed. Two candidate LTRs in *S. cerevisiae* may significantly increase expression of their adjacent host genes, which may in turn have an impact upon host fitness. Although the insertion preference and ability of *Ty* insertions to alter gene expression is unlikely to have evolved

as a method of increasing host fitness, insertions are clearly linked to the survival of their hosts, particularly under stressful conditions.

Chapter 4

Ty transposable element diversity in the *Saccharomyces sensu stricto* complex

S. cerevisiae was considered a model organism long before it became the first eukaryote to have its genome sequenced by Goffeau *et al.* (1996), due to molecular biology work performed by research teams in the 1970s and 1980s (e.g. Hartwell *et al.*, 1970; Rytka, 1975; Ogden *et al.*, 1979; Prakash and Higgins, 1982; Proffitt *et al.*, 1984; Beggs, 1978). Its transposable elements were thoroughly investigated by the independent teams of Kim *et al.* (1998) and Hani and Feldmann (1998), who inspired numerous analyses with the more recent availability of the genomes of multiple strains (e.g. Liti *et al.*, 2009; Akao *et al.*, 2011; Bleykasten-Grosshans *et al.*, 2013; Istace *et al.*, 2017; Yue *et al.*, 2017).

The analyses of *Ty* families of *S. cerevisiae* were followed by brief investigations into the elements of *S. paradoxus* (e.g. Fingerman *et al.*, 2003), *S. kudriavzevii, S. mikatae* and *S. uvarum* (prev. *S. bayanus*; Liti *et al.*, 2005). Since this time, the additions of *S. arboricola* and *S. eubayanus* genomes into available data have suggested the need for an updated investigation into the *Ty* content of *Saccharomyces* species.

In this chapter, the available genomes of each *Saccharomyces* species are systematically screened for insertions, with the *Saccharomyces* Genome Resequencing Project (SGRP) strains of *S. cerevisiae* and *S. paradoxus* given priority in these species. In the remaining species, the state of families was often unknown until the work conducted here. These investigations are designed to characterise and quantify *Ty* insertions in preparation for the phylogenetic analyses of Chapter 5.

Although there is still some debate as to whether *S. cariocanus* should be classed as a separate species to *S. paradoxus* (G Naumov, pers. comm., 2016; Naumov *et al.*, 2000; Liti *et al.*, 2006, 2009; Hittinger, 2013), it is considered as such here for the purposes of the investigations into its

Ty families.

4.1 *Ty* families are not homogeneous across the genomes of *Saccharomyces* species

Summary information for the reference strains of species in the *Saccharomyces sensu stricto* complex is displayed in Table 4.1. *S. cariocanus* has not been designated a reference strain, therefore the assembly of UFRJ50816 (Yue *et al.*, 2017) was selected. The NCBI trace archives, nucleotide and WGS databases were searched with LTR queries to obtain hits from each *Ty* family in all *Saccharomyces* species. To obtain genomic GC and TE content (%; Table 4.1), genomes were screened with RepeatMasker using a custom library, to which consensus sequences were added of insertions extracted from each species. The genomes were then screened again with the complete library to check for any sequences that may have been incomplete during a BLAST search. While GC% for all species is relatively consistent, a four-fold difference in genomic TE fraction is observed between the lowest content (*S. eubayanus*) and the highest (*S. cerevisiae*).

The search criteria are different to those employed Carr *et al.* (2012), as partial/truncated LTRs were excluded from analysis here. The datasets generated were used to compile the phylogenetic trees in Chapter 5. Family names previously assigned by Neuvéglise *et al.* (2002) are used consistently here, and newly described families are named according to the suggestions made by these authors. Names are assigned based upon species name and most closely related *Ty*-family, e.g. *Tse4* of *S. eubayanus* is homologous to *Ty4* in *S. cerevisiae*.

Species	Reference strain	TE fraction of genome (%)	GC content (%)
S. arboricola	H-6	1.12	38.8
S. cariocanus*	UFRJ50816	3.70	38.2
S. cerevisiae*	S288c	3.80	38.3
S. eubayanus	FM1318	0.80	39.9
S. kudriavzevii	IFO 1802	1.46	39.8
S. mikatae	IFO 1815	2.47	38.1
S. paradoxus	NRRLY-17217	1.61	38.5
S. uvarum	MCYC 623	1.35	40.2

Table 4.1: ***Ty* contents of *Saccharomyces* species reference strains.** *indicates high-throughput 3^{rd} generation genome sequencing. *S. cariocanus* has not been designated a reference strain, so the PacBio assembly of UFRJ50816 was selected due to its high quality and low rate of gaps.

4.1.1 A negative correlation exists between genomic TE and GC content

GC content (%) and the TE fraction of genomic content (%) were generated with RepeatMasker and the custom library and collated for all available strains for all species. The relationship is unclear, but a slight negative correlation between the two is observed when data from all species are collated (Figure 4.1). No significant relationships are observed for the data of individual species, as for example, genomic TE fraction in *S. cerevisiae* and *S. paradoxus* does not appear to influence GC%. The data for *S. kudriavzevii, S. eubayanus* and *S. uvarum* do not follow this pattern however.

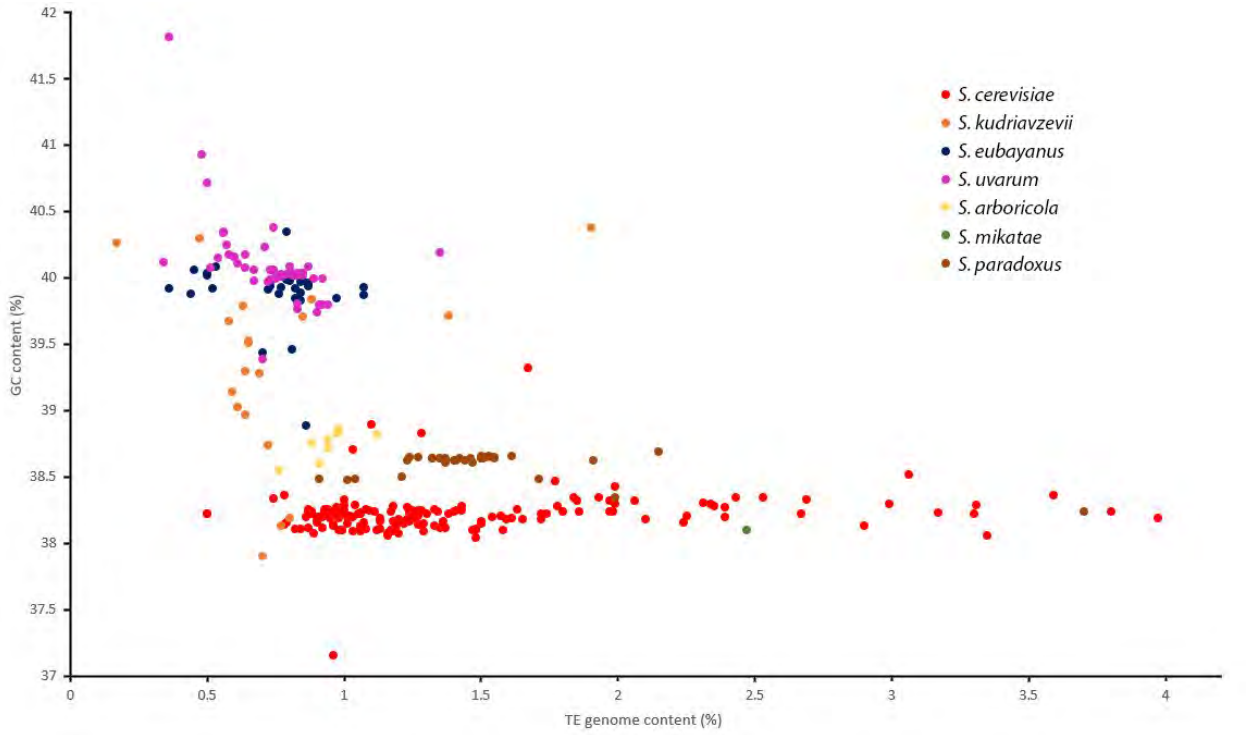

Figure 4.1: **Negative correlation between genomic GC and TE content.** Both values were generated by RepeatMasker using the custom library. *S. boulardii*, Brazilian and Peterhof strains of *S. cerevisiae* (Chapter 6) are included with SGRP *S. cerevisiae*, and *S. cariocanus* with *S. paradoxus*, respectively, as there is no correlation in these strains when analysed separately.

4.1.2 *Ty* copy numbers differ in the reference strains of *Saccharomyces*

Copy numbers were determined by screening reference strains (Figure 4.2) with RepeatMasker using a custom library. Only full-length LTRs - i.e. those possessing boundary sequences - are included in the totals. For full-length elements (FLEs), elements are counted if they are a pseudoelement but flanking LTRs are intact and most *pol* domains functional, such as the *Ty5* relic in *S. cerevisiae*. Fragments of elements, including those that are cut off by sequencing reads, are

not counted in the totals unless the full element was able to be confidently reconstructed, therefore it should be noted that copy numbers here are not directly comparable to those of other studies. All available strains of species were screened for *Ty* content, the results of which are detailed in Section 4.2 onwards.

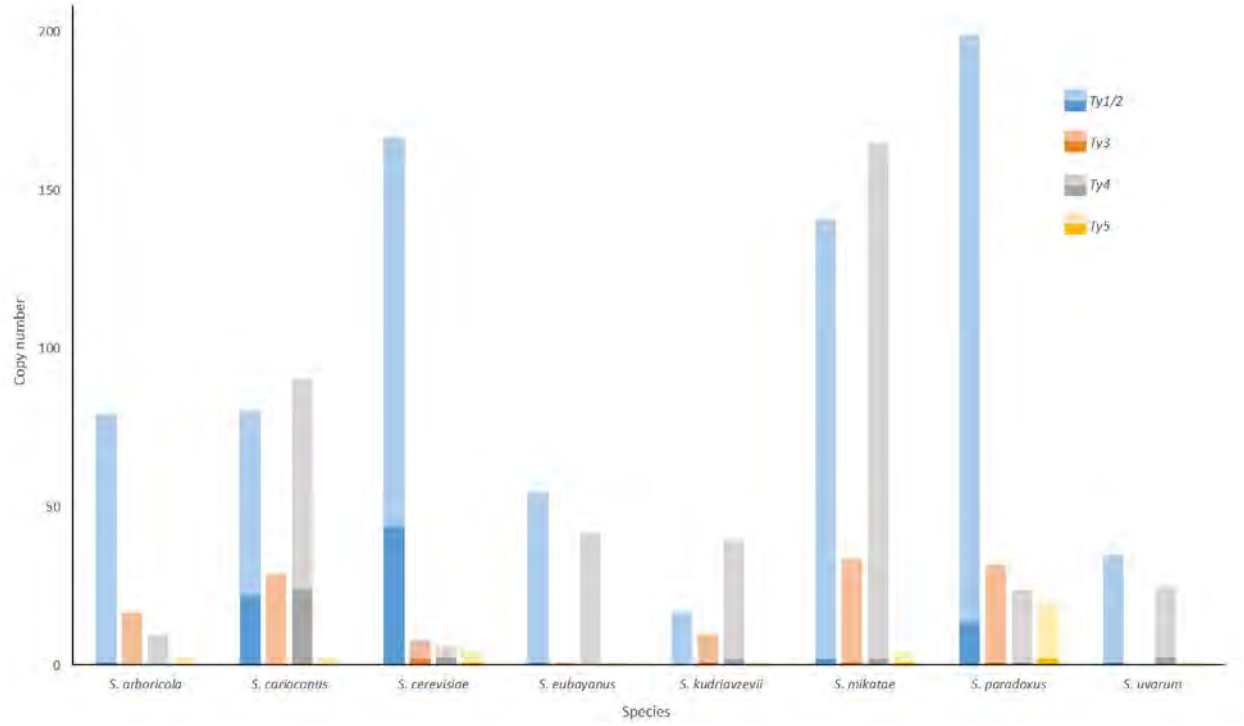

Figure 4.2: **Copy number in *Saccharomyces* species.** Proportion of FLEs to solo LTRs are represented by darkened and lighter proportions of bars, respectively.

4.2 SGRP strains of *S. cerevisiae*

Table 4.2 displays the characteristics of all LTRs in the families of the SGRP strains of *S. cerevisiae*. Carr *et al.* (2012) previously analysed the reference strain and reported π values of all families and Tajima's *D* values for active sequences (i.e. those that possessed paralogous copies on short phylogenetic branches), which differ to those reported here. Excluding *Ty2*, the π values recorded here are lower than those reported by Carr *et al.* (2012). Although in this work sequences were included from multiple strains (*n*=38), this is unlikely to have an impact upon π values. Therefore the differences observed are at least in part due to the decision of Carr *et al.* (2012) to include partial - and likely far older - sequences in their analyses. Significantly negative values of *D* are consistent with recent familial ancestry, which is also observed in the corresponding phylogenies of Chapter 5.

	Family							
	Ty1	*Ty1'*	*Ty2*	hybrid	*Ty3*	*Ty4E*	*Ty4A**	*Ty5*
Nucleotide diversity (π)	0.131	0.078	0.087	0.194	0.050	0.021	0.004	0.128
Tajima's *D*	**-1.841**	**-2.401**	**-2.360**	-1.407	**-2.612**	**-2.544**	0.592	**-1.866**

Table 4.2: **Characteristics of *Ty* families in *S. cerevisiae*.** *family present in strain L-1528 only. Bold formatting indicates those statistically significant values.

Although significantly negative, the Tajima's *D* values for *Ty1* and *Ty5* were lower than those of *Ty2-4* (excluding American *Ty4*, section 4.2.3). Carr *et al.* (2012) analysed the active *Ty1* sequences and recorded a positive value of Tajima's *D* (1.176), which was likely due to the inclusion of subfamilies *Ty1'* and *Ty1/Ty2* hybrids in the test, as this substructure (as with population subdivision) can result in a positive *D* value.

4.2.1 LTR sequences in the *Ty1/2* superfamily contain variable 3' boundaries

During the generation of LTR datasets, approximately 4,200 sequences appeared to be truncated at ∼30bp from the 3' end and so were separated for further analysis. Upon alignment, the LTRs were conserved up to point of the black vertical line in Figure 4.3. Duplicates and poor quality reads were removed, leaving a total of 372 unique insertions. All identifiable *Ty1'* (*n*=74) and *Ty1/2* hybrids (*n*=24) were also removed, leaving 274 unique variant insertions across the SGRP strains. As the sequences do not possess the characteristic deletion associated with *Ty2* sequences, they may be a variant form of *Ty1* (hereafter referred to as *Ty1v*). This is confirmed by 34% (*n*=93) of *Ty1v* sequences being associated with *Ty1* internal coding regions. Upon exploration of each LTR containing contig, it was discovered that these particular LTRs have variant 3' boundaries when compared to canonical LTRs from the *Ty1/2* superfamily, and had not been returned by the BLAST search, instead appearing as if truncated. The missing ∼30bp was manually extracted from the contigs and added to the end of each of these sequences. It then became clear that these unusual 3' boundaries are abundant across multiple strains and insertions. Figure 4.3 displays the diversity across all the boundary sequences, once duplicates and poor quality reads were excluded. TSDs are identical in 17% of LTRs (*n*=47), indicating recent intra-element recombination.

Variable 3' ends were also unlikely due to recombination, as DnaSP and TOPALi recombination tests, the Phi test within SplitsTree and visual inspection of sequences all failed to find significant evidence of recombination. To ensure that these variable ends are not simply due to degradation, both full LTR sequences and the variable ∼30bp of the LTRs were used as BLAST queries to

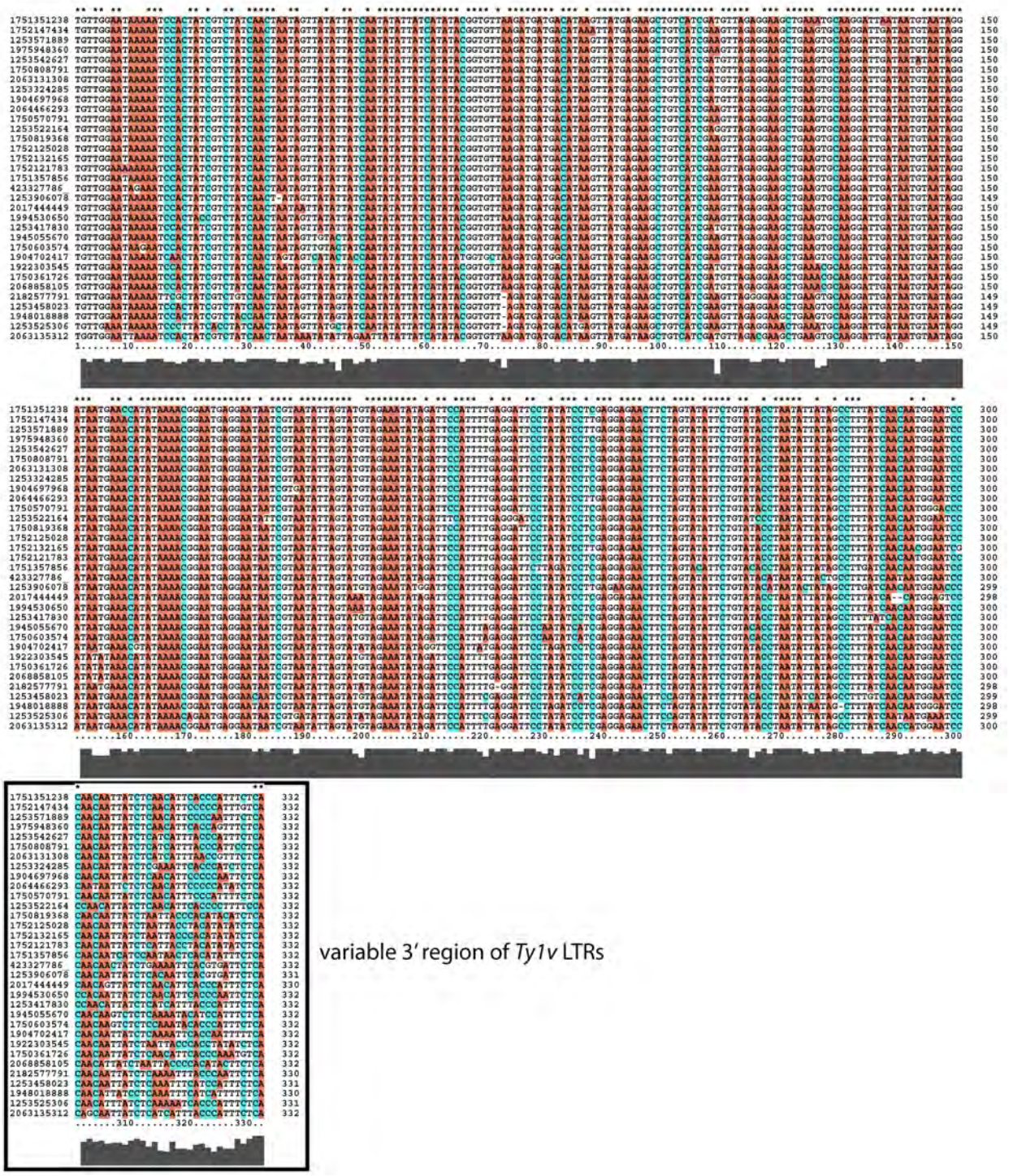

Figure 4.3: **Alignment of *Ty1v* LTRs in *S. cerevisiae* illustrating the highly variable 3' boundaries.** The boxed section of the alignment shows the point at which the sequences typically lost conservation. The graphic at the base of the alignments indicates the level of conservation at each position.

search all available *Saccharomyces* species. Both were found to be widespread throughout other strains of *S. cerevisiae*. Interestingly, eight full-length LTRs were present with 100% identity in the Japanese saké strain, Kyokai No. 7 (also referred to as K7). This strain also possesses more of the variable 3' ends than any other strain when these were used as queries without the rest of the

LTR.

Nucleotide diversity was calculated in 30bp sliding windows across the alignment (Figure 4.4). The region of conservation (sites 291-320) immediately before the variable 3' boundary produced the lowest π value of the sliding windows (π=0.04). Interestingly, regions of higher nucleotide diversity coincide with *GCN4* binding sites and TATA boxes (Servant *et al.*, 2008).

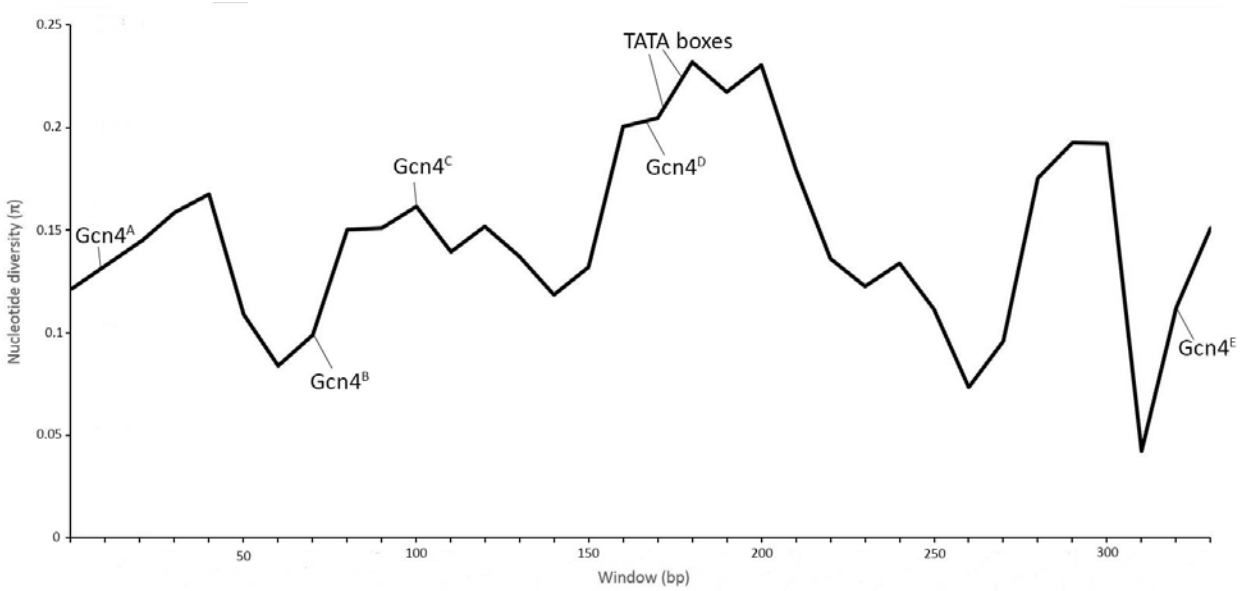

Figure 4.4: **Nucleotide diversity of *Ty1v* LTRs.**. Nucleotide diversity (π) calculated across 30bp sliding windows of the alignment of 227 *Ty1v* LTRs across 332 sites.

4.2.2 Recombination breakpoints of *Ty1/2* LTRs are variable

20,000 short reads of SGRP strains containing LTRs from the *Ty1/2* superfamily were downloaded from the NCBI trace archives. Once duplicates, poor quality and *Ty1v* sequences were removed, this left 453 unique LTR sequences, 33% (*n*=152) of which are associated with coding regions of full length elements. 97 (21%) of the total are classified at *Ty1'* LTRs; 261 (57%) are classified as *Ty1*; the characteristic deletion in *Ty2* LTRs is identifiable in 24 (5%) sequences. Of the remaining 75 (17%) sequences, four distinct groupings of hybrids are visually observed, examples of which are displayed in Figure 4.5.

Interestingly, no statistically significant evidence of recombination was found by the TOPALi recombination and SplitsTree Phi tests on sequences from the *Ty1/2* superfamily (*P*=0.9) despite the possible recombination breakpoints being visually apparent. DnaSP however, found evidence

of five possible recombination regions (between sites 22-83, 102-123, 131-216, 216-233 and 233-237). However, searches for recombination in *Ty1/2* sequences have previously failed (Kupiec and Petes, 1988*b*; Wilke *et al.*, 1989; Kim *et al.*, 1998; Sandmeyer, 1998).

4.2.3 *Ty3* and *Ty4* are widespread in *S. cerevisiae*

The SGRP strains possess unique *Ty3* (*n*=212) LTRs, indicating strain-specific activity occurring since the last common ancestor. No evidence of recombination was found (*P*=1.0), and the LTRs of this family possess the second lowest nucleotide diversity in this species (Table 4.2).

A subdivision is observed in the sequences of *Ty4* when query sequences of *S. paradoxus* were used to search the SGRP genomes of *S. cerevisiae*. The distinction was made in the sequences of *S. paradoxus*, as strains contain differing types of *Ty4* depending on geographical origin (Section 4.4.4). However, similar distinctions are not observed in *S. cerevisiae*; instead all but one strain contain *Ty4* elements possessing LTRs of the canonical *S. cerevisiae* length regardless of geographical origin ($\sim$370bp, *n*=37, Hug and Feldmann, 1996, hereafter referred to as European).

The *S. cerevisiae* SGRP strains contain unique *Ty4* insertions (*n*=113), indicative of strain-specific activity occurring since the divergence of populations. The Tajima's *D* value for this family (Table 4.2) is significantly negative, consistent with relatively recent ancestry. No evidence of recombination was found between European *Ty4* LTRs (*P*=0.8; Phi test). In addition, *Ty4* also displays the lowest nucleotide diversity of the widespread families (Table 4.2, π=0.021).

The Chilean strain L-1528 is alone in containing additional insertions (*n*=4) that share higher identity with those of *S. uvarum* and *S. paradoxus* than the elements typically observed in *S. cerevisiae* as a species (hereafter referred to as the American *Ty4* subtype). The presence of this specific type of *Ty4* prompted the analysis of the genomes of both Chilean strains (L-1528 and L-1378) for evidence of introgression from other species. Whereas the former contains evidence of introgression from *S. paradoxus* (or hybridisation, or even contamination), the latter is "pure" *S. cerevisiae* (data not shown). The mosaic state of the genome of L-1528 was not observed by Liti *et al.* (2009) however.

The low quality assembly of L-1528 – both that released by Liti *et al.* (2009) and the independent construction performed here – prevented FLEs from being extracted in full. A BLAST search of other *S. cerevisiae* strains was conducted in order to obtain the element sequence and survey the extent to which the American *Ty4* subfamily is present in this species. American-type insertions are

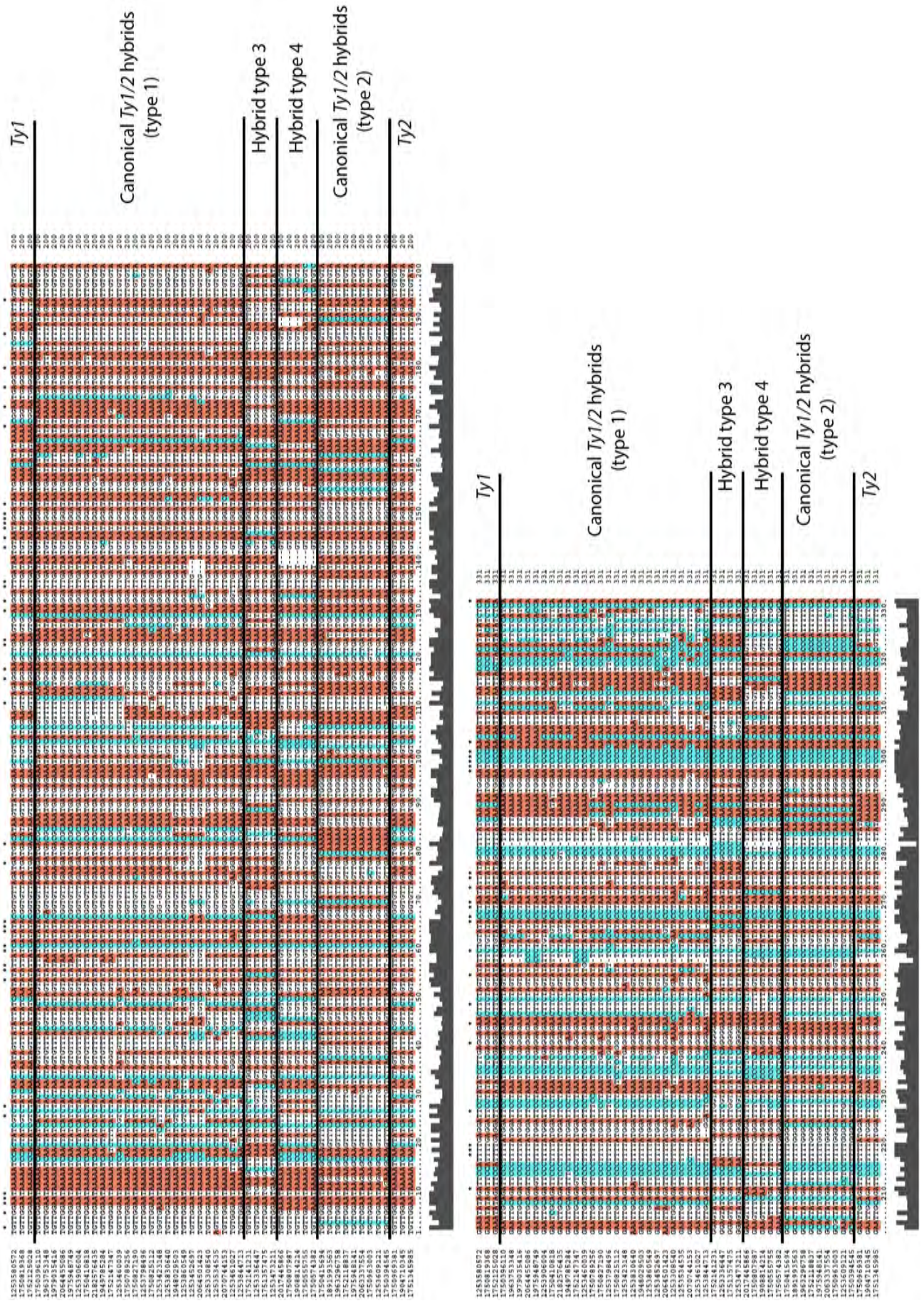

Figure 4.5: **Alignment of *Ty1/Ty2* LTRs and potentially new hybrid sequences.**. Four distinct groupings of hybrids outside the reference strain. The full number of type 3 (n=4) and type 4 (n=5) hybrids are displayed, but due to the number of sequences, only a selection of the canonical hybrids are included to demonstrate their variability. Canonical hybrids are those observed by Carr *et al.* (2012).

present in only two other currently sequenced strains isolated in the Americas. Strain 245 (Legras *et al.*, 2018) contains a full-length copy and multiple solo LTRs (*n*=3) while only solos are present in strain 460 (Legras *et al.*, 2018; *n*=2).

Figure 4.6 displays the comparison of both *Ty4* elements' nucleotide sequences. The sequences share 74% identity over 84% of the internal DNA, with similarity of residues lower at 66% over the entirety of the elements.

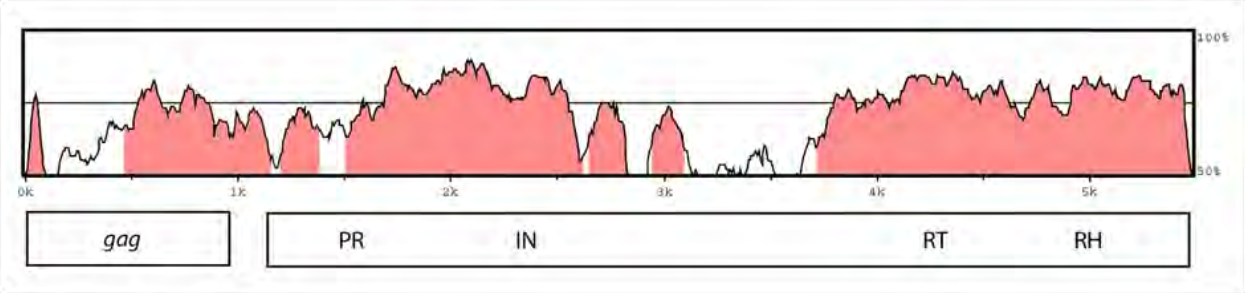

Figure 4.6: **Shared identity between *Ty4* elements in European and American *S. cerevisiae*.** Sequences from strains S288c (European) and 245 (American) were aligned and visualised with mVISTA (Frazer *et al.*, 2004). The LTRs were excluded from the comparison as the 3' LTR in the American type element was cut off by the end of the read. Both coding regions are 5.5kb in length.

4.2.4 Improved sequencing technique allows multiple *Ty5* relic elements to be identified in *S. cerevisiae*

Genomes of *S. cerevisiae* strains, including a selection of SGRP strains (*n*=8), were recently resequenced by Yue *et al.* (2017), Istace *et al.* (2017) and Matheson *et al.* (2017). The techniques of PacBio and Oxford Nanopore are a vast improvement on the previous Illumina-style sequencing methods, with the caveat that coverage must be increased in order to counteract the higher rate of sequencing mistakes (Feng *et al.*, 2015; Rhoads and Au, 2015; Salazar *et al.*, 2017). Using short-read sequencing methods such as Illumina can prove difficult when sequencing and assembling repeated sequences as single reads do not span an entire full-length element (Treangen and Salzberg, 2011; Hoban *et al.*, 2016). The *Ty* contents of these newly sequenced assemblies were compared with those of the previous sequencing method. Table 4.3 summarises the differences in genomic TE fraction (%) between both sequencing techniques.

There is a significant difference in genomic TE fraction depending on the sequencing method used (*P*=0.001; two-tailed paired t-test). Previously Bleykasten-Grosshans *et al.* (2013) did not observe a significant difference between genomic TE content depending on sequencing method, but it should be noted that different strains were analysed. Genomic TE content for the reference

Strain	Illumina		PacBio/Nanopore	
	Genome content %	Reference	Genome content %	Reference
CLIB324	1.69	Fay *et al.*, 2011	1.99	Istace *et al.*, 2017
T73	1.07	Fay *et al.*, 2011	1.59	Istace *et al.*, 2017
S288c	2.91	Liti *et al.*, 2009	3.80	Yue *et al.*, 2017
W303	2.62	Ralser *et al.*, 2012	3.97	Matheson *et al.*, 2017
Y12	0.92	Liti *et al.*, 2009	2.35	Yue *et al.*, 2017
DBVPG6044	1.37	Liti *et al.*, 2009	2.34	Yue *et al.*, 2017
DBVPG6575	1.31	Liti *et al.*, 2009	1.78	Yue *et al.*, 2017
SK1	1.66	Liti *et al.*, 2009	2.43	Yue *et al.*, 2017
YPS128	1.13	Liti *et al.*, 2009	2.06	Yue *et al.*, 2017
UWOPS03-461.4	0.87	Liti *et al.*, 2009	1.27	Yue *et al.*, 2017

Table 4.3: **Improving sequencing technique increased genomic *Ty* content.** Increased genomic TE content (%) is observed when strains of *S. cerevisiae* are sequenced with a long read technique. Strains from S288c onwards are from the SGRP collection.

strain S288c increased from 2.91% to 3.80% with the improved sequencing technique. The original reference strain was manually created from several derivative strains, and that the new sequencing method was performed solely on S228c, which has since been reported to contain a greater number of insertions than previously identified (Shibata *et al.*, 2009).

Resequencing of these strains allows conclusions based on Illumina assemblies (Liti *et al.*, 2009) to be reassessed. For example, Bleykasten-Grosshans *et al.* (2013) reported that *Ty2* was extinct in strain Y12, yet this genome contains an autonomous copy in the PacBio assembly. Additionally, it was believed that the *Ty5* relic observed on chromosome III of the reference strain (Kim *et al.*, 1998) is present in the majority of *S. cerevisiae* genomes (Bleykasten-Grosshans *et al.*, 2013). Using the Liti *et al.* (2009) assemblies, it was discovered that only $\sim$55% of the SGRP strains possessed the relic element (n=21), while the remaining strains contain a solo LTR at this locus (n=17). Furthermore, strain RM11-1A, from a differing lineage than S288c (Drozdova *et al.*, 2016), is alone in containing a non-autonomous FLE in addition to the relic. Along with unique *Ty5* solo LTRs found across the SGRP strains (n=42), this not only indicates that strain-specific activity has occurred, but that *Ty5* may have become extinct on a strain-independent basis.

The improved assemblies of Istace *et al.* (2017) (Table 4.3) allowed the identification of further multiple relics - usually on chrIII and XI - in 30% of strains (n=3). The relic is missing the IN region, and is only $\sim$2.7kb in length. *Ty5* has escaped extinction in non-SGRP strains only (YJM1078, YJM270, PW5, wine070, wine010 and ZP611).

4.3 Variation in TE content across strains, sources and origins of *S. cerevisiae*

The 1343 currently sequenced strains of *S. cerevisiae* from varied sources/origins were screened with RepeatMasker and the custom library to ascertain their genomic TE content. Figure 4.7 displays the results of this analysis. Wild and laboratory strains possess the lowest and highest ranges, respectively, while those of bakery and unknown origins display the smallest ranges of genomic TE content.

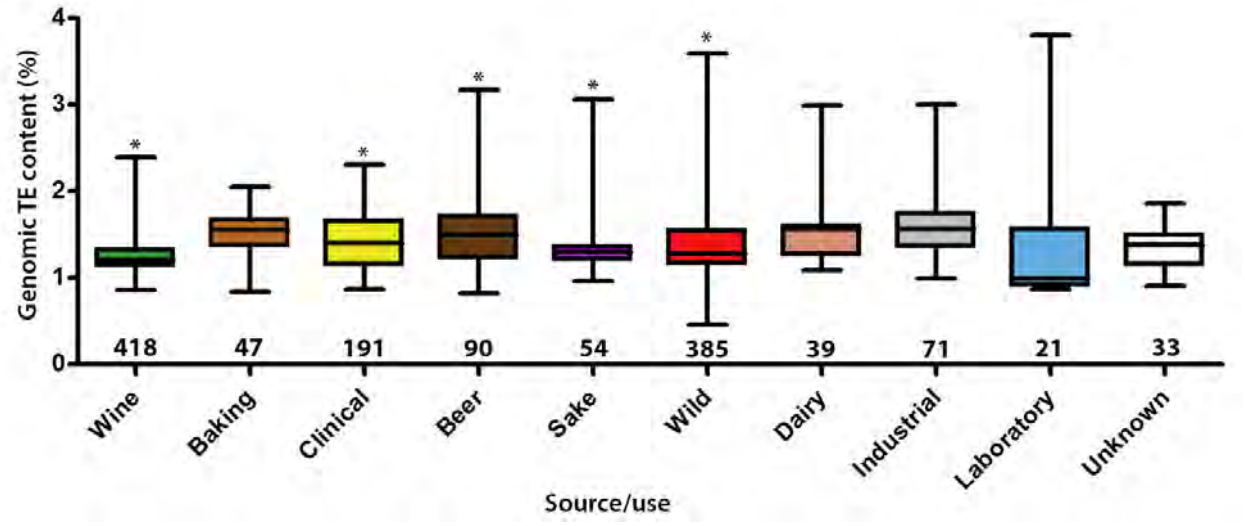

Figure 4.7: **Genomic TE content of *S. cerevisiae* strains by source/origin.** Strains are grouped according to source/use as classified by respective research teams. Number of strains screened are displayed below each box plot. *indicates those categories whose mean genomic TE content significantly differs.

Wild strains as a group represent the majority of sequenced isolates (31%), closely followed by wine strains (29%). The industrial group of strains possess the highest mean genomic TE content (1.56%), followed by beer, baking and dairy strains. Other authors have found that laboratory strains tend to have the highest TE content (Wilke and Adams, 1992; Liti *et al.*, 2009; Bleykasten-Grosshans *et al.*, 2013). A one-way analysis of variance (ANOVA) test on the ten groups indicates that there is a significant difference between the genomic TE content % values (P=0.0001), which has not been observed previously, likely due to the far larger dataset used here. It is possible that the close phylogenetic relationships shared by these strains by source may also influence their genomic TE content, rather than it being largely dependent on source/use.

4.4 *S. paradoxus*

Table 4.4 displays the characteristics of the *Ty* families in the SGRP strains of *S. paradoxus*. All families returned significantly negative values of Tajima's *D*, excluding *Ty3* which was represented by too few unique insertions to complete the test (*n*=3). The strongly negative *D* values are consistent with recent ancestry of these families, which is also observed in the corresponding phylogenies of Chapter 5.

	Family						
	Ty1p	*Ty1*	*Ty3p*	*Ty3*	*Ty4E*	*Ty4A*	*Ty5*
Nucleotide diversity (π)	0.119	0.056	0.038	0.038	0.069	0.033	0.091
Tajima's *D*	**-2.120**	**-2.008**	**-2.494**	-	**-2.503**	**-2.254**	**-1.816**

Table 4.4: **Characteristics of *Ty* families in *S. paradoxus*.** European - *Ty4*E; American - *Ty4*A, detailed in section 4.4.4.

The reference genome of *S. paradoxus*, NRRLY-17217, like that of *S. cerevisiae*, was constructed from a number of strains, most of which were obtained in the UK, to obtain a consensus genome (Johnson *et al.*, 2003). This strain contains the highest genomic TE content of the sequenced strains (1.61%) which is likely due to the compiled nature of the genome. The other British strains also show a tendency towards a higher genomic TE content than those isolated elsewhere in the world.

4.4.1 *Ty* content is underestimated in the SGRP *S. paradoxus* assemblies

As seen in the *S. cerevisiae* SGRP assemblies, those *S. paradoxus* SGRP assemblies (Liti *et al.*, 2009)underestimate the *Ty* content, especially FLEs, of many strains. Comparisons made between those strains resequenced by Yue *et al.* (2017) and the corresponding Liti *et al.* (2009) assemblies prove the latter vastly underestimated the *Ty* content of this species (Table 4.5). As in *S. cerevisiae*, there is a significant difference in the genomic TE content % depending on the sequencing method used (*P*=0.01; two-tailed paired t-test). Regions (>5kb) of indeterminate bases (Ns) were consistent throughout strains. Examination of these masked regions in the four resequenced strains (Table 4.5) revealed FLEs, meaning the RepeatMasker screening of the original assemblies was inaccurate as to the actual content of the genomes. For example, Hawaiian strain UWOPS91-917.1 contains more than double the genomic TE fraction when the assembly was improved, corresponding to multiple FLEs (*n*=5), including the presence of *Ty2* (Section 4.4.2).

Strain	Illumina		PacBio	
	Genome content %	Reference	Genome content %	Reference
CBS342	1.55	Liti *et al.*, 2009	2.15	Yue *et al.*, 2017
N44	1.35	Liti *et al.*, 2009	1.91	Yue *et al.*, 2017
UWOPS91-917.1	0.82	Liti *et al.*, 2009	1.99	Yue *et al.*, 2017
YPS138	1.03	Liti *et al.*, 2009	1.21	Yue *et al.*, 2017

Table 4.5: **Improving sequencing technique increased genomic *Ty* content.** Increased genomic TE content % is observed when strains of *S. paradoxus* are sequenced with a long read technique.

4.4.2 The *Ty1/2* superfamily is represented by widespread *Ty1p* and rare *Ty2*

Solo *Ty1p* LTRs are present in all strains, whereas autonomous elements are present in ∼70% of SGRP strains (*n*=22). Only Hawaiian strain UWOPS91-917.1 contains evidence of *Ty2*. Partial elements (*n*=2) were extracted from the PacBio assembly, therefore this family is extinct in *S. paradoxus*. Liti *et al.* (2005) previously discovered *Ty2* in *S. paradoxus* strains that showed evidence of hybridisation with *S. cerevisiae*. Raw reads of the Hawaiian strain were mapped onto reference genomes of *S. cerevisiae* and *S. paradoxus* in order to screen for introgression from *S. cerevisiae* genomes. However, reads failed to map to the *S. cerevisiae* genome, therefore this strain is unlikely to have undergone recent - if any - hybridisation.

4.4.3 The presence of *Ty5p* in *S. paradoxus* is variable

Potentially up to 40% of SGRP strains contain autonomous copies of *Ty5p* (*n*=13), but due to the sequencing quality of the SGRP strains, copy number could not confidently be ascertained. Strain CBS342 however was recently resequenced by Yue *et al.* (2017), and found to contain multiple functional copies of *Ty5p* (*n*=9). The remaining strains in this resequencing project contain only solo LTRs however, therefore autonomous copies could actually be rare, contrary to the findings of Liti *et al.* (2005).

4.4.4 *Ty4* displays evidence of divergence with isolation of American and European populations

A divergence in element sequences is found in the SGRP genomes of *S. paradoxus* depending on isolate origin. These are primarily distinguished by two types of *Ty4* LTR, similarly observed in other *Saccharomyces* species (Sections 4.2.3, 4.6.3, 4.7.4, 4.9.3 and 4.10.3). The American type (291bp, *S. eubayanus*-like) appears to have been active more recently than the European type

(∼370bp, *S. cerevisiae*-like), as they are often found to disrupt previous European insertions in the genomes of strains isolated from American and Hawaiian populations. Currently three further populations of *S. paradoxus* have been identified (Liti *et al.*, 2009; Leducq *et al.*, 2014), yet no evidence of American *Ty4* is found in these populations, indicating that this subfamily may be confined to the American continent. Comparisons of European and American elements, and American with that of *Tsu4* of *S. uvarum* is displayed in Figure 4.8. Strains YPS138 and UWOPS91-917.1,

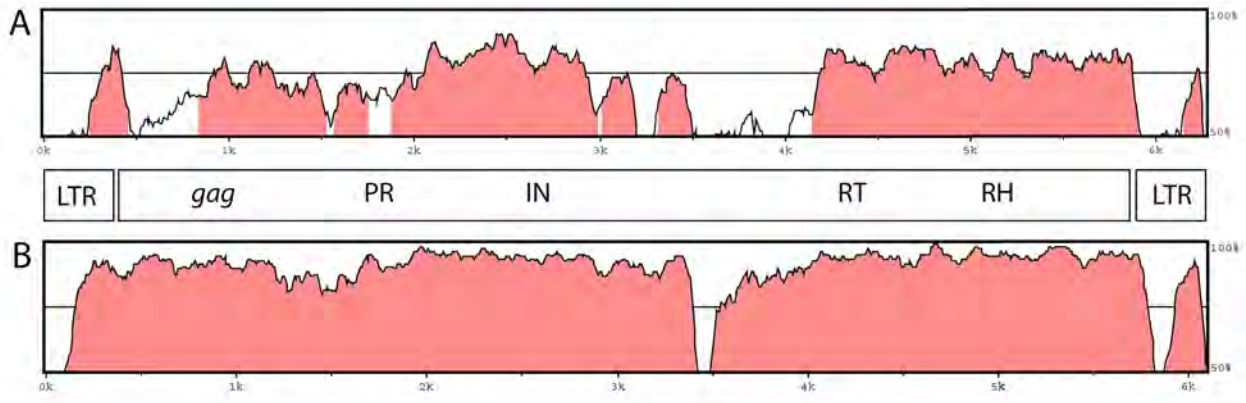

Figure 4.8: **Shared identity between *Ty4* elements.** European and American elements (A) share 73% identity over 82% of the sequences. LTRs are not conserved, and identity is lost between IN and RT domains. American *Ty4* and *Tsu4* of *S. uvarum* (B) share higher identity (∼90%) across coding regions.

isolated in the USA and Hawaii, respectively, are the only strains to contain both European and American *Ty4*, based upon the improved assemblies of Yue *et al.* (2017). In the Hawaiian strain however, the two copies of American *Ty4* are both non-autonomous as they contain a stop codon between IN and RT domains (chromosome VII), or are interrupted by a *Ty1p* insertion that has since undergone recombination (chromosome XIV). Few European solo LTRs are present intact, as more recent *Ty1p* insertions have disrupted the sequences.

4.5 *S. cariocanus*

Table 4.6 displays the nucleotide diversity and Tajima's *D* values for the unique *Ty4* families of *S. cariocanus*. These are the only families in this species to possess the required amount of insertions independent of those of *S. paradoxus* to complete the tests. A significantly negative value of *D* is consistent with a family possessing recent ancestry, which is also observed in the corresponding phylogenies of Chapter 5.

	Family	
	*Ty4*E	*Ty4*A
Nucleotide diversity (π)	0.018	0.030
Tajima's *D*	-0.383	**-2.451**

Table 4.6: **Characteristics of *S. cariocanus*-specific *Ty4* families.** Characteristics of unique insertions only, i.e. not those shared with *S. paradoxus*.

4.5.1 *Ty4* shows high levels of activity in the reference strain of *S. cariocanus*

Liti *et al.* (2009) analysed the raw reads of the two sequenced strains of *S. cariocanus* and reported that both possessed a higher abundance of insertions than strains of *S. paradoxus*, a result that was not observed here (data not shown). In the Liti *et al.* (2009) Illumina assembly of strain URFJ50816, there appear to be only remnants of *Ty4* coding regions (<500bp, *n*=3) and solo LTRs (*n*=11). As with the *S. paradoxus* and *S. cerevisiae* SGRP assemblies, the majority of regions containing *Ty* elements are masked with undetermined bases (N). Examining the PacBio assembly of Yue *et al.* (2017) reveals full-length unique copies of *Ty4* (*n*=24, Table 4.7) and solo LTRs (*n*=66; American type *n*=59) that shared a high rate of identity (>98%).

Of the European insertions, the majority of solo LTRs (*n*=5) predate divergence from *S. paradoxus* as they are shared with the parental species. Despite being unique to *S. cariocanus*, the FLE appears to be non-autonomous however, as the RT domain is unrecognisable in a conserved domain search and fails to align with RT regions of other elements. ∼25% of American type elements are also likely non-autonomous (Table 4.7).

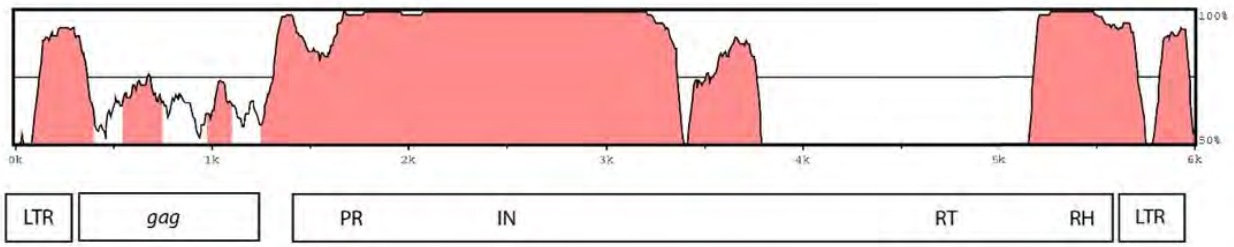

Figure 4.9: **Shared identity between European and American *Ty4* elements in *S. cariocanus*.** Sequences were aligned and visualised with mVISTA (Frazer *et al.*, 2004). Comparison of the two different types of *Ty4* in *S. cariocanus*, American (endogenous; *n*=23) and European (foreign; *n*=1). In the European copy, the RT is degraded, which is not due to a frameshift in/del mutation. However, similarity is also below 75% for much of the *gag* region. The IN region is 100% conserved between the two.

Comparison of the types of elements revealed 88% similarity over 75% of the element, with less than 75% similarity in the *gag* region and none in the RT region (Figure 4.9). Almost 70% of the American FLEs in *S. cariocanus* are unique to this subspecies (*n*=16), consistent with activity since the split from *S. paradoxus*.

Chromosomal location	TSDs 5'	TSDs 3'	Length (bp)	LTR identity (%)	Notes
II: 554570-560566	TATAC	TATAC	5998	100	
IV: 503431-509427	GAAAG	GAAAG	5998	100	
IV: 1017158-1023474	AATTA	AATTA	5964*	98	*Ty4* solo insertion in *gag*
V: 498859 504855	AATTC	AAGGA	5998	99	Recombinant
VII: 538638-544707	GTTGA	TGTGT	5998*	99	Recombinant, 5' LTR disrupted by solo *Ty4* LTR
VII: 699278-705352	GTTTC	GTTTC	5995	100	
VII: 778685-784682	CAGTT	GATTG	5998	99	Recombinant
VII: 874927-880922	GTAAG	ATAAT	5996	99	Recombinant
VII: 932265-938261	ATTCT	ATTCT	5997	99	
VIII: 373186-377824^{E}	ATTAC	ATTAC	4788	99	
IX: 253236-259231	GGCTG	GTTCC	5996	99	Recombinant
X: 96565-102558	CTATA	CTATA	5994	98	Stop codon before RT
X: 185145-191143	ATTAT	ATTAT	5999	100	
X: 409062-414977	TAACC	TAACC	5915	-	Disrupts European *Ty4* solo LTR; disrupted by *Ty4* solo in 3' LTR
XII: 104356-110352	TTAGA	TTAGA	5997	99	
XII: 647012-653005	AGAAG	AGAAG	5994	99	
XII: 952681-958679	GTAAT	ATTAT	5999	99	Recombinant
XIII: 66203-72199	ATTAT	ATTAT	5997	100	
XIII: 266453-272447	ATTTT	ATTTT	5995	98	
XIII: 664357-670350	AAACC	AAACC	5995	98	Stop codon before RT
XIV: 608490-614486	AGATG	AGATG	5997	100	
XV: 322306-328300	GTTCC	CAATT	5995	98	Recombinant; 10bp insertion in LTRs
XV: 1068902-1075215	TTATC	TTATC	5997*	98	*Ty4* solo insertion in 5' LTR
XVI: 52499-58491	ATATG	ATATG	5993	99	

Table 4.7: **Details of *Ty4* elements in *S. cariocanus*.** *length excludes insertion/disruption. ^{E}European type element.

4.5.2 *Ty5* may be extinct in *S. cariocanus*

Only two LTRs are unique to the currently sequenced strains of *S. cariocanus*, both of which are solo copies, suggesting that *Ty5* has been minimally active in the time since the subspecies has split from *S. paradoxus*. However, as no evidence of coding regions were discovered in either strain, it appears that *Ty5* has since become extinct in the currently sequenced strains of *S. cariocanus*. These results are consistent with previous investigations (Liti *et al.*, 2005).

4.5.3 *Ty* insertions are present at translocation and inversion breakpoints in *S. cariocanus*

Liti *et al.* (2006) previously identified four translocations in *S. cariocanus* that were likely the cause of mating incompatibilities with *S. paradoxus*. However, the co-ordinates were not previously reported. The genomic alignment program Mauve (Darling *et al.*, 2010) was used to determine

the locations of translocations (Table 4.8) by comparing the genome assemblies of *S. cariocanus* UFRJ50791 and *S. paradoxus* YPS138. An additional translocation (Table 4.8), as well as inversions were also discovered during this process (Table 4.9; Figure 4.10).

Co-ordinates	Total size (bp)		Co-ordinates	Total size (bp)
II: 486568-706593	220025	⟺	XVI: 693420-1006788	313368
IV: 20599-327300	306701	⟺	XI: 10614-423993	413379
XV: 28192-308223	280031	⟺	IX: 69589-242941	173352
XIV: 626984-847008	220024	⟺	XIII: 23253-156601	133348
XIV: 26918-73590	46672	⟺	XII: 24883-78222	53339

Table 4.8: **Co-ordinates of translocations in *S. cariocanus*.** Co-ordinates of translocations identified in *S. cariocanus* based upon *S. paradoxus* American strain YPS138.

Co-ordinates	Total size (bp)
VII: 705180-750140	44960
VII: 845454-870632	25178
X: 343235-354026	10791
XII: 677821-810902	133081
XIII: 15844-60804	44960
XIV: 615001-857784	242783

Table 4.9: **Co-ordinates of inversions in *S. cariocanus*.** Co-ordinates of inversions identified in *S. cariocanus* based upon *S. paradoxus* American strain YPS138.

The whole genome alignment of *S. cariocanus* (Figure 4.10) allows the visualisation of these breakpoints (Tables 4.8 and 4.9) and the difference in genome length (>20kb) compared to that of *S. paradoxus* YPS138.

This is most likely due to the difference in number of full-length - particularly *Ty4*-like - elements. Due to the lower quality of the strain UFRJ50791 assembly, it was not possible to confidently produce an alignment against the genome of *S. paradoxus*.

The locations of breakpoints were correlated with the output file of RepeatMasker using the custom library in order to search for nearby *Ty* insertions. Full-length recombinant LTRs (i.e. possessing differing TSDs that were unlikely to be the result of mutations) were found within $\sim$1kb of 27% of breakpoints (n=3). Partial LTRs are also present within $\sim$4kb of 45% of breakpoints (n=5).

4.6 *S. mikatae*

Represented by a single strain sequenced by Kellis *et al.* (2003), characteristics of the families in *S. mikatae* are displayed in Table 4.10.

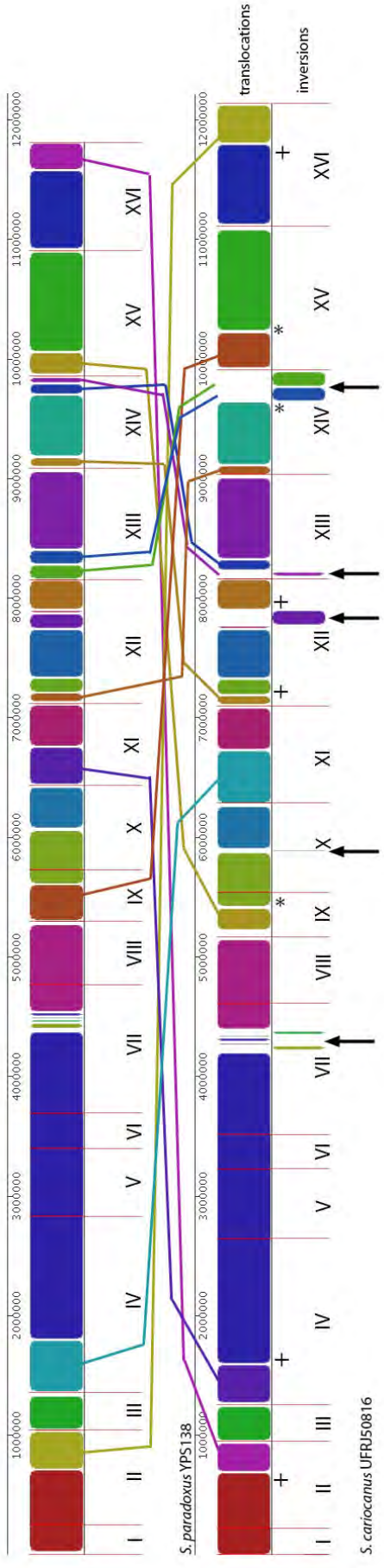

Figure 4.10: **Translocations and inversions in the genome of *S. cariocanus*.** The whole genome alignment of *S. paradoxus* strain YPS138 (top) and *S. cariocanus* UFRJ50816 (bottom) allowed the visualisation of the breakpoints of translocations (connected by coloured lines) and inversions (arrows). Chromosome lengths between the two species differ due to the rearrangements and the presence of an increased number of *Ty4* elements in *S. cariocanus*. Vertical red lines separate chromosomes. *indicates the presence of a potential recombinant LTR (i.e. possessing very different TSDs); +indicates the presence of partial LTR(s).

	Family					
	Ty1	*Ty2*	*Ty3*	*Ty4*L	*Ty4*S	*Ty5*
Nucleotide diversity (π)	0.097	0.078	0.055	0.071	0.118	0.297
Tajima's *D*	**-2.420**	**-2.244**	**-2.201**	**-2.219**	**-2.003**	-0.278

Table 4.10: **Characteristics of *Ty* families in *S. mikatae*.** Two types of *Ty4* LTRs were present, designated long (*Ty4*L) and short (*Ty4*S) (Section 4.6.3).

All families received significantly negative values of Tajima's *D* except for the LTRs of *Ty5*, indicating that this is the only family that is unlikely to have recent ancestry, which is also observed in the corresponding phylogenies of Chapter 5.

4.6.1 The *Ty1/2* superfamily and *Ty5* may be extinct in the reference strain of *S. mikatae*

The reference strain of *S. mikatae* contains both *Ty1/2*-like elements, differentiated primarily using an alignment of coding regions. It is unclear however, as to whether the elements of these families remain functional due to the fractionation of coding regions over multiple contigs (Table 4.11). The full-length *Ty1* element (*Ty1*.1, Table 4.11) contains frameshift mutations that result in the coding region spanning three ORFs. Furthermore, *Ty2* appears be extinct in the reference strain of *S. mikatae*, as all elements are lost by truncation and/or degradation.

Element	Contig	Region(s)	TSDs		Length (bp)
			5'	3'	
Ty1.1	AACH01000018	all	CAAAT	TATAT	5252
Ty1.2	AACH01000199	*gag*; PR; IN; RT	ATAAG	-	4321
Ty1.3	AACH01000084	*gag*; PR; IN	ATACT	-	3129
Ty1.4	AACH01000112	RT; RH	-	ATTAC	2360
Ty1.5	AACH01001507	RH	-	ATTTA	261
Ty2.1	AACH01000570	*gag*; RT; RH	ATTAG	ACACT	2743
Ty2.2	AACH01000109	*gag*; IN	*	*	2297
Ty2.3	AACH01000196	*gag*	-	AATGA	2108
Ty2.4	AACH01001496	RH	-	AAATT	465
Ty2.5	AACH01002226	*gag*	GTTAT	-	740

Table 4.11: **Details of *Ty1/2* elements in *S. mikatae*.** –indicates the region and/or LTR was disrupted by the end of a contig read. *pseudoelement without intact LTRs.

Due to the multiple coding regions present in *S. mikatae* and differing TSDs, they could not reliably be reconstructed. Additionally, the two FLEs (*Ty1*.1 and *Ty2*.1, Table 4.11.) both possess differing TSDs, suggesting that either recombination or the accumulation of mutations in TSDs has occurred in order to cause the non-identical TSDs.

The coding regions of *Ty1/2* elements in *S. mikatae* were analysed for shared nucleotide identity, displayed in Figure 4.11.

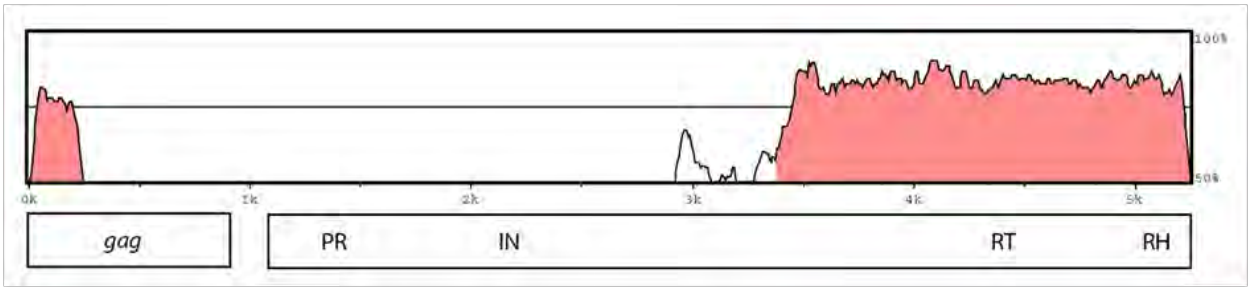

Figure 4.11: **Comparison of *S. mikatae Ty1* and *Ty2* coding regions.** Nucleotide sequences were aligned and shared identity visualised with mVISTA (Frazer *et al.*, 2004). The *Ty2* element is shorter than *Ty1*, lacking PR and IN regions. Elements share 84% identity over 43% of the coding region. However, AA similarity is much higher at 87% over 85% of the translated coding region.

The *S. mikatae* reference strain contains both *Ty1* and *Ty2* LTRs, which share on average 81% identity. Characteristic deletions enabled the LTRs of each family to be distinguished (Figure N.1, Appendix N). Recombination between the two, as seen in *S. cerevisiae*, was not observed in FLEs of *S. mikatae*, but perhaps only in solo LTRs, as only limited evidence was documented (Figure N.1, Appendix N). Furthermore, solo LTRs that could not be reliably assigned to either *Ty1* or *Ty2* subfamilies were also extracted from the reference genome. These may be the remnants of an additional – and now extinct – subfamily, further evidence of hybridisation, or simply degradation of relatively ancient *Ty1/2* LTRs.

Furthermore, a single copy of a *Ty5* element is disrupted by the end of a contig. The 3' end of the element containing IN, RT and RH domains and 3' LTR was recovered from the genome. Although other partial coding regions are present on smaller contigs, the 5' end of the element was not discovered. An alignment of LTRs (Figure N.2, Appendix N) shows that the sequences of this family are particularly susceptible to mutations, with the highest nucleotide diversity of all families in *S. mikatae* (Table 4.11).

4.6.2 *Ty3* is present as a single FLE

A single copy of a *Ty3* FLE is present in the reference strain of *S. mikatae*, but split over two contigs. When aligned to the elements of other *Saccharomyces* species, it appears that approximately 50bp, covering the C-terminus of the IN domain, may be missing from the sequence. LTRs and TSDs are identical (TTAAC), suggesting this is likely to be a relatively new copy. Interestingly, over

60% of solo LTRs possess identical TSDs, indicating that intra-element recombination is the most common reason for loss of elements in the *S. mikatae* genome.

4.6.3 The reference strain of *S. mikatae* contains two subtypes of *Ty4*

Two full-length *Ty4* elements are present in *S. mikatae*, of which only one appears to be autonomous. The autonomous copy, 6.3kb in length, possesses differing TSDs (5'-TACCT-TATAT-3'), either the result of nucleotide changes or LTR-LTR recombination between multiple elements.

The 5'-LTR of the non-autonomous copy is degraded after 220bp, along with the first ∼300bp of *gag*. TSDs (5'-GTATA-CTATG-3') suggest a possible recombination event between multiple elements or mutations in one or both TSDs. In addition, this copy is shorter at 5.2kb, perhaps a result of recombination or assembly error. Figure 4.12 displays the comparison of the two *Ty4* elements in *S. mikatae*.

Although these copies differ within *gag* and PR regions (Figure 4.6.3), their LTRs also differ in length, designated *Ty4*L (long; ∼381bp) and *Ty4*S (short; ∼330bp). Both types are associated with coding regions, with the autonomous element possessing the short type. Many of the short type LTRs possess identical TSDs (*n*=8, 22%), suggesting that despite appearing truncated in comparison to the longer type, they likely retain their functionality as part of FLEs, and that the longer type is the result of an insertion into the LTR of an ancestral element that subsequently spread this type of LTR throughout the genome. Figure 4.13 displays an alignment of both types of LTRs, highlighting the indel region that distinguishes the two.

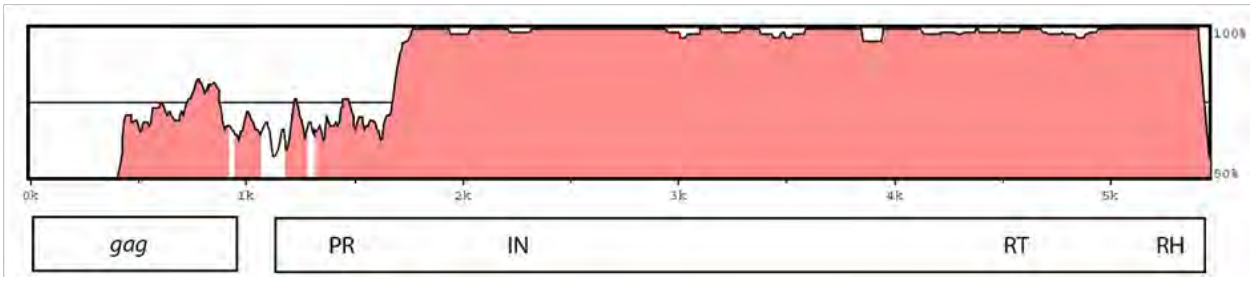

Figure 4.12: **Comparison of *Ty4* elements in *S. mikatae*.** Sequences were aligned and visualised with mVISTA (Frazer *et al.*, 2004). Identity is 91% over 92% of the coding regions. AA similarity was 86% over 92% of the coding regions. Two residue changes are observed between the RT sequences of the two copies. Identity and similarity are mainly lost within *gag* and PR regions.

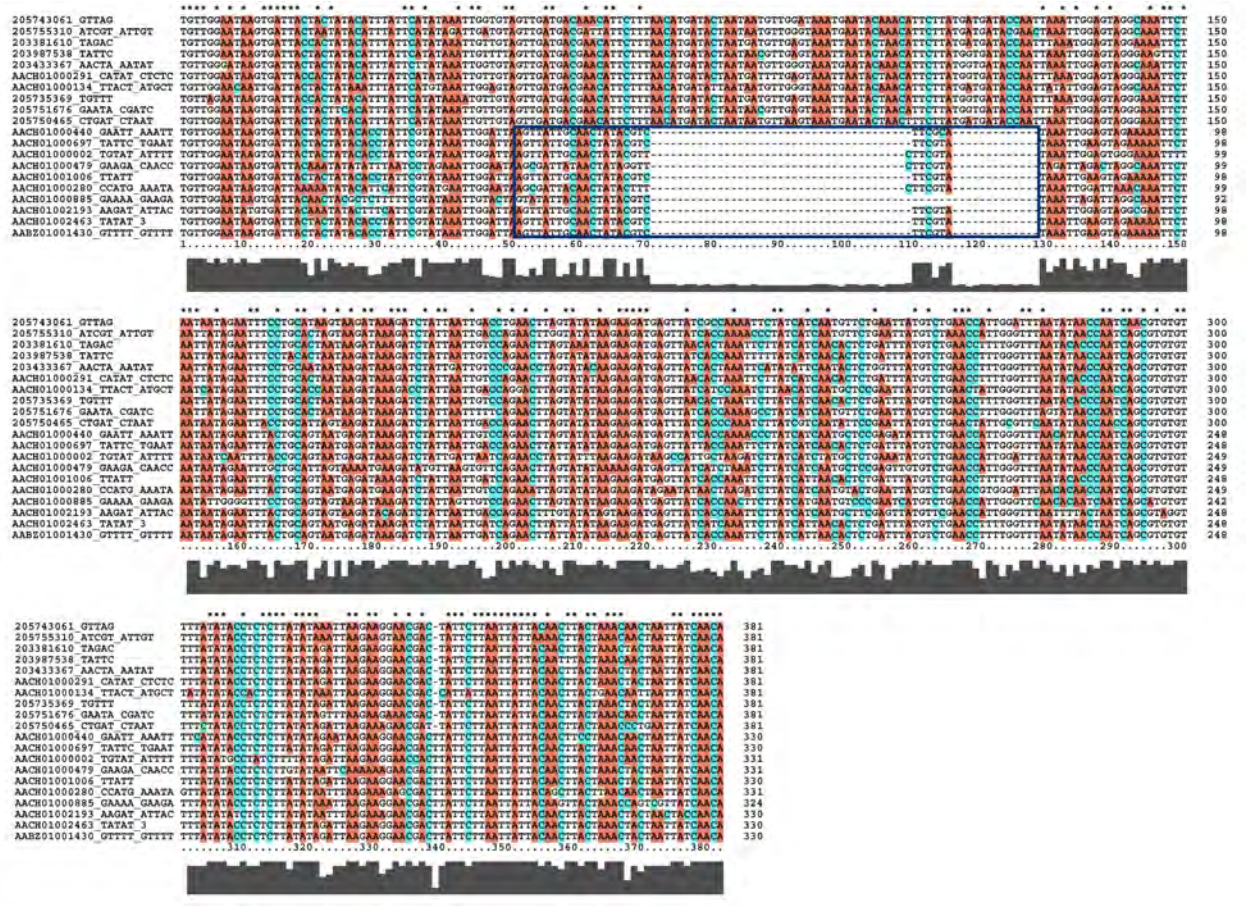

Figure 4.13: **Alignment of *Ty4* LTRs in *S. mikatae*.** The alignment illustrates the indel region that differentiates between the two types of *Ty4* LTRs observed in the reference strain of *S. mikatae*.

4.7 *S. kudriavzevii*

The reference genome of *S. kudriavzevii* (Cliften *et al.*, 2003) contains the highest TE fraction (∼1.5%), as the remaining 19 sequenced isolates (Hittinger *et al.*, 2010) contain on average <0.5%. Table 4.12 displays the characteristics of the *Ty* families present in *S. kudriavzevii*.

	Family						
	Ty1	*Ty3*	*Ty3p*	*Ty4*A	*Ty4*E1	*Ty4*E2	*Ty5*
Nucleotide diversity (π)	0.060	0.227	0.055	0.101	0.082	0.068	0.012
Tajima's *D*	-1.769	-0.750	**-1.693**	-1.022	-1.786	**-1.888**	1.124

Table 4.12: **Characteristics of *Ty* families in *S. kudriavzevii*.** *Ty3p* of *S. paradoxus* was present in *S. kudriavzevii*, and the LTRs of *Ty4* assumed three subgroups: European (*Ty4*E1-2) and American (*Ty4*A).

All families, excluding *Ty5*, received negative values of Tajima's *D*, yet only two are significantly negative. This suggested that all but *Ty3p* and a type of American *Ty4* have recent ancestry in the genomes of *S. kudriavzevii*, which is also observed in the corresponding phylogenies of Chapter 5.

4.7.1 *Ty* content varies depending on genome sequencing quality

The reference strain of *S. kudriavzevii* was isolated in Japan (Cliften *et al.*, 2003) with additional strains isolated in Europe and Japan and sequenced at a lower coverage (Hittinger *et al.*, 2010). No significant difference is observed between the genomic TE fraction of European and Japanese isolates (P=0.19; two-tailed t-test). However, as in the SGRP assemblies of *S. cerevisiae* and *S. paradoxus*, a difference in *Ty* content is observed in the sequenced strains of *S. kudriavzevii*. Reference strain IFO1802 and additional strain ZP591 both possess genomic TE fractions of >1% and are of high quality with few undesignated bases (Ns), whereas the more recently sequenced strains contain far lower genome TE fractions. Inspection of the contigs reveals regions >1kb of entirely Ns, likely the result of unfilled gaps or the masking process, which may account for the low genomic TE content. Raw reads were also obtained from NCBI in order to reconstruct the genomes to a higher quality and therefore obtain their TE sequences. However, this did not increase genomic TE fraction, and the majority of LTRs are found disrupted on short contigs (<0.2kb). Despite the attempts to improve assembly quality, TSDs and flanking regions show that no unique insertions are present outside of the reference strain, therefore sequenced *S. kudriavzevii* genomes do not currently display evidence of strain-specific activity.

4.7.2 *Ty1* and *Ty5* are extinct in *S. kudriavzevii*

No complete *Ty1*-like elements are found in the two high quality genomes of *S. kudriavzevii*, as coding regions are disrupted by more recent solo LTRs. The 18 lower quality genomes possess only partial (<0.5kb) coding regions, however this may be attributed to poor assembly quality. Furthermore, evidence of this species having come into contact with *Ty2* was not found.

Remnants of *Ty5p*-like coding regions were recovered in the reference strain (n=3), but only RH domains are recognisable. A single solo *Ty5* LTR is fixed in all strains with identical TSDs, but the actual LTR sequences vary by $\sim$8bp across strains.

4.7.3 The *S. kudriavzevii* reference strain may contain multiple *Ty3* subfamilies

Evidence of *Ty3* elements is present in each of the two high quality genome sequences of *S. kudriavzevii*, the details of which are displayed in Table 4.13. In the reference strain, element *Ty3*.1 is the only FLE, whereas the second element, *Ty3*.2, is lost after $\sim$3kb due to a later *Ty1* solo insertion. The final evidence of coding regions containing RT and RH domains is present on

a short contig and the remainder of the element could not be located (*Ty3*.3, Table 4.13). This partial element shares 52% similarity with the FLE in RT domains, and 56% similarity with that of *Tnd3* of *Naumovozyma dairenensis* (Appendix P). Coding regions without flanking DNA have been treated cautiously throughout this work, but as *Ty3*.3 is also found in other strains of *S. kudriavzevii* (*n*=2; IFO10990 and IFO10991), this indicates it is unlikely the result of contamination from another species. As in the reference strain, the coding region is unfortunately present on a short contig (∼1kb) in these two strains and further regions of this copy could not be located.

Element	TSD(s)	Contig	Domains	Length (bp)	Notes
Ty3.1	5'-CTCTC TTATC-3'	AACI03000452	all	4754	potentially recombinant
Ty3.2	5'-GCCTT	AACI03000505	gag; PR; RT	3604	degraded 3' end
Ty3.3	-	AACI03000958	RT; RH	1581	spans full contig

Table 4.13: **Details of *Ty3* elements in the reference strain of *S. kudriavzevii***

4.7.4 Multiple subtypes of *Ty4* are present in the populations of *S. kudriavzevii*

S. kudriavzevii genomes contain up to three subtypes of *Ty4*, depending on geographical origin. While the European type of *Ty4* is present across the genomes of both populations, those strains isolated in Japan contain an additional subfamily with LTRs more similar to those of the American type of other *Saccharomyces* species (Figure 4.14).

All American type LTRs are solos, and are easily distinguished via multiple indels and sequence differences. The LTRs of the European type elements are divided further based upon length (*Ty4*E1-2, Figure 4.14). Both subtypes of European LTRs are associated with coding regions, of which only the longer type (*Ty4*E1) is autonomous. Although a 5' LTR of *Ty4*E2 is present in 20% of strains (*n*=5), the remainder of the element could not be located in the genomes. In the remaining strains, this insertion is present as a solo LTR.

4.8 *S. arboricola*

The reference strain of *S. arboricola* was isolated in China and sequenced by Liti *et al.* (2013). A further nine genomes were recently sequenced by Gayevskiy and Goddard (2016), having discovered a population in New Zealand (NZ) which shares its habitat with strains of *S. eubayanus* and *S. cerevisiae*. Although Liti *et al.* (2013) briefly commented on the presence of a *Ty2*-like element

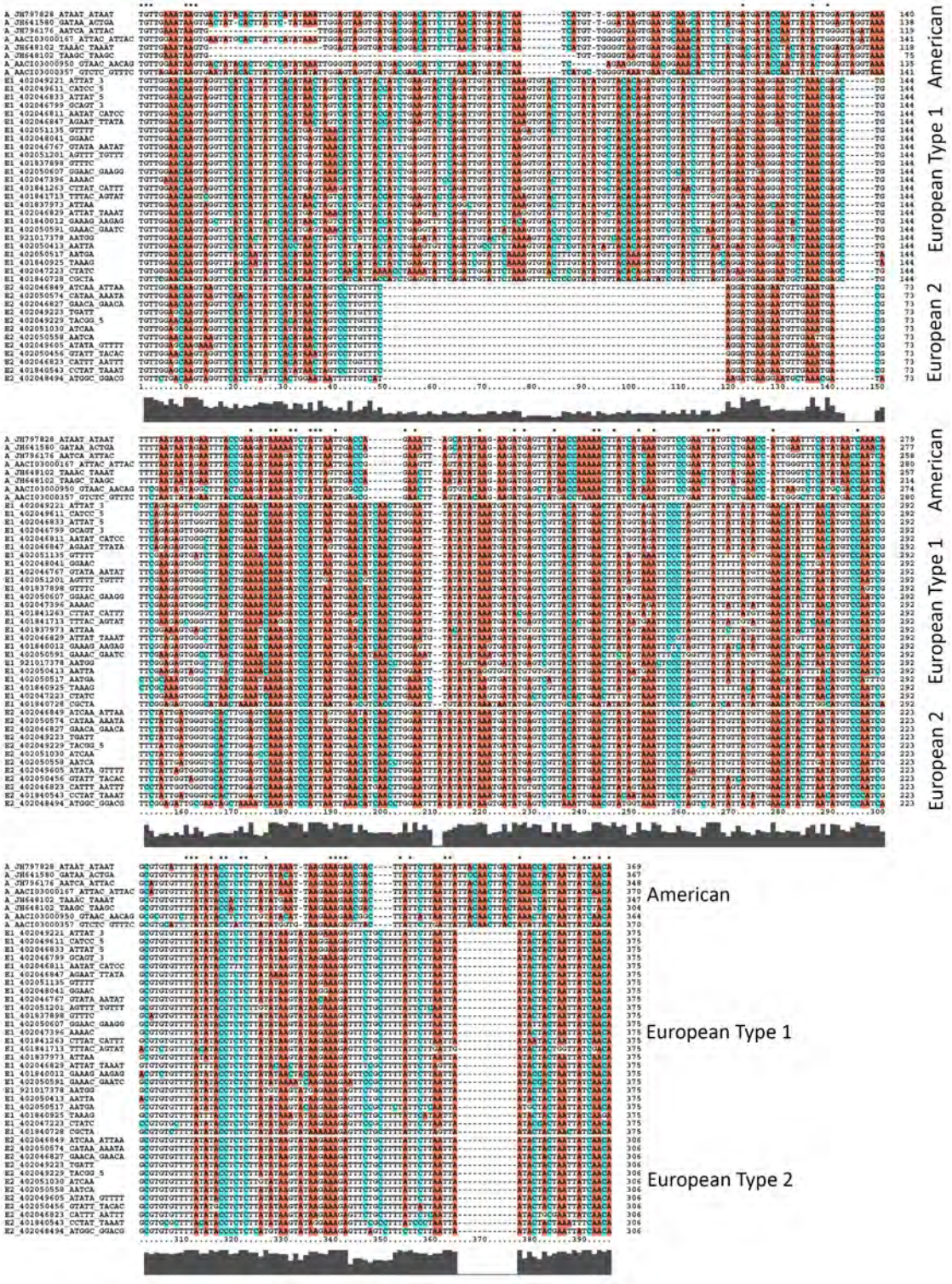

Figure 4.14: **Alignment of *Ty4* LTRs in *S. kudriavzevii*.** The LTRs designated "American" share identity with those types of *Ty4*A of other *Saccharomyces* species. The three types of *Ty4* LTRs are separated by horizontal white lines.

in the reference strain of *S. arboricola*, the study did not investigate the copy numbers or dispersal of solo LTRs of *Ty2* or any other families.

Only *Ty2* received a significantly negative value of Tajima's *D* and therefore is consistent with

	Family				
	Ty1	*Ty2*	*Ty3*	*Ty3p*	*Ty4*
Nucleotide diversity (π)	0.241	0.085	0.136	0.033	0.291
Tajima's *D*	-0.511	**-2.181**	-1.444	-1.398	-1.418

Table 4.14: **Characteristics of *Ty* families in *S. arboricola*.**

this being the only family with recent ancestry in the genomes of *S. arboricola*, an observation which is further supported in the corresponding phylogenies of Chapter 5.

4.8.1 The *Ty1/2* superfamily is present in both populations of *S. arboricola*

Ty1 sequences are found in the genomes of *S. arboricola*, but the family has undergone stochastic loss as only solo LTRs remain. No insertions are shared by both populations (Figure N.3, Appendix N), suggesting that activity in this family occurred after the population split. In addition, no LTRs possess identical TSDs, indicating that LTR-LTR recombination occurred between multiple elements, or that the age of the family has allowed mutations in the TSDs to accumulate.

In contrast to *Ty1*, *Ty2* is widespread throughout both populations, and all strains contain at least one FLE in addition to multiple solo LTRs. NZ strain P3H5 is alone in containing an additional, partial element. Coding regions of elements in all strains share >90% similarity, with that of the reference strain being most degraded. As the FLEs of the NZ strains are highly conserved and contain no stop codons, they are likely to be autonomous. Unique TSDs of FLEs and solo LTRs in each genome indicate strain-specific activity. Sequence identity of *Ty2* LTRs is higher with the sequences of *S. cerevisiae* (94%) than those of *S. mikatae* (83%). Intra-element recombination is observed in 10% of solo *Ty1* LTRs (n=1), while this is increased in *Ty2*, with $\sim$28% of solo LTRs possessing identical TSDs (n=55). No evidence of recombination is observed between *Ty1/2* LTRs (Figure N.3, Appendix N).

4.8.2 Population isolation affects *Ty3* in *S. arboricola*

Seven solo insertions are shared across all NZ strains, indicating they were present in the last common ancestor of the population (Figure 4.15) but not of the species, as these are not fixed in both NZ and Chinese populations. The reference strain however, contains unique insertions (n=14), of which $\sim$80% appear to be distinct, possessing a higher shared identity with *Ty3p* of *S. paradoxus* than the predominant type of *Ty3* LTRs in this species.

The two subtypes of LTRs share 78% nucleotide identity over 71% of the ∼360bp alignment, and should therefore be considered different families according to the 80-80-80 rule proposed by Wicker *et al.* (2007). Recombination between the two types was not observed in the reference strain. All sequenced strains contain consistent regions (<0.3kb) of coding sequences, but *Ty3* may ultimately be extinct in both known populations of *S. arboricola*.

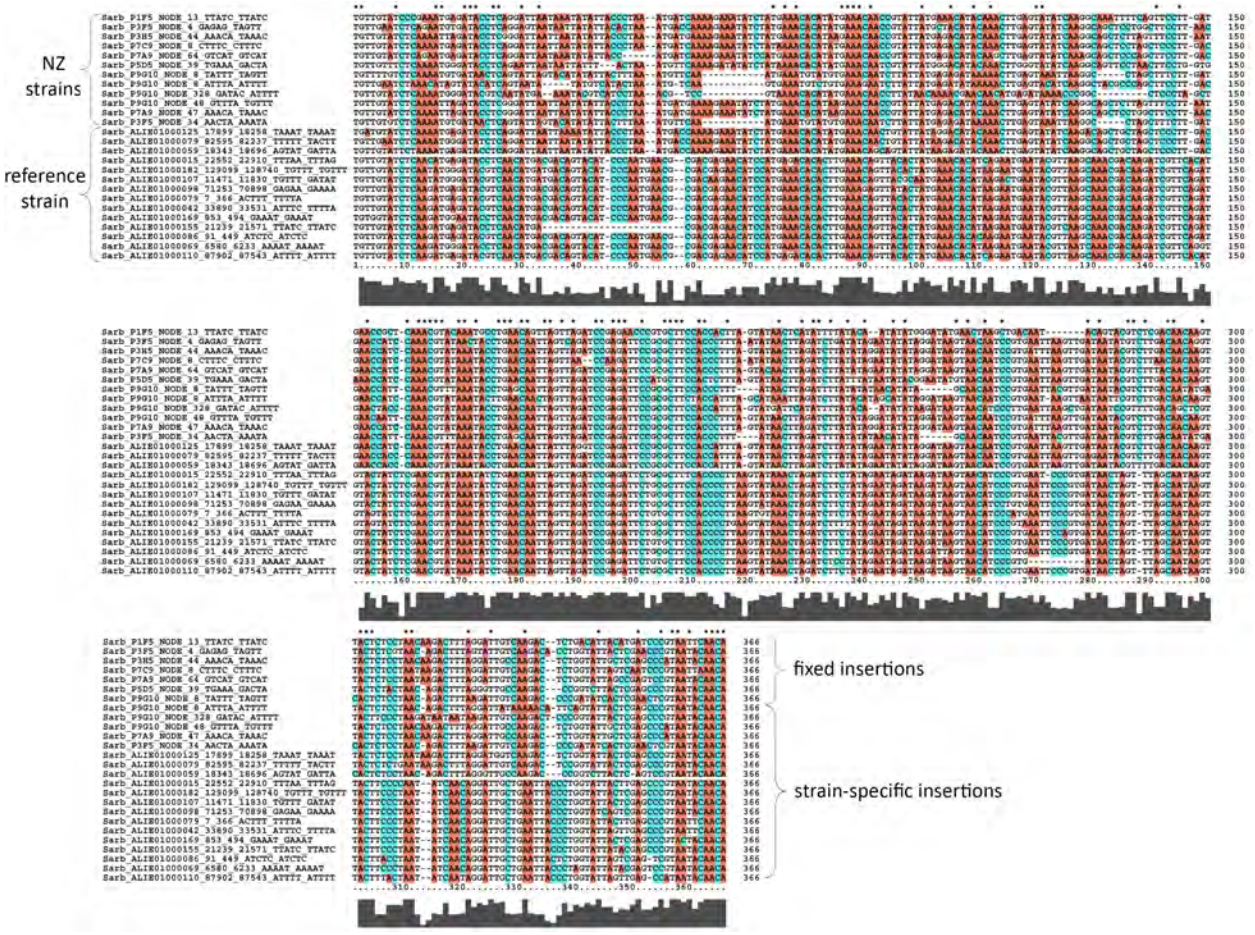

Figure 4.15: **Alignment of *Ty3* LTRs in *S. arboricola***. A horizontal white line separates the two distinct types of LTRs in the alignment. The first group of insertions (*n*=15) is more closely related to those of *S. kudriavzevii* and shared between both populations, therefore more likely to be the ancestral *Ty3* family. The second group, confined to the Chinese reference strain (*n*=11), shares identity with LTRs of *S. paradoxus*. Insertions within NZ strains and the reference strain, and fixed and strain-specific are distinguished with parentheses.

4.8.3 *Ty4-5* are lost in both populations of *S. arboricola*

Partial *Ty4* RT regions are present in both the reference strain and P5D5 of the NZ population. In all other strains, only solo LTRs are recovered from the genomes. It is therefore most likely that *Ty4* has been lost from *S. arboricola* as a species. Additionally, degraded *Ty5* coding regions are

present in the reference strain and almost half of the NZ strains (*n*=4). Whereas a solo *Ty5* LTR is present in the reference strain of *S. arboricola*, no NZ strains contain solo LTRs.

4.9 *S. eubayanus*

The most recent species to have its genome sequenced (Baker *et al.*, 2015), *S. eubayanus* is the previously unknown parental species of hybrid *S. pastorianus* along with *S. cerevisiae* (Libkind et al., 2011). The work here is the only examination of its *Ty* families to date, collecting 32 currently sequenced genomes, isolated across the Americas and NZ (Baker *et al.*, 2015; Gayevskiy and Goddard, 2016; Peris *et al.*, 2016). Table 4.15 displays the characteristics of the families in *S. eubayanus*.

	Family			
	Tse1	*Tse4*	*Ty4*A	*Tsu4*
Nucleotide diversity (π)	0.046	0.237	0.029	0.164
Tajima's *D*	-1.785	-1.183	-0.802	-1.602

Table 4.15: **Characteristics of *Ty* families in *S. eubayanus*.**

No families received significantly negative values of Tajima's *D*, which is somewhat at odds with the short-branched phylogenies of Chapter 5. Insignificant negative values of *D* are consistent with neutral coalescence. Furthermore, nucleotide diversity differs by an order of magnitude.

4.9.1 *Tse1* is an autonomous *Ty1*-like family in *S. eubayanus*

Coding regions are present in >70% of genomes (*n*=23) and autonomous FLEs were extracted where assembly quality allowed. Elements share >90% similarity, with those of strains CBS12357 and yHKS210, both isolated in South America, containing putative envelope protein domains ($5.35e^{-04}$) located between IN and RT regions.

Tse1 is likely to be the most recently active family, as in all strains FLEs or solo LTRs are found to disrupt previous insertions of the *Tse4* family.

4.9.2 *Ty5* and *Ty3* are predominantly absent in *S. eubayanus*

The reference strain FM1318 contains single solo LTRs of the *Ty3* and *Ty5* families, which are not observed as full-length LTRs in any other strain surveyed. As the *Ty5* insertion, along with its

flanking DNA, is shared by sister species *S. uvarum*, it suggests that *Ty5* became extinct in their last common ancestor, and that it has never been active in either of the extant species.

In addition to partial *Ty3* LTRs in each *S. eubayanus* genome, four strains (reference FM1318, CRUB1761, yHRVM108 and NZ strain P1C1) all contain multiple regions of >0.6kb that appear to be remnants of partial RH, PR and *gag* domains, each sharing ∼50% similarity with *Ty3p* of *S. paradoxus*. The partial RH domains and regions downstream were aligned to check for the presence of an LTR. However, when this sequence is used as a BLAST query, no further full-length sequences are found, therefore it was concluded that this putative LTR may be the only copy remaining in the genomes of *S. eubayanus*, and the state of this family is similar to that of *Ty3p* in *S. cerevisiae*.

4.9.3 *Ty4* possesses a complex history in *S. eubayanus*

Evidence of multiple *Ty4*-like families was discovered across the genomes of *S. eubayanus*. All strains contain *Tse4*, therefore it is likely that this is the ancestral family in this species. However, *Tse4* is lost to recombination, mutation or disruption by other insertions in more than half of strains surveyed (*n*=20). Figure 4.16 illustrates a copy of *Tse4* in the Argentinian strain CRUB1971, disrupted by more recent *Tse1* insertions. Disruptions similar to this are also observed in a further 25% of strains (*n*=13).

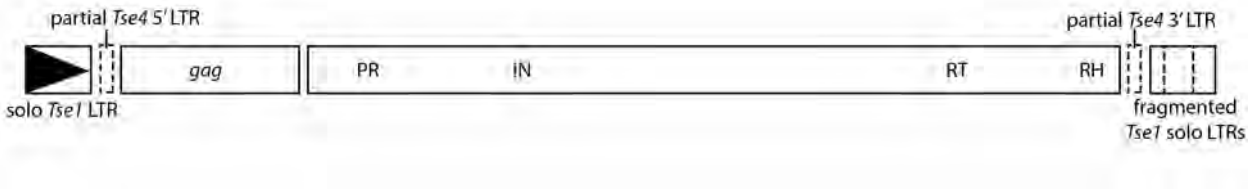

Figure 4.16: **Commonly observed disrupted state of *Tse4* elements in *S. eubayanus*.** Strain CRUB1971 contained nesting of insertions which was also similarly observed in other strains of *S. eubayanus*. The internal *Tse4* coding regions were free of mutations, suggesting that *Tse1* transposition into the existing *Tse4* elements occurred relatively recently, as mutations had not had chance to accumulate.

Multiple partial and autonomous copies of *Tse4* were recovered from NZ strain P1C1 and Argentinian strains. Interestingly, the only regions remaining in strain CRUB1977 are that of *gag* and the associated 5' LTR (*n*=7). Elements of two strains, CRUB1971 and yHCT92, contain putative extra domains that share similarity with the COG3942 superfamily, a surface antigen protein (e values 2.10e^{-04} and 4.76e^{-06}, respectively).

A high number of strains (*n*=17) also possess solo LTRs of *Tsu4*, the *Ty4*-like family of sister species *S. uvarum*. This may be the result of the tendency for both species to share the same habitat (Gayevskiy and Goddard, 2016), resulting in frequent hybridisation. The reference strain contains two *Tsu4* RT pseudogenes on chromosome XIV, but no full-length *Tsu4* elements were recovered from any surveyed strain.

Furthermore, evidence of the American *Ty4* seen in *S. paradoxus* and *S. cariocanus* is also present in *S. eubayanus*. This family is widespread, as all surveyed strains contained at least partial LTRs, with >80% of strains containing a minimum of one solo LTR (*n*=28). LTRs of *S. eubayanus* and *S. paradoxus* share >93% identity, differing primarily by a 12bp indel. The strain yHRVM108, isolated in the USA, contains the highest copy number of this subfamily (*n*=127). However, due to sequencing and assembly quality, and disruptions by later (usually *Tse1*) insertions, no American-type FLEs are able to be extracted from any *S. eubayanus* strain. It therefore remains to be seen if this family is still functional in this species as a whole.

Evidence of the European *Ty4* is confined to Argentinian strain CRUB1971, in the shape of a single solo LTR that possesses *S. eubayanus* flanking DNA. As flanking DNA from another species would have strongly suggested the occurrence of introgression, it is therefore unclear as to where this insertion originated.

4.10 *S. uvarum*

Unlike *S. eubayanus*, *S. uvarum* has been isolated worldwide (Sylvester *et al.*, 2015). Despite the large number of strains surveyed (*n*=51), little variability in *Ty* insertions is observed between the genomes. The reference strain, MCYC623, contains the highest copy number and its genomic fraction of insertions is double that of all other isolates (∼1.4%). The characteristics of families in this species are displayed in Table 4.16.

	Family			
	Tsu1	*Tsu4*	*Tse4*	*Ty4*A
Nucleotide diversity (π)	0.164	0.212	0.104	0.107
Tajima's *D*	**-2.095**	-1.067	-1.287	1.212

Table 4.16: **Characteristics of *Ty* families in *S. uvarum*.**

Tsu1 is the only family to receive a significantly negative value of Tajima's *D*, consistent with

recent ancestry, which is also observed in the corresponding phylogenies of Chapter 5. The American type of *Ty4* returned a positive value of *D*, possibly the result of further subdivision in *Ty4*.

4.10.1 *Tsu1* is widespread across all populations of *S. uvarum*

A single copy of a FLE is present in multiple strains (*n*=12), which share >98% similarity. The Argentinian strain, CRUB1784, lacks any evidence of coding regions, while only partial elements are present in the remaining strains (*n*=36), possibly due to assembly quality. Strain-specific solo insertions (*n*=27) show that this family has been active since the last common ancestor.

4.10.2 *Ty5* and *Ty3* are absent in *S. uvarum*

DNA flanking a single solo *Ty5* LTR in the genomes of *S. uvarum* is shared with *S. eubayanus* (Section4.9.2). As the state of *Ty5* in *S. uvarum* strains is highly similar to that of *S. eubayanus*, it is indicative that the loss of this family most likely occurred in their common ancestor.

Unlike *S. eubayanus* however, strains of *S. uvarum* contain partial *Ty3*-like LTR sequences, but no evidence of coding regions. The loss of *Ty3* in *S. uvarum* therefore likely occurred independently to that of *S. eubayanus*. Potential partial LTR sequences were collected for phylogenetic analysis (Chapter 5).

4.10.3 *Tsu4* in *S. uvarum* differs from *Tse4* in *S. eubayanus*

As in *S. eubayanus*, evidence of multiple *Ty4*-like families is present in strains of *S. uvarum*. The reference strain MCYC623 contains three full-length *Tsu4* elements, which differ from the *Tse4* family of *S. eubayanus* (Figure 4.17). Further strains (*n*=28) contain a single FLE. Strain-specific activity is evident (*n*=30 insertions), indicating that this family is likely still active in the populations of *S. uvarum*. Only solo LTR evidence of *S. eubayanus Tse4* is found across most strains (*n*=44), therefore this “swapping” of families between the two species has not succeeded in the long-term in either direction.

The American type of *Ty4* observed in *S. eubayanus*, *S. cariocanus* and *S. paradoxus* is also present in strains of *S. uvarum* isolated in the Americas only, supporting the hypothesis that this subfamily is likely confined to the American continent. As in *S. eubayanus* however, due to sequencing and assembly quality, and disruptions by later *Tsu1* insertions, no American-type FLEs

Figure 4.17: **Comparison of nucleotide coding regions of *Tse4* and *Tsu4*.** Sequences were aligned and visualised with mVISTA (Frazer *et al.*, 2004). The element in *S. eubayanus* CRUB1971 is compared with canonical *Tsu4* from reference strain of *S. uvarum.*

were extracted from any *S. uvarum* strain. It therefore remains to be seen if this family is still functional in this species as a whole.

4.11 The insertion preference of the *Ty5* family into telomeric regions is *Saccharomyces*-specific

The distance between chromosome ends and *Ty5* insertion sites for all *Saccharomyces* species was recorded during genome analysis. *Ty5* has previously been documented to prefer insertion sites within telomeric regions in *S. cerevisiae* (Kim et al., 1998) and *S. paradoxus* (Zou et al., 1996). Figure 4.18 displays the insertion preference of *Ty5* in all species of *Saccharomyces*, showing that the preference is not limited to *S. cerevisiae* and *S. paradoxus*, but may be a family-specific attribute.

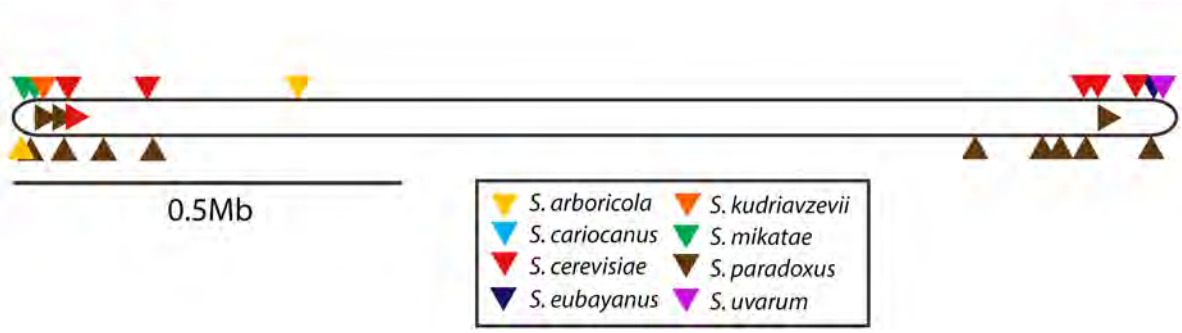

Figure 4.18: ***Ty5* insertion site preference.** The distances from chromosome ends of *Ty5* insertions in all *Saccharomyces* species is plotted onto a hypothetical chromosome in order to visualise insertion site preference. Both upper and lower edges of the hypothetical chromosome are utilised for clarity only. Horizontal arrows represent full-length insertions, where as vertical arrows are representative of solo LTRs. The figure is drawn to scale.

Conversely, insertion preferences are not observed in the families of any *sensu lato* species. However, *Saccharomyces* assemblies are unusual in that most are constructed into chromosomes, whereas this is not the case for the majority of *sensu lato* species surveyed in this study (data not shown).

4.12 Discussion

Few species of the *Saccharomyces sensu stricto* complex have previously been examined for *Ty* content (Fingerman *et al.*, 2003; Liti *et al.*, 2005; Carr *et al.*, 2012). Therefore, this chapter details one of the first explorations of the genomes of previously unexplored *Saccharomyces* species. This forms part of a broader analysis of *Ty*-like insertions in budding yeast, the RT and LTR sequences of which provides the basis of the phylogenetic analyses of Chapter 5. This section discusses briefly the findings of this chapter, while an in depth discussion of the evolutionary relationships between the *Ty* elements of all surveyed species is presented at the end of Chapter 5.

A previous investigation into genomic *Ty* content depending on sequencing technique failed to find a significant difference in *S. cerevisiae* (Bleykasten-Grosshans *et al.*, 2013). However, in both *S. cerevisiae* and *S. paradoxus* strains here, a significant difference was observed when genomes were sequenced using Illumina and PacBio/Nanopore techniques. This may be due to the more extensive custom library used when screening genomes, and the fact that individual strains, sequenced with both techniques, were directly compared, rather than the 41 strains sequenced by one of two techniques as in the study by Bleykasten-Grosshans *et al.* (2013).

Differing genomic *Ty* content was also previously investigated in relation to strain source and usage (Bleykasten-Grosshans *et al.*, 2013). Analyses of individual strains of varying backgrounds have also been conducted, with the result of wine and laboratory strains typically displaying the highest *Ty* content (Lesage and Todeschini, 2005; Wei *et al.*, 2007; Borneman *et al.*, 2008; Argueso *et al.*, 2009; Fritsch *et al.*, 2009; Novo *et al.*, 2009; Bleykasten-Grosshans *et al.*, 2013). Investigations into a far larger dataset here identified significant differences between strains grouped by usage and isolation source.

Hybridisation and the *Ty1/2* superfamily

The *Ty1/2* superfamily was present in all *Saccharomyces* species to varying degrees. A number of new observations, such as variation in *Ty1* 3' boundaries and alternative recombination points in *Ty1/2* of *S. cerevisiae* suggest that the superfamily is still structurally evolving. Furthermore, the transposition of this family into the highly active *Ty4* of *S. cariocanus* indicates there may be multiple active families across the species.

To date, recombination between the elements of the *Ty1/2* superfamily has only been observed in *S. cerevisiae* (Jordan and McDonald, 1998; Carr *et al.*, 2012). Despite the thorough analysis

of insertions in all *Saccharomyces* species, evidence of recombination was not observed in the elements of other species. Additional recombination breakpoints in *S. cerevisiae* were however discovered, leading to the conclusion that breakpoints are not confined to particular regions of LTRs as previously seen in the work of Jordan and McDonald (1998) and Carr *et al.* (2012). Recombination may occur ectopically between DNA sequences or during the complex template switching event between RNAs in the VLP (Temin, 1991; Jordan and McDonald, 1998). Recombination may therefore require elements of each subfamily to be active at the same time, in order for transportation to the VLP to occur. This may in part account for the lack of recombination in for example *S. mikatae*, as it appears that the superfamily may be extinct in this species, and may not have been active at the same time in the past. However, this is based purely on the sequenced genome of a single strain, and may not reflect the state of this superfamily in the species as a whole. Similarly in *S. arboricola, Ty1* shows so little past activity that the chances of this undergoing recombination with *Ty2* are particularly low. Recombination has been observed between *Ty1* and *Ty2* in *S. cerevisiae* most likely due to the high rates of activity in both of these subfamilies (Carr *et al.*, 2012; Kim *et al.*, 1998).

Ty3 differs between species of *Saccharomyces*

The presence of *Ty3p*, a distinct yet homologous element to *Ty3* of *S. cerevisiae*, was reported in *S. paradoxus* by Fingerman *et al.* (2003). As expected, evidence of species-specific *Ty3*-like families were discovered here in the remaining species of *Saccharomyces*. However, two species contained evidence of multiple *Ty3*-like families: *S. kudriavzevii* and *S. arboricola*. The former appears to contain distinct copies of elements, one of which was more *Ty3p*-like, whereas population isolation observed in *S. arboricola*, represented by Chinese and New Zealand strains, affects the family of *Ty3* by way of divergence between the two populations, which is not observed in other species. However, unlike the other *Ty* families, *Ty3* is predominantly absent in *S. uvarum* and *S. eubayanus*.

Unexpected *Ty4* activity

Rather than simply the presence of *Ty* families in *Saccharomyces* varying according to geographical origin (Liti *et al.*, 2005; Bleykasten-Grosshans *et al.*, 2013), the unusual case of the *Ty4* family recorded here was perhaps the result of sequence divergence or emergence of subtypes. Bergman

(2018) independently reported the horizontal transfer of *Tsu4* of *S. uvarum* into populations of *S. paradoxus* and *S. cariocanus*. Although the work here supports the presence of a *Tsu4*-like family in these species, conclusions drawn here differ from those offered by Bergman (2018). All *Ty4*-like families share some degree of identity, yet the recent invader of American populations of *S. cerevisiae, S. paradoxus* and *S. cariocanus* does not appear to be that of *Tsu4* from *S. uvarum* for a number of reasons. Firstly, multiple types of *Ty4*-like families were recovered from both *S. eubayanus* (Section 4.9.3) and *S. uvarum* (Section 4.10.3). Unfortunately the potentially full-length American type copies were unable to be extracted in full. Secondly, the American elements share $\sim$90% identity with those of *Tsu4* (Figure 4.8), which is similar to the relationship between *Ty1* of *S. cerevisiae* and *Ty1p* of *S. paradoxus*. Thirdly, LTRs of *Ty4A* and *Tsu4* only share identity towards the 3' boundary (Figure 4.8 B), yet *Ty4A* LTRs are present in *S. eubayanus* and *S. uvarum* alongside their endogenous *Tse4* and *Tsu4*, respectively. These reasons indicate that although the HT event reported by Bergman (2018) is very likely to have occurred, it was unlikely to have involved *Tsu4*, but instead the HT of a different subfamily of *Ty4*.

The divergence of *Ty* families into subtypes with population isolation has not previously been reported in the species of *Saccharomyces*. Whereas the impact of insertions driving population divergence is documented in a variety of organisms (e.g. Begin and Schoen, 2007; Senerchia *et al.*, 2015; Oppold *et al.*, 2017), a review of the current literature revealed only two instances of divergence of TEs with host population. *Arabidopsis* (Lockton *et al.*, 2008; Lockton and Gaut, 2010) and Mediterranean grass show divergence in TE families, where polymorphisms are specific to populations (Stritt *et al.*, 2018). Lerat *et al.* (2003) found that retrotransposons are far less likely to diverge than DNA transposons, indicating that both differing levels of selection and the method of transposition may affect their tendency to diverge. Divergence of TE families is possibly a method by which insertions can escape deletion via LTR-LTR recombination of copies at differing sites (Charlesworth and Langley, 1989; Petrov *et al.*, 2003) and small RNA targeting in eukaryotes (reviewed in Castillo and Moyle, 2012). Therefore, the divergence of *Ty4* in this manner may well be an alternative method of escaping genome defences, should this family be targeted in a similar fashion to that of *Ty1* (Tucker *et al.*, 2015).

Ty5

The extinction of *Ty5* in the majority of *Saccharomyces* species has previously been attributed to the age of this family (Neuvéglise *et al.*, 2002), yet this does not fully explain its tendency for stochastic loss. Enough insertions were present in these species to allow the insertion site preference of *Ty5* to be surveyed however. While previously determined in *S. cerevisiae* and *S. paradoxus* (Zou *et al.*, 1995, 1996*a*,*b*), the work here provides evidence for the insertion site preference of telomeres and heterochromatin as being *Ty5* family-specific.

Insertion site preference of TE families is also observed in other species, such as the *P* and *Galileo* elements of *Drosophila* (Spradling *et al.*, 2011; Gonçalves *et al.*, 2014) and *Tos17* of *Oryza* (Miyao *et al.*, 2003). Location preference of *Ty5* insertions in *S. cerevisiae* and *S. paradoxus* was established in the 1990s (Zou *et al.*, 1996*a*,*b*; Zou and Voytas, 1997; Ke *et al.*, 1997; Xie *et al.*, 2001; Gai and Voytas, 1998), and now with increased data, the insertion preference across the genus could be examined. It is now clear that this insertion site preference is *Ty5* specific, and only apparent in *Saccharomyces* species, as a preference was not observed in the *sensu lato* species surveyed for the phylogenetic analyses of Chapter 5. However, this may be due in part to the fact that few *sensu lato* species assemblies are completed to chromosome standard, meaning the distance between telomere and insertion site was not always able to be determined.

4.13 Summary and conclusions

This chapter explores the genomic TE content of all available *Saccharomyces* genomes. While some results are consistent with previous findings, new insights were gained into the effects of population isolation upon *Ty* families, including the division into subfamilies of varying activity levels. In the preparation for the phylogenetic analyses of Chapter 5, the work here documents the first exploration of elements in *S. eubayanus, S. uvarum* and *S. arboricola*. These new genomes provide evidence for the insertion site preference of *Ty5* elements as family-specific, having predominantly targeted telomeric regions before extinction.

Chapter 5

Phylogenetic analysis of TEs in *Saccharomyces sensu lato* and Saccharomycetaceae species

Studies concerning the TEs in and beyond *Saccharomyces* species typically focus upon small-scale analyses of species such as *S. cerevisiae* (Kim *et al.*, 1998; Carr *et al.*, 2012; Bleykasten-Grosshans *et al.*, 2013), *Schizosaccharomyces* (Bowen *et al.*, 2003; Rhind *et al.*, 2011), *Candida* (Goodwin and Poulter, 2000; Goodwin *et al.*, 2003) and *hAT* transposons in *Lachancea* (Sarilar *et al.*, 2015). Neuvéglise *et al.* (2002) produced one of the first phylogenetic analyses of elements in hemiascomycete yeasts, which surveyed the genomes of 13 species sequenced by the Génolevures Consortium (Souciet *et al.*, 2000). Muszewska *et al.* (2011) expanded this to 49 genomes of diverse fungal species available at the time, and more recently, a thorough investigation of *Ty1/copia* elements in filamentous fungi was published by Donnart *et al.* (2017). A focussed update of the elements within yeast is now required in order to utilise the data of recent large-scale genome sequencing projects such as those of Riley *et al.* (2016) and Vakirlis *et al.* (2016), among others.

The extent of the horizontal transfer (HT) of elements across species barriers has not been elucidated. Most reports have so far been serendipitous findings, such as the *mdg3* family in *Drosophila* (Syomin *et al.*, 2002). In yeast, Neuvéglise *et al.* (2002) identified the potential transfer of a *Ty1*-like family into *Lachancea*, but could not identify the *Saccharomyces* source. Liti *et al.* (2005) noted discrepancies in the distribution of *Ty2*, and suggested an event between *S. cerevisiae* and *S. mikatae*, with *S. cerevisiae* later confirmed as the recipient (Carr *et al.*, 2012). No systematic screening for HT across yeast has been performed, which is the aim of this chapter.

Chapter 4 documents the collection of element sequences from *Ty*-like families in *Saccharomyces* species. TE sequences were also collected from all *sensu lato* species containing *Ty*-like elements. In this chapter, two methods of phylogenetic analysis, Maximum Likelihood (ML) and

Bayesian Inference (BI), are used to examine and establish the evolutionary relationships of the *Ty*-like RT and LTR sequences, which also allow for a systematic screening for HT events to be performed. The proportion of families that have undergone stochastic loss in the available data is also recorded, and visualised by the presence of LTRs but not RT sequences in the corresponding family trees. Finally, the potential histories of the *Ty*-like families are discussed, with a focus on the elements in the *Saccharomyces sensu stricto* complex.

5.1 Background to RT and LTR Phylogenetics

5.1.1 Maximum Likelihood and Bayesian Inference

Two complementary methods of phylogenetic inference were chosen to conduct the analyses of RT and LTR sequences, as correlation between the topologies of both ML (Felsenstein, 1985) and BI (Rannala and Yang, 1996) provide far more confident conclusions than that of a single method.

ML relies upon bootstrapping (BP) to infer confidence, selecting the most probable tree topology on the basis of the highest logarithmic likelihood scores, given the data and specified model of evolution (Wiley and Lieberman, 2011). The Bayesian method combines the likelihood scores and prior probabilities of a phylogeny to produce support values known as posterior probability (PP), using the Markov chains Monte Carlo (MCMC) algorithm (Huelsenbeck *et al.*, 2001). Trees are created over generations and sampled at fixed intervals, with the best estimate of the phylogeny possessing the highest PP (Yang and Rannala, 1997).

Both are highly conservative methods and allow for complex evolutionary models, but the bootstrapping of ML is less susceptible to strongly supporting false topologies (Douady *et al.*, 2003). In contrast, the Bayesian method tends to inflate support (Rannala and Yang, 1996), but may provide closer estimates of the true probabilities of clades (Wilcox *et al.*, 2002). Support values are not interchangeable or directly comparable (Table 5.1; Douady *et al.*, 2003).

Support	Inference method	
	ML	BI
Strong	≥70%	≥0.95
Moderate	50-69%	0.70-0.94
Weak	<50%	<0.70

Table 5.1: **Support value thresholds for ML and BI methods.** The ranges considered for strong, moderate and weak inference support determined for ML (Hillis and Bull, 1993) and BI (Rannala and Yang, 1996).

In all trees, the ML topology was used as the basis of the phylogeny with added BI support values. This was an aesthetic preference and to avoid the use of misleading trees, as artificial branch-lengths for identical sequences are used by the BI methodology. Additionally, long-branch attraction (LBA) is more apparent in phylogenies generated by BI. Discrepancies in inferred topologies between methods are highlighted and discussed.

5.1.2 RT data collection and construction of phylogenies

Each familial dataset of RT sequences was constructed with a combination of BLAST (pBLAST and tBLASTn) searches and sequences from RepeatMasker output files with translation by ExPASy Translate where applicable. Query sequences of 300 residues were used to compile RT datasets of all available species possessing *Ty*-like elements. Sequences with clear null mutations were identified as pseudogenes and excluded from analyses but were included if not truncated, or the state of an element was unclear. RT sequences present in a pseudoelement, in which other domains were the most likely reason the element was non-autonomous, were included. A *Ty* RT query was initially used in searches to identify elements in other species with reasonable similarity. With increasing phylogenetic distance from *Saccharomyces* species, elements were identified by a combination of BLASTp and tBLASTn searches with translated RT query sequences from species other than *Saccharomyces*. Coding regions of distantly related elements were also identified by RepeatMasker screenings using both the custom library described here and the standard RepBase fungi library.

In all families, sequences were included until a dramatic drop in e-value was reached (Table 5.2), signifying a lack of full-length RT domains or loss of similarity in the BLAST hits.

Family	e value	No. of species
Ty1/2	1e-47	30
Gypsy	3e-69	47
Ty4	7e-92	14
Ty5	8e-53	43

Table 5.2: **e-value cut-off points for RT searches.** The e-values and total number of species included in RT phylogenies were recorded when using a query sequence from *Saccharomyces* elements. The sequences from the most distantly related species were then used as a further query to ensure elements from species were included where *Saccharomyces* failed to return hits.

Furthermore, the phylogenetic placement of elements was established primarily with RT sequences, particularly when LTRs shared identity too low to produce reliable phylogenetic relationships.

5.1.3 LTR data collection and construction of phylogenies

Full-length LTRs were used for phylogenies, with the exception of partial LTRs of potentially degraded families such as *Ty3p* detailed in the appropriate sections. Searches were performed with a combination of BLAST searching all available species and genomes, and screening with RepeatMasker using a custom library and extracting hits from genomes. Canonical LTRs from each element family in each species were added to the RepeatMasker custom library and genomes screened an additional time to ensure all potential HT events and divergent LTRs were collected.

As the elements from most species could be reliably placed with RT sequences, their corresponding LTRs were excluded from alignment if identity with other those of other species was poor (Table 5.3). LTR trees allowed the evolutionary history to be estimated where coding regions were lost or unreliably placed by RT, to confirm relationships proposed by the corresponding RT phylogeny and to highlight the existence of families in which coding regions had been stochastically lost or unsuccessfully horizontally transferred. Additionally, this approach revealed the donor and recipients that RT phylogenies alone could not identify.

E-value cut off points were not recorded for LTRs, as higher values were achieved by partial LTRs that shared a higher nucleotide identity than more divergent full-length copies. An e-value cut-off point that was too early would exclude full length LTRs but keep the partial LTRs that would be unnecessary in these analyses. E-values also varied for each family in each species. Family-specific alignments and RAxML phylogenies of LTR sequences in each species were visually inspected and dubious copies removed before being added to the main familial alignments. Additionally, at least one preliminary version of each phylogeny was created with RAxML (rapid bootstrap) to identify further questionable sequences before the final versions were created.

Although a BLAST search of all available strains of each species was conducted, only *Saccharomyces* Genome Resequencing Project (SGRP) sequences from *S. cerevisiae* and *S. paradoxus* were used in phylogenies unless specified in each familial section. Sequences from *S. cariocanus* were included in the phylogenies only when these insertions were unique to this species and not shared with parental *S. paradoxus*.

Family	Species (family) excluded
Ty1/2	*K. exigua*, K. naganishii*, K. saulgeensis**
	N. castellii, N. dairenensis (*Tnd1*)
	T. blattae, T. phaffii
	V. polyspora
	T. delbrueckii
	L. dasiensis, L. quebecensis*, L. thermotolerans*, L. waltii (Tlw1-2)**
	K. dobzhanskii, K. lactis*, K. marxianus* (*Tkm2*), *K. wickerhamii**
Ty3/gypsy	*K. africana, K. exigua, K. saulgeensis*
	N. castellii (*Tnc3*), *N. dairenensis* (*Tnd3*)
	Nk. glabrata
	V. polyspora
	T. delbrueckii
Ty4	*K. saulgeensis, K. servazzii*
	Nk. bacillisporus
	V. polyspora
Ty5	*K. exigua*
	*Nk. glabrata**
	V. polyspora
	O. polymorpha

Table 5.3: **Species with LTR sequences removed from final analysis.** The sequences possessed poor identity with those of other species, causing disruption of alignments and/or biologically implausible topologies in LTR phylogenies. *explored in separate trees and detailed in appropriate sections below.

5.1.4 Identifying horizontal transfer

TE phylogenies are often incongruent with those of their host species, indicating that vertical inheritance is not the only method by which elements are transmitted. Patchy distribution of families commonly observed across genera can be explained only in part by stochastic loss, by either LTR-LTR recombination or the mutation of coding regions. Divergence of families in an ancestor can also cause incongruence in phylogenies (Silva *et al.*, 2004; Wallau *et al.*, 2012). While studying the *P* element in *Drosophila*, Daniels *et al.* (1990) established that HT was the mechanism by which TEs were able to cross reproductive species barriers. HT allows TE families to escape the almost inevitable extinction of confinement in their original hosts, increasing their likelihood of survival in new genomes. Figure 5.1 illustrates the way in which potential HT events were identified in both RT and LTR phylogenies.

Highly similar, shared short-branched sequences in RT trees, incongruent with host species phylogenies, are indicative of HT, but are unlikely to reveal the direction of transfer (Figure 5.1 A). Corresponding LTR phylogenies can however reveal the direction, as the sequences of the recipient typically nest within those of the donor (Figure 5.1 B). Single LTRs nested in the sequences of another species, unless part of a full-length element (FLE), were counted as a single unsuccessful transfer. Propagation in a new species was counted as a single successful transfer, unless the

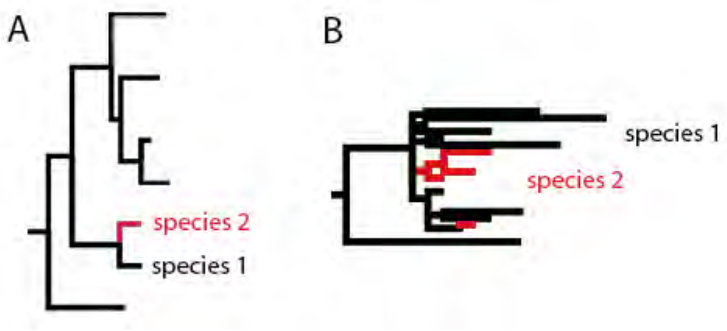

Figure 5.1: **Illustration of potential horizontal transfer in phylogenies.** Possible HT events in RT (A) and LTR (B) phylogenies were identifiable where relationships were incongruent with those of their host species. In these examples, highly similar RT sequences indicated a potential HT event (A) with the direction confirmed with species 2 as the recipient, as LTR sequences were nested within those of donor species 1 (B).

lineage had become extinct. Additionally, nucleotide diversity and branch lengths are both characteristically greater in the donor, as lower nucleotide diversity, shorter terminal branch lengths and lack of fixed copies in comparison to other families are all indicative of a more recent invasion of the recipient genome (Carr *et al.*, 2012). Patchy distribution resulting in a family's presence in only one of two sister species can also suggest HT, such as *S. paradoxus* predominantly lacking *Ty2*, whereas both sister species *S. cerevisiae* (recipient) and *S. mikatae* (donor) possess it (Liti *et al.*, 2005).

Although both RT and LTR sequences could indicate possible HT, the successful:unsuccessful HT event ratio was determined ultimately by LTRs, as they signify both extinction and propagation in the new genome. Table 5.4 displays potential HT events documented throughout phylogenetic analysis, and relationships are detailed in the familial sections below, alongside RT and LTR phylogenies.

5.1.5 Tajima's *D* and phylogenetics estimate recent evolutionary history

Tajima's *D* test was performed on familial alignments in each species. Those that possessed a significant result are displayed in Table 5.5. The full tables of characteristics are located in Appendix P.

Employing Tajima's test in this way can identify those family possessing the signature of recent transposition activity (Maside *et al.*, 2003; Sánchez-Gracia *et al.*, 2005; Bartolome *et al.*, 2009; Carr *et al.*, 2012; Carr and Suga, 2014). In all cases, Tajima's *D* results were evaluated using corresponding phylogenies. Families whose elements have a recent common ancestry and nucleotide variants at a low frequency may return negative values of *D*. While the majority of families (80%)

Family	Species involved		RT/LTR Evidence	Figure
	Donor	Recipient(s)		
Ty1	*S. paradoxus*	*L. waltii; L. kluyveri; L. fermentati* *S. arboricola; S. eubayanus*	both LTRs	6.2 & 6.3
	L. kluyveri	*N. castellii*	LTRs	6.3
Ty2	*S. mikatae*	*S. cerevisiae; S. arboricola*	both	6.2 & 6.3
	S. cerevisiae	*S. paradoxus*	both	6.2 & 6.3
Ty3/gypsy	*S. paradoxus*	*S. kudriavzevii; L. waltii*	both	6.9, 6.10 & 6.12
	between *S. mikatae* and *S. cerevisiae*		RT	6.9 & 6.10
	N. dairenensis	*N. castellii*	LTRs	6.12
	L. quebecensis	*L. thermotolerans*	LTRs	6.11 & 6.12
	E. coryli	*E. cymbalariae*	both	6.9-6.12
	between *E. aceri* and *E. gossypii*		RT	6.9 & 6.10
Ty4	*S. cariocanus*	*S. cerevisiae*	both	6.16-6.18
	between *S. paradoxus* and *S. cerevisiae*		LTRs	6.17-6.19
	S. eubayanus	*S. uvarum; S. cariocanus; S. mikatae; S. paradoxus; S. arboricola; S. kudriavzevii*	LTRs	6.17 & 6.18
	S. mikatae	*S. arboricola*	LTRs	6.17 & 6.18
	S. paradoxus	*L. waltii; S. eubayanus*	LTRs	6.17 & 6.19
Ty5	*S. paradoxus*	*S. cerevisiae*	LTRs	6.21
	S. cerevisiae	*S. cariocanus* *S. mikatae*	LTRs both	6.20 & 6.21

Table 5.4: **Species involved in HT of TE families.** A comprehensive list of the donor and recipient yeast species involved in potential HT events for each family, and figure(s) in which the event was identified.

	Species	Family	Tajima's *D*
Ty1/2	*N. dairenensis*	*Tnd2*	-2.01629
	T. blattae	*Ttb1*	-2.53199
	T. phaffii	*Ttp1*	-2.19733
	T. delphensis	*Ttd1*	-1.97698
	L. dasiensis	*Tld1*	-2.30971
	L. fermentati	*Tlfe1*	-1.91668
	L. kluyveri	*Tsk1*	-2.28109
	L. waltii	*Tlw2*	-2.27988
		Ty1p	-1.85843
Ty3	*N. castellii*	*Tnc3*	-2.21807
	T. blattae	*Ttb3*	-2.45781
	L. thermotolerans	*Tlt3*	-1.91813
	Sz. japonicus	*Tj3*	-1.93778
	Sz. pombe	*Tf2*	-1.99130
Ty4	*Nk. bacillisporus*	*Tnkb4*	-1.94715
	T. blattae	*Ttb4*	-1.99580
Ty5	*K. servazzii*	*Tkse5*	-2.20653
	Nk. glabrata	*Tnkg5*	1.82421
	T. blattae	*Ttb5*	-2.13189

Table 5.5: **TE families possessing a significant value of Tajima's *D*.**

do indeed possess a negative value of *D*, only those in Table 5.5 are significantly so (14%). Families whose elements have recently transposed are expected to display short terminal branches as copies have not had time to accumulate mutations. Therefore, identical copies of elements are strong evidence for recent transposition. Older insertions present as long-branched termini as mutations have had time to accumulate.

Additionally, 6% of families possess positive values of *D*, with *Tnkg5* of *Nk. glabrata* the only

family to be significantly so. However, insignificant values or those close to zero are likely indicative of families evolving neutrally, or that conflicting signals of multiple lineages and recent ancestry caused this clouding of signal.

5.2 *Ty1/2* superfamily phylogenies

Ty1 and *Ty2* were named as they were discovered in *S. cerevisiae*, and form a *Ty1/2* superfamily due to high similarity between the elements and their subfamilies, with a propensity to undergo recombination within LTRs to form hybrid elements (Kim *et al.*, 1998; Jordan and McDonald, 1998). A minority of species within and beyond the *Saccharomyces sensu stricto* complex contain *Ty1*-like subfamilies, but most possess a single representative of a *Ty1*-like family. Table 5.6 summarises the potential HT events and stochastic losses observed in the *Ty1/2* superfamily.

Horizontal transfer			Stochastic loss	
Potential events		Success	*n* of	Proportion
RT	LTR	ratio	families	lost
5	30	0.20	21 of 44	0.48

Table 5.6: **Potential HT events and stochastic loss in the *Ty1/2* superfamily.** The number of HT events in RT and LTR phylogenies were counted as detailed in Section 5.1.4 and Figure 5.1. HT success was determined by LTRs, i.e. if elements propagated and contained coding regions. For stochastic loss, subfamilies were counted as part of the same family outside *sensu stricto* species unless clear distinctions were made, e.g. *Ty1*-like families in *L. waltii* and *K. marxianus*. Stochastic loss is recorded as the proportion of lost:autonomous families.

5.2.1 *Ty1/2* superfamily RT phylogeny

Figure 5.2 displays the *Ty1/2* RT phylogeny of sequences collected from 30 species, which is rooted with *S. cerevisiae Ty4*. The *Saccharomyces* branch is collapsed in order to better visualise the tree; the boxed region encompasses the sequences within these clades. BI and ML topologies were relatively consistent, excluding discrepancies in the *Lachancea* clade and KNTV group which ae discussed below. Two *Lachancea* sequences from *Ty1/copia* pseudoelements of undetermined identity (*L. waltii and L. quebecensis*) clustered with the outgroup sequence (*S. cerevisiae Ty4*) and so were removed from the final phylogeny.

The *Ty1/2* RT phylogeny reflects that of the host species outside of the *Saccharomyces* branch. For example, the basal position is assumed by sequences from the *Lachancea* clade (*n*=7), but maximally supported by BI only (<70%mlBP; 1.0biPP). Within this clade however, the monophyly

of the group was recovered with strong ML support (90%mlBP; 0.82biPP), but neither method was able to resolve the internal branches.

The position of *Kluyveromyces* RT sequences next to diverge is in line with vertical inheritance within hosts. The remaining sequences, excluding the position of *Vanderwaltozyma polyspora* RT, also reflect the species phylogeny. However, the position of *V. polyspora* RT, plus the majority of internal branches, are unsupported by both methods. There are discrepancies regarding placement of sequences within the KNTV group, depending on inference method. BI topology forms two sister relationships with high to strong support: one between *N. dairenensis Tnd2* and *K. africana* and a second between *K. exigua* and *K. saulgeensis* RT sequences. ML corroborates only the *K. exigua* and *K. saulgeensis* relationship (97%mlBP).

The *Saccharomyces* clade (Figure 5.2, inset) forms two main sister groups: *Ty1*-like (containing *Tsu1*) and *Ty2*-like, with maximum support from both ML and BI. RT sequences from *S. eubayanus*, *S. uvarum* and *S. kudriavzevii* differentiate into their own *Ty1*-like cluster (68%mlBP; 0.99biPP), named *Tsu1* for clarity. The relationships within the *Tsu1* clade are all very high to maximally supported by both methods (88-93%mlBP; 0.99-1.0biPP, respectively).

Ty2 sequences found in *S. cerevisiae, S. mikatae, S. arboricola* and a Hawaiian isolate of *S. paradoxus* (UWOPS91-917.1) cluster together, with two HT events a possibility: between *S. mikatae* and *S. arboricola* (<50%mlBP; 0.90biPP) and *S. cerevisiae* and *S. paradoxus* (<70%mlBP; 0.96biPP).

Within the main *Ty1*-like clade, three *Lachancea* sequences cluster with *Saccharomyces*, rather than the remaining *Lachancea* sequences. RT sequences from *L. fermentati* and *L. kluyveri* cluster together (76%mlBP; 1.0biPP) as a sister group to the other *Ty1* sequences (<50%mlBP; 0.99biPP). The short branch lengths and close relationship between these sequences are suggestive of an HT event as these species are actually distantly related. Surprisingly, an additional sequence from *L. waltii* shares a branch with a *S. paradoxus* sequence, also with very high to maximum support (93%mlBP; 1.0biPP), a strong indication of HT. Similarly, the second *S. paradoxus* sequence shares a branch with *S. cariocanus* (98%mlBP; 1.0biPP). Given their parental and subspecies relationship, this is indicative of vertical inheritance rather than HT. Collectively, these are designated *Ty1p*, to which *S. cerevisiae Ty1* RT forms the outgroup (83%mlBP; 1.0biPP).

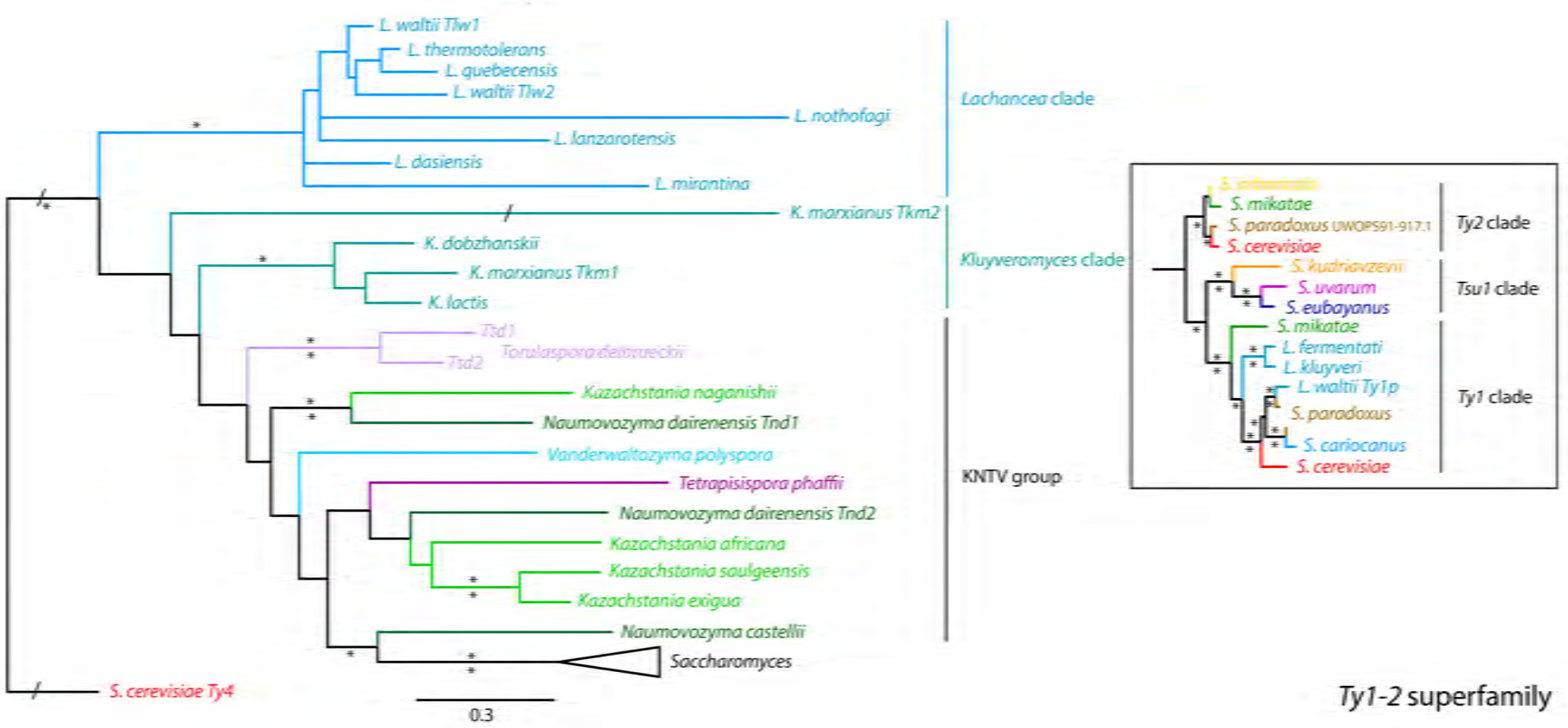

Figure 5.2: ***Ty1/2* superfamily RT phylogeny.** The tree is based upon an alignment of 300 AA positions and rooted with *S. cerevisiae Ty4* RT. ML topology is displayed with BI support values where topologies are consistent. *indicates ≥70%mlBP above the branch and ≥0.95biPP below the branch. The branches are drawn to scale, with the scale bar representing number of substitutions per site. The boxed region on the right of the figure displays the collapsed *Saccharomyces* branch from the main tree.

5.2.2 Multiple HT events in the *Ty1/2* superfamily LTR phylogeny

Figure 5.3 displays the *Ty1/2* LTR phylogeny of *Saccharomyces* species, along with sequences from three *Lachancea* species and *N. castellii*, which share identity with *Ty1/2*. An LTR phylogeny containing sequences from all available species was originally produced, with LTRs from *L. dasiensis* selected as the outgroup as these LTRs were the most diverged within the alignment upon visual inspection. Due to the biologically implausible placement of a number of species on long branches, they were removed in order to improve tree topology (Table 5.3). Figure 5.3 may therefore be the most plausible evolutionary history, given the available data. The separate phylogenies of *Kazachstania, Kluyveromyces* and *Lachancea* are described separately in Sections 5.2.4-5.2.6. Additionally, those families listed in Table 5.5 and *Saccharomyces* families (*S. cerevisiae, S. paradoxus, S. arboricola* and *S. uvarum*; Chapter 4) all display recent common ancestry and activity by way of short terminal branches, reflecting the results of the Tajima's *D* tests (data for *Ty1*-like families of *Tetrapisispora blattae, T. phaffii* and *Torluaspora delphensis* not shown due to lack of shared identity).

The *Ty1/2* RT and LTR trees share some similarities, strengthening relationships and possible HT events suggested by the RT phylogeny. *Ty1* and *Ty2* RT sequences share a sister relationship (Figure 5.2), the divergence of which may have occurred in a *S. mikatae* ancestor (Figure 5.4). Regardless of the size of the alignment and number of species analysed, the sequences of *S. mikatae Ty2* were consistently placed as a subgroup to *Ty1* sequences, suggesting that the divergence into the distinct *Ty2* element may have taken place in its ancestor. In addition to the HT event of *Ty2* from *S. mikatae* into *S. cerevisiae* (Carr *et al.*, 2012), *S. arboricola* appears to be a further recipient, determined by the ML method only. The BI topology places *S. mikatae Ty2* within *S. arboricola Ty2* sequences, with *S. mikatae Ty1* a further descendant. Despite both methods producing poor support values (<50%mlBP; <0.5biPP), the BI topology is implausible due to the greater age of *Ty1* (Kim *et al.*, 1998). Additional *Ty2* transfers from *S. cerevisiae* into Hawaiian strain UWOPS91-917.1 of *S. paradoxus* may have failed, as these were solo LTRs and partial relic elements upon inspection (detailed in Chapter 4).

S. cerevisiae Ty1 contains a number of possible HT events into *S. paradoxus* (*n*=10), two of which are associated with FLEs. Although this transfer does not appear to be reciprocal, *S. paradoxus* appears to have donated elements to *S. arboricola* (*n*=2) which have since become solo LTRs. In all cases, the flanking DNA is that of the recipient host, indicating that transposition

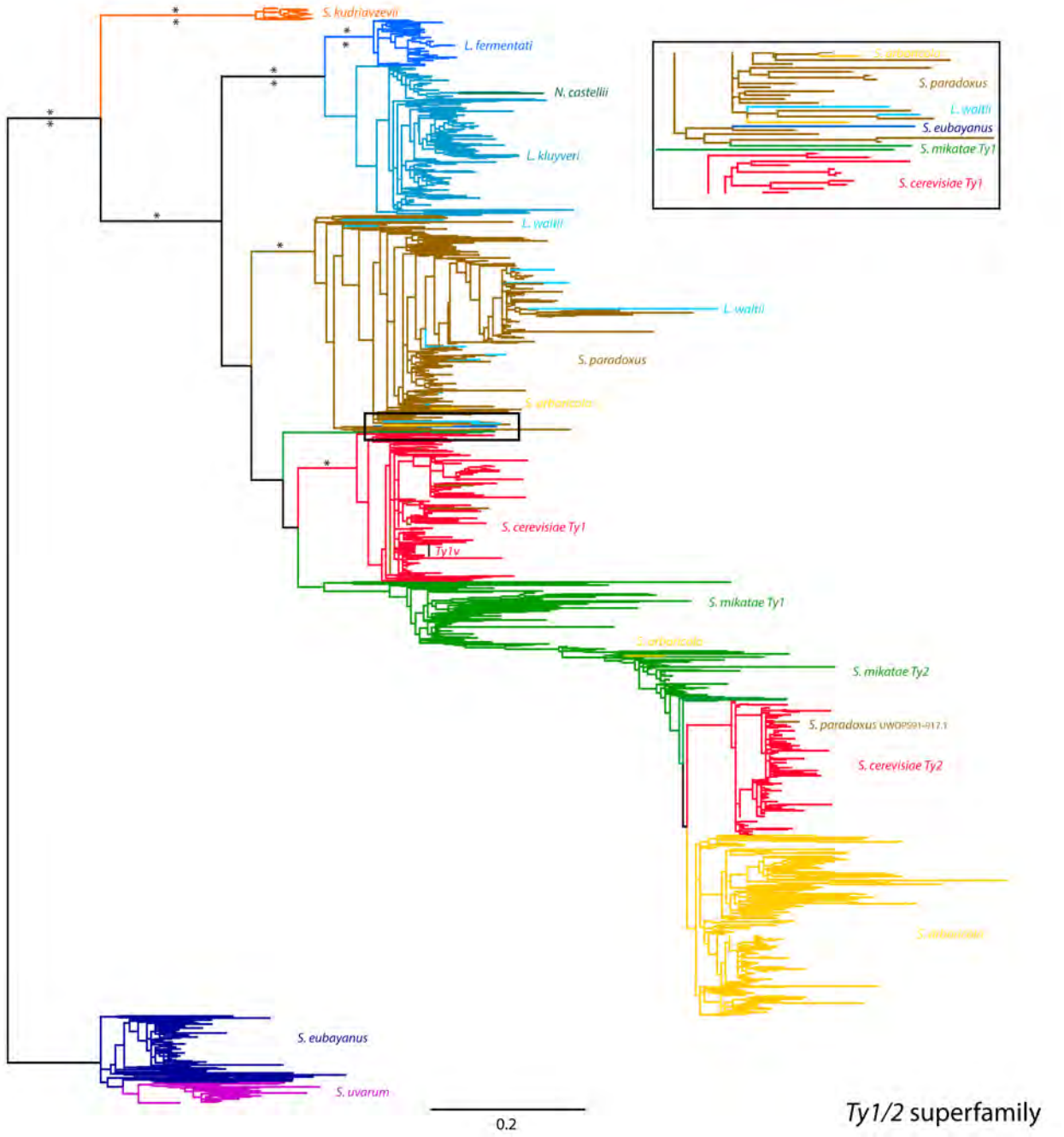

Figure 5.3: ***Ty1/2* superfamily LTR phylogeny.** The tree is based upon an alignment of 388 nucleotide positions and rooted with *S. eubayanus* and *S. uvarum* sequences. Layout and labels are as in Figure 5.2. LTRs in *S. mikatae* are explored in more detail in Figure 5.4. The boxed area of the phylogeny is magnified at the top right of the figure. The positioning of sequences from the subfamily *Ty1v*, described in Chapter 4 are also highlighted.

occurred since the HT event.

It was interesting to note that with the exception of a single *S. eubayanus* solo LTR sequence positioned within *S. paradoxus Ty1*, the families of *S. eubayanus, S. uvarum* and *S. kudriavzevii* are completely isolated, lacking infiltrations of elements from other species.

Additionally, *S. cerevisiae Ty1*-like LTRs discovered in *N. castellii* fall within *L. kluyveri* sequences, rather than the expected donor of a *Saccharomyces* species. Figure 5.3 shows that three *Lachancea* species, as seen in the RT phylogeny (Figure 5.2), contain the corresponding LTRs of *Ty1*-like elements. The sister relationship of *L. fermentati* and *L. kluyveri* elements (100%mlBP; 1.0biPP) shared an ancestor with *Saccharomyces Ty1*-like elements, after the divergence of the family in *S. kudriavzevii* (82%mlBP; 0.91biPP). The positions of *L. waltii* LTRs within *S. paradoxus* sequences (*n*=19) strongly support the suggestion of HT by the RT phylogeny. The relationships between the LTRs of the remaining *Lachancea* species are displayed in Section 5.2.5.

5.2.3 *S. mikatae* LTRs suggest a potential point of divergence for *Ty2*

A separate phylogeny of *S. mikatae* LTR sequences was produced in order to establish the relationship between *Ty1-2* and long-branched sequences (Figure 5.4).

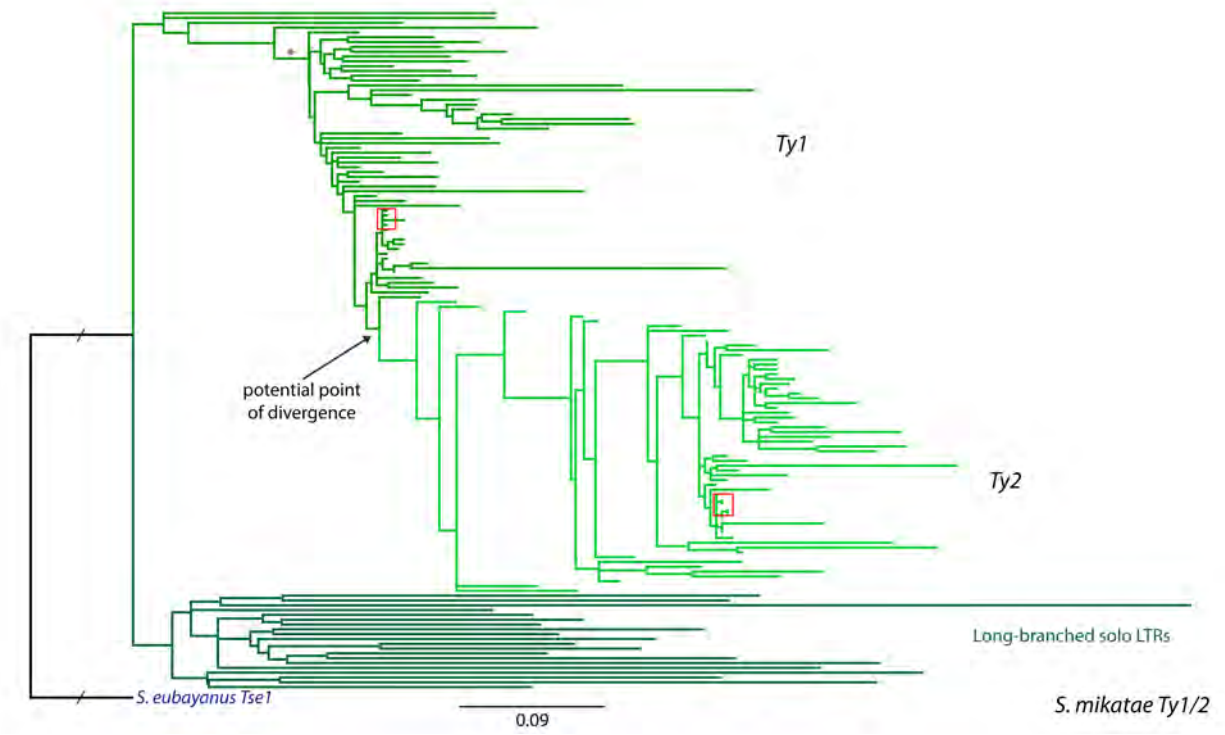

Figure 5.4: ***Ty1-2* LTR phylogeny of sequences in *S. mikatae*.** The tree is rooted with *S. eubayanus* sequence, as sequences from any other species tended to cluster with the long-branched solo *S. mikatae* sequences. Areas of LTRs associated with FLEs are boxed in red. Layout and labels are as in Figure 5.2.

Despite aligning the sequences with those of other species in turn, alternative topologies, although consistently poorly supported, could not be achieved. The long-branched *S. mikatae* sequences caused disruption to a preliminary version of Figure 5.3, clustering with *S. cerevisiae Ty1*

and assuming an implausible root. This was likely a reflection of the age of these sequences, interfering with the phylogenetic signal. Separately however, they form a sister group to both the current *Ty1* and *Ty2* families in *S. mikatae* (Figure 5.4).

As in *S. cerevisiae*, differentiation of *Ty1-2* sequences in *S. mikatae* could only reliably be achieved in the LTR phylogeny, rather than in an alignment. *Ty2* appears to diverge from a *Ty1*-like ancestor, before its subsequent transfer to other species.

5.2.4 *Kazachstania Ty1*-like LTRs display vertical inheritance

Figure 5.5 displays the *Kazachstania* LTR sequences that were excluded from the main *Ty1/2* phylogeny due to low shared identity.

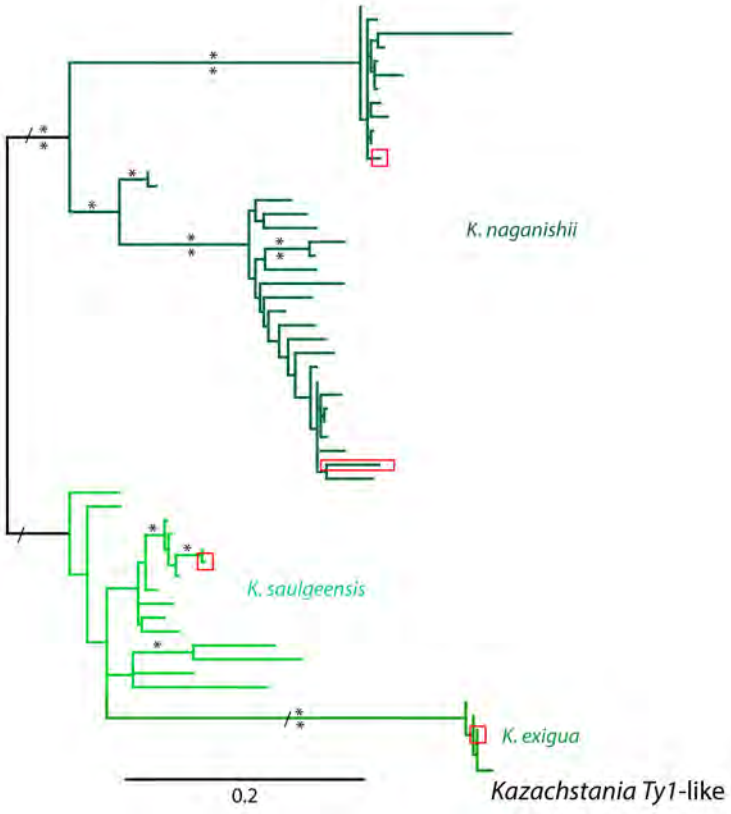

Figure 5.5: **LTR phylogeny of *Kazachstania Ty1*-like sequences.** The tree is based upon an alignment of 352 nucleotide positions and rooted with sequences of *K. naganishii*. Layout and labels are as in Figure 5.2. Red boxes indicate LTRs associated with FLEs.

The relationships between the sequences reflect both the RT phylogeny, with the elements of *K. naganishii* as more distantly related, and the species tree, in that the *Ty1*-like family was likely vertically inherited, based upon the data available. Sequences from *K. naganishii* split into two sublineages, both of which contain LTRs associated with FLEs. The full dataset of LTRs was

not available from *K. exigua*, as its genome was not fully sequenced. Whereas ML placed these sequences within those of *K. saulgeensis* (<50%mlBP), BI's positioning as the sister group was more likely (1.0biPP), due to the negligible ML support. All species show evidence of recent activity in their elements by the way of short-branched sequences.

5.2.5 LTRs in the *Lachancea Ty1/2* superfamily split into two clades

The only *Ty1/copia* elements of *L. fermentati* and *L. kluyveri* are *Ty1*-like families, whereas the other *Lachancea* species instead contain elements from a *Lachancea*-specific clade (Figure 5.6). *L. waltii* is unusual in that it is the only species to contain sequences in both *Saccharomyces*-like and *Lachancea*-specific clades. Sequences were extracted from up to three *Ty1*-like families, depending on the host strain's geographical origin: endogenous *Tlw1-2* are present across the species, with *Ty1p* confined to European strains, likely gained as a *S. paradoxus* ancestral element (Figures 5.2 and 5.3). Aside from those in *L. waltii*, sequences from *Lachancea* species are confined to only one of the two clades, which is not observed in the elements of any other species. Evidence of *Lachancea*-like families, rather than *Saccharomyces*-like, is not observed in *L. kluyveri* and *L. fermentati*.

In Figure 5.6, the divide between the *Saccharomyces*-like and endogenous *Lachancea* families is maximally supported by both methods. In the *Lachancea* clade, *L. dasiensis* forms the basal position, leaving the sequences of *L. thermotolerans* and *L. waltii* to form sister groups (100%mlBP; 1.0biPP). As the host species also share a sister relationship, the LTRs show evidence of vertical inheritance which is also seen in the RT phylogeny (Figure 5.2). Nested within the sequences of *L. thermotolerans* are those of *L. quebecensis*, poorly supported by both methods (<50%mlBP; <0.5biPP) but likely the result of vertical inheritance. The sequences of all species display older activity in long-branched clades together with shorter branches, consistent with more recent activity.

In the *Saccharomyces*-like clade, sequences from *L. fermentati* and *Ty1p* of *L. waltii* are nested within those of *L. kluyveri*, likely the result of the extensive copy number in the latter species causing the disruption of the phylogenetic signal and implausible placement by the inference methods. As seen previously in Figure 5.3, *L. kluyveri* and *L. fermentati* families were likely gained as a *Saccharomyces*-like ancestral element, while *L. waltii* received *Ty1p* more recently, perhaps directly from *S. paradoxus*.

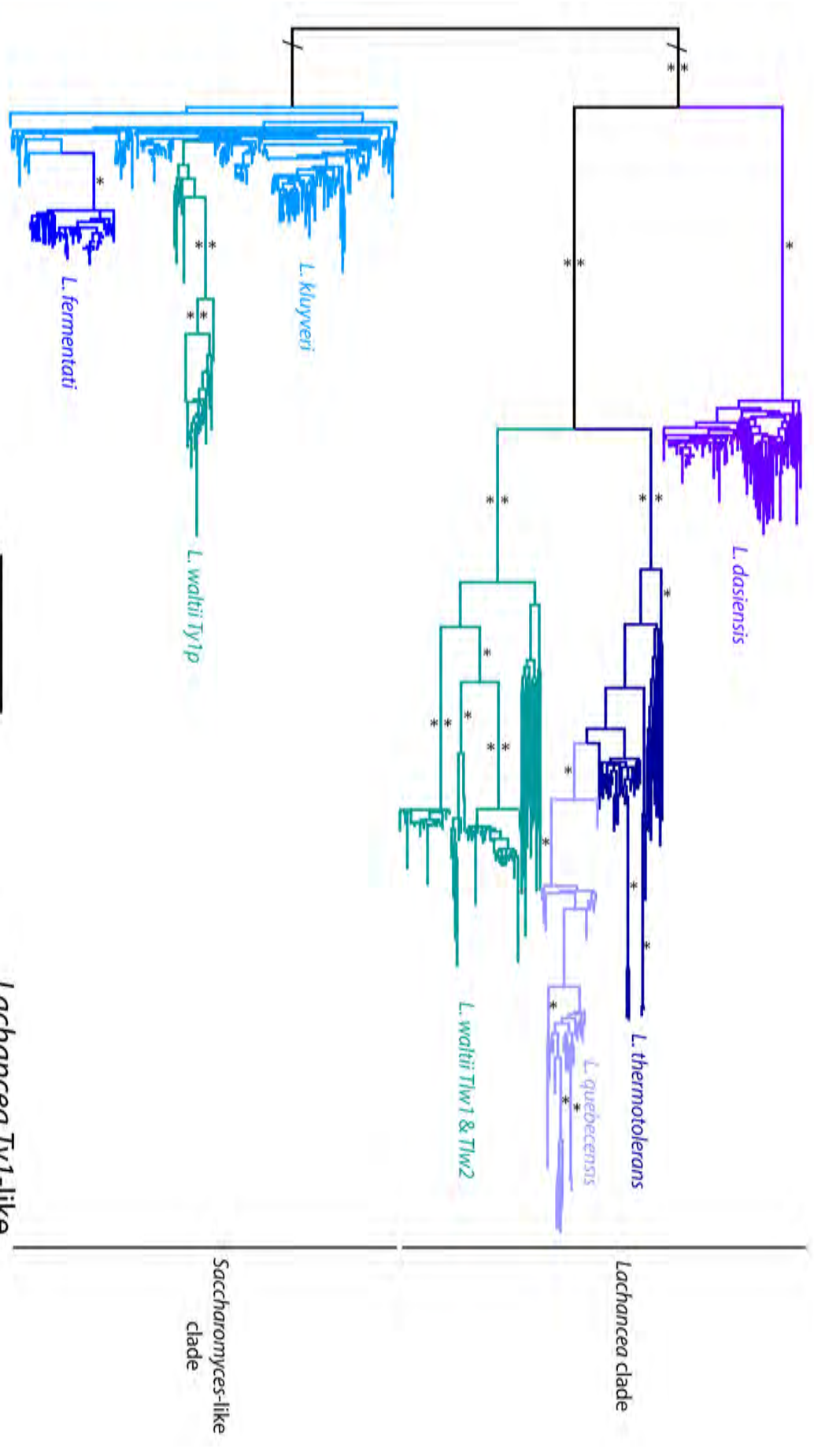

Figure 5.6: **LTR phylogeny of *Ty1/copia* sequences in *Lachancea* species.** The tree is based upon an alignment of 367 nucleotide positions and rooted with the *Saccharomyces*-like LTRs in *L. kluyveri*, *L. fermentati* and *L. waltii*. Layout and labels are as in Figure 5.2. Short-branched insertions of *Ty1*-like families in *L. dasiensis*, *L. fermentati*, *L. kluyveri* and *L. waltii Tlw2* and *Ty1p* support the significantly negative values of Tajima's *D* (Table 5.5).

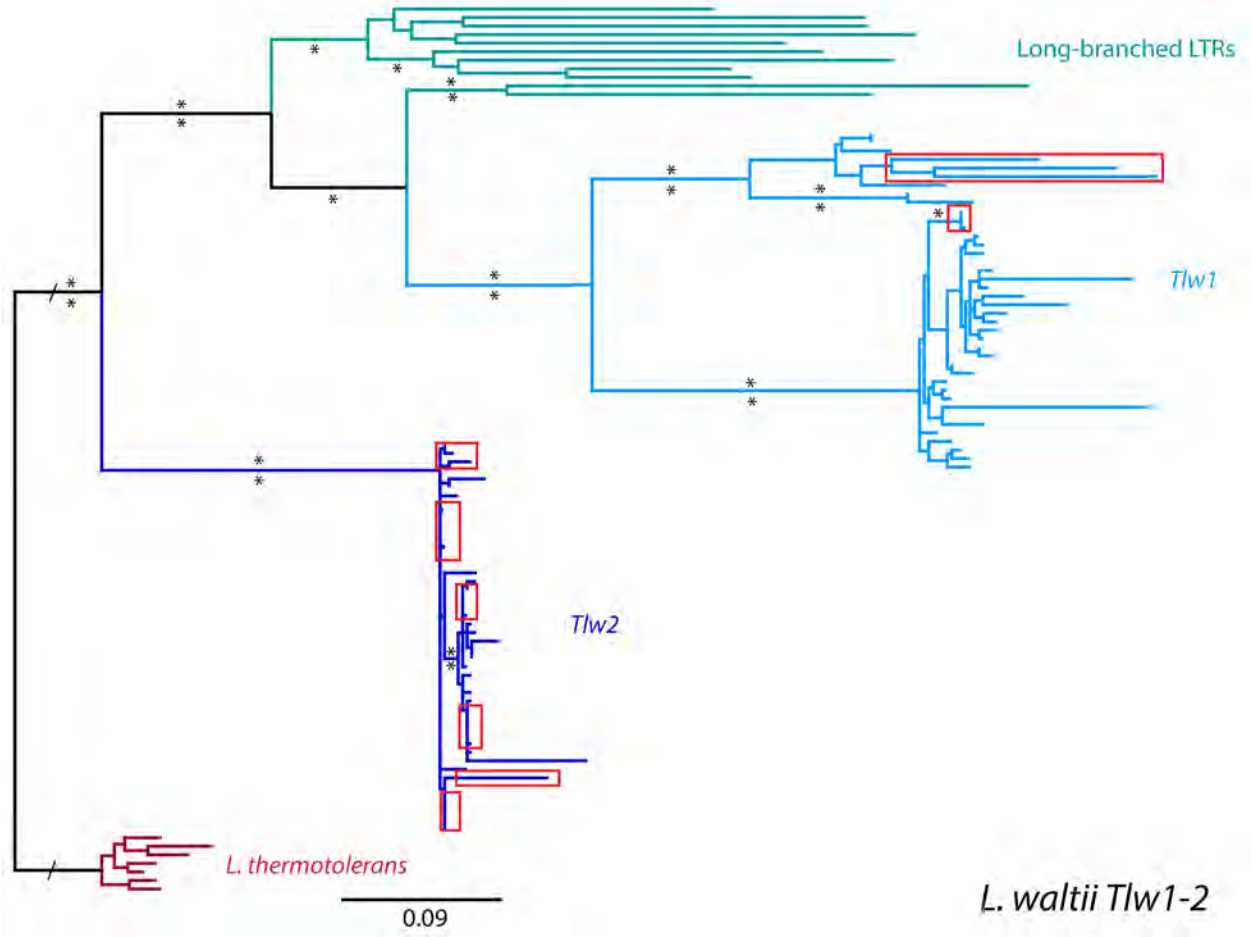

Figure 5.7: **LTR phylogeny of the two main endogenous families, *Tlw1-2*, in *L. waltii*.** The tree is based upon an alignment of 399 nucleotide positions and rooted with sequences of *L. thermotolerans*. Red boxes indicate LTR sequences associated with FLEs. Layout and labels are as in Figure 5.2.

The exact relationship between *Tlw1* and *Tlw2* in *L. waltii* is unclear in Figure 5.6 – ML determined the two as closely related sister groups but with poor support (<50%mlBP), whereas BI placed the long-branched clade of sequences between *Tlw1* and *Tlw2*, distancing the separation (0.76biPP). As the long-branched sequences clustering together is likely the result of long-branch attraction acting upon the BI topology in particular, the *L. waltii* sequences were further analysed alone in order to establish the relationship between the two major families in this species (Figure 5.7). Sequences of the two main families in *L. waltii* form maximally supported sister groups, with only *Tlw1* diverging from the long-branched ancestral sequences. The active *Tlw2* family contains autonomous FLEs, whereas only pseudoelements remain in the likely older family of *Tlw1*. The copy numbers of solo LTRs also vary in each subfamily (*n*=40 *Tlw1*; n=27 *Tlw2*). Although the relationship is reminiscent of that seen between *Ty1-2* in *S. mikatae, Tlw1-2* in *L. waltii* are unlikely to be subfamilies, and instead share a more distant (and now unidentified) ancestor. Alternatively,

the possibility that *Tlw2* was gained from an unknown source cannot be discounted.

5.2.6 *Kluyveromyces Ty1*-like family

Figure 5.8 displays the LTR phylogeny of *Ty1*-like sequences in the *Kluyveromyces* genus, which excludes the sequences of the *Tkm2* family of *K. marxianus* due to poor shared identity. The phylogeny suggests that the relationship between elements in these species is that of vertical inheritance. The species *K. marxianus* and *K. wickerhamii* are themselves the most distantly related, which is also reflected in the positioning of their LTRs. *K. dobzhanskii* and *K. lactis* LTRs are more closely related, likely a result of their host species possessing a sister relationship. The additional long-branched LTR of *K. wickerhamii* placed within *K. lactis* sequences is more than likely misplaced, with no support for this position (<50%mlBP; <0.5biPP).

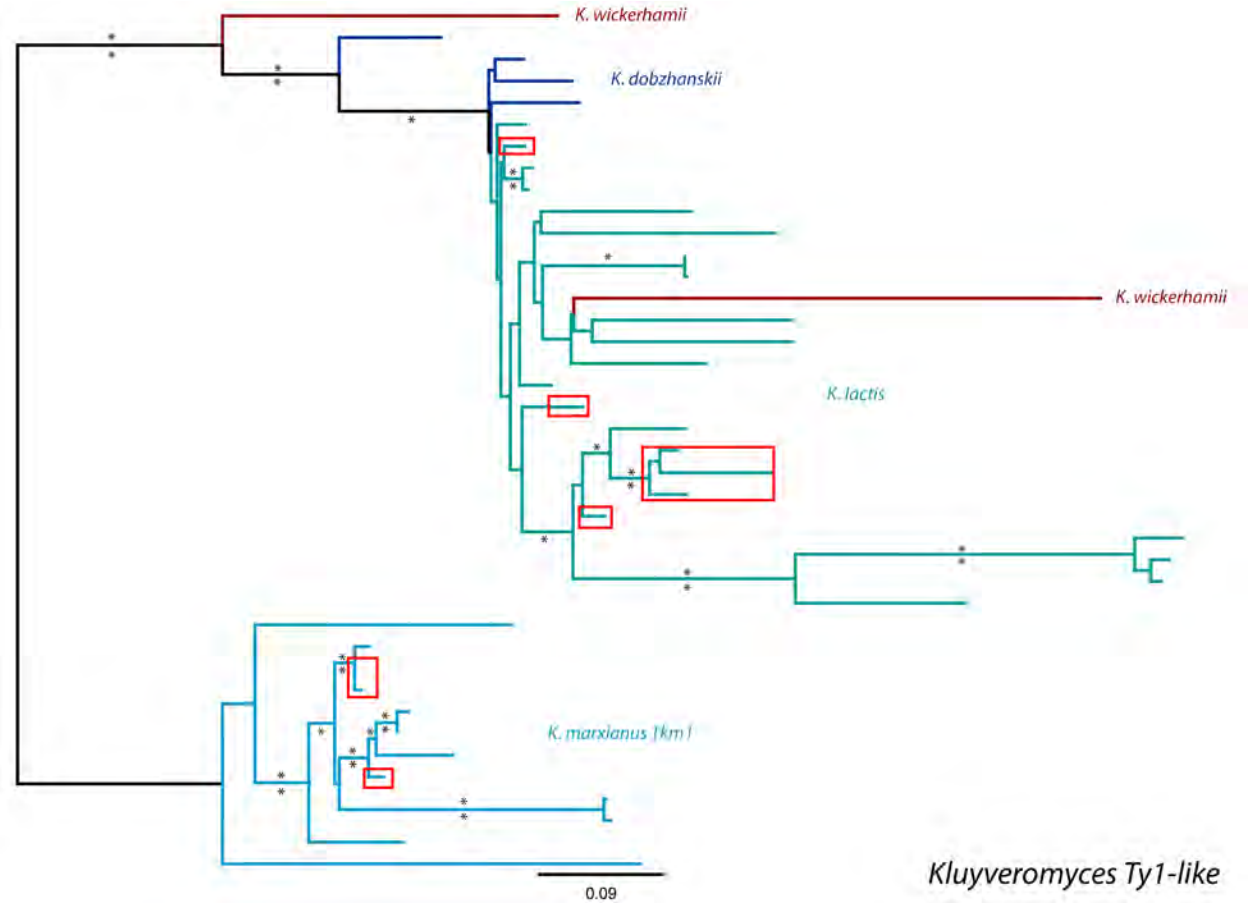

Figure 5.8: ***Ty1*-like LTR phylogeny of sequences in *Kluyveromyces* species.** The tree is generated using an alignment of 396 nucleotide positions and rooted with sequences of *K. marxianus Tkm1*. *K. marxianus Tkm2* LTRs were excluded due to poor identity, resulting in an unreliable alignment. Layout and labels are as in Figure 5.2. Red boxes indicate LTRs associated with FLEs.

5.3 *Ty3/gypsy* phylogenies

As one of the most widely distributed families, *Ty3*-like elements are present in most surveyed species. Table 5.7 summarises the potential HT events and stochastic losses observed in the *Ty3/gypsy* family.

Horizontal transfer			Stochastic loss*	
Potential events		Success	*n* of	Proportion
RT	LTR	ratio	families	lost
4	10	0.5	6 of 52	0.12

Table 5.7: **Potential HT events and stochastic loss in the *Ty3* family.** Stochastic loss and the occurrence of HT events in RT and LTR phylogenies were counted as in Table 5.6. *families possessed by *Sz. japonicus* were not counted towards stochastic loss as the state of most families could not be confidently determined.

5.3.1 Two *Ty3/gypsy* lineages in the RT phylogenies

Due to the wide distribution of *Ty3/gypsy* elements, RT sequences were collected from species beyond *sensu lato* species, to allow the position within fungal sequences to be examined (Figure 5.9). RT sequences are well conserved, considering the diversity of species surveyed.

Upon construction of the *Ty3/gypsy* phylogeny, it became apparent that two main lineages form, which were assigned the names *Ty3A* and *Ty3B* (see also Figure 5.10). The split between species does not coincide with the whole genome duplication (WGD) event and instead, inheritance and loss of the lineages are complex. Within the *Ty3A* lineage are the RT sequences from post-WGD species of *Saccharomyces, Torulaspora, Tetrapisispora* and *Vanderwaltozyma*, plus the pre-WGD *Eremothecium, Lachancea* and *Kluyveromyces* species. Given the currently available data, *K. africana* is the only *Kazachstania* species to possess the *Ty3A* lineage, whereas the remaining species possess *Ty3B*. Species from *Nakaseomyces*, the sister group to *Saccharomyces*, also possess *Ty3B* which is present in the minority of *sensu lato* species. Interestingly, *N. dairenensis* possesses elements from both lineages, but its sister species *N. castellii* only contains evidence of *Ty3B*. *Schizosaccharomyces japonicus* contains RT sequences that fall within the *Ty3B* lineage (*n*=7) but the remaining sequences and those of the other *Schizosaccharomyces* species fall further from both *Ty3A* and *Ty3B*. Relationships within this genus are explored in Section 5.3.4.

RT sequences from *Wickerhamomyces* species were initially included Figure 5.9, but as they placed on long branches, they were found to be too distantly related to even *Schizosaccharomyces* sequences to be phylogenetically informative. In a preliminary version of the phylogeny containing

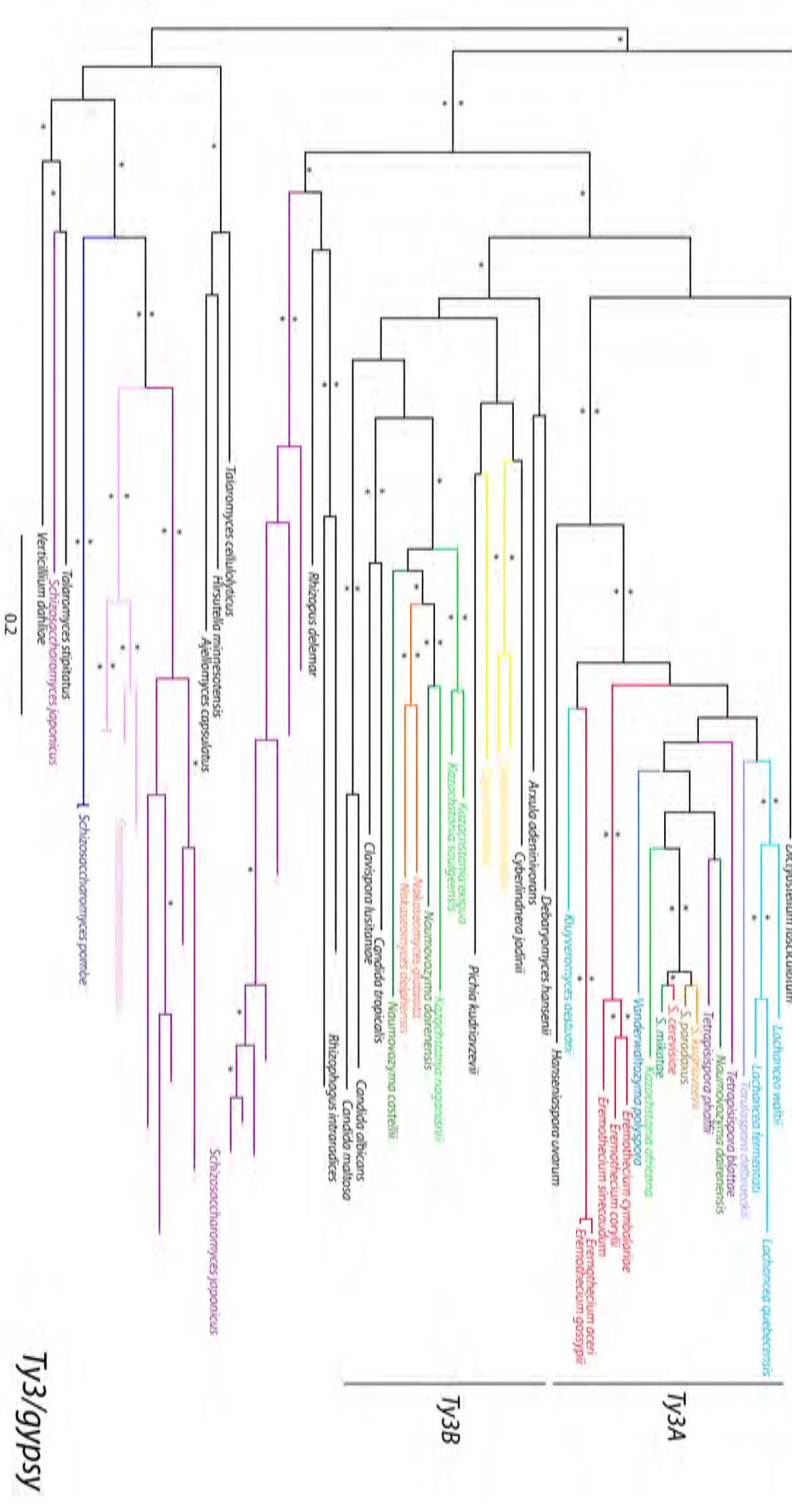

Figure 5.9: ***Ty3/gypsy* RT phylogeny of elements in fungal species.** The tree is based upon an alignment of 266 AA positions and midpoint rooted. Two lineages, *Ty3A* and *Ty3B* are distinguished. Layout and labels are as in Figure 5.2. Elements found in species in black were not investigated beyond the collection of RT sequences.

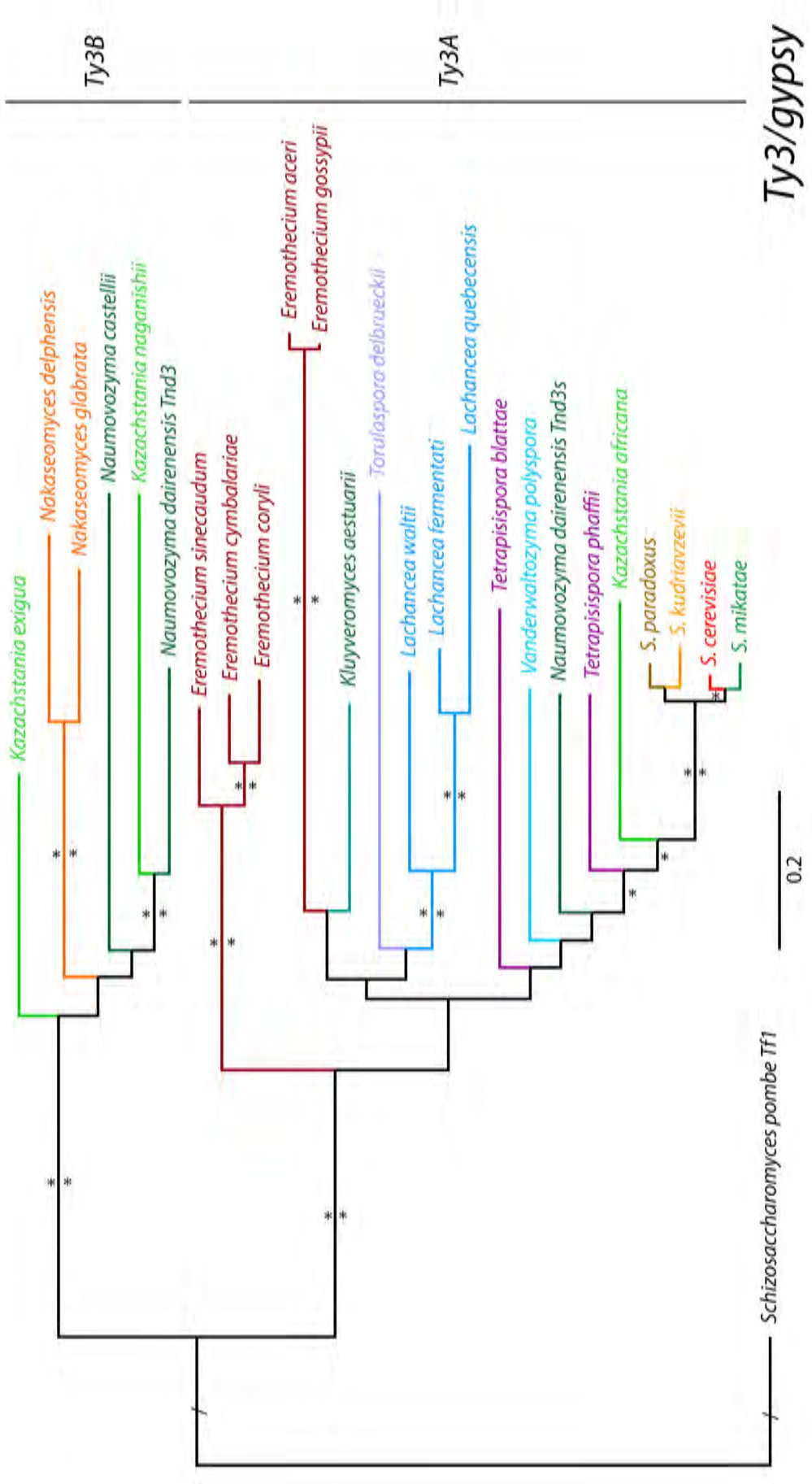

Figure 5.10: **RT tree of *Ty3/gypsy* elements in *Saccharomycetaceae* containing the *Ty3A* and *Ty3B* lineages.** The tree is generated from an alignment of 278 AA positions and rooted with *Sz. pombe Tf1*. Layout and labels are as in Figure 5.2. / indicates arbitrarily shortened root.

105 RT sequences, those from *Wickerhamomyces* assumed the strongly supported basal position, indicating a far more ancient divergence for the elements in this genus (data not shown).

The relationship between the RT sequences of *Talaromyces stipitatus* and an undesignated element from *Schizosaccharomyces japonicus* most likely displays convergent evolution due to the branch lengths, but an ancient HT event cannot be discounted. The pairings of sequences from *Nk. glabrata* and *Nk. delphensis*, and *C. albicans* and *C. maltosa* follow the vertical inheritance of these two groups of species and are therefore not indicative of HT. Two *Ty3/gypsy* sequences were extracted from *V. polyspora* and *C. albicans* strains: *Ty3*-like and *Tca3*. The *Tca3* sequences formed the outgroup to the entire phylogeny (data not shown) and so were excluded from Figure 5.9, whereas the *Ty3*-like sequences fell within *Ty3A* and *Ty3B*, respectively.

Those close relationships of *sensu lato* sequences are detailed in Figure 5.10. Three potential HT events are suggested by this phylogeny: between elements of *S. kudriavzevii* and *S. paradoxus* (<70%mlBP; 0.75biPP); *S. cerevisiae* and *S. mikatae* (95%mlBP; 0.95biPP) and potentially an older event between *E. coryli* and *E. cymbalariae* (90%mlBP; 1.0biPP), as each of these pairings differ from their respective host species phylogenies. Additionally, vertical inheritance is observed between elements in the pairings of sister species *E. gossypii* and *E. aceri*, and *Nk. glabrata* and *Nk. delphensis*. Excluding the unsupported position of *T. delbrueckii* RT in *Ty3A*, the topology of this clade in Figure 5.10 predominantly follows that of the species phylogeny. The distinction between elements of *Ty3A* and *Ty3B* clades in closely related species, such as *Saccharomyces* and *Nakaseomyces*, suggests a complex evolutionary history of this family.

Beyond those in *Schizosaccharomyces* (section 5.3.4), RTs from *Saccharomyces* and *Eremothecium* are the only relatively short-branched sequences, indicating recent activity and/or a closer common ancestor of the elements in those genera, with the remaining species' element sequences placed upon far longer branches, representing far more ancient divergences and distancing from the elements of other species.

5.3.2 Unclear relationships between *sensu lato Ty3/gypsy* LTRs

The topology of Figure 5.11 is unsupported by both methods, and few comparisons could be made with that of RT trees. Additionally, the BI topology is unresolved, with sequences branching from a single internal node in a pectinate-like formation. Due to the diversity of LTRs, a number of species' sequences were not included in the LTR phylogeny due to biologically implausible placements

(*n*=10; Table 5.3). The phylogeny was also split due to the impracticality of such a large number of sequences: Figure 5.11 contains LTRs from all *sensu lato* species able to be aligned, with close relationships detailed in Figure 5.12. A tree of *Saccharomyces sensu stricto* species is displayed in Figure 5.13 (Section 5.3.3). The *Ty3/gypsy* LTR sequences of *T. blattae* are all short-branched,

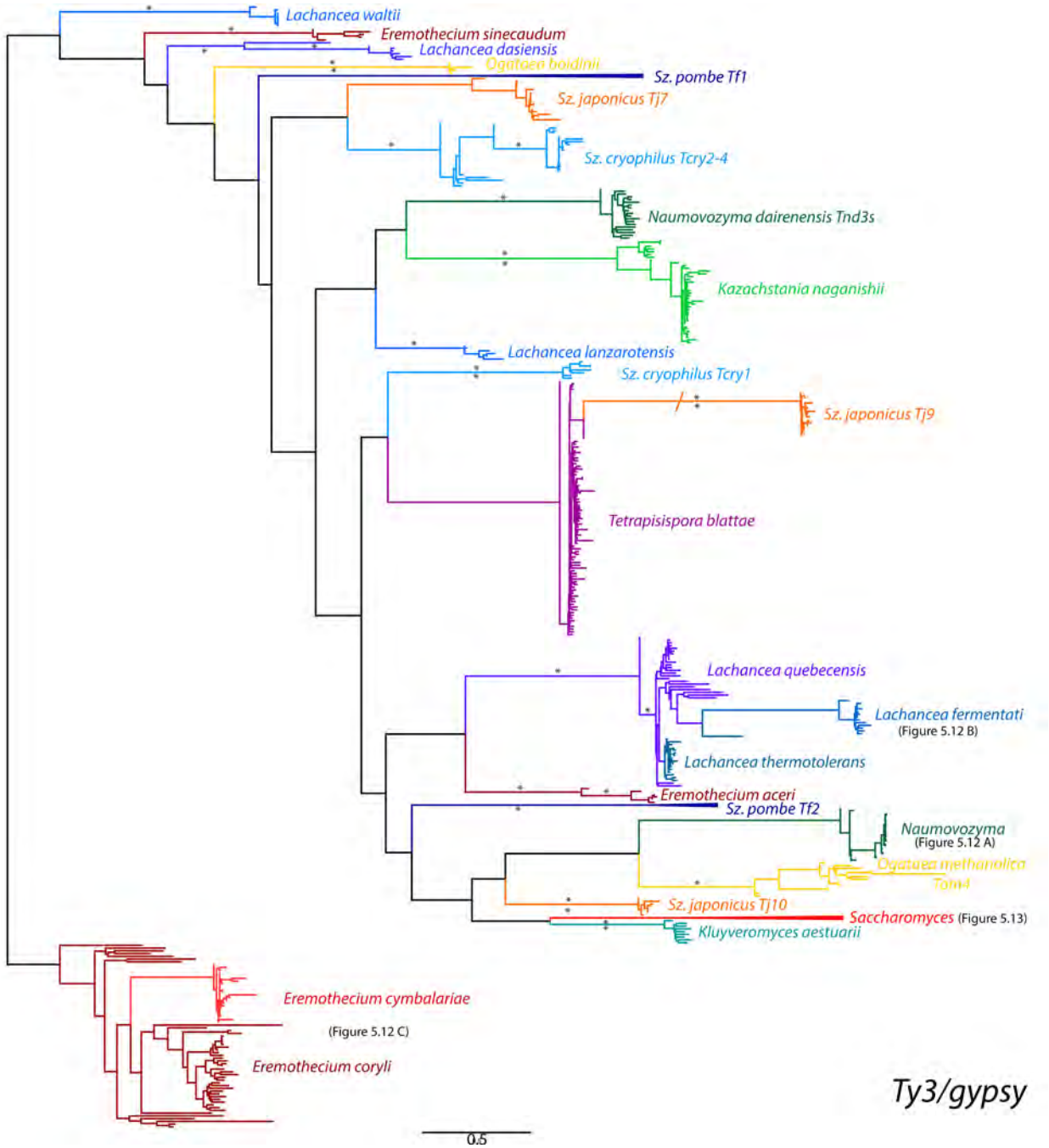

Figure 5.11: **Phylogeny of *Ty3/gypsy* LTR sequences in yeast species.** The tree is based upon an alignment of 473 nucleotide positions and rooted with sequences from *E. coryli* and *E. cymbalariae*. Potential HT events in were detailed in Figure 5.12. The *Saccharomyces* phylogeny was collapsed due to space limitations and so displayed separately in Figure 5.13. Layout and labels are as in Figure 5.2.

indicative of a relatively young family displaying recent ancestry and recombination of elements

resulting in solo LTRs (*n*=88), alongside the four FLEs in this species. This is also indicated by the significantly negative value of Tajima's *D* for this family (Table 5.5). In contrast are the small clades of sequences in *E. aceri* and *E. sinecaudum*. Although relatively short-branched, the LTRs of these elements are remnants of extinct families, suggesting that the families underwent only a small burst of activity before they were lost in their host species. Similarly observed in *K. aestuarii, L. waltii* and *L. lanzarotensis*, little activity is evidenced by low frequency insertions. Unlike the extinct elements of *Eremothecium* and *L. lanzarotensis* however, the *Ty3/gypsy* families in *K. aestuarii* and *L. waltii* consist of autonomous copies with low levels of activity. A number of sister relationships are observed in Figure 5.11, including that of *Schizosaccharomyces japonicus Tj7* and LTRs from *Sz. cryophilus Tcry2-4* which echo that of the RT sequences (detailed in Section 5.3.4). A further sister relationship is observed between the sequences of *N. dairenensis Tnd3s* and *K. naganishii*, reflecting that of the sequences in the RT phylogeny, suggesting that the entirety of the elements share identity despite diverging in an ancient ancestor. The LTR sequences of the more distantly related *N. dairenensis* family, *Tnd3*, were not included in the final phylogeny due to poor identity (Table 5.3).

The nesting of sequences of one species within those of another may be indicative of HT if the relationship between elements differs to that of the species phylogeny. Figure 5.12 illustrates these nested relationships in the sequences of *Naumovozyma* (A), *Lachancea* (B) and *Eremothecium* (C) species. Prior to the extinction of the *Tnc3* family in *N. dairenensis*, it appears that the family may have been transferred into *N. castellii*, which went on to successfully transpose (Figure 5.12 A). The position of the recipient sequences is however strongly supported by ML only (97%mlBP; <0.7biPP).

A further nesting relationship in Figure 5.11 is that between three *Lachancea* species: *L. quebecensis, L. thermotolerans* and *L. fermentati* (detailed in Figure 5.12 B). However, the former two species are closely related and so their respective elements likely follow that of vertical inheritance rather than HT. *L. fermentati* sequences are more distantly related, the long-branched, unsupported placement of which is likely incorrect and not indicative of HT. This shows that each relationship must be individually evaluated before concluding that elements have undergone HT.

Finally, the nesting of *E. cymbalariae* LTRs within *E. coryli* sequences (Figure 5.12 C) support the possible older HT event observed in the RT phylogeny. Neither method provides support for these positions however. The LTR sequences of these two species are the only evidence of

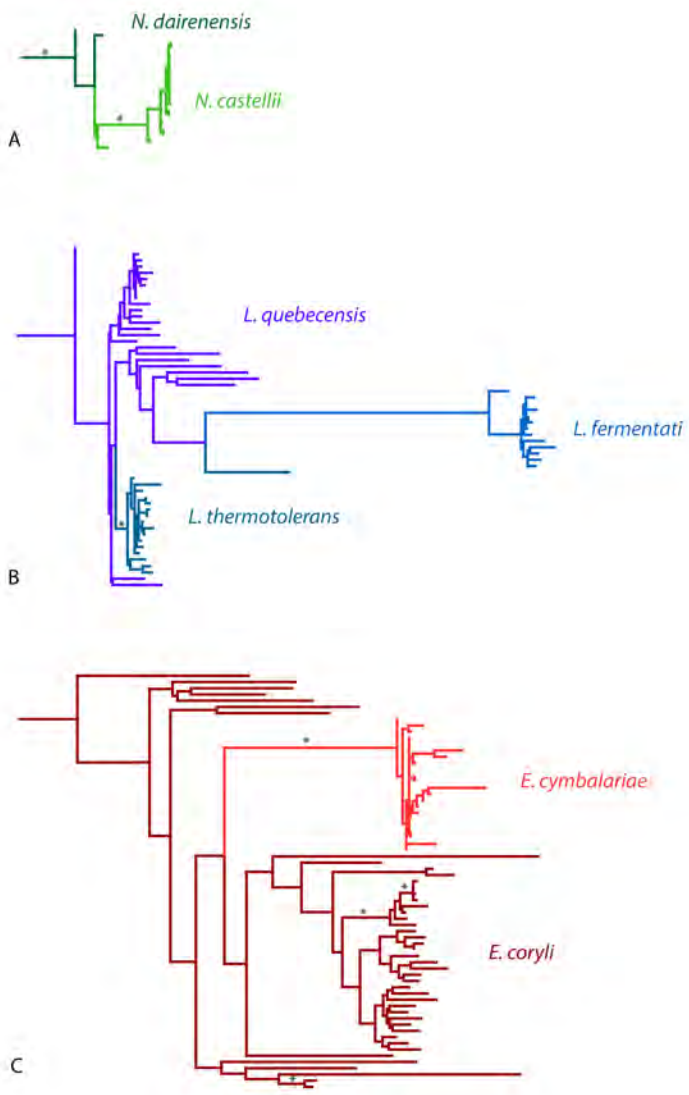

Figure 5.12: ***Ty3*-like LTR sequences illustrating nested sequences.** Relationships such as these must be individually examined for evidence of potential HT. A: *N. dairenensis* (now present as solos and therefore extinct) into *N. castellii* (functional FLEs); B: *Lachancea* LTRs; C: *Eremothecium* LTRs. Layout and labels are as in Figure 5.2. The significantly negative values of Tajima's *D* (Table 5.5) generated by *N. castellii* (A) and *L. thermotolerans* (B) LTR alignments are consistent with recent ancestry, as also seen in the short terminal branches of these sequences.

relatively extensive past activity in the *Eremothecium* genus.

5.3.3 Well-supported relationships between the *Ty3* LTRs of *sensu stricto* species

Relationships between *Ty3* LTRs in *Saccharomyces* species are displayed in Figure 5.13. The species fall into a series of sister relationships: between *S. arboricola* group 1 and *S. kudriavzevii*, which is maximally supported by both methods, and the placement of *S. arboricola* group 2 with *S. paradoxus* sequences as very high to maximum (81%mlBP; 1.0biPP). The positioning of *S.*

mikatae and *S. cerevisiae* is supported by ML only (96%mlBP), but likely correct, as BI places *S. mikatae* as the outgroup to *S. paradoxus* and *S. arboricola* group 2 sequences with less reliable support (<0.7biPP).

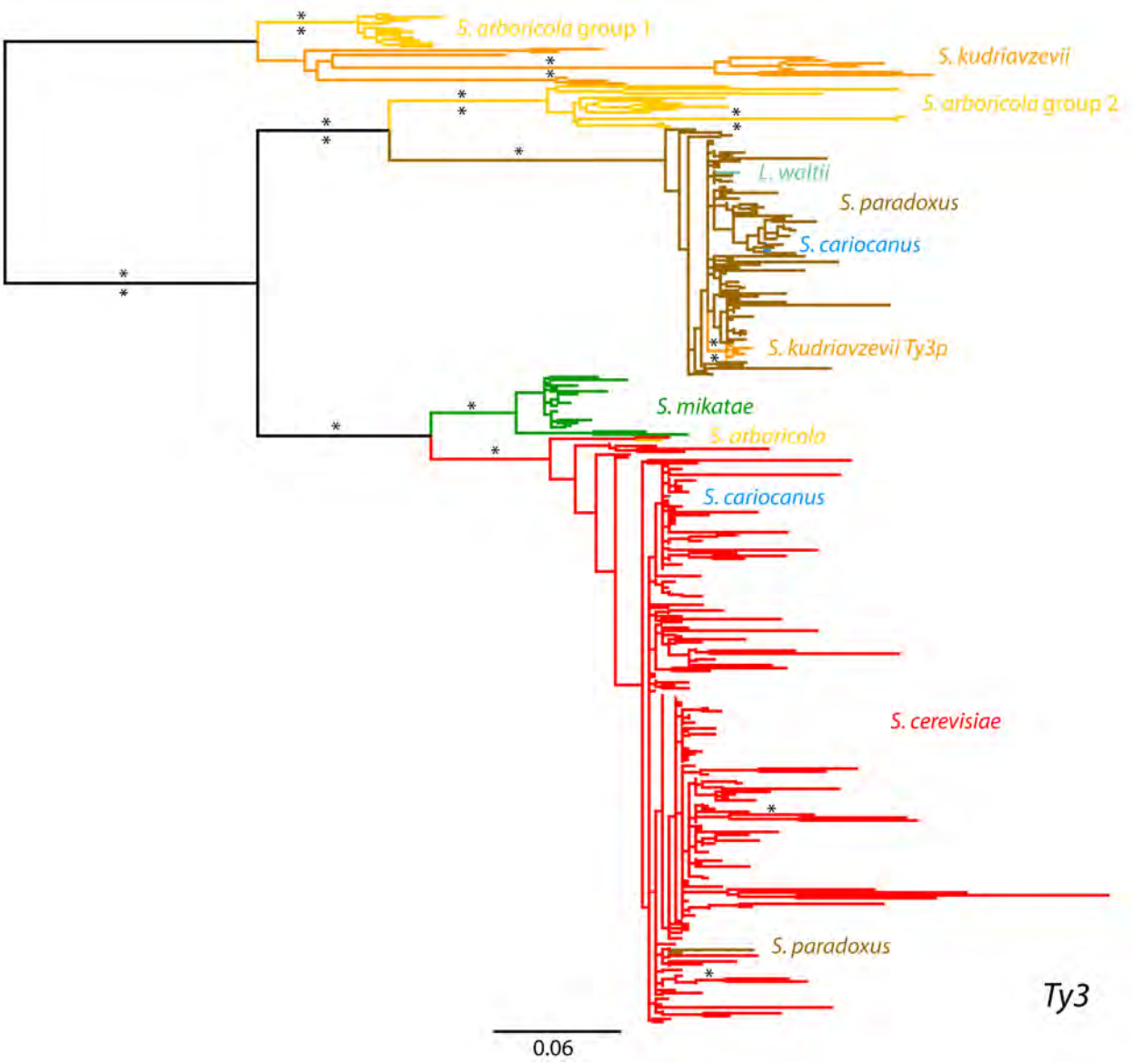

Figure 5.13: ***Ty3* phylogeny of LTR sequences from *Saccharomyces* species.** The tree is based upon an alignment of 367 nucleotide positions and rooted with *S. kudriavzevii* and *S. arboricola* sequences. Layout and labels are as in Figure 5.2. *S. cerevisiae, S. paradoxus, S. kudriavzevii* (*Ty3p*) and *S. mikatae* all possessed short terminal branched sequences, supporting the Tajima's *D* results of significantly negative values (Chapter 4) consistent with recent ancestry and possible activity.

Partial *Ty3p* sequences from *S. cerevisiae* were excluded from the final tree as Carr *et al.* (2012) previously demonstrated the likely relationship between this extinct family and the donor *S. paradoxus* sequences. The potential *Ty3* sequence from *S. eubayanus* (Chapter 4) was also excluded from the alignment due to poor shared identity. Seemingly partial *Ty3*-like LTR sequences in *S.*

eubayanus and *S. uvarum* were aligned with those of other species to ascertain as to whether an event similar to the extinction of *Ty3p* has occurred in these species as it has in *S. cerevisiae*. However, the partial sequences formed a clade with zero support away from the sequences of the other species (data not shown as the potential LTR from chrIV was long-branched in comparison to the other species and clustered with the partial sequences). It was therefore concluded that although *S. eubayanus* contains evidence of degraded coding regions, if LTRs remain within the genome, lack of identity with those of other species prevent them from being identified and phylogenetically analysed.

The endogenous *Ty3* family in *S. kudriavzevii* consists of solo LTRs and presents as a clade of long-branched sequences. *Ty3p* is also present in *S. kudriavzevii* (70%mlBP; 1.0biPP), seemingly a result of a successful HT from a *S. paradoxus* source. *S. paradoxus* itself contains solo LTR evidence of failed HT events from *S. cerevisiae* (*n*=2). Unique *S. cariocanus* LTRs were identified (*n*=3), one of which was apparently donated by *S. cerevisiae* as a FLE but has since undergone LTR-LTR recombination and is now present as a solo LTR. The unique LTRs nested within the *S. paradoxus* sequences are evidence of *Ty3* activity since the divergence of *S. cariocanus*, and a single solo LTR in *L. waltii*, representative of a further unsuccessful HT event.

5.3.4 Complex relationships between *Ty3/gypsy* families of *Schizosaccharomyces*

Figures 5.14 and 5.15 display RT and LTR phylogenies of sequences in *Schizosaccharomyces* species, respectively.

Phylogenetic relationships of RT sequences in *Sz. japonicus* are similar to those reported by Rhind *et al.* (2011) but as the authors noted the presence of fewer *Ty3/gypsy* families (*n*=10), Figure 5.14 contains the additional families discovered here (*n*=4). With the exception of *Tj7*, all families in *Sz. japonicus* are represented by a minimum of two RT sequences, which indicates previous activity of the elements in each family. Additionally, the clades of *Tj3* and *Tj9* were not fully resolved by either method. Little variation was observed between topologies suggested by ML and BI, with strong support consistent throughout the majority of internal branches. As in Figure 5.9, RT sequences of *Schizosaccharomyces* split into two main clades, with the one more closely related to *Ty3* of *S. cerevisiae* containing the majority of sequences from *Sz. japonicus* (*n*=23). The remaining sequences of this species (*n*=15) along with RT sequences from sister species *Sz. cryophilus* and *Sz. pombe* fall into the second clade. The RT sequences of *Tf1* and *Tf2* in *Sz.*

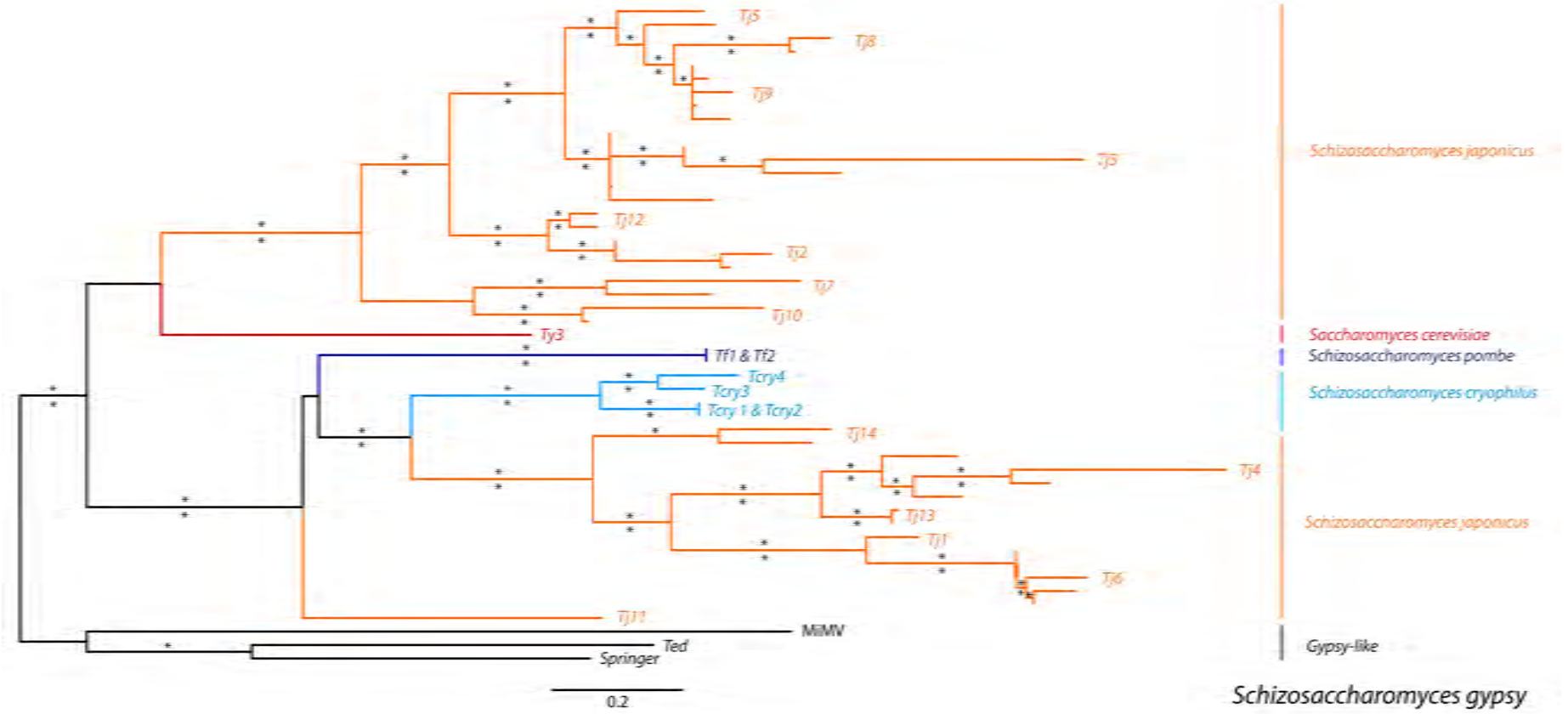

Figure 5.14: **RT phylogeny of elements within *Schizosaccharomyces* species.** The tree is based upon an alignment of 286 AA positions and rooted with *Ty3/gypsy*-like RT sequences from viral families: *MtMV* of *Nicotiana tabacum*, *Ted* of *Autographa californica* and *Springer* of *Drosophila melanogaster*. *S. cerevisiae Ty3* RT was included for comparison. ML and BI topologies had little variation and positions were generally well supported by both methods. Layout and labels are as in Figure 5.2

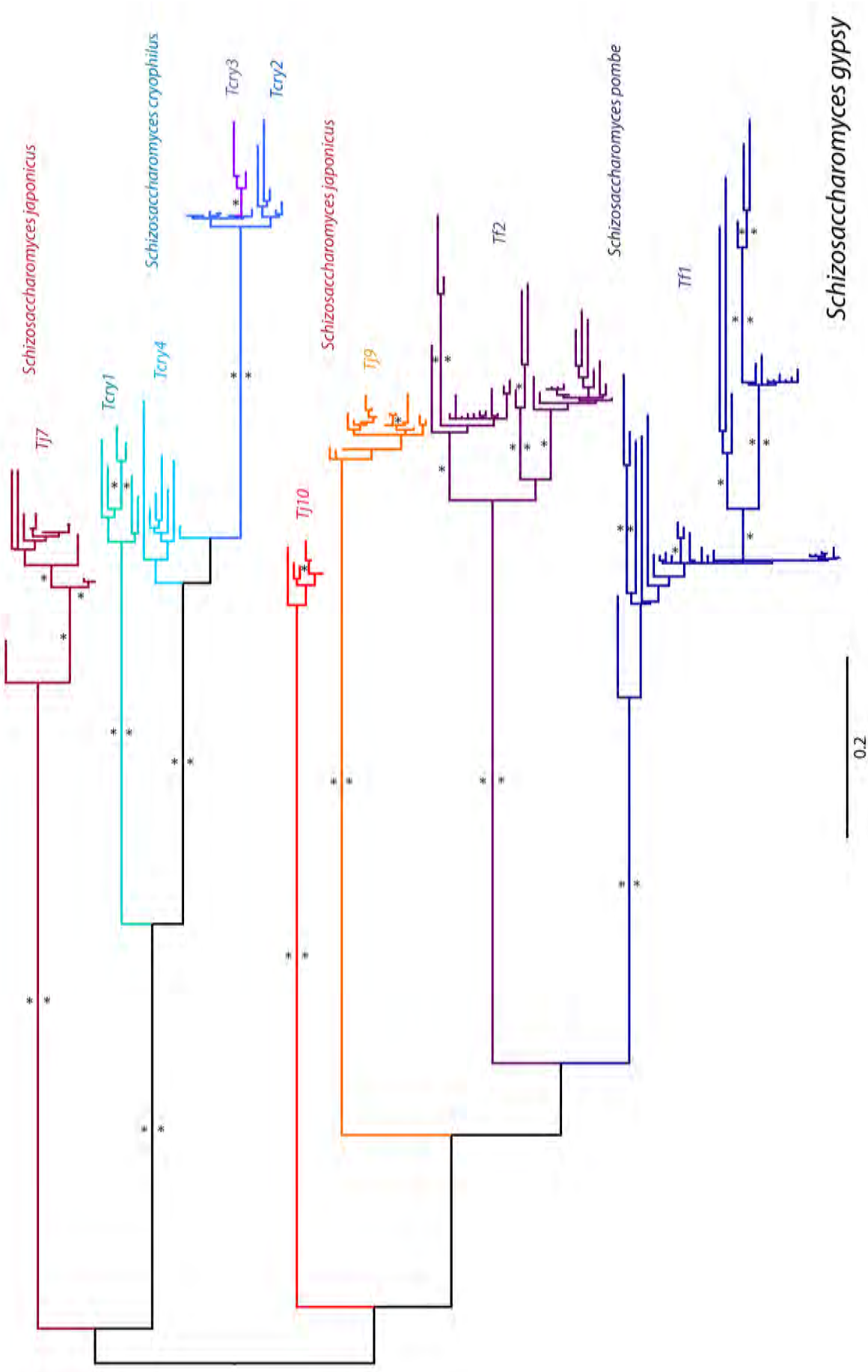

Figure 5.15: **LTR phylogeny of *Schizosaccharomyces* element sequences.** The tree is based upon an alignment of 385 nucleotide positions and midpoint rooted. ML and BI topologies had little variation and positions were generally well supported by both methods. Layout and labels are as in Figure 5.2

pombe are identical, observed previously by Rhind *et al.* (2011). Interestingly, this is mirrored in the RT sequences of *Tcry1-2* in *Sz. cryophilus*, which was not recorded by Rhind *et al.* (2011), as the authors failed to identify families beyond *Tcry1*.

The LTR sequences of three families (*Tj7, Tj9-10*) in *Sz. japonicus* are included in the phylogeny of Figure 5.15. The LTRs of the remaining families lack nucleotide identity with those included in the figure, causing the alignment to become unreliable. The *Tj7* family possesses only one FLE but has been active as evidenced by multiple solo LTRs (*n*=13). *Tj9* appears to be a relatively young, active family with short-branched LTRs (*n*=16). With the fewest sequences, *Tj10* is the least active family (*n*=5). An alignment of *Tj3* sequences returned a significantly negative value of Tajima's *D* (Table 5.5), a phylogeny of which form a clade with short terminal branches (data not shown).

The close relationships of RT sequences in *Sz. cryophilus* are reflected in their corresponding LTRs. *Tcry3* LTRs (*n*=3) are nested within those of *Tcry2*, but with no support (<50%mlBP; <0.5biPP). The single copy of *Tcry4* is degenerate and additional solo LTR copies (*n*=5) attest to previous activity of this family, reflected in its long-branched sequences. *Tcry2* contains multiple lineages which are strong to maximally supported (99%mlBP; 1.0biPP), with long-branched sequences consistent with older activity as well as regions of short-branched, highly similar sequences indicating far more recent activity within this family. *Tcry1* also contains the relatively long-branched sequences of past activity (*n*=6). Despite *Tcry1-2* possessing identical RT sequences, their LTRs display a far more distant relationship, which is also observed in the LTRs of *Tf1-2* in *Sz. pombe*. The elements of this species display more extensive activity than those of its sister species, with *Tf1* (*n*=20) and *Tf2* (*n*=18) maintaining similar numbers of short-branched sequences. Despite the separation of *Tf2* into two sublineages, this family possesses a significantly negative value of Tajima's *D*, which is consistent with recent ancestry and activity illustrated by short terminal branches. The long-branched sequences in both of these families illustrate past activity and subsequent degradation of solo LTRs.

The topologies of the RT and LTR phylogenies share similarities in that the sequences of *Sz. pombe* and *Sz. cryophilus* are closely related, yet few conclusions can be made regarding those of *Sz. japonicus*. As the relationships between the families that were able to be included are poorly supported in Figure 5.15, the evolutionary history of the elements should therefore be interpreted primarily from the RT phylogeny (Figure 5.14).

5.4 *Ty4*-like phylogenies

Neuvéglise *et al.* (2002) proposed that *Ty4* was gained shortly before the divergence of the ancestor of *Saccharomyces*. Here, the family was discovered beyond the *sensu stricto* complex in other post-WGD genera, and the dataset contains the fewest intact coding sequences. Table 5.8 summarises the potential HT events and stochastic losses observed in the *Ty4* family.

Horizontal transfer			Stochastic loss	
Potential events		Success	*n* of	Proportion
RT	LTR	ratio	families	lost
3	26	0.23	5 of 20	0.25

Table 5.8: **Potential HT events and stochastic loss in the *Ty4* family.** Stochastic loss and occurrence of HT events in RT and LTR phylogenies were counted as in Table 5.6. Subfamilies determined by population isolation were counted separately.

5.4.1 A geographical split in the *Ty4* RT phylogeny

The RT phylogeny for *Ty4*-like sequences is displayed in Figure 5.16. Overall, the RT topology is congruent with that of the species phylogeny, however the relationships between the sequences of the non-*Saccharomyces* group of species - with the exception of the sister relationship of *T. blattae* and *V. polyspora* (91%mlBP; 0.98biPP) - are moderately supported at best. The methods disagree regarding the positioning of the *K. servazzii* sequence, as BI instead places it as the outgroup to this smaller clade, but its position has very little support from either method (<50%mlBP; <0.7biPP). It was interesting to note that, as in the *Ty3/gypsy* phylogeny, RT sequences from *Tetrapisispora* perhaps did not share an immediate ancestor.

Nakaseomyces bacillisporus RT falls as the outgroup to sequences from the *Saccharomyces* species (99%mlBP; 1.0biPP) which is congruent with the species phylogeny and therefore vertical inheritance of *Ty4*. No other *Ty4*-like sequences were located in the genomes of this genus. An apparent *Ty4*-like RT sequence was however located in a contig of *Spathaspora hagerdaliae*, sharing more than 90% similarity with that of *Saccharomyces* elements. However, upon examination of the element's flanking DNA, it proved to be *Saccharomyces* contamination in the sequencing project of *Spathaspora* species by Lopes *et al.* (2016).

Strikingly, RT sequences from *Saccharomyces* species are split by geographical origin (74%mlBP; 0.99biPP): those isolated from Europe and Russia (simplified to the European clade) and those from the Americas and East Asia (simplified to the American clade). Internal and terminal branch

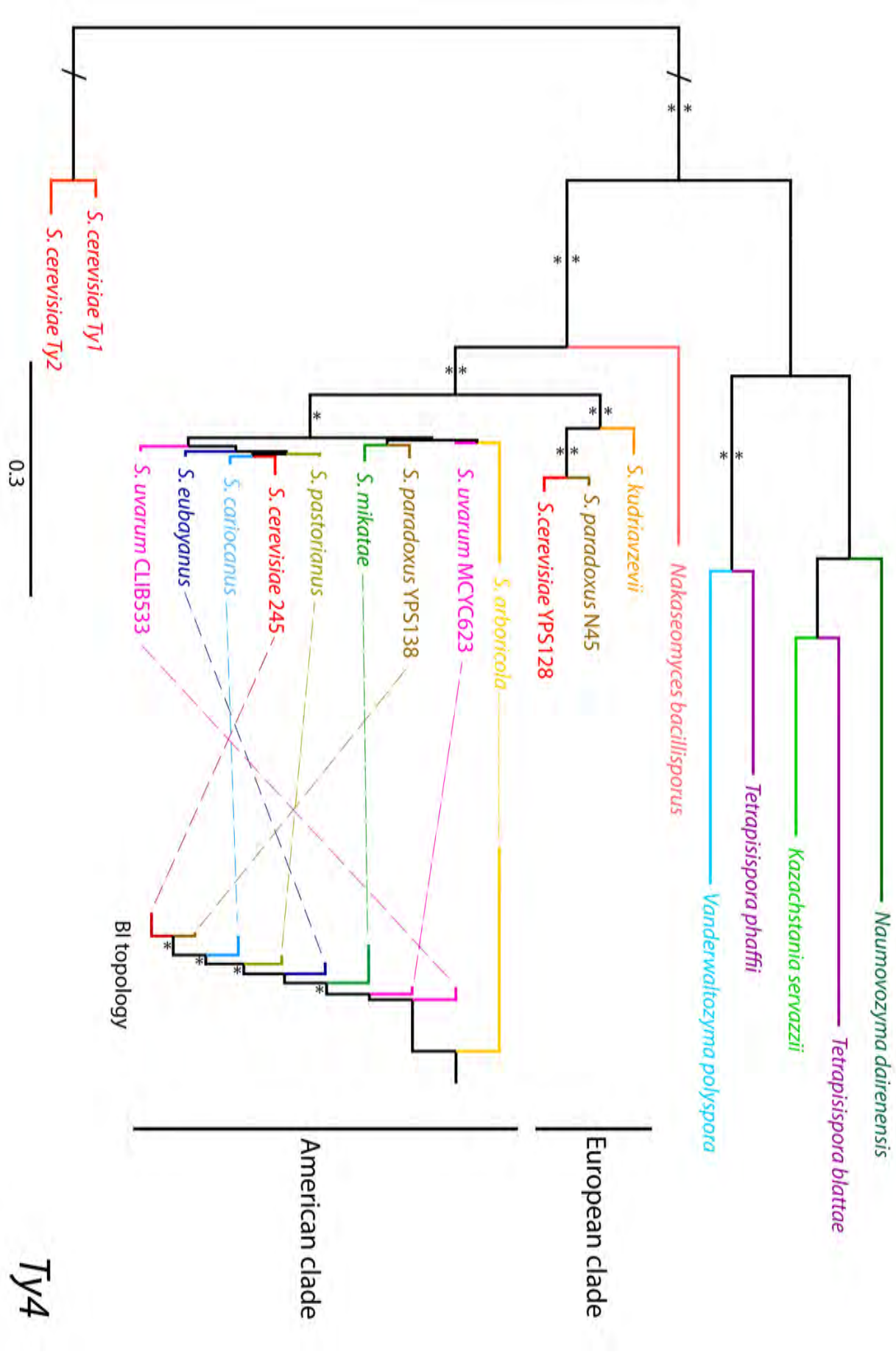

Figure 5.16: **RT phylogeny of *Ty4*-like sequences in *sensu lato* species.** The tree is based upon an alignment of 300 AA positions and rooted with *S. cerevisiae Ty1/2* sequences. Tree layout and labels are as in Figure 5.2. The BI topology of the American *Saccharomyces* clade is included as it differed from the ML topology. / indicates root shortened arbitrarily. Strain names were included to differentiate between sequences of different populations.

lengths for the *Saccharomyces* sequences, with the exception of *S. arboricola*, are all very short, indicating little divergence/recent ancestry and perhaps HT events.

The European clade consists of sequences from just three species: *S. kudriavzevii* and the closely related RT sequences of *S. paradoxus* and *S. cerevisiae*. The positions of these sequences is very high to maximally supported by both methods (83-95%mlBP; 1.0biPP, respectively). The relationship between RT sequences of *S. paradoxus* and *S. cerevisiae* reflects that of the species phylogeny. Interestingly, the South American species *S. cariocanus* contains a single European-like FLE, but as its RT sequence has degraded, it was excluded from the phylogeny.

The American clade ML topology produced poor support values, the highest of which was <70%mlBP for the monophyly of the American clade. BI topology differed (see right side of American clade in Figure 5.16) but unlike ML, generated very high support for four branches. However, this Bayesian topology suggests that all American RT sequences evolved from a *S. uvarum*-like ancestral sequence, followed by a series of divergences that are incongruent with the species phylogeny. Prior to the addition of *S. cerevisiae* strain 245 and *S. pastorianus* sequences, those of *S. cariocanus* and *S. eubayanus* clustered together with very high support, suggesting an HT event (data not shown). The partial RT sequence of *S. arboricola* was recovered from a poor quality contig and shares an unsupported branch with that of *S. uvarum* MCYC623 in the ML topology. As the only *S. uvarum* and *S. eubayanus* RT sequences were recovered from American strains, they are not necessarily representative of the elements in these species. However, no intact *Ty4* RT sequences were found in European strains of *S. uvarum*. Additionally, no intact American type of *S. kudriavzevii* RT was discovered, despite this species' isolation in Japan.

The almost identical sequences of *S. mikatae* and *S. paradoxus* YPS138 may also be indicative of a HT event, as this relationship differs from the species phylogeny. Interestingly, hybrid *S. pastorianus* (represented by strain M14) possesses an RT sequence distinct from that of its parental species *S. cerevisiae* and *S. eubayanus*, indicating active evolution. The *S. pastorianus* sequence falls within the American clade, suggesting its family was gained from its *S. eubayanus* parent as opposed to European *S. cerevisiae*, but is distinct from the sequences of either parent. Another potential HT event suggested by ML is that between elements of *S. cerevisiae* 245 and *S. cariocanus*, which in turn share a recent ancestor with RT sequences in *S. eubayanus* and *S. uvarum* CLIB533. Due to the differing topologies offered by the two methods, as well as the low level of divergence and/or recent evolutionary history in the American RT sequences represented

by very short branch lengths, the true relationships between most elements cannot be discerned with the currently available data.

5.4.2 Geographical distinctions in the *Ty4* LTR phylogeny

The geographical split in the LTR phylogeny (Figure 5.17) is more evident than that of the RT phylogeny. Sequences clearly separate into the European and American clades, which are examined in Figures 5.18 and 5.19 separately below. Although LTR sequences from *T. blattae* and *Nk. bacillisporus* share enough identity with *Saccharomyces* LTRs to allow alignment, they were removed from the final phylogeny as both ML and BI placed them within *S. mikatae* sequences on a long branch – an implausible topology.

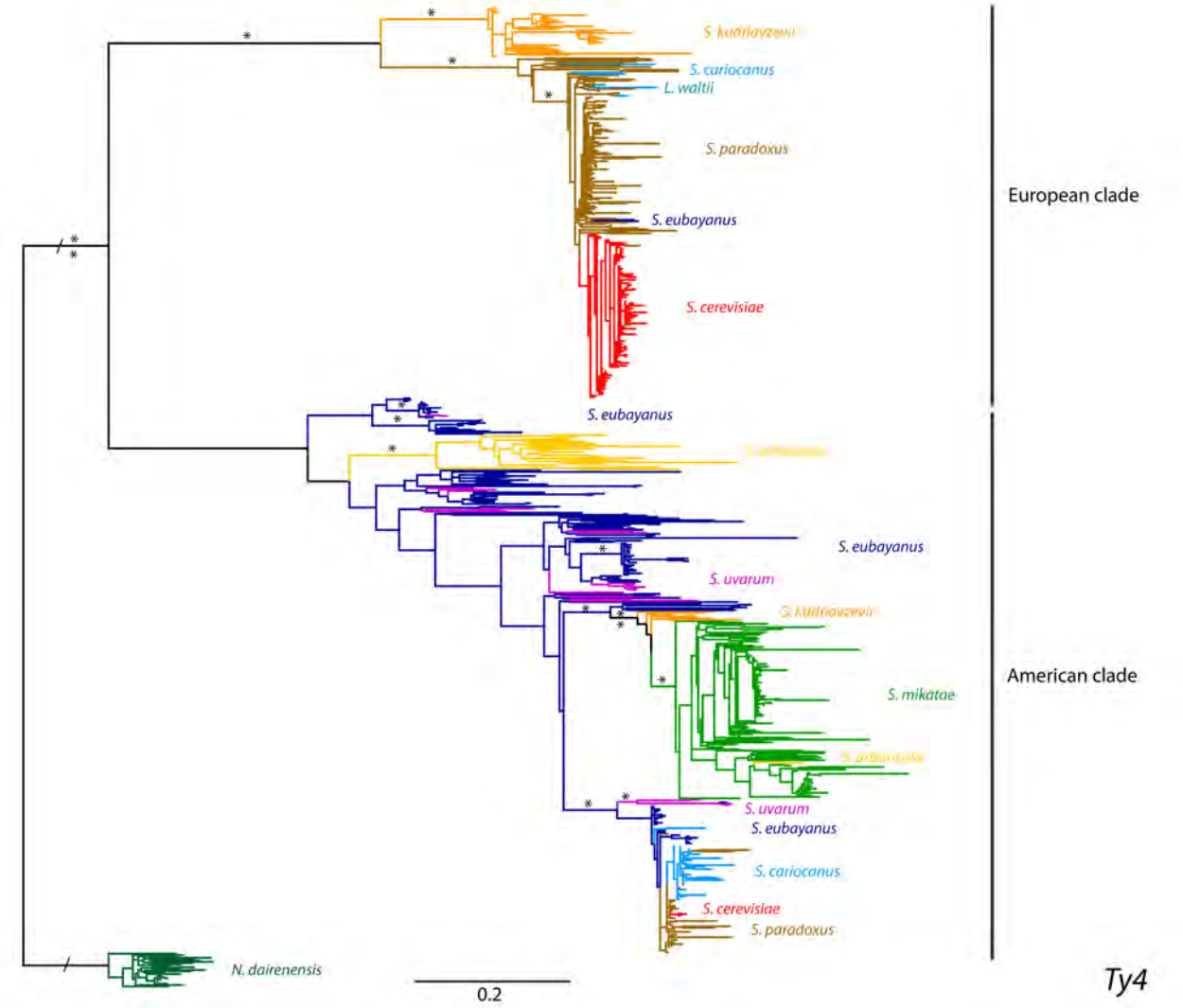

Figure 5.17: **The two clades of the *Ty4* LTR phylogeny in *sensu stricto* species.** The tree is based upon an alignment of 429 nucleotide positions and rooted with sequences from *N. dairenensis*. The two clades were explored separately in more detail in Sections 5.4.3 and 5.4.4. Layout and labels are as in Figure 5.2.

Their clades however displayed short terminal branches consistent with recent ancestry as suggested by significantly negative Tajima's *D* results (Table 5.5). Furthermore, LTR sequences from *T. phaffii, K. servazzii* and *V. polyspora* did not share enough identity to create a reliable alignment and so were rejected (Table 5.3).

Surprisingly, sequences from the SGRP strains of *S. cerevisiae* are almost entirely confined the European clade of *Ty4*, with only sequences from the American isolate L-1528 present within the American grouping. Additionally, non-SGRP strain 245 contains a single copy of an American-type FLE, the RT sequence of which shares a branch with *S. cariocanus* in Figure 5.16, suggesting that this species may be the origin of a HT event into this population of *S. cerevisiae*. Discovery of this foreign element was independently reported by Legras *et al.* (2018). No further evidence of the American sequences was discovered in *S. cerevisiae*, suggesting that this HT event may currently be confined to these small populations.

5.4.3 Evident HT events between species in the American clade

Figure 5.18 displays the LTRs in the American clade of *Ty4*, in which three distinctions, also visible in the alignments, are made between sequences. A grouping of divergent LTRs contains sequences of *S. eubayanus, S. uvarum* and *S. arboricola* (Figure 5.18, top). A second grouping of *Tse4/Tsu4*-like sequences contains those from *S. kudriavzevii* and *S. mikatae* in addition to the three previous species (Figure 5.18, middle). Finally, a very distinct grouping of the shorter LTRs consists of *S. eubayanus* and *S. uvarum* sequences with those of *S. paradoxus, S. cariocanus* and *S. cerevisiae* nested within (Figure 5.18, bottom and inset). The maximally supported rooting of this phylogeny within *S. eubayanus* sequences is likely a reflection of the age and diversity within this species, rather than it being the true source of elements in the other species. Two smaller, unsupported groupings of *S. eubayanus/S. uvarum* sequences in the divergent and *Tse4/Tsu4*-like clades likely represent phylogenetic artefacts, in that these should be two clades each with a single root, but have separated onto multiple branches. In the divergent grouping, the vast majority of *S. arboricola* sequences (*n*=26) are nested within those of *S. eubayanus* in the Bayesian tree on a long branch (<0.7biPP), but this relationship was not inferred in the ML phylogeny, and instead the sequences cluster together between branching *S. eubayanus* clades, again with poor support (<50%mlBP). The remaining LTRs (*n*=2) fall within *S. mikatae* sequences. All *S. arboricola* LTRs are solos and therefore representative of extinct lineages, possibly gained from *S. eubayanus* and

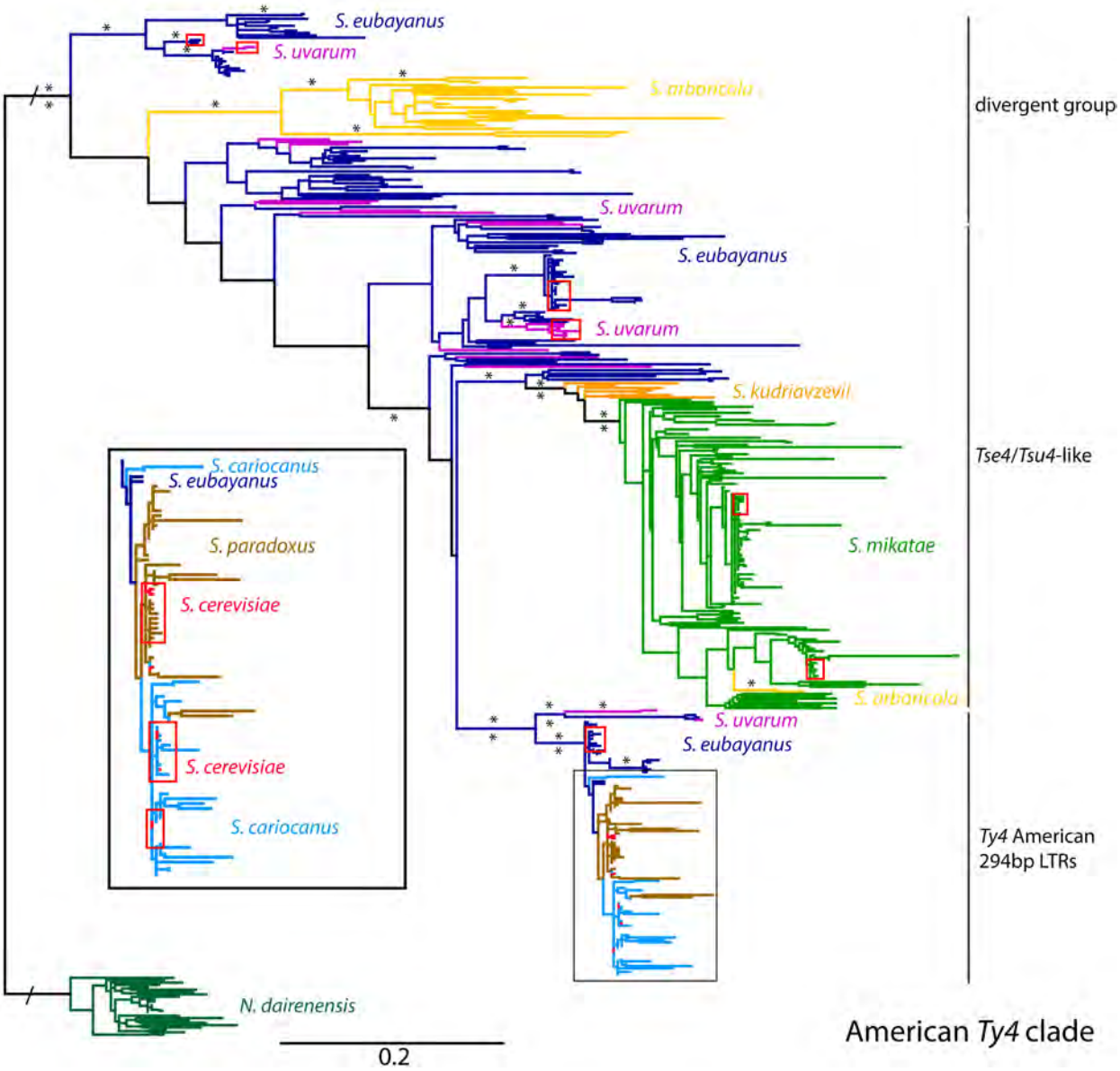

Figure 5.18: **LTRs in the American clade of *Ty4*.** The alignment upon which this tree was based is taken from that of Figure 5.17 and rooted with *N. dairenensis* sequences. The boxed region is magnified to display potential HT events. Small red boxes indicate LTRs associated with FLEs. The three main distinctions between sequences were also visible in the alignment. Layout and labels are as in Figure 5.2. Significantly negative Tajima's *D* values for *S. paradoxus, S. cariocanus* and *S. mikatae* (Chapter 4) supported the presence of short terminal branches and recent ancestry of the LTRs in these species.

a single unsuccessful transfer of a solo LTR from *S. mikatae*. It is unclear as to whether this family is lost in all strains of *S. arboricola*, as although a partial RT sequence was recovered from a New Zealand strain (Chapter 4), the state of the element could not be elucidated due to the short length of the sequencing reads and low quality assembly. As the majority of LTRs are long-branched, it would suggest this family has been present in *S. arboricola* for a relatively long time, and due to the lack of short-branched LTRs, has not been recently active.

Both the divergent and *Tse4/Tsu4* groups contain a greater number of long-branched LTRs in strains of *S. eubayanus* and *S. uvarum*, implying that these copies have not been active recently as time has allowed the accumulation of mutations. At least one LTR associated with a FLE is found in each of the main groupings of sequences from these species (red boxes in Figure 5.18), suggesting a complex and ancient history. The close relationship between the LTRs of *S. uvarum* and *S. eubayanus* is potentially the result of hybridisation events and the sharing of the *Ty4*-like family. 20 solo LTR sequences from *S. uvarum* are nested within *S. eubayanus*, and LTRs from FLEs (*n*=5) are also present in two clusters, likely the result of at least two independent HT events and subsequent transposition in the new *S. uvarum* genome.

Although *S. kudriavzevii* sequences (*n*=7) branch immediately before those of *S. mikatae*, it is unlikely that the former was the donor in a transfer, as this relationship is instead indicative of vertical inheritance. ML topology is similar to that of the BI tree, but only the position of the initial *S. kudriavzevii* sequences is highly supported by both methods (76%mlBP; 1.0biPP). The branch placing *S. kudriavzevii* within *S. eubayanus* sequences is strong to maximally supported by both (74%mlBP; 1.0biPP), suggesting a source for the HT into *S. kudriavzevii* as this relationship is incongruent with that of the species phylogeny. All *S. kudriavzevii* LTRs in this clade are solos, therefore the transfer of this American-type family was likely unsuccessful.

S. mikatae sequences split into two sublineages immediately in the BI topology but further towards the terminal clades in the ML phylogeny. LTRs associated with FLEs are found in both groups, the smaller of which also contains the unsuccessful additional transfer into *S. arboricola*. Long-branched sequences (*n*=29) are common in *S. mikatae*, as in *S. arboricola* and *S. eubayanus*, indicative of past activity and subsequent accumulation of mutations.

In the lower region of the tree containing the 294bp LTRs, the positions of *S. paradoxus* and *S. cariocanus* within a *S. eubayanus* clade received very high to maximum support from both methods (99%mlBP; 1.0biPP). This reflects the close relationships observed in the RT tree between the species and confirms *S. eubayanus* as the most likely original donor. The nesting of the majority of *S. cariocanus* sequences (*n*=38) within *S. paradoxus* indicates vertical transmission as opposed to HT, with what is known about these species undergoing the speciation process. The *S. cariocanus* sequences represent the high numbers of activity unique to this subspecies. This family likely originated ultimately in *S. eubayanus*, however, as it is sequences from this species in which *S. cariocanus* and *S. paradoxus* are nested. A direct but unsuccessful transfer of LTRs from

S. eubayanus into *S. cariocanus* (*n*=2) also appears to have occurred without involving *S. paradoxus* as an intermediate (top of the magnified box region), therefore occurring in the time since the divergence of *S. cariocanus* from *S. paradoxus*. Most LTRs in *S. cariocanus* are from FLEs and therefore discounted from phylogenetic analysis as they were identical, resulting in apparently fewer sequences in this species. The species does however contain a number of long-branched solo LTR sequences (*n*=8), indicating that this family has been present long enough for mutations to accumulate. A similar number in *S. paradoxus* (*n*=7) illustrates an analogous story. The acquisition of the *Ty4* family from *S. eubayanus* is likely to have occurred shortly prior to the speciation event in order for both species to gain it and for the LTRs to diverge in a similar fashion. *S. paradoxus* however does contain a larger number of short-branched LTRs (*n*=22), but this may simply be a reflection of the number of sequenced strains available, as opposed to its level of activity, i.e. minimally active in multiple strains. Sequences in *S. paradoxus* within this clade are confined to just five strains, all of which were isolated in the USA, Canada or Hawaii.

Additional LTRs (*n*=3) from two strains, 245 and 460, of *S. cerevisiae* nesting within the *S. paradoxus/S. eubayanus*-like 294bp LTRs are indicative of HT. Both ML and BI placed these within *S. paradoxus sequences*, indicating the successful transfer of a FLE into strain 245 and but the failed HT into strain 460, by the presence of a single solo LTR. All SGRP *S. cerevisiae* strains were screened for partial/degenerate American *Ty4* sequences. Most SGRP strains contain a 147bp stretch with 76% identity shared with the American LTR query sequence. However, when aligned with the other sequences the partial copies fell on their own branch and did not nest within potential donor sequences (data not shown), forcing the conclusion that the hits were a false positive and the strains did not contain evidence of a degenerate family akin to *Ty3p*.

5.4.4 Fewer examples of HT in the European clade

Figure 5.19 displays the sequences of the European clade, the topology of which is relatively consistent between both inference methods. While the relationship between *S. cerevisiae* and *S. paradoxus* RT sequences is consistent with vertical inheritance (Figure 5.16), the nesting of *S. cerevisiae* LTR sequences within those of *S. paradoxus* suggests that HT has occurred between these species. The possibility of this positioning being the result of vertical transfer cannot be discounted however. *S. paradoxus* contains long-branched LTRs indicating relative age, whereas these are absent in *S. cerevisiae*. Additionally, *S. cerevisiae* contains solo LTR evidence of failed

transfers back into *S. paradoxus* (*n*=7). *S. cariocanus* sequences are placed throughout *S. para-*

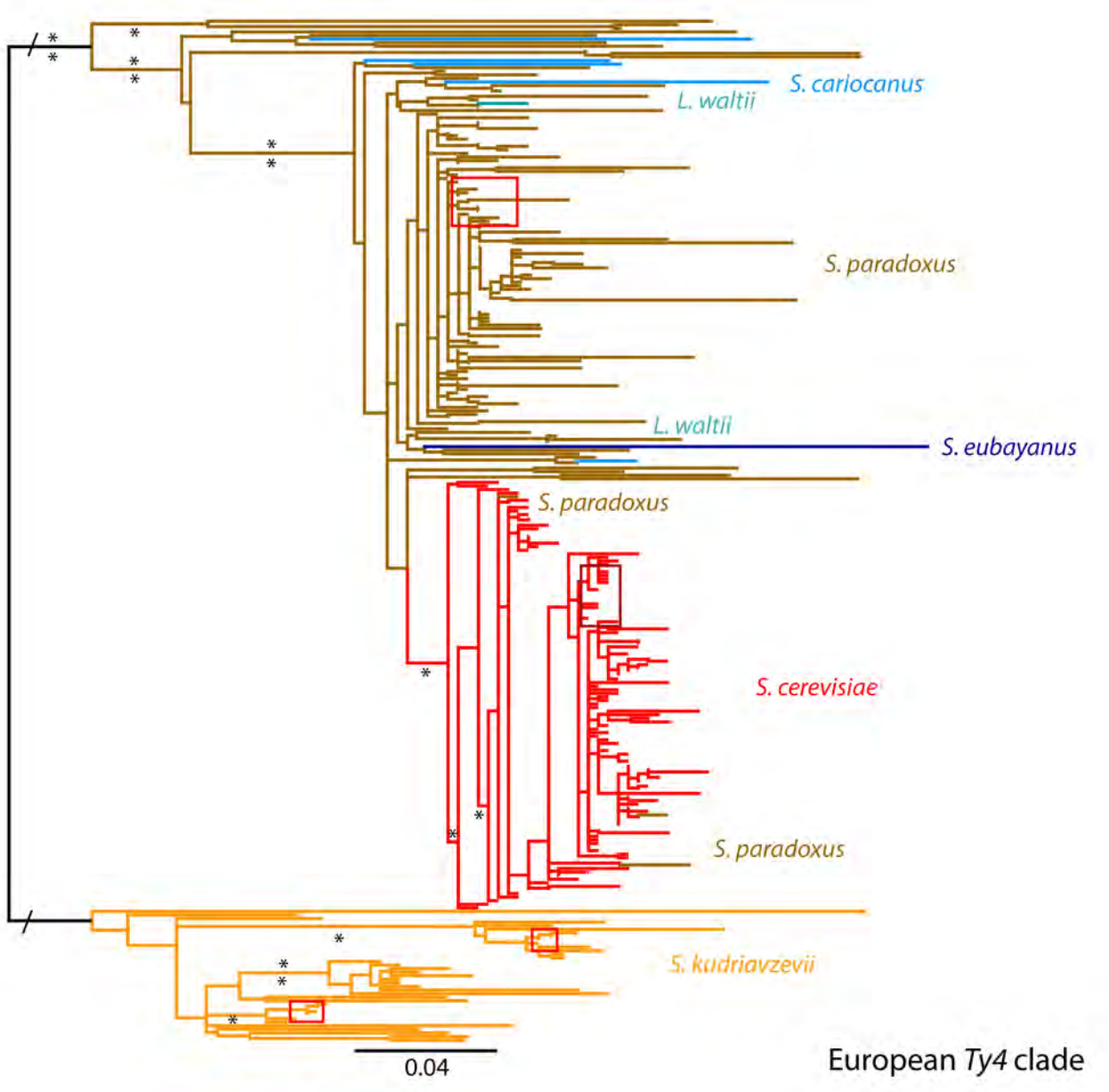

Figure 5.19: **LTRs in the European clade of *Ty4*.** The alignment upon which this tree was based was taken from that of Figure 5.17 and rooted with *S. kudriavzevii* sequences. Layout and labels are as in Figure 5.2. Red boxes indicate those LTRs associated with FLEs.

doxus (*n*=5), all of which are solos, which represent activity in *S. cariocanus* post-speciation. *S. paradoxus* also contains nested *L. waltii* (*n*=3) and *S. eubayanus* (*n*=1) sequences, all of which are solo LTRs and therefore representative of failed HT events. Additionally, two sublineages of *S. kudriavzevii* sequences emerge, both of which contain LTRs associated with FLEs, but this separation is weakly supported (<50%mlBP; <0.7biPP).

5.5 *Ty5*-like phylogenies

Like *Ty3*, the family was found to be widespread in many species, beyond fungi and into plants. Although not as limited as the *Ty4* family, RT sequences are confined to three *sensu stricto* species. Phylogenetically analysed here are the LTRs of all *sensu stricto* species, but the sequences of few *sensu lato* species are included as lack of identity meant that alignments became unreliable. Table 5.9 summarises the potential HT events and stochastic losses observed in the *Ty5* family.

Horizontal transfer			Stochastic loss	
Potential events		Success	*n* of	Proportion
RT	LTR	ratio	families	lost
2	9	0.29	10 of 22	0.45

Table 5.9: **Potential HT events and stochastic loss in the *Ty5* family.** Stochastic loss and the occurrence of HT events in RT and LTR phylogenies were counted as in Table 5.6. In this family, LTR numbers may not reflect the true frequency of HT events, as discussed below.

5.5.1 Ancient divergences in the *Ty5* RT phylogeny

The RT sequences in the *Ty5* phylogeny (Figure 5.20) split into yeast and other fungi, both of which subdivide further in a fashion congruent with the species phylogenies. The *Tca5* clade, named for a *Candida albicans* element, consists of sequences from yeast distantly related to the *sensu lato* species. Sequences within the *Ogataea* clade are next to diverge, the positioning of which is only supported by BI (<50%mlBP; 0.97biPP). The *Ogataea* RT sequences (*n*=8) cluster together with high or maximum support on all but one branch, immediately before the divergence of *O. parapolymorpha Top5* (<50%mlBP; <0.7biPP). The general layout of the *Ogataea* sequences is consistent with vertical transfer, with the only suggestion of HT between the sequences of *O. polymorpha* and *O. angusta* (100%mlBP; 1.0biPP).

The *sensu lato* clade for the most part reflects that of the host species phylogeny. As in previous families, placement of sequences from *T. blattae* and *T. phaffii* suggest the elements do not share an immediate ancestor. A similar ancestry is observed between the sequences of *Kazachstania* species, as they are split into two groupings.

The short branch lengths and close relationship of RT sequences in *S. cerevisiae* and *S. mikatae* (77%mlBP; 0.94biPP) are suggestive of an HT event as opposed to vertical inheritance, as this positioning is incongruent with that of the species phylogeny. Short-branched sequences

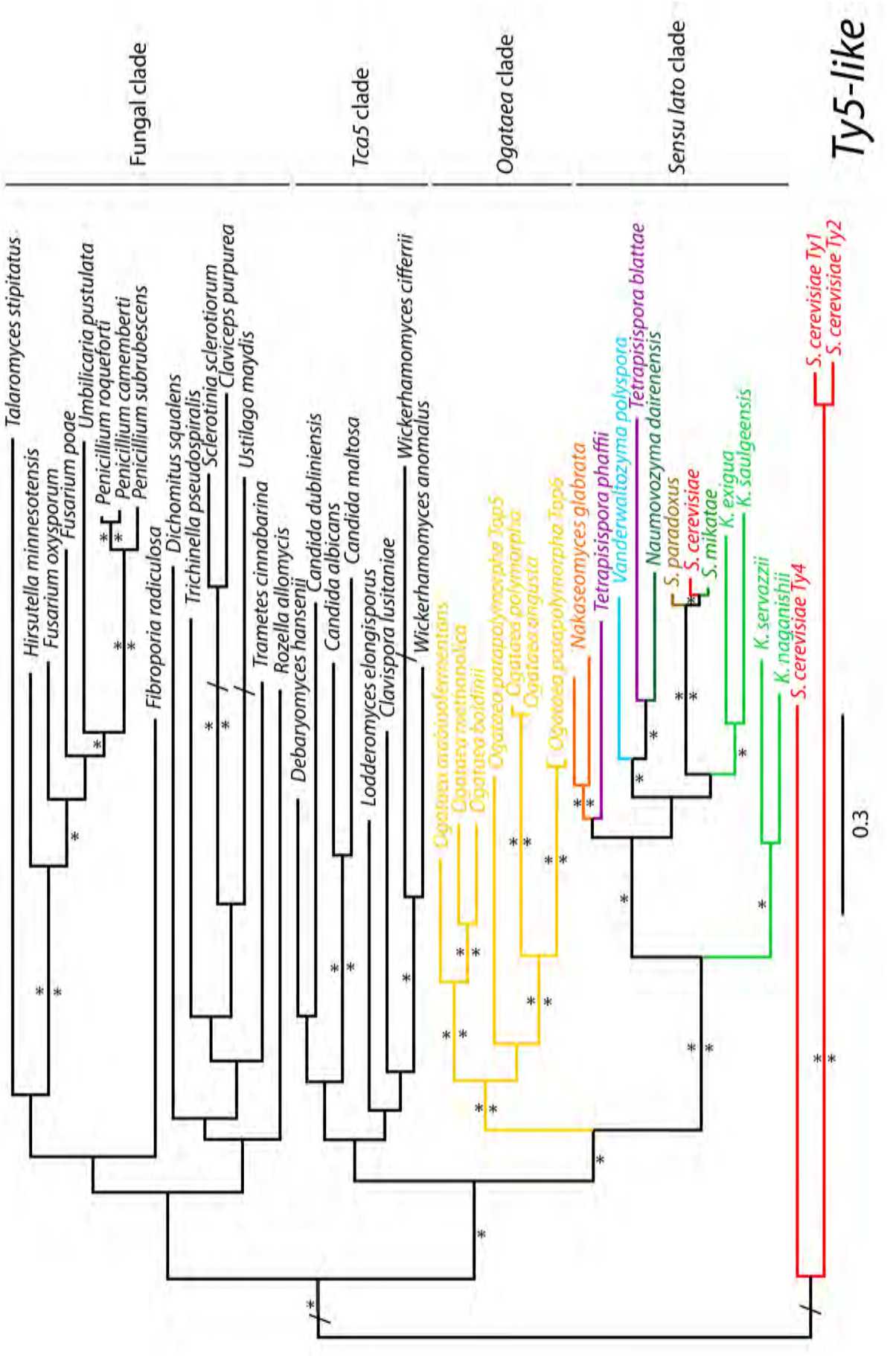

Figure 5.20: ***Ty5*-like RT phylogeny of yeast and other fungal species.** The tree is based on an alignment of 297 AA positions and rooted with *Ty1-2* and *Ty4* from *S. cerevisiae*. Tree layout and labels are as in Figure 5.2. / indicates arbitrarily shortened branches and root.

are confined to sequences of *Saccharomyces, Penicillium* and *Ogataea*, reflecting the likely recent history within these species, and the relative divergence of the remaining sequences.

5.5.2 *Ty5* LTR phylogeny

In contrast to the RT phylogeny, the LTR sequences of fewer species were confidently aligned (*n*=13; Figure 5.21). However, the achieved topology resembles that of Figure 5.20, and is congruent with species phylogenies. As they placed upon biologically implausible long branches within *Saccharomyces* sequences with zero support, LTR sequences of *K. exigua, V. polyspora, Nk. glabrata* and *Wickerhamomyces* species were removed from the final phylogeny (Table 5.3). Sequences from *K. servazzii* fall into the basal position, but only with strong ML support (78%mlBP), as the BI method instead places LTRs from *N. dairenensis* in the basal position (1.0biPP). The BI topology as a whole is poorly supported, with biologically implausible placements, such as the sequences of *T. blattae* on a long branch within *S. paradoxus* sequences. Further discrepancies between BI and ML topologies include the sister relationship of *K. servazzii* and *K. naganishii* sequences, again with poor support (<0.7biPP). ML does not significantly support three internal branches, but the remaining branches returned strong to maximum support.

The LTR sequences of the *sensu lato* species *K. naganishii, K. servazzii* and *N. dairenensis* display a mix of older, long-branched sequences together with relatively short-branched insertions with a recent common ancestry. *T. blattae* however, possesses only short-branched sequences, with fewer degraded LTRs discovered in its genome. The activity of this family is therefore likely to have been very recent, yet has since become extinct via a combination of LTR-LTR recombination and the accumulation of null mutations resulting in pseudoelements.

Sequences from all *Saccharomyces* species cluster together and nest within those of *S. paradoxus*, likely a reflection upon the age and diversity of sequences observed in these species. The direction of the potential HT event between *S. cerevisiae* and *S. mikatae* seen in the RT phylogeny remains unclear in the LTR tree. *S. mikatae* sequences (*n*=3) nest within those of *S. cerevisiae* and *S. paradoxus*, suggesting that *S. mikatae* is the recipient. However, support values are poor and branch lengths indicate any potential transfers were ancient, allowing nucleotide changes to occur in the sequences of *S. mikatae*.

Five families returned significant Tajima's *D* results (Table 5.5; Chapter 4) including those of *S. cerevisiae, S. paradoxus, K. servazzii* and *T. blattae*, which are all significantly negative. In

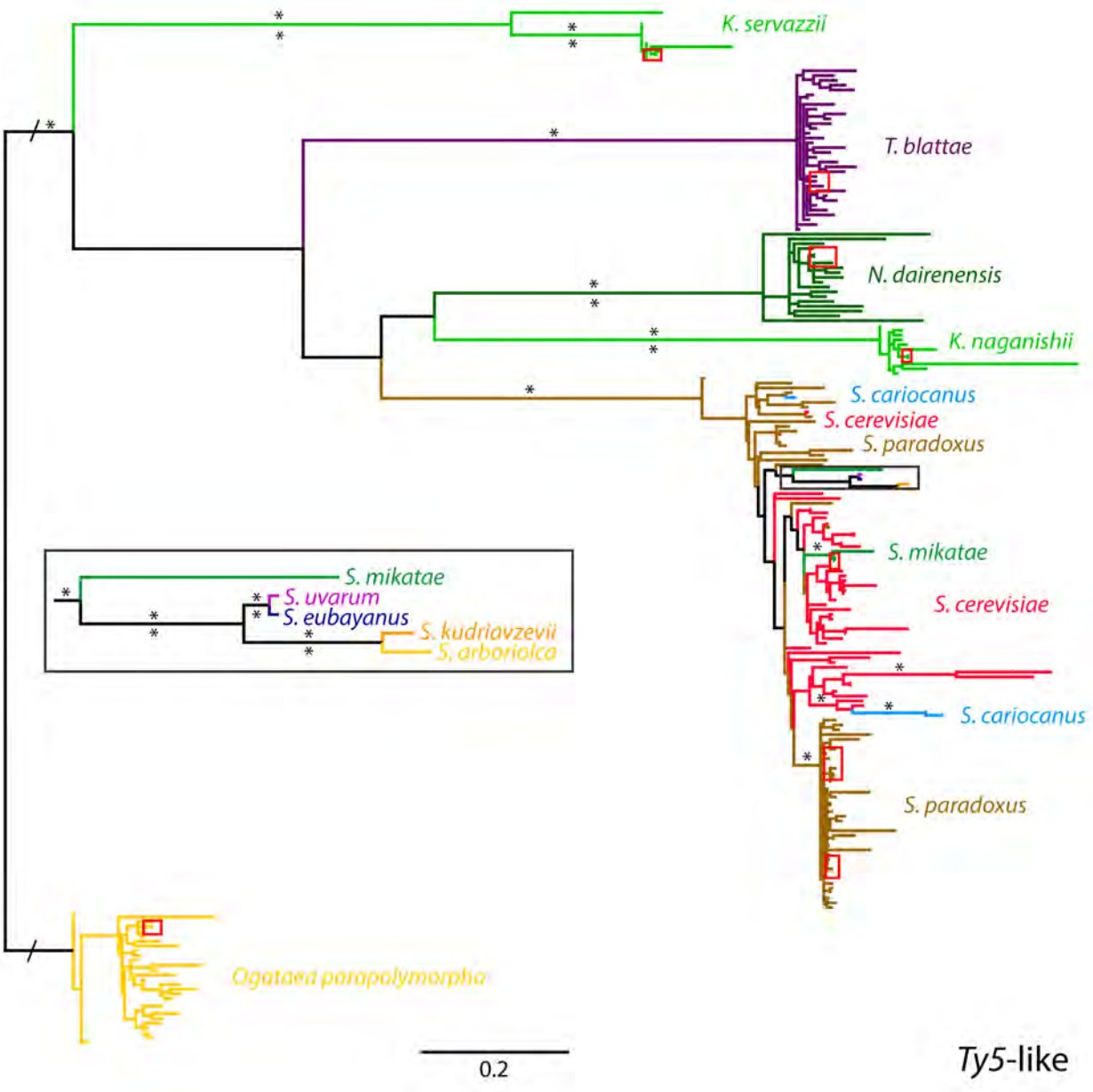

Figure 5.21: ***Ty5* LTR phylogeny of sequences in *sensu lato* species.** The tree is based upon an alignment of 235 nucleotide positions and rooted with *O. parapolymorpha* sequences. The boxed region is magnified in the middle of the figure. Sequences of *S. uvarum* and *S. eubayanus* possess highly similar flanking DNA, indicating this insertion pre-dated their speciation. A recombination event resulting in the solo LTR may have occurred in the ancestor or independently in each species before any transposition events. Layout and labels are as in Figure 5.2. Red boxes indicate those LTRs associated with FLEs.

these four families, recent ancestry and potential activity are indicated by short terminal branches, consistent with the results of Tajima's tests. The final family, *Tnkg5* of *Nk. glabrata*, returned a significantly positive value of *D* (Table 5.5). This suggests that multiple lineages are present in the genomes of this species, which is visualised in the phylogeny of Figure 5.22. Within the *Tnkg5* clade, a division between the sequences is observed, with each containing LTRs associated with FLEs. This is the stronger signal, as although the identical and highly similar sequences with short

terminal branch lengths indicative of recent activity and common shared ancestry would cause the *D* value to be more negative, Tajima's test resulted in a significantly positive *D* value.

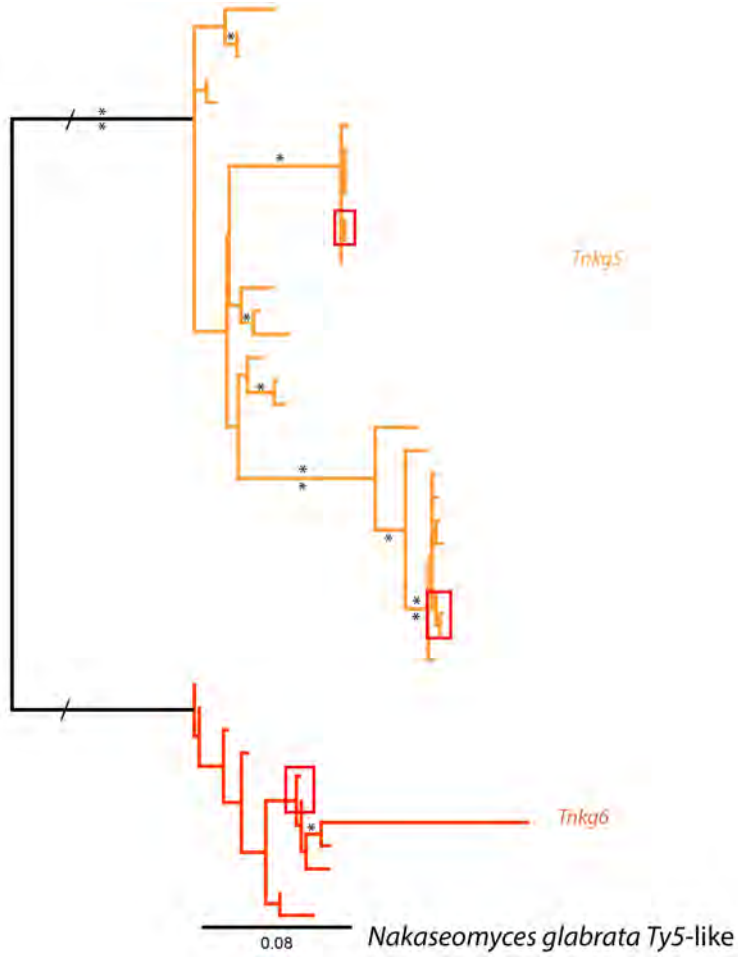

Figure 5.22: **LTR phylogeny of sequences in the two *Ty5*-like families in *Nk. glabrata*.** The tree was based upon an alignment of 512 nucleotide positions and midpoint rooted. Layout and labels are as in Figure 5.2. Red boxes indicate those LTRs associated with FLEs.

5.6 Discussion

This chapter presents the possible phylogenetic relationships and HT events of *Ty*-like sequences in *Saccharomycetaceae* yeast. Protein phylogenies were created for each family using the RT domain as this is well conserved across the elements of different species (Eickbush and Jamburuthugoda, 2008). Previous studies showed that RH (Malik and Eickbush, 2001), IN (Malik and Eickbush, 1999; Llorens and Marin, 2001) and even PR (the least conserved domain of *pol*; Llorens *et al.*, 2009) all support RT phylogenies and therefore evolutionary history. Corresponding LTR phylogenies were created to establish possible direction of HT events, and to examine the complex evolution of TEs. LTR phylogenies are particularly useful for species in which a particular family has undergone loss of coding regions, thereby establishing its possible position in the evolutionary history of the elements where RT sequences are unavailable.

Yeast and other fungi are some of the most sequenced microorganisms to date, yet gaps in data are still apparent. Although the results here support a number of findings by Neuvéglise *et al.* (2002) and Liti *et al.* (2005), both research teams were working with very limited data. Sampling bias, evident in the high number of genomes assemblies available for *S. cerevisiae*, and undersampling were both apparent in this study, with the latter potentially contributing to discrepancies between RT and LTR phylogenies. Although the amount of genomic data is increasing, many species outside of *Saccharomyces* are represented by a single strain. Where some conclusions may be drawn about the elements from these genome assemblies, the likelihood of other strains and populations of any given species containing very different TE landscapes should be taken into consideration. An excellent example of this is the diversity observed across populations of *S. cerevisiae* (Bleykasten-Grosshans *et al.*, 2013). Surprisingly few species were found to contain as many TE families as observed in *S. cerevisiae*. It has been previously noted that the *S. cerevisiae* genome contains more full-length elements than any other known yeast species (reviewed by Bleykasten-Grosshans and Neuvéglise, 2011). These differences in genomic landscapes and interactions with their elements, even across genera, highlight the fact that defence mechanisms employed are very much individual to the host species (Neuvéglise *et al.*, 2002).

Sequences in *Saccharomyces* are far less diverse than those of other genera, illustrating the relatively short history of *Ty* elements in this genus. The potential history of each family in *Saccharomyces sensu stricto* species is summarised in Figure 5.23 and detailed, along with *Saccharomyces*-specific family histories, in familial sections below.

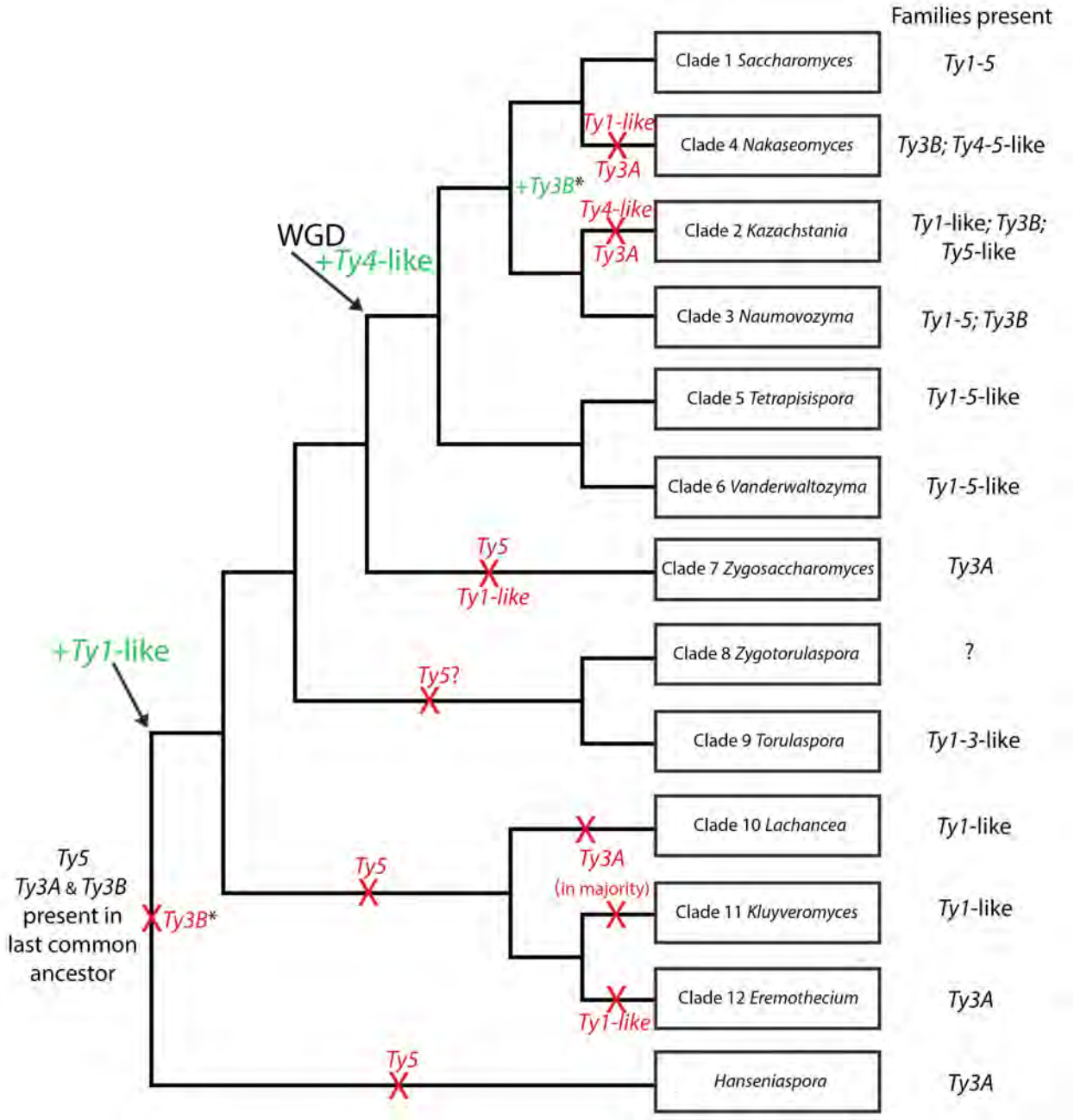

Figure 5.23: **Summary cladogram of the likely losses and gains of *Ty*-like families.** The history of *Ty*-like element families in relation to divergences of the 12 clades of *Saccharomyces sensu lato* species and the WGD event. Gains (green +) and losses (red x) primarily occurred in an ancestor and are applicable to the majority of species in the clade, e.g. *Ty4*-like is present only in a minority of *Kazachstania*, indicating multiple losses, rather than a single loss in an ancestor. *Ty5* and *Ty3* were likely present in the last common ancestor as families were discovered in species pre-dating the divergence of *Hanseniaspora*. **Ty3B* has two possible options: 1) the loss of *Ty3B* pre-dates the *sensu lato* ancestor, and was regained by the ancestor of clades 1-4 before being lost again in *Saccharomyces* (illustrated in the figure); or 2) *Ty3B* was lost sporadically throughout, but retained in clades 2-4. The loss of *Ty3A* is confined to a minority of clades (2, 4, 10-11). The source of *Ty4* is unknown, and was likely gained around the time of the WGD, as it does not appear in species that did not undergo this event. A series of losses of the *Ty5*-like family account for its sporadic dispersal across the *sensu lato* complex. As no *Zygotorulaspora* species have been sequenced, the TEs in this genus are unknown, so the loss of *Ty5* in the ancestor of clades 8-9 is based speculatively on the loss in *Torulaspora*.

History of the *Ty1/2* superfamily

Ty1-2 were named for the order in which they were discovered in *S. cerevisiae* (Boeke and Sandmeyer, 1991), and along with a number of subfamilies, constitute the *Ty1/2* superfamily. By examining the dispersal of these elements across species, they were most likely gained by the last common ancestor after the divergence of the *Hanseniaspora* clade (Figure 5.23), as a single *Ty1*-like family, which then underwent species-specific divergences. These conclusions are in agreement with those drawn by Neuvéglise *et al.* (2002). According to this theory, independent losses therefore occurred in *Nakaseomyces* (Clade 4), *Zygosaccharomyces* (Clade 7) and *Eremothecium* (Clade 12), as the family remains in species of the surrounding clades. Sequencing of any species of *Zygotorulaspora* may shed light on the possibility of further independent loss of this superfamily.

Based upon a probe hybridisation technique using elements from *S. cerevisiae*, Liti *et al.* (2005) hypothesised that the *Ty1* family emerged or was gained shortly before the divergence of *S. mikatae*, as the remaining species displayed no hybridisation signals. However, it is now known that all *Saccharomyces* species possess the *Ty1/2* superfamily. The authors' conclusions were likely the result of the limited data produced by the technique used and lack of shared identity in the elements of *S. kudriavzevii* and *S. uvarum* (*S. eubayanus* and *S. arboricola* were discovered almost a decade later). The ancestral *Ty1*-like element, present in an early *Saccharomyces* ancestor, diverged with the species to form the distinct elements observed. Although *Ty2* shares this same *Ty1*-like ancestor (Figure 5.2), *Ty2* was once thought to be a recently arisen subfamily of *Ty1* (e.g. Kim *et al.*, 1998) until Carr *et al.* (2012) demonstrated the HT event from *S. mikatae* into *S. cerevisiae*. However, as no major evolutionary events such as host speciation occurred during the existence of *Ty2* within *S. mikatae*, this species as *Ty2*'s true origin is questionable. Divergence from a *Ty1*-like element in a *S. mikatae* ancestor is possible as genomic turnover in *Saccharomyces* is high (Jordan and McDonald, 1999*c*). Alternatively, recombination with host DNA or with elements of another family (Jordan and McDonald, 1998) may have introduced enough variability for *Ty2* to emerge. Given the propensity for *Ty2* to undergo HT, it may have been gained from a currently unknown source. However, according to this theory, *Ty2* would then assume a sister relationship with *Ty1*, rather than nest within, as seen in Figure 5.4. The possible evolutionary history of the *Ty1/2* superfamily within *Saccharomyces* species is displayed in Figure 5.24. This cladogram, as with those created for the remaining families, are intended to infer time only in relation to speciation of host species, rather than the exact timing of element-related events.

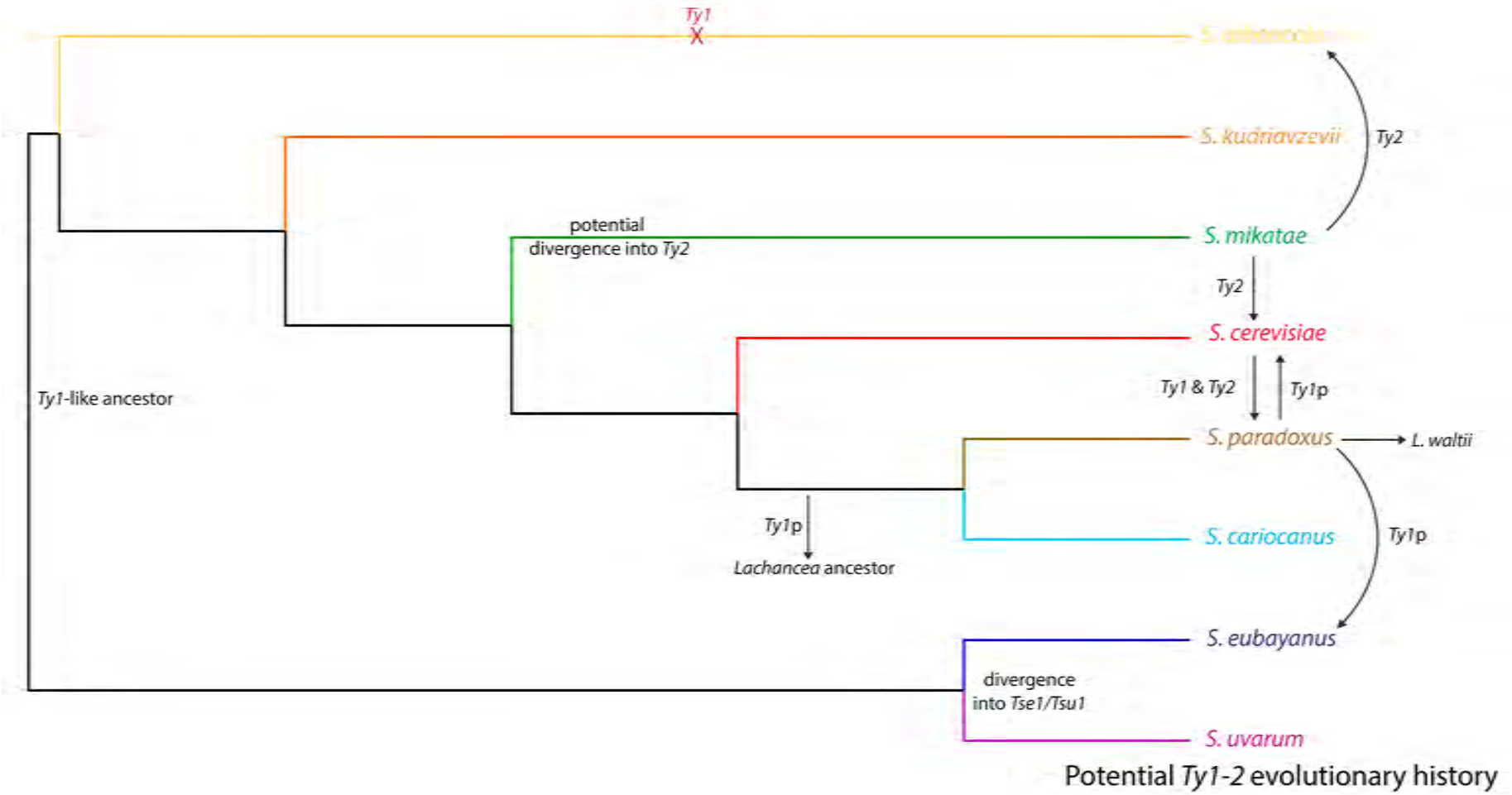

Figure 5.24: **Summary cladogram of the possible evolutionary history of the *Ty1/2* superfamily.** In *Saccharomyces* species, the family likely originated with a *Ty1*-like ancestor only, with the possible divergence into *Ty2* in an ancestor of *S. mikatae*. The possibility of *Ty2* being donated to *S. mikatae* from an unknown, external donor cannot be discounted at this stage. Evidence recorded here suggests the HT of *Ty1p* in *S. paradoxus* into a *Lachancea* ancestor, *L. waltii*, *S. eubayanus* and *S. cerevisiae*. Furthermore, the transfer of *Ty2* from *S. mikatae* into *S. arboricola* and *S. cerevisiae* (with a further HT into *S. paradoxus*) and *Ty1* from *S. cerevisiae* into *S. paradoxus* were documented here. The divergence into *Tse1* and *Tsu1* occurred with the speciation of *S. eubayanus* and *S. uvarum*, respectively.

History of *Ty3/gypsy*

Present in all eukaryotic supergroups and known to long pre-date the divergence of the Ascomycota ancestor (Llorens *et al.*, 2008, 2009), *Ty3/gypsy* is the most widespread of families. In addition to its age, it may be more persistent than other families. Neuvéglise *et al.* (2002) noted phylogenetic divisions of *Ty3/gypsy* lineages were possible, given an increase in data. Their work failed to produce clades containing only yeast sequences, but found *Ty3/gypsy* elements across all branches. Supporting this finding, almost all species surveyed here contained at least pseudoelement evidence of *Ty3/gypsy* families and with the increase in used data, the monophyletic clade of *Ty3A* is presented, containing sequences from *Saccharomyces* to the more distantly related *Eremothecium* yeast. Drinnenberg *et al.* (2009) noticed a divide in *Ty3/gypsy* elements between *Ty3*-like and *C. albicans Tca3*-like. However, this may have been a result of the misidentification of *Ty3/gypsy* elements within *C. albicans*, as the *Tca3* RTs shared so little similarity with *Ty3*-like that they were not included in the phylogenetic analyses here. Including a greater number of domains in the phylogenetic analyses may improve resolution.

The simplest explanation for the complex dispersal of subtypes observed in the *sensu lato* yeasts indicates that both *Ty3A* and *Ty3B* were present in their last common ancestor (Figure 5.23). Clades then underwent sporadic stochastic loss of one or both subtypes: *Ty3B* was independently lost throughout *Saccharomyces*, clades 5-12 and *Hanseniaspora*, but was retained by the species of clades 2-4. Additionally, *Ty3A* was lost in *Nakaseomyces*, and most *Kazachstania*, *Lachancea* and *Kluyveromyces* species. Alternatively, the loss of *Ty3B* pre-dated the *sensu lato* ancestor, but was later regained by the ancestor of clades 1-4 before being lost again in *Saccharomyces* only.

Although *Ty3* is a long-term inhabitant of *Saccharomyces* (Fingerman *et al.*, 2003), a reliable conclusion regarding the evolutionary history of this family cannot be made with currently available genomic data. The work here built on that of Carr *et al.* (2012), supporting their suggestion that at least two possible scenarios could account for the dispersal of the distinct *Ty3* and *Ty3p* types in *Saccharomyces* (Figure 5.25). Evidently once active in most species of the *Saccharomyces* clade, *Ty3* is not apparent in the *S. eubayanus/S. uvarum* lineage. Carr *et al.* (2012) used partial *Ty3p* sequences of *S. cerevisiae* to show the insertions were possibly gained from *S. paradoxus*. A similar approach here using the sequences of *S. eubayanus* and *S. uvarum* was unsuccessful (data not shown). However, if this lost family was present in the *Saccharomyces* ancestor, remnants in *S. eubayanus* and *S. uvarum* would not be expected to nest within sequences of another species.

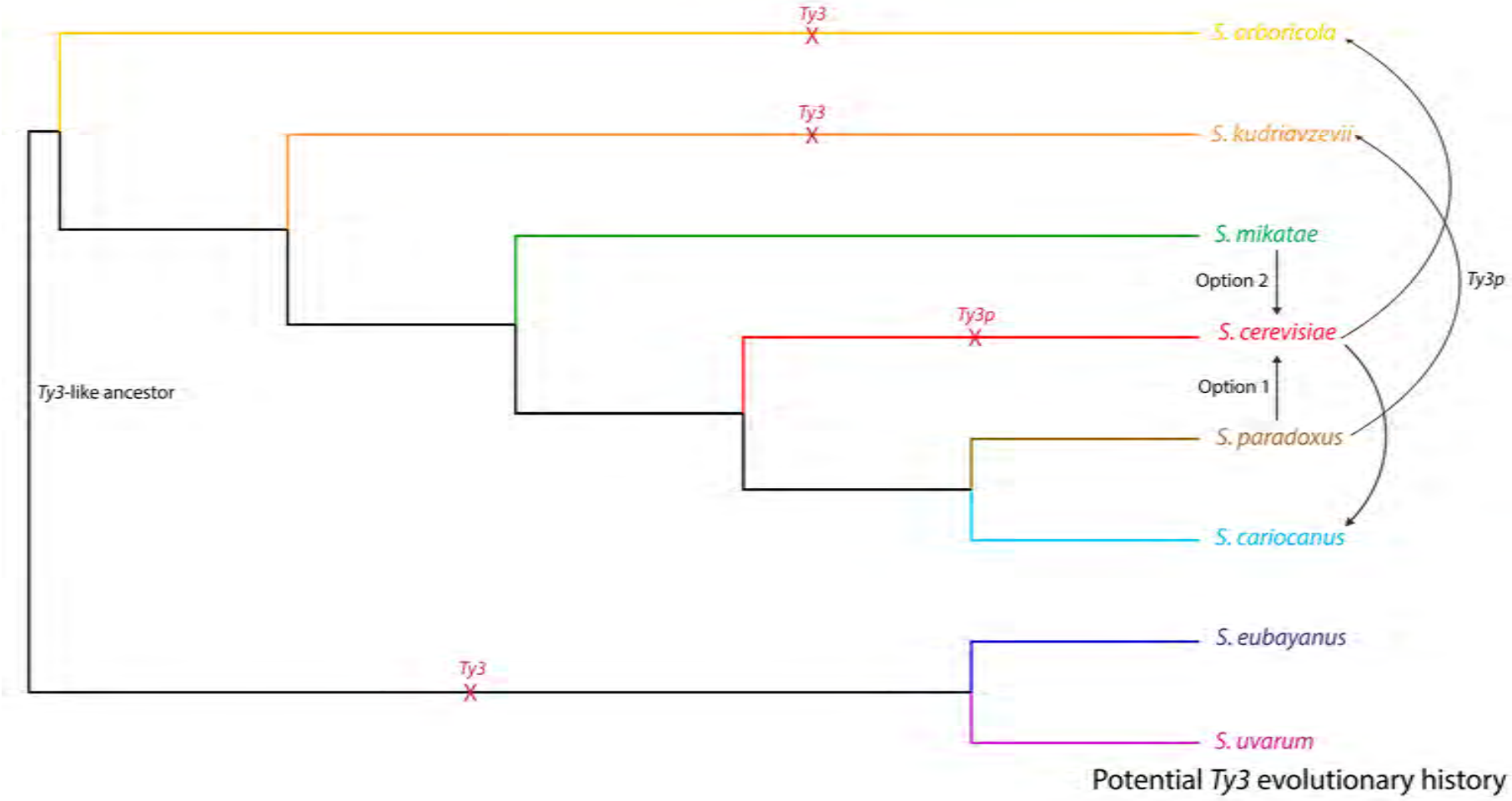

Figure 5.25: **Summary cladogram of the possible evolutionary history of the *Ty3* family.** Two plausible options for *Ty3* are illustrated, in both of which the family was lost in the lineage of *S. eubayanus* and *S. uvarum* prior to their speciation. Option 1: *Ty3* is the ancestral copy and diverged into species-specific elements (e.g. *Ty3p* in *S. paradoxus*) via vertical inheritance. It was lost independently in *S. arboricola* (twice) and *S. kudriavzevii* but remained functional in the others. Additionally, single unsuccessful transfers occurred from *S. cerevisiae* into *S. arboricola* and *S. cariocanus*. *Ty3p* from *S. paradoxus* was donated to *S. cerevisiae* and *S. kudriavzevii*, but became extinct in the former and spread in the latter. A weakness to this option is the poorly supported phylogenetic positioning of the heavily degraded *Ty3p* sequences in *S. cerevisiae* (Carr *et al.*, 2012). Option 2: The elements again diverged into species-specific lineages, with *Ty3p* as the ancestral copy. *Ty3p* became extinct in all but *S. paradoxus* and *S. mikatae*, with the latter successfully donating its copy of *Ty3* to *S. cerevisiae*. Additionally, *S. kudriavzevii* gained a subfamily from *S. paradoxus*. However, this option has a number of weaknesses, in that *S. mikatae* and *S. cerevisiae* LTRs do not support this theory, and two lineages of *Ty3* exist in *S. arboricola*, which would suggest further divergence in this species not attributable to population isolation. Alternatively, the second lineage was the result of the transfer from a currently unknown source.

History of *Ty4* and connection to the whole genome duplication event

It was proved conclusively here that *Ty4* was gained before the last common *Saccharomyces* ancestor, contrary to previous hypotheses made using the limited data available (Neuvéglise *et al.*, 2002; Liti *et al.*, 2005). Although *Ty4* is uncommon within Saccharomycetaceae, this family was not observed in pre-WGD species such as *Lachancea* and *Kluyveromyces*, occurring only after the speciation of *Zygosaccharomyces* (Clade 7). *Vanderwaltozyma* is the most distant relative of *Saccharomyces* to contain a *Ty4*-like family, and also happens to be the earliest post-WGD genus to diverge, suggesting a clear point of acquisition (Figure 5.23). The WGD was an interspecies hybridisation event, most likely between the ancestral eight-chromosome members of the *Kluyveromyces-Lachancea-Eremothecium* (KLE; Clades 10-12) and *Zygosaccharomyces-Torulaspora* (ZT; Clades 7-9) clades, resulting in the 16-chromosome polyploidy descendants of the post-WGD clades (Marcet-Houben and Gabaldon, 2015). A number of options can explain the origin of *Ty4*. The first hypothesis is that the *Ty4*-like ancestor was once present in one or both of the parental ancestors of the WGD event and was since lost, accounting for its absence in both ZT and KLE descendent clades. A second option is that the *Ty4*-like ancestor could have been gained around the time of the WGD from an unknown (or now extinct) species. Phylogenetic analysis of all *Ty1/copia* RTs showed that additional sequences from elements of *Brettanomyces* and *Ogataea* cluster closer to *Ty4* than *Ty5* (data not shown), suggesting that familial differences may not be as firm away from the species of the *sensu lato* complex. A third possibility is that an existing *Ty1/copia* element was prompted by the WGD and subsequent evolutionary innovation to undergo divergence alongside host speciation, thus resulting in the new *Ty4* family. However, as the WGD did not cause the divergence of other families, this may have been the result of population isolation.

Further divergence may have recently occurred in *Ty4*, corresponding with the isolation of different populations of *Saccharomyces* species. This family has clearly separated into main two lineages, referred to here as European (representing strains and species isolated in Russia and European countries) and American (which encompasses isolates from East Asia alongside those from the Americas). However, the evolutionary history of these lineages is complex, and can only be explained by multiple events. Neither American nor European lineages themselves appear to be ancestral, but existence of the American type is evident in all species, but not all populations of each species.

Figure 5.26 illustrates the possible evolutionary history of this complex family. The divergence into the European type element is likely to have occurred after the speciation of *S. arboricola*, as this accounts for the existence of European *Ty4* in *S. kudriavzevii, S. cerevisiae* and *S. paradoxus*. Being confined to these three species does not reflect that of vertical inheritance, as *S. mikatae* does not contain European-like insertions, and so presents a flaw in this theory. However, *S. mikatae* has not yet been isolated outside of Japan, therefore if European populations of this species do not exist, the point of divergence into the European-type element was not after the speciation of *S. arboricola*.

Although species-specific divergences of elements were observed throughout this work, a second and very noticeable divergence occurred at the point of distinction between *S. eubayanus* and *S. uvarum* into *Tse4* and *Tsu4*, respectively (Figure 5.26). This would suggest that up until the point of divergence, it may have been these *Tse4/Tsu4* types of element that most closely resembled the ancestral copy of *Ty4*. The greater nucleotide diversity observed in the LTRs of these elements would also add weight to this theory (Chapter 4).

Divergence into the American type of *Ty4* may have occurred, most likely in *S. eubayanus* (Figure 5.26), due to an isolation of a population in the Americas. Alternatively, the American *Ty4* originated from an unknown source and was gained via HT by *S. eubayanus*. As a species, *S. eubayanus* is alone in containing evidence of all four subfamilies: European *Ty4* from *S. paradoxus* (one solo LTR); endogenous *Tse4* (now extinct); *Tsu4* gained from *S. uvarum* (extinct or failed transfer); *Ty4* American (autonomous). No single strain of *S. eubayanus* contains evidence of all four subfamilies however, but a selection instead, presumably a result of population and geographical differences.

The American copies appear to either originate in *S. eubayanus* or have been present in this species far longer than the others, considering the high nucleotide diversity (Chapter 4). Given the current data, subsequent and direct HT events are observed from the *S. eubayanus* populations into all other species except *S. cerevisiae*, which appeared to gain the subfamily via American populations of *S. paradoxus*. As LTR branch lengths are short, the transfer into *S. paradoxus* may have been relatively recent, and/or representative of the burst of activity of a newly acquired family soon after an HT event. Additionally, these high levels of activity experienced by *S. cariocanus* in particular coincided with the speciation from parental *S. paradoxus* (discussed below). *S. kudriavzevii* has since lost the American type, as only long-branched solo LTRs are present, positioned

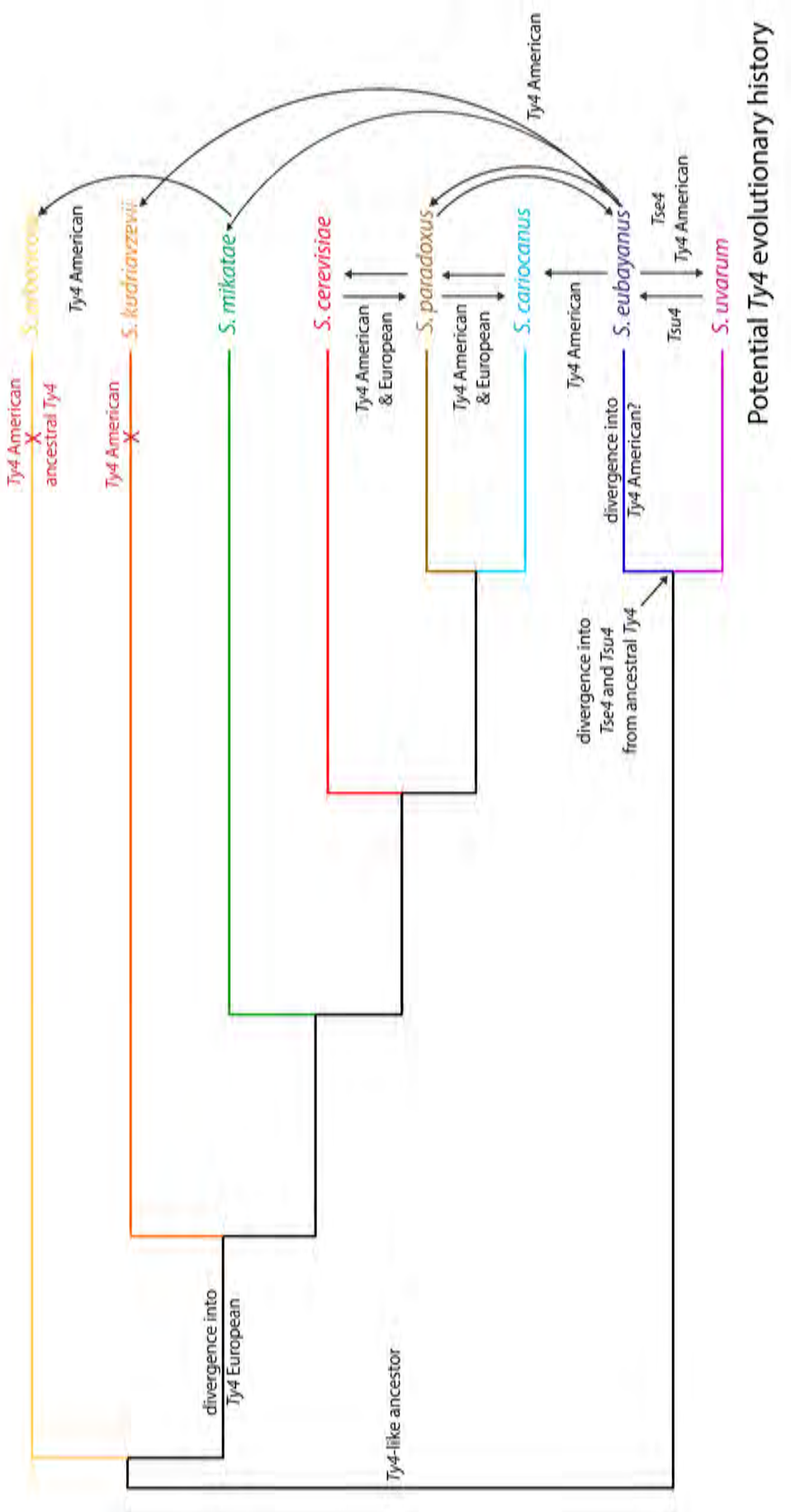

Figure 5.26: **Summary cladogram of the possible evolutionary history of the *Ty4* family.** The simplest potential history of *Ty4* is illustrated. Ancestral *Ty4* diverged at least twice: once after the speciation of *S. arboricola* into the more European-type, and again in the *S. eubayanus* and *S. uvarum* lineage. A further potential divergence occurred in *S. eubayanus*, likely due to population isolation, unless the American *Ty4* was gained from another source. It is possible that they all diverged from one ancestral *Ty4*, but the donation of the subfamilies by unknown sources cannot be discounted. As populations of each species did not consistently contain the same subfamilies, HT events have likely occurred since their isolation.

in the phylogeny as the potential origin of *S. mikatae* sequences - or representing vertical inheritance. However, without the *S. kudriavzevii* RT to corroborate, the true origin of this subfamily of *Ty4* is unknown.

European elements are primarily confined to populations of *S. kudriavzevii, S. paradoxus* and *S. cerevisiae* isolated upon this continent, with possible HT events into *S. eubayanus* and *L. waltii* (discussed in the HT section). The LTR phylogeny suggests the European type was gained by *S. cerevisiae* from *S. paradoxus* (Figure 5.19). This possibility is strengthened by the lack of older, long-branched copies in *S. cerevisiae* as compared with *S. paradoxus*, along with relatively recent activity in *S. cerevisiae*, accounting for the shorter branch lengths observed in this species.

Saccharomyces species predominantly contain the type of *Ty4* appropriate to their population. As the isolated European populations of *S. cerevisiae, S. paradoxus* and *S. kudriavzevii* do not contain evidence of the American type of *Ty4*, the evolution into the two lineages therefore most likely occurred after this major population isolation event and migrations to the Americas. Limited interactions between the two populations occurred, as strains of *S. paradoxus* only contain both main types of *Ty4* if they were isolated in the USA, Canada or Hawaii, adding weight to this theory.

The multiple possible divergences of *Ty4* was surprising, given the short age relative to the other families in *Saccharomyces*. The ancestral WGD and genus-specific population isolation and speciation events cannot account for these divergences, as similar changes in the other *Ty* families, or in other species, have not occurred. Therefore, other events - or perhaps combinations of events - caused the changes in the *Ty4* family to occur.

History of *Ty5*

Ty5 is widespread across genera and may therefore be one of the oldest *Ty1/copia* families, as Neuvéglise *et al.* (2002) suggested the characteristic fused ORFs of this family are a result of age, as also observed in non-LTR retrotransposons. Also testament to its age, *Ty5* is sporadically lost throughout clades 7-12 of *sensu lato* species (Figure 5.23). Based upon their limited data, Neuvéglise *et al.* (2002) further believed that *Ty5* was gained after the divergence of yeast *Yarrowia lipolytica*. However, remnants of a *Ty5*-like family were recovered in this species in addition to autonomous copies in sister species *Y. keelengensis* (data not shown). The existence of *Ty5* significantly pre-dates the divergence of ascomycete yeast from other fungi in a lineage-specific form, long before the emergence of the *sensu lato* ancestor (Figure 5.20). Within ascomycete

yeast, genera- and species-specific divergences of this family are very clear, such as the distinct *Ty5* of extant *Saccharomyces* species (Figure 5.27).

As with the other families in *Saccharomyces*, the evolutionary history of *Ty5* is complex. Liti *et al.* (2005) suggested that *Ty5* was gained shortly before the speciation of *S. mikatae*, which explained the apparent confinement to just three species. However, this conclusion was due to the limitations of the technique used, and that in searching only for coding regions, the authors failed to recover the solo LTR evidence of *Ty5* in *S. kudriavzevii*. Now with the availability of additional data such as the *S. arboricola* genome, it is far more likely that the family existed in the *Saccharomyces* ancestor and underwent LTR-LTR recombination to become extinct in most species (Figure 5.27). Alternatively, *Ty5* was gained from an unknown source before *S. mikatae* diverged, as suggested by Liti *et al.* (2005) or perhaps later by *S. paradoxus*, with subsequent HT events into the other species. Given the relative rarity of HT and ease with which families are lost however, the ancestral theory with multiple losses is far more likely. Regardless of the original source of *Ty5*, more recent HT events from *S. paradoxus* into *S. mikatae* and *S. cerevisiae* were uncovered here (Figure 5.21).

A single solo LTR was present in each of the early branching species. *S. kudriavzevii* and *S. arboricola* experienced independent stochastic loss of *Ty5*, whereas the shared TSDs and flanking DNA of the single insertion within *S. eubayanus* and *S. uvarum* indicates this family was far more likely to have been lost in their last common ancestor, rather than more recent, independent losses.

Extensive horizontal transfer in the TE families of yeast

Prior to the availability of genome sequencing data, most discoveries of HT of TEs were serendipitous, such as those in *Drosophila* (Bartolome *et al.*, 2009), and many plants (El Baidouri *et al.*, 2014) such as *Rider* in tomato (Cheng *et al.*, 2009) and *MULE* in maize (Diao *et al.*, 2006). Detection of potential HTs has increased across kingdoms due to the recent focus on genomic sequencing. Large-scale screenings for HT of TEs have generally been confined to multicellular eukaryotes (Thomas *et al.*, 2010; Wallau *et al.*, 2012, 2016; Ivancevic *et al.*, 2013; Dupeyron *et al.*, 2014; El Baidouri *et al.*, 2014; Peccoud *et al.*, 2017), therefore this study represents one of the first in yeast.

Processes such as the transfer of episomes (O'Brochta *et al.*, 2009), formation of VLPs (Kim *et al.*, 1994), viral infection (Dupuy *et al.*, 2011) and parasite- and symbiont-mediated transmission

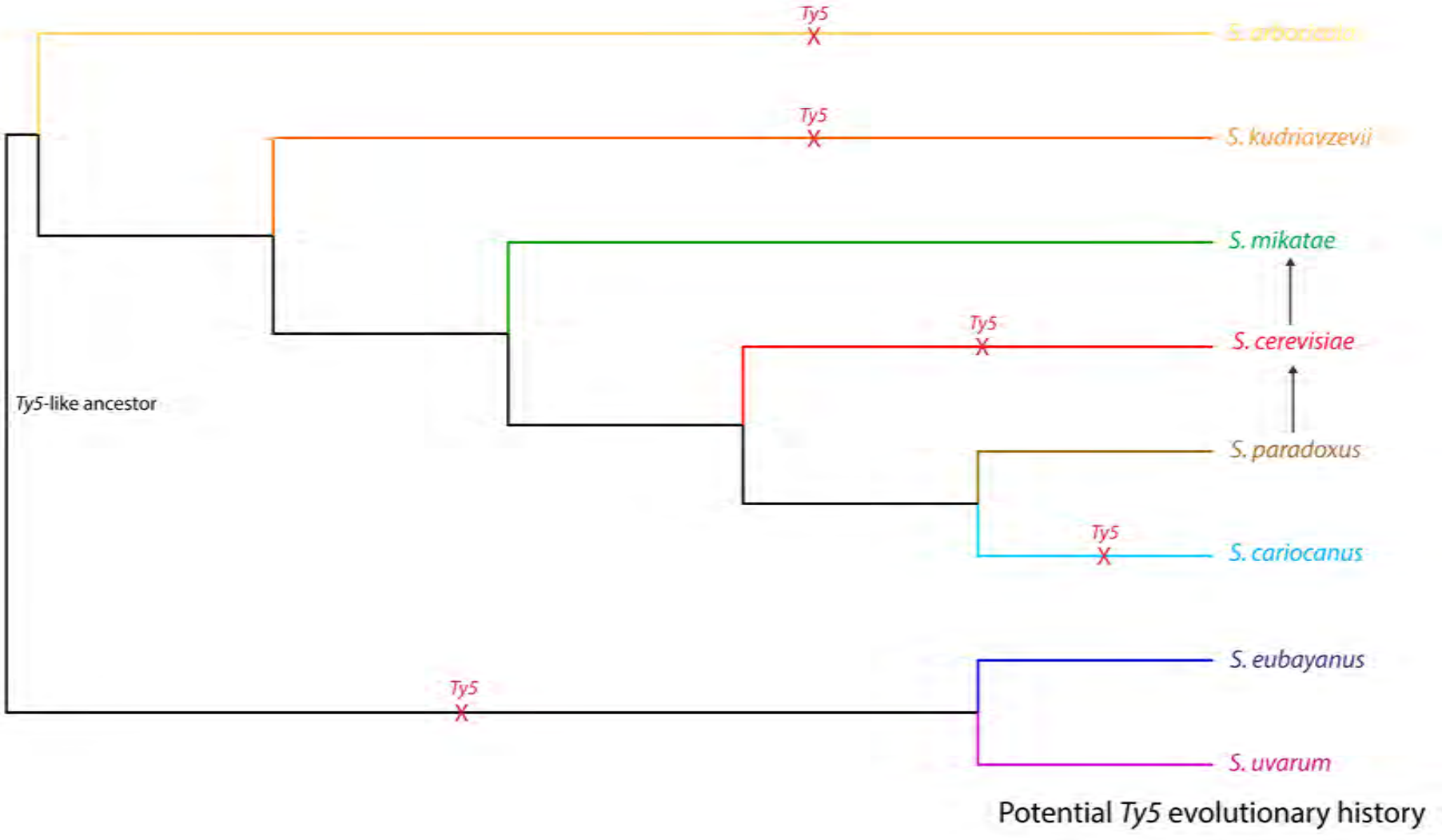

Figure 5.27: **Summary cladogram of the possible evolutionary history of the *Ty5* family.** In *Saccharomyces* species, the current *Ty5* may have been gained in the ancestor of *S. mikatae, S. paradoxus* and *S. cerevisiae*, with transfers into each of the other species (Liti *et al.*, 2005). However, given the dispersal of the family across almost all species, it is far more likely that an ancestor of all species possessed *Ty5*, and then a series of stochastic losses ensued.

(Gilbert *et al.*, 2010; Brown and Lloyd, 2015) have all been suggested for various eukaryotes. Interspecies hybridisation is the most likely mechanism behind HT in yeast (reviewed by Morales and Dujon, 2012). Interspecific hybrids are usually infertile, requiring backcrossing with parental species in order to become viable (Greig *et al.*, 2002). Although the transfer and introgression of genomic DNA between *Saccharomyces* species is regularly reported (e.g. Paques and Haber, 1999; Naumova *et al.*, 2005, 2011; Wei *et al.*, 2007; Muller and McCusker, 2009; Dunn *et al.*, 2013; Almeida *et al.*, 2017), HT of TEs is considered far rarer. Previously documented events were limited to *Ty2* between *S. mikatae* and *S. cerevisiae* (Liti *et al.*, 2005; Carr *et al.*, 2012), the degraded *Ty3p* in *S. cerevisiae* from *S. paradoxus* (Carr *et al.*, 2012), and *Ty1* in *Lachancea* from an unknown *Saccharomyces* species (Neuvéglise *et al.*, 2002). All of these events are supported here, with the identification of the donor to *Lachancea* as an ancestral *Saccharomyces* element. This study discovered an additional 75 HT events involving 19 species in all five families (Tables 5.4-5.9). The analyses here present a conservative estimate of HT events in yeast, as although the phylogenetic methods employed here are robust, they likely underestimate the true frequency of events (Andersson, 2012). A slight tendency towards *Ty1/2* was observed, most likely a reflection on the prevalence of this superfamily in the *Saccharomyces sensu lato* complex. Despite previous reports of low activity (Hug and Feldmann, 1996) and its confinement to only 14 species, *Ty4* surprisingly displays a high number of potential transfers.

Data bias and limitations in HT screenings

Few HT events were observed outside *Saccharomyces* species, which may in part be due to their propensity to hybridise, whereas this is not commonly investigated in other genera (reviewed by Morales and Dujon, 2012). Reports indicate that species of *Zygosaccharomyces* (Watanabe *et al.*, 2017; Wrent *et al.*, 2017), *Candida* (Pujol *et al.*, 2004; Schröder *et al.*, 2016) and *Pichia* (Smukowski Heil *et al.*, 2017) may undergo hybridisation, but no evidence was found to support the HT of TEs in these species. Whereas the ability to hybridise is now known to depend upon the level of genetic compatibility (reviewed by Maheshwari and Barbash, 2011), TEs do not appear to experience such difficulties, demonstrated by the success of *LINE*s from *Candida albicans* introduced into *S. cerevisiae* (Dong *et al.*, 2009; Horn and Han, 2016).

HT may also appear to be widespread in *Saccharomyces* species due to sequencing bias, in that high numbers of these species that have been sequenced (Hittinger, 2013), particularly

in comparison to those of other genera. Excluding highly successful families such as *Ty2* in *S. cerevisiae*, transfers were usually found to be confined to a single or limited number of strains. In these analyses, most species outside *Saccharomyces sensu stricto* were represented by a single type strain, and the chances of a single given strain displaying evidence of HT, already a relatively uncommon event, is rarer still.

Only relatively recent HT events were able to be documented here, as phylogenetic signals become weaker over time, and ancient events may therefore go undiscovered. In ancestral species where similarity was previously very high (i.e. before the current species arose), any HT signals may have been completely eroded. Additionally, the ability to differentiate between HT events and ancestral variation, poorly resolved phylogenies and stochastic loss becomes more unclear with time due to the erosion of signals.

Extinction of families across budding yeast

Although pseudoelements are common (Chapter 4, Appendix P), the majority of families were lost as a result of LTR-LTR recombination, with solo LTR evidence of a once active family. Solo LTRs of a family were located in genomes by using those associated with FLEs as queries. If coding regions cannot be located in a genome due to deletion by LTR-LTR recombination, any remaining solo LTRs go undiscovered should they lack identity with those of a closely related species. The occurrence of stochastic loss, ranging from 0.12-0.48 throughout the families (Tables 5.6-5.9), is undoubtedly an underestimation due to these undiscoverable solo LTRs.

Throughout *Saccharomyces sensu stricto* and *sensu lato* species, *Ty5*- and *Ty3*-like were found to be the families most susceptible to stochastic loss, each with a minimum of four independent extinctions, depending on evolutionary history (Figure 5.23). This is likely a result of the combination of relative age of these two families and their widespread presence in the genomes of species in the *sensu lato* complex. In the *sensu stricto* species, *S. paradoxus* and *S. cerevisiae* experienced the fewest losses, whereas all families except *Ty2* are lost in *S. arboricola*. *Ty3* was present in the last common ancestor of *Saccharomyces*, but lost in the *S. eubayanus/S. uvarum* lineage, with potentially only partial LTRs remaining. Given the available genomic data, *Ty5* may also have suffered a similar fate in these species, as no coding regions remain associated with the single solo LTR. As the insertions share TSDs and flanking DNA, the family was likely lost in their common ancestor.

Phylogenetic analyses support patterns of Tajima's *D* and nucleotide diversity

TEs are considered to have their own life cycle within their hosts (Brookfield, 2005). Their phylogenies, nucleotide diversity and Tajima's *D* values all reflect the evolutionary history in each family. Whereas this has previously been established for the model organism *S. cerevisiae* (Carr *et al.*, 2012), the families of other species of yeast have not been examined in this respect.

Elements with a recent common ancestry will show lower nucleotide diversity with variants present at low frequency, therefore resulting in a negative value of *D*. This may also indicate recent transposition activity (Maside *et al.*, 2003; Sánchez-Gracia *et al.*, 2005; Bartolome *et al.*, 2009; Carr *et al.*, 2012; Carr and Suga, 2014), which would generate short terminal branches in phylogenies as relatively few mutations can accumulate in this relatively short time. Identical paralogous copies are also indicative of recent activity.

Nucleotide diversity and Tajima's *D* values were calculated for LTRs in *Saccharomyces* (Chapter 4) and other species (Table 5.5; Chapter 4; Appendix P). LTRs of *Saccharomyces* (*n*=21) and *sensu lato* (*n*=15) families obtained significantly negative *D* values, consistent with recent transposition. In all cases, these families possess short terminal branch lengths in the LTR phylogenies.

A minority of families received a positive value of *D* (*n*=8; Chapter 4; Appendix P) but only that of *Tnkg5* in *Nk. glabrata* is significant. Population substructure, resulting in variants at intermediate frequency may result in positive values of *D*. This is observed noticeably in *Tnkg5* of *Nk. glabrata* as the development of possible active sublineages (Figure 5.22).

The remaining families returned negative values of *D*, but not significantly so (Chapter 4; Appendix P). Of these, *S. eubayanus* families did not return significant values of *D*, despite the presence of multiple identical copies, and FLEs in *Tse1* and *Ty4*. The signal of recent ancestry was likely disrupted by the presence of relatively long-branched insertions. In addition, a lower level of activity would account for the less significant, yet still negative, *D* values seen in *S. eubayanus* and the families of other species.

Limitations of phylogenetics in the analysis of TEs

In comparison to the phylogenetic analysis of host genes, that of TEs is notoriously difficult due in part to the necessity of repeated TE evolution to avoid host defences. Phylogenetic analysis of TEs is comparable to that of pseudogenes, as individual TE sequences are highly subject to drift and negative selection (Biémont *et al.*, 1997; Charlesworth *et al.*, 1997). Support can vary for

phylogenies of coding regions (Malik and Eickbush, 1999, 2001), whereas those of LTRs typically lack support (Sánchez-Gracia *et al.*, 2005; Benachenhou *et al.*, 2009, 2013; Carr *et al.*, 2012; Carr and Suga, 2014). Recombination between subfamilies can also result in conflicting phylogenetic signals (Jordan and McDonald, 1998). Observed here in particular was the propensity for internal branches to lack support, with support steadily increasing towards the terminal branches of both LTR and RT trees. Both phylogenetic methods have their drawbacks: poorly supported short internal branches are a recorded issue with ML (Anderson and Swofford, 2004; Kück and Wägele, 2015), whereas BI has a tendency to overinflate support values (Rannala and Yang, 1996). However, BI and ML are far more suitable than parsimony-based methods for datasets containing long-branched sequences and those that undergo recombination (Hillis, 1996, 1998; Hedtke *et al.*, 2006).

Long-branch attraction, the clustering of older sequences to the point of distorting topology, is well documented in phylogenetics, yet the reasons behind the phenomenon are not fully understood (reviewed by Bergsten, 2005). While care was taken here to produce strongly supported estimations of evolutionary relationships, phylogenies were nonetheless susceptible to issues such as long-branch attraction, as independent mutations cause sequence divergence and excessive phylogenetic noise. However, there is evidence to suggest that long-branch attraction may be less of an issue with larger phylogenies such as the LTR datasets used here (Hillis, 1996). Additionally, the practice of including only active lineages or removing branches over an arbitrary length (Carr *et al.*, 2012) did not improve tree topology, or allow resolution (data not shown).

A number of discrepancies between the placement of sequences in corresponding RT and LTR trees was observed throughout the families. In the *Ty3/gypsy* RT phylogeny, there was a suggestion of HT between *S. mikatae* and *S. cerevisiae*, whereas this was not supported by the LTR phylogeny, instead inferring a sister relationship between the two. In *Kazachstania, Tetrapisispora* and *Naumovozyma*, RT sequences often shared close similarity with those of *Saccharomyces*, yet their LTR sequences possessed very little shared identity. Similarly, *V. polyspora* LTRs were excluded from all phylogenies due to poor shared identity with other species, despite the often high similarity in RT regions, particularly with *Saccharomyces* species. Instances such as these aptly demonstrate the tendency for LTRs to diverge, as opposed to the functionally constrained internal coding regions (Benachenhou *et al.*, 2009).

5.7 Summary and conclusions

In this chapter, the most likely evolutionary relationships of four superfamilies of retrotransposons in yeast are presented using two highly robust methods of phylogenetic inference. Analyses used the highly conserved RT domain and LTR sequences to screen for possible HT events between species. HT allows elements to escape the almost inevitable eradication within their host genome in order to propagate in those of other closely related species. In excess of 75 potential HT events between 19 species are documented, with around half involving the successful transposition of families in new genomes, indicating that HT is far more common in yeast than previously believed.

Chapter 6

Isolated *Saccharomyces cerevisiae* populations share TE insertion profiles

Geographical isolation of *Saccharomyces* species and populations has previously been documented by various authors (Johnson *et al.*, 2003; Fay and Benavides, 2005; Aa *et al.*, 2006; Sampaio and Gonçalves, 2008; Liti *et al.*, 2009; Schacherer *et al.*, 2009; Goddard *et al.*, 2010; Legras *et al.*, 2014; Peris *et al.*, 2014, 2016; Almeida *et al.*, 2015; Clowers *et al.*, 2015). There are currently >1000 *S. cerevisiae* strains of differing origins whose genomes have now been sequenced (Liti *et al.*, 2009; Borneman and Pretorius, 2015; Strope *et al.*, 2015; Peter *et al.*, 2018). Their comparison allows the evolutionary history of yeast populations to be determined, and the identification of TEs that persist in the genomes under specific environmental conditions. TEs are now known to play a complex role in the evolution of their hosts, rather than existing simply as genomic parasites (Bonchev and Parisod, 2013). Liti *et al.* (2005) were the first to conduct a major investigation into elemental differences between populations and *Saccharomyces* species, identifying conserved coding regions with hybridisation probes. Bleykasten-Grosshans *et al.* (2011) discovered variation in the *Ty1* sequences of populations of *S. cerevisiae* and further analyses by Carr *et al.* (2012) were predominantly confined to the *S. cerevisiae* reference strain. Most recently, Bleykasten-Grosshans *et al.* (2013) surveyed the TE content of 41 strains of *S. cerevisiae* to establish distribution patterns of insertions but did not determine element evolutionary relationships.

In this chapter, two distinct populations of *S. cerevisiae* are explored for their TE content. As publicly available data, the two isolated populations were chosen to highlight common patterns of TE distributions, while tracking the spread of a newly acquired *Ty1*-like family. The Peterhof collection, originating in Russia and distinct from commonly used S288c-related strains, are of an industrial background, used in distillation (Drozdova *et al.*, 2016). The Brazilian population underwent

an environmental relocation from Europe before frequent hybridisation with South American populations of *S. paradoxus* (Barbosa *et al.*, 2016). Despite possessing very different levels of human association and geographical backgrounds, their *Ty* families show remarkably similar evolutionary histories. Phylogenetic analyses show that LTRs are distinct from those of the *Saccharomyces* Genome Resequencing Project (SGRP) strains, reflecting the isolation of their populations.

6.1 The *Saccharomyces cerevisiae* Peterhof Genetic Collection

The Peterhof Genetic Collection (PGC) of laboratory yeast strains originated from a single Russian industrial distillery population, thought to be distant from the Carbondale stocks that produced the *S. cerevisiae* reference strain S288c and commonly used laboratory derivatives (Lindegren, 1949; Mortimer and Johnston, 1986). The PGC strains were used extensively in early work with translation termination (Zhouravleva *et al.*, 1995) and more recently with prions (Du and Li, 2014).

Drozdova *et al.* (2016) sequenced five examples from the collection and found that their relatedness to SGRP strains is comparable to that of geographically isolated populations of *S. cerevisiae*. Phylogenetic analysis of >800 genes showed that PGC yeast share more similarity with bakery strains than S288c-derived laboratory or European strains. Two strains from the progenitor lineage (P), 15V-P4 and P3982, are considered 'pure' and have had no association with other strains outside the Peterhof distillery population. The remaining three strains are maintained as diploids obtained from the mating of progenitor and hybrid (mosaic) isolates (D; Figure 6.1).

The authors discovered that the strains contain unique combinations of up to 11 ORFs which are absent from S288c-derived strains, but are not confined to the Peterhof genomes. These included gene clusters specific to wine strains (Borneman *et al.*, 2011), the *RTM1* cluster and *KHR1*, a gene encoding a heat-resistant killer toxin (Goto *et al.*, 1990) which is also present in strain YJM789 (Wei *et al.*, 2007).

The authors assembled each of the genomes into scaffolds but exact chromosomal locations could not be elucidated due to genomic rearrangements in the strains. Here, TE landscapes are explored in all five of the PGC strains in preparation for phylogenetic analyses.

6.1.1 TE variation in the Peterhof strains

A summary of the PGC strains is displayed in Table 6.1. Genomic GC content is similar across all isolates, whereas they show slight variation in TE content. The mosaic strains are closely

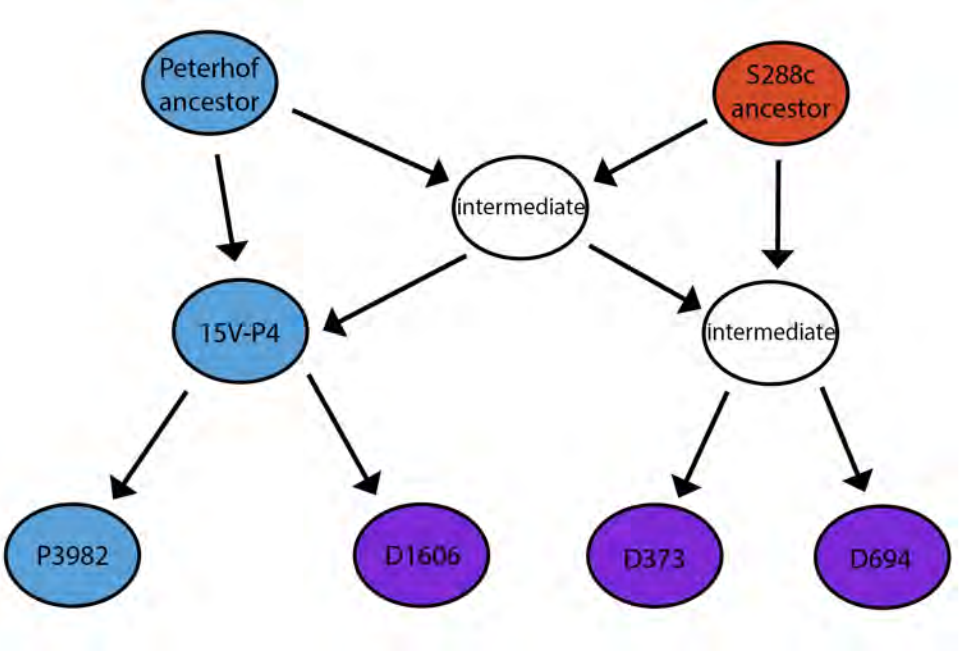

Figure 6.1: **Simplified pedigree of the Peterhof strains.** Blue - 'pure' strains; purple – mixed heritage with S288c-derived strains; red – original S288c progenitor strain. Adapted from Drozdova *et al.* (2016).

related, yet possess a very different distribution of insertions. Copy numbers are also lower than in comparison to many SGRP strains with a similar industrial background. It was hypothesised that the progenitor strains would contain evidence of *Ty* activity unique to this population, whereas the mosaic strains would contain a greater number of insertions from both populations as a reflection of their mixed heritage. Interestingly strain D694 contains the lowest number of insertions in all families.

Strain	GC content (%)	TE genome content %	Family				Origin
			Ty1/2	*Ty3*	*Ty4*	*Ty5*	
15V_P4	38.22	1.13	2(44)	1(13)	1(5)	1(2)	Pure Peterhof; distillation
P3982	38.18	1.20	3(65)	1(9)	1(7)	1(2)	Pure Peterhof; laboratory
D373	38.71	1.03	3(31)	1(11)	2*(9)	1(3)	Peterhof/S288c-derived hybrid
D1606	38.17	1.36	3(76)	1(14)	2(4)	1(2)	Peterhof/S288c-derived hybrid
D694	38.14	0.78	2(25)	1(0)	1(4)	1(1)	Peterhof/S288c-derived hybrid

Table 6.1: **Summary of the Peterhof strains.** Element numbers: FLE(solo LTRs); *tandem.

Nucleotide diversity was calculated for LTRs unique to the Peterhof collection (Table 6.2) and not present in SGRP strains, which show that elements have likely been active since the isolation of the PGC population. The majority of insertions are fixed across the strains as solo LTRs of all families. *Ty1* insertions prove to be the most diverse and also the most abundant, whereas *Ty2* is both the most conserved and more than ten times less abundant than *Ty1* in the genomes of the

PGC strains, likely the result of the relatively short age of *Ty2* in this species (Carr *et al.*, 2012).

	Family				
	Ty1	*Ty2*	*Ty3*	*Ty4*	*Ty5*
Nucleotide diversity (π)	0.24418	0.05677	0.09659	0.14328	0.12790
No. of unique LTRs	109	10	32	18	12

Table 6.2: **Nucleotide diversity calculated on the number of insertions unique to the PGC strains.**

6.1.2 *Ty1/2*: diversity and extinction of subfamilies

An alignment of the *Ty1/2* LTR sequences was inspected for the presence of *Ty1*', *Ty1/2* recombinant hybrids and the *Ty1v* subfamily (documented in Chapter 3), as almost half of the sequences show divergence from the canonical *Ty1* and *Ty2* sequences. No common features were identified, and so it was concluded that the state of these insertions is the result of an accumulation of nucleotide changes due to age. Figure 6.2 displays the main differences between *Ty1* and *Ty2* LTR sequences, including two deletions not commonly seen in SGRP strains of *S. cerevisiae*.

All coding regions from elements in the *Ty1/2* superfamily were extracted, translated and aligned. Due to the presence of multiple copies in three strains, *Ty1* elements were accurately constructed only in 15V-P4 and D694. The copy count in D373 is based on LTRs with short regions of associated coding DNA and TSDs, as the *gag* and *pol* ORFs are too fragmented over contigs to construct (<500bp). If *Ty1/2* hybridisation has occurred, the short contigs (<1kb) prevented breakpoints from being located. To determine the subfamilies of the *Ty1/2* elements, Gag regions were aligned (Figure 6.3). *Ty1*' is present in four strains (D1606, 15V-P4, P3982 and D373, the latter not shown due to short contigs), whereas 15V-P4 is the only strain lacking *Ty1*. A *Ty1* relic is also present, fixed in all five strains, which possesses the 5' LTR but is degraded from the *gag* region onwards.

Each strain possesses a single *Ty2* copy that spans an entire contig in each assembly. All elements contain at least one stop codon, likely rendering this family non-functional in the Peterhof collection. Unique *Ty2* LTR sequences were recovered from the strains (*n*=10), half of which are solo insertions, indicating strain-specific activity. Interestingly, occurrences of recombination differ between populations, as the insertions that are associated with *Ty2* coding regions are all present as solo LTRs in the SGRP strains of *S. cerevisiae*, Conversely, one solo LTR in common across all Peterhof strains is present as a full-length element (FLE) in the SGRP *S. cerevisiae* strains.

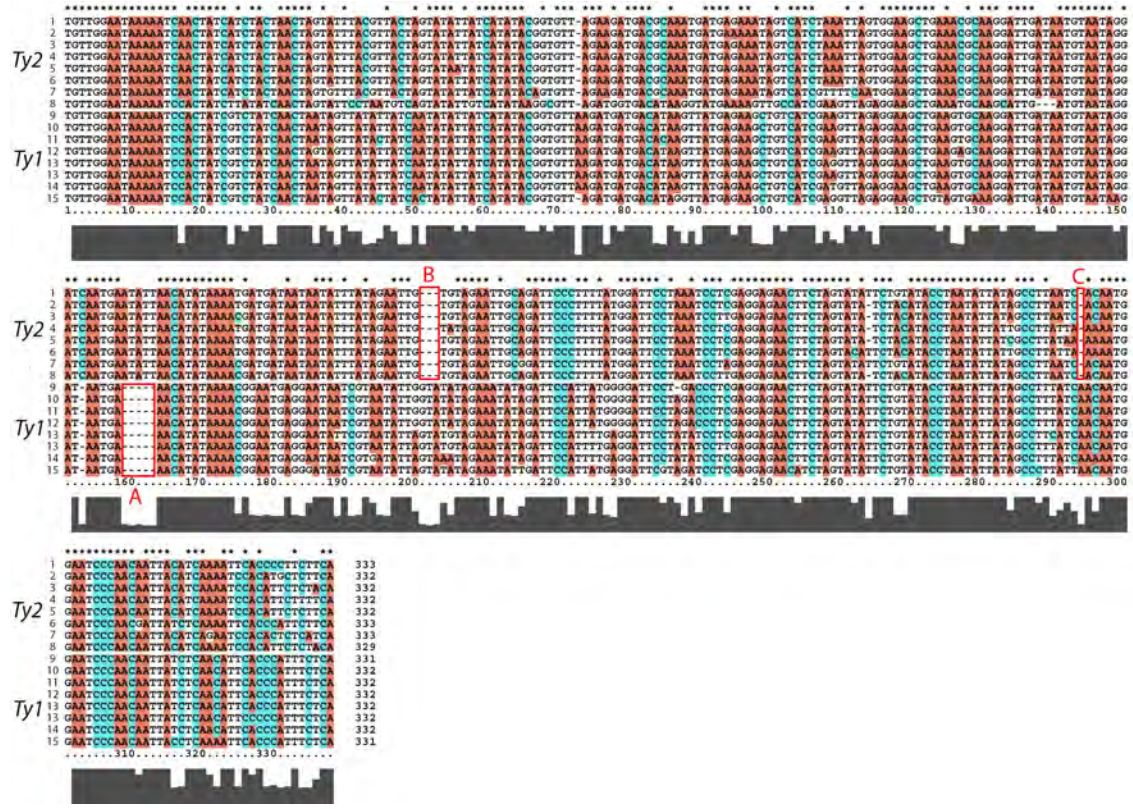

Figure 6.2: **Alignment of *Ty1-2* LTR sequences in the Peterhof strains.** Although sequence 8 shows divergence from the canonical *Ty2* sequence, no evidence of hybridisation with *Ty1* was noted. A: 5bp deletion in *Ty1* sequences; B: 3bp deletion in *Ty2* sequences; C: characteristic 1bp deletion in *Ty2* sequences used to distinguish between subfamilies. *denotes position is fully conserved. Graphic below the alignment represents the number of sequences in which the position is conserved. Numerals below the alignment denote position in the alignment.

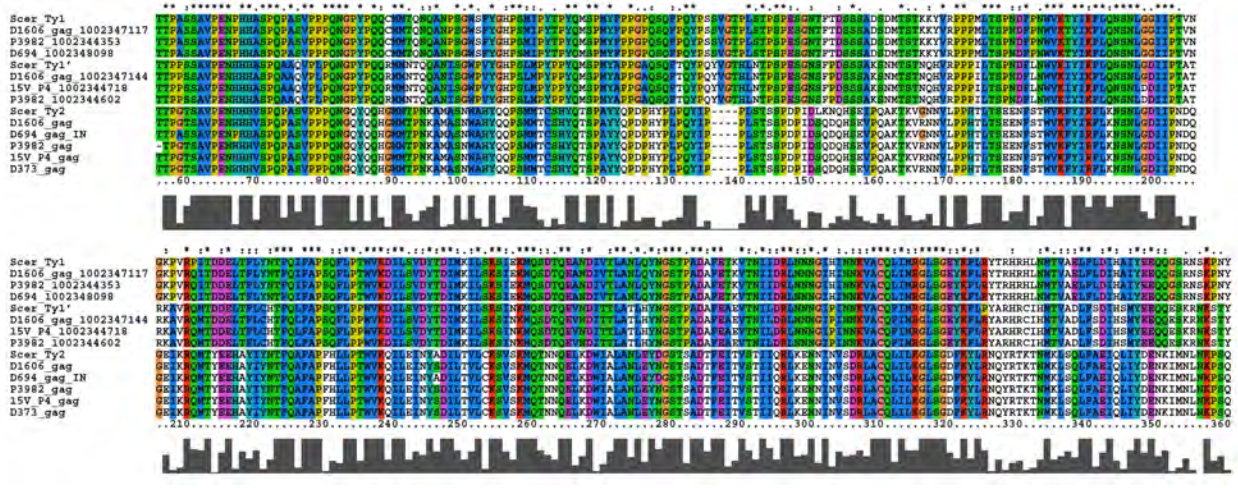

Figure 6.3: **Alignment of *Ty1-2* Gag sequences in S288c and the Peterhof strains.** Layout and conservation indicators are as in Figure 6.2, with the addition of : denoting strong conservation and . denoting weak conservation.

6.1.3 Strain-specific activity of *Ty3* in the Peterhof strains

LTR copy numbers in four of the strains are relatively consistent (Table 6.1), but D694 is unusual in that it possesses only the single autonomous FLE consistent across all strains, but no intact *Ty3* solo LTRs. It did however contain remnants of solo LTRs (*n*=19), which shared more identity (∼70%) with those from *S. kudriavzevii*. The partial solo LTRs are either interrupted by a more recent *Ty1* insertion (*n*=4) or degraded at the 5' boundary (*n*=15). Additionally, a single partial *S. paradoxus*-like insertion (95% identity with *Ty3p* canonical LTR) shared with an insertion in S288c on chromosome IV was recovered from D373. No further evidence of *Ty3p*-like sequences was found.

6.1.4 Variant *Ty4* frequency and duplication resulting in a tandem formation

As with *Ty3*, copy numbers in the *Ty4* family are relatively consistent in the five PGC strains (Table 6.1). Unfortunately, in all assemblies, LTRs are cut off by the ends of the contigs, and so the possible identities of these insertions were manually determined from sequencing reads. Three strains likely contain a single copy of *Ty4*, while two copies assumed a tandem formation in D373 (Figure 6.4), unlikely to be the result of an assembly error as the RT domains differed (∼300bp at 90% identity; 62% similarity in 64 residues of RT). Due to the short sequencing reads (∼400bp) it remains unclear as to whether the tandem formation was the result of duplication or transposition. Furthermore, the LTR associated with the FLE in strain D694 could not be located in the reads, perhaps due to low coverage.

The D1606 assembly contains a single FLE, yet the flanking regions of LTRs suggested two copies. Rebuilding the genome from sequencing reads confirmed the presence of two copies, but unfortunately placed only one within a scaffold (5'-GAGAG-GAGAG-3'), which is also shared by progenitor strain P3982. Both progenitor strains each possess an autonomous copy with differing TSDs, but the family is lost in the mosaic strains due to the presence of multiple stop codons (Figure 6.4). It is unclear as to whether both the RT domains of D1606's elements contain stop codons, Furthermore, only European insertions are present in the PGC strains, as they are presumably isolated from populations containing the American type of *Ty4*.

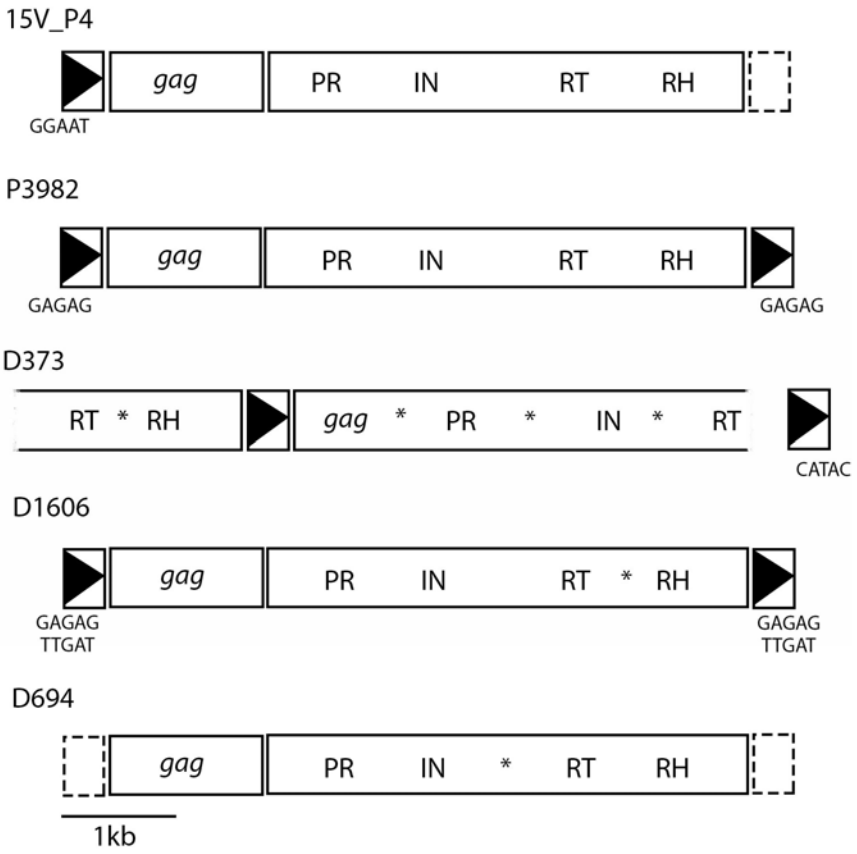

Figure 6.4: **Diagrammatic representation of the *Ty4* elements in the PGC strains.** *denotes the presence of a stop codon. As the stop codons are in differing positions depending on the strain, they represent the independent loss of this family in each strain. LTRs cut off by the ends of contig reads are represented by dotted outlines. In D373, the ends of contigs prevented the full-length elements being recovered. Elements are drawn to scale.

6.1.5 The *Ty5* relic is fixed in PGC strains

The *S. cerevisiae Ty5* relic (Chapter 4) is present in all strains, with identical TSDs and flanking DNA suggesting this insertion predates the isolation of the Peterhof population. When aligned against the full-length copy found in *S. paradoxus* strains, the 2.7kb relics are missing the IN and full RH domains (data not shown). The copy in D373 is the most heavily degraded, but large deletions are present in the copies of all strains except D1606. Interestingly, the copies in the progenitor strains 15V-P4 and P3982 are more closely related to that of S288c, whereas the mosaic strains are further diverged (62 nucleotide changes resulting in 24 non-synonymous changes, plus stop codons spread throughout the domains). Mutations have therefore accumulated in the generations between the progenitor and mosaic strains.

Slight variation in copy numbers in addition to the relic element meant that, as seen in the majority of *S. cerevisiae* strains, *Ty5* is extinct in the PGC strains. As in the other families, unique

solo LTRs are the result of strain-specific activity or lineage sorting through drift. Due to its fixed position, the relic must pre-date the independent activity, therefore *Ty5* became inactive after the isolation of the Peterhof strains.

6.2 Peterhof *S. cerevisiae* phylogenetics

The relationships between the LTRs of Peterhof and SGRP strains of *S. cerevisiae* are explored in a series of phylogenetic trees below. Analyses were conducted only on LTRs as RT sequences are highly conserved or identical to those in S288c-derived *S. cerevisiae* (data not shown). As in Chapter 5, only full-length LTR sequences were used in analysis.

6.2.1 Large numbers of long-branched sequences in the *Ty1/2* superfamily

The *Ty1-2* superfamily phylogeny, displayed in Figure 6.5, shows the main groupings of *Ty1* and *Ty2* in the SGRP strains, and that the majority of the Peterhof sequences fall outside these.

The Peterhof sequences present within the *Ty2* grouping are all short-branched (as defined by Carr *et al.*, 2012; *n*=10), consistent with the presence of young copies and therefore potentially recent activity. The position of all *Ty2* sequences may be the result of an artefact, causing the root to be placed within *Ty1* sequences, as *Ty2* is known to have been gained from *S. mikatae* via horizontal transfer (Carr *et al.*, 2012) rather than descent from *S. cerevisiae Ty1*. PGC *Ty1* insertions (*n*=109) are far more numerous than *Ty2* in these strains, 95% of which are solo LTRs (*n*=103). 24% of all Peterhof *Ty1* sequences fall within the SGRP *Ty1* grouping (*n*=26) and of these, eight are short-branched (31%). The remaining 76% of *Ty1* LTRs (*n*=83) fall outside the main SGRP *Ty1* group on long branches, indicative of their degradation. These long-branched LTR sequences share <97% identity with LTRs in the SGRP strains. Despite this high shared identity, the Peterhof insertions are distinct, with a small number of SGRP sequences also falling within this long-branched group (*n*=3). Upon examination of these SGRP sequences, they could not be identified as belonging to another subfamily such as *Ty1*', but appeared simply to be degenerate insertions. The long-branched insertions are not only predominantly confined to the Peterhof strains, but form their own grouping which caused issues with outgroup selection. LTR sequences from *S. eubayanus* were selected as the outgroup as sequences from other *Saccharomyces* species fall within those of the SGRP sequences. All long-branched Peterhof sequences are clearly inactive,

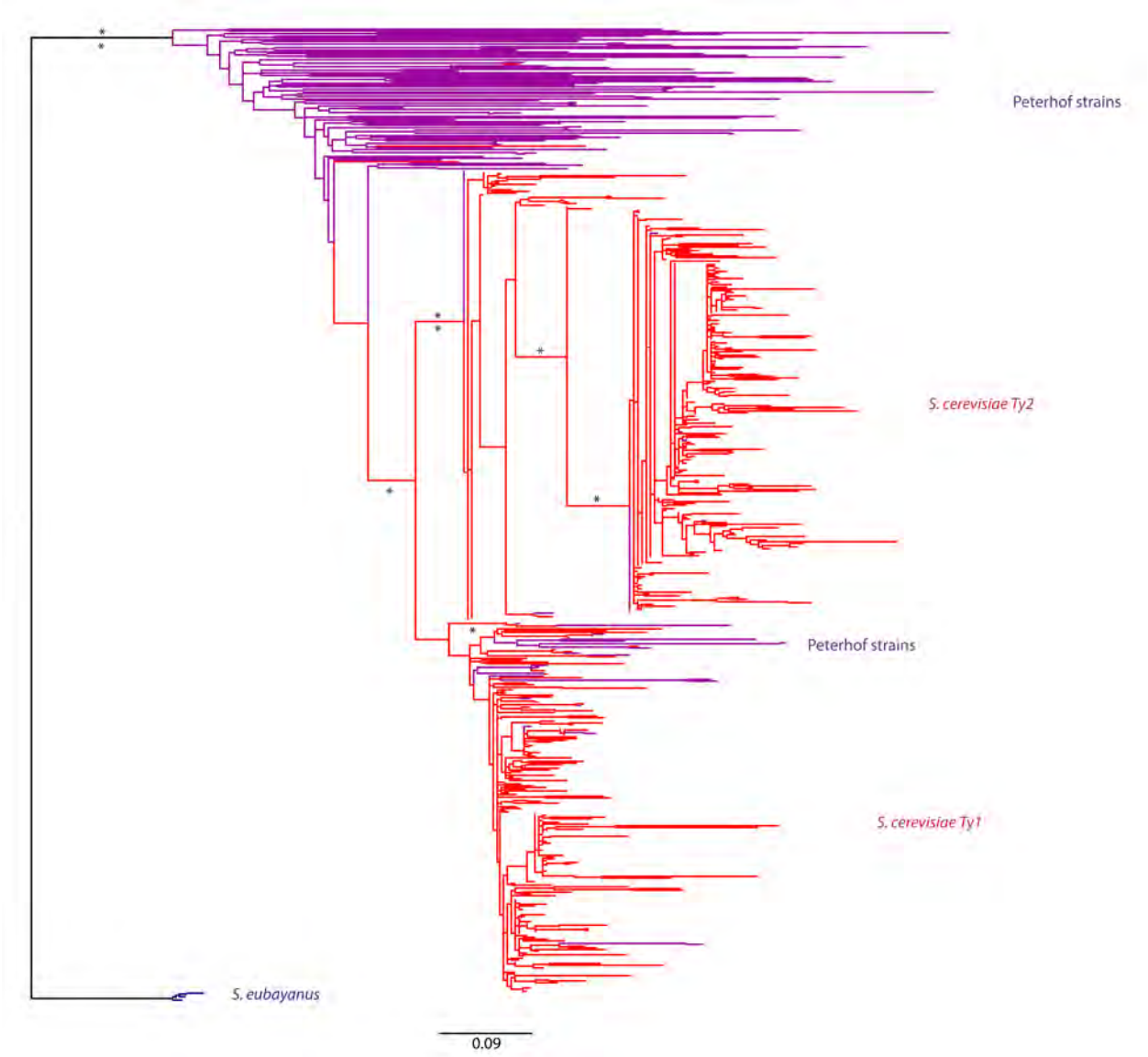

Figure 6.5: ***Ty1/2* LTR sequences in the SGRP and Peterhof strains of *S. cerevisiae*.** The phylogeny was produced from an alignment of 338 nucleotide positions and rooted with sequences from *S. eubayanus*. Topology is that determined by Maximum Likelihood (ML). ML support values are indicated by * above the branch (≥70%mlBP) and Bayesian Inference (BI) below the branch (≥0.95biPP). Scale bar represents substitutions per site.

with the majority existing as solo LTRs, with only one present as a 5' LTR belonging to the degenerate relic element shared by all strains (section 6.1.2). The long-branched insertions are the only evidence of ancient activity unique to the Peterhof strains.

It was suspected that the Peterhof sequences that are positioned within the SGRP group would be confined to the three mosaic strains that were crossed with the S288c-derived strains, thus providing a recent source for the short-branched shared insertions. However, this is not the case as the insertions are also present in the two progenitor strains, suggesting these insertions predate

the isolation of the PGC.

An additional phylogeny was created incorporating sequences of other *Saccharomyces* species to ensure that the origin of the Peterhof sequences is not a result of another source. However, this failed to change the relationships between Peterhof and SGRP sequences, and caused further rooting issues, including the placement of all *S. cerevisiae* sequences within those of *S. mikatae* with no support (data not shown).

6.2.2 Peterhof *Ty3* sequences cluster with SGRP insertions

Figure 6.6 displays the phylogeny of *Ty3* LTR sequences in the Peterhof and SGRP strains of *S. cerevisiae*.

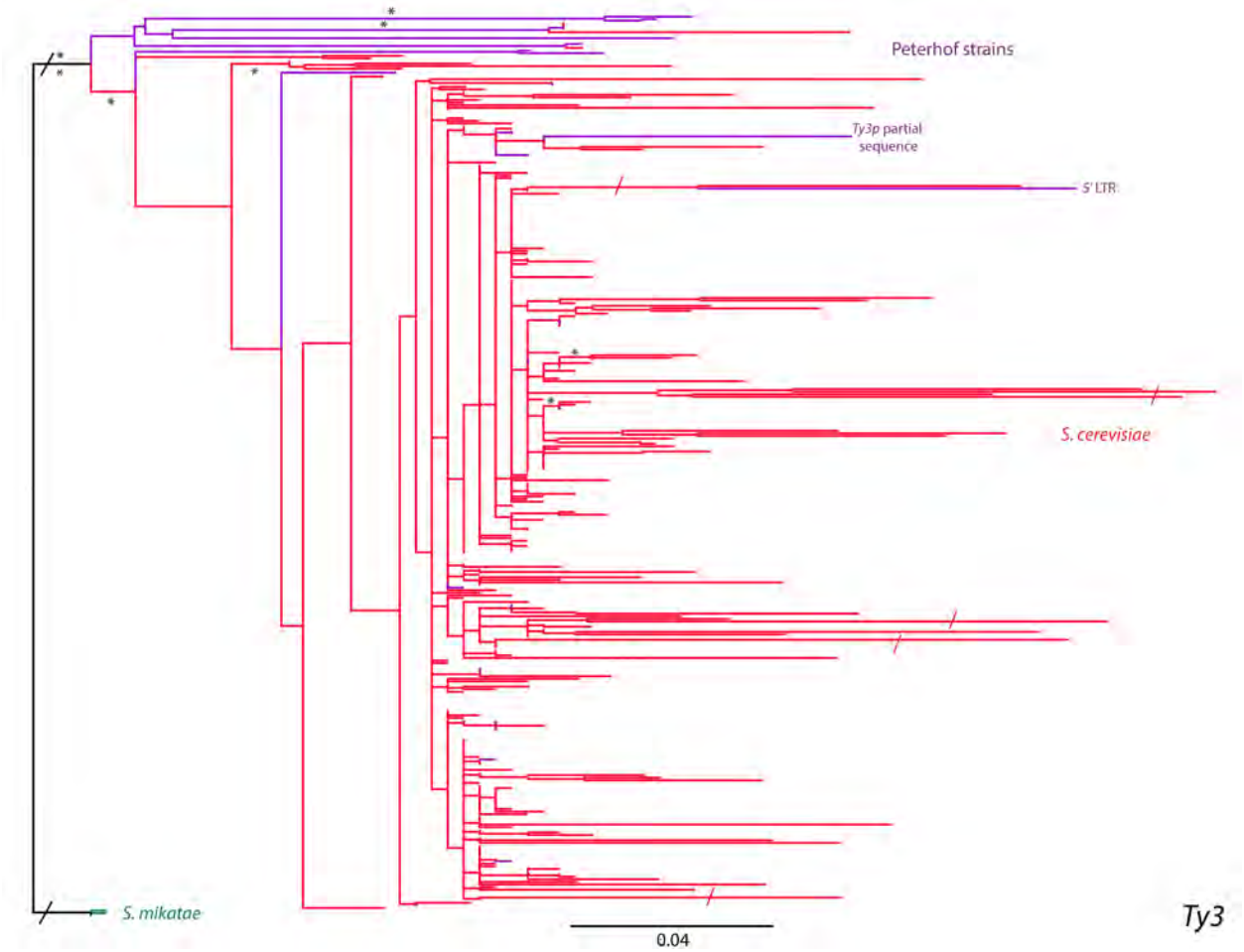

Figure 6.6: ***Ty3* LTR sequences in the SGRP and Peterhof strains of *S. cerevisiae*.** Formatting is the same as in Figure 6.5, but based on an alignment of 350 nucleotide positions and rooted with sequences from *S. mikatae*. / indicates arbitrarily shortened long-branched sequences and root. The position of potential *Ty3p* sequence is weakly supported by ML (29%mlBP) and with other long-branched SGRP sequences according to BI (0.95biPP).

The Peterhof sequences form a small grouping of long-branched insertions (*n*=9) in the basal position, which is supported by ML only (100%mlBP; 0.44biPP). A third of these insertions are shared with the SGRP strains and are therefore likely to be ancestral, persisting only in the PGC strains after the isolation of this population.

This Peterhof grouping shares a small number of insertions with the SGRP sequences which share a similar rate of degradation (*n*=4). The remaining Peterhof sequences (*n*=31) fall within the main group of *S. cerevisiae* sequences, 25% of which are short-branched and therefore consistent with recent activity in the PGC strains. A minority of Peterhof insertions are shared with those of the SGRP strains (*n*=13), but the majority differ in TSDs and status of the elements, e.g. FLE in the SGRP strains but solo in Peterhof or vice versa. The *Ty3p*-like sequence falls within the main SGRP sequences in an unsupported position by ML (29%mlBP) and at an earlier branching position according to BI (0.95biPP). The *Ty3p*-like sequence also failed to cluster with *S. paradoxus* sequences when added to the phylogeny, so its identity remains unknown.

6.2.3 PGC *Ty4* LTRs form the basal position

Figure 6.7 displays the phylogeny of *Ty4* LTR sequences in the Peterhof and SGRP strains of *S. cerevisiae*.

As duplicates were removed from alignments, the multiple identical LTRs associated with FLEs are represented by a single sequence within the boxed region of Figure 6.7. As with the *Ty1/2* and *Ty3* phylogenies, *Ty4* shows a similar topology in that almost 60% of the Peterhof sequences form the basal position and an early branching separate group on long branches (*n*=11) with only one sequence in common with the SGRP *S. cerevisiae* strains. All LTR sequences from strain D694 (*n*=4) fall within the long-branched group, indicating their relatively ancient loss and subsequent sequence mutations. The remaining long-branched sequences are spread across the other strains. In the SGRP grouping, Peterhof LTR sequences are short-branched (*n*=7) and cluster with the other *S. cerevisiae* sequences, four of which are insertions unique to the Peterhof strains, indicative of activity post-Peterhof isolation. Despite the pseudoelement state of the tandem elements in strain D373, the shared middle LTR has not undergone mutation as it clusters with 5' LTRs of both Peterhof and SGRP sequences.

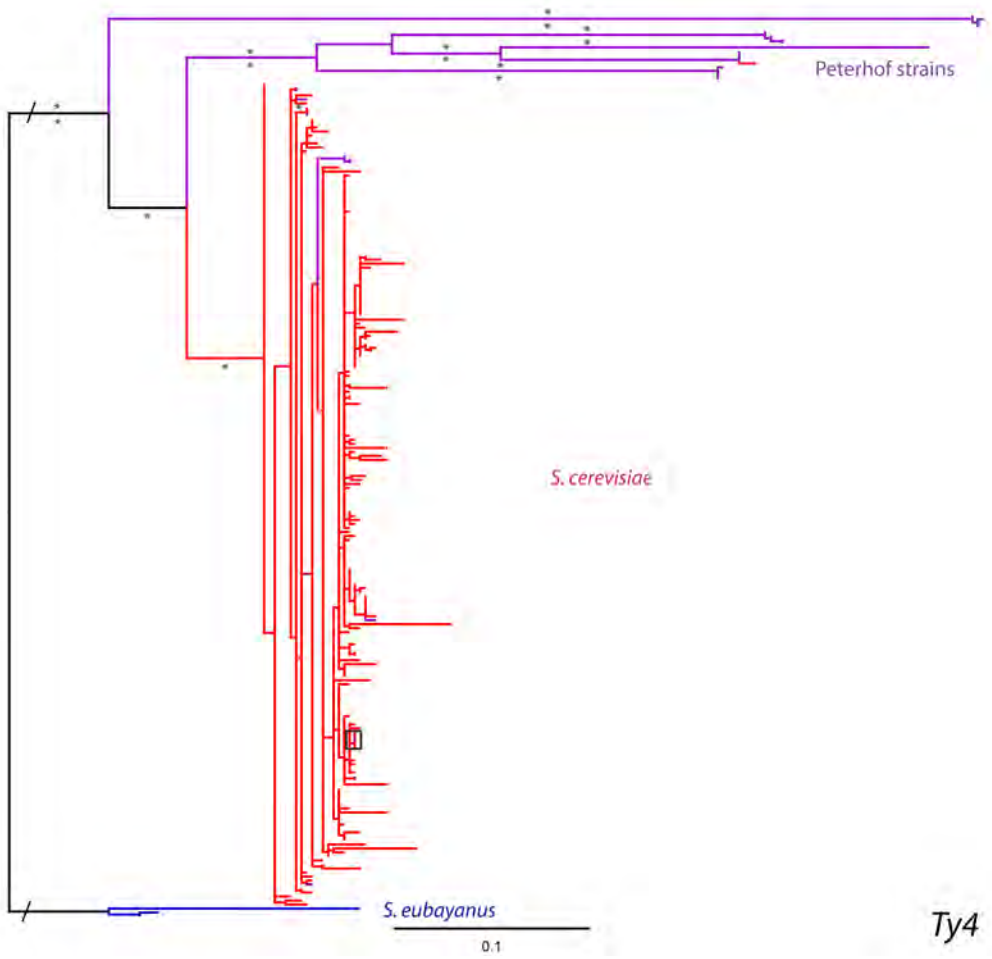

Figure 6.7: ***Ty4* LTR sequences in the SGRP and Peterhof strains of *S. cerevisiae*.** Formatting is the same as in Figure 6.5, but based on an alignment of 383 nucleotide positions and rooted with *S. eubayanus* sequences. / indicates arbitrarily shortened root branch. The boxed region contains Peterhof 5' and tandem LTRs.

6.2.4 The loss of *Ty5* in the Peterhof strains may have been relatively recent

Figure 6.8 displays the phylogeny of *Ty5* LTR sequences in the Peterhof and SGRP strains of *S. cerevisiae*. Although a single Peterhof sequence forms the basal position, the remaining *Ty5* sequences (*n*=17) fall with the SGRP *S. cerevisiae* LTRs. Around half of the sequences (*n*=8) are associated with relic elements and positioned on relatively short branches. This suggests that although the coding regions are degraded (section 6.1.5), the LTRs are less affected by drift as conservation is higher across strains. The remaining sequences are those of solo LTRs and also positioned on short branches, suggesting the possibility of recent activity shortly followed by recombination in this family. Insertions unique to the PGC strains represent over half of the sequences (*n*=10), whereas the remainder are shared with the SGRP strains (*n*=7). The topology is relatively well supported on terminal branches by both methods, but not within internal branches.

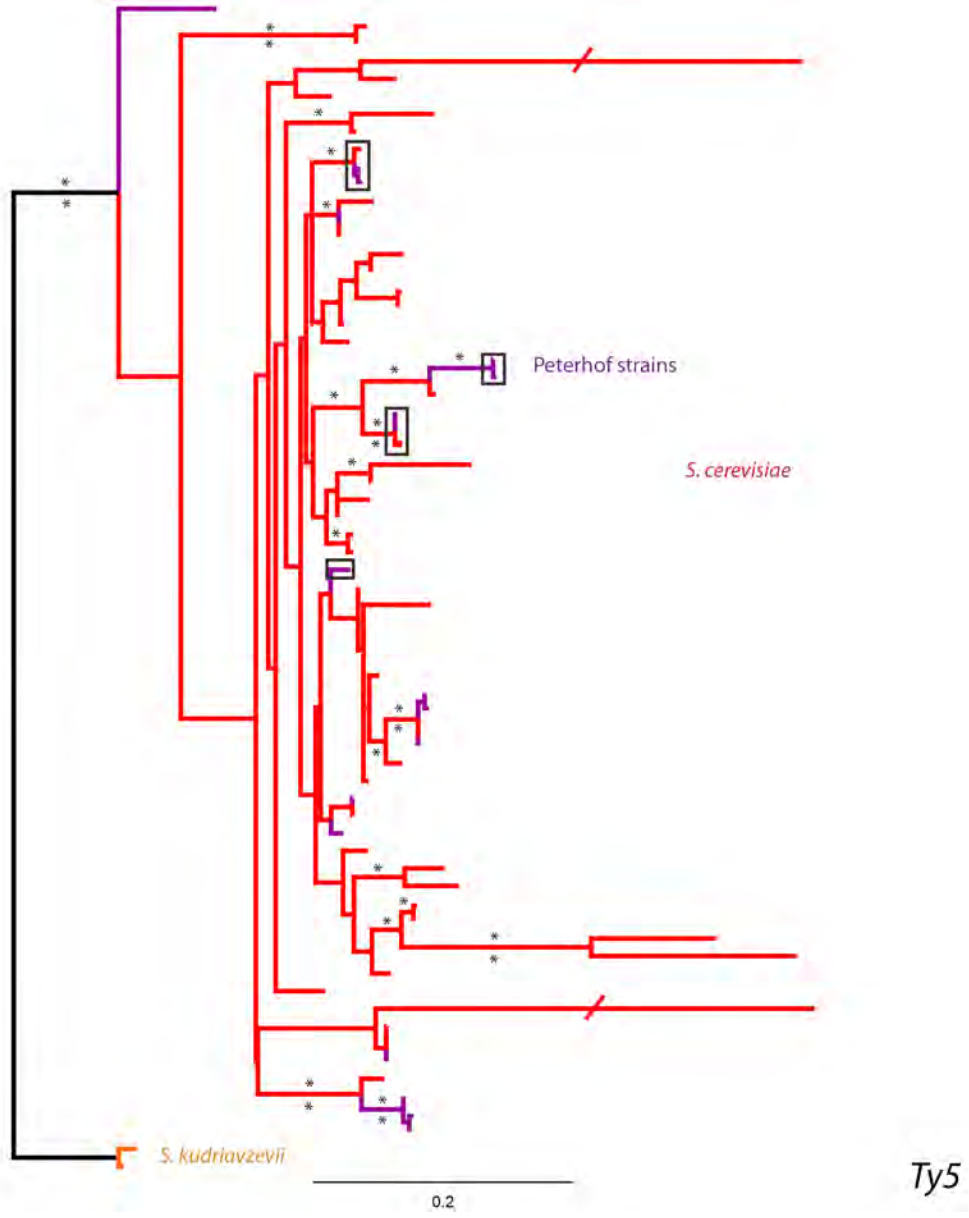

Figure 6.8: ***Ty5* LTR sequences in the SGRP and Peterhof strains of *S. cerevisiae*.** Formatting is the same as in Figure 6.5, but based on an alignment of 265 nucleotide positions and rooted with *S. kudriavzevii* sequences. / indicates arbitrarily shortened branches. Boxes indicate the LTRs associated with relics of elements.

6.3 Brazilian strains of *S. cerevisiae* contain widespread introgression from *S. paradoxus*

Barbosa *et al.* (2016) isolated 28 strains of wild *S. cerevisiae* in Brazil, two of which proved to be hybrids with *S. paradoxus*. The remaining strains contain introgressed regions of varying sizes. Figure 6.9 displays the ancestry of the 28 Brazilian strains as determined by Barbosa *et al.* (2016). The majority of the strains share a common ancestor within clade B1. For brevity the strains are referred to as Y*nnn*, where *n*=strain number, which is a shortened version of those used in Barbosa *et al.* (2016).

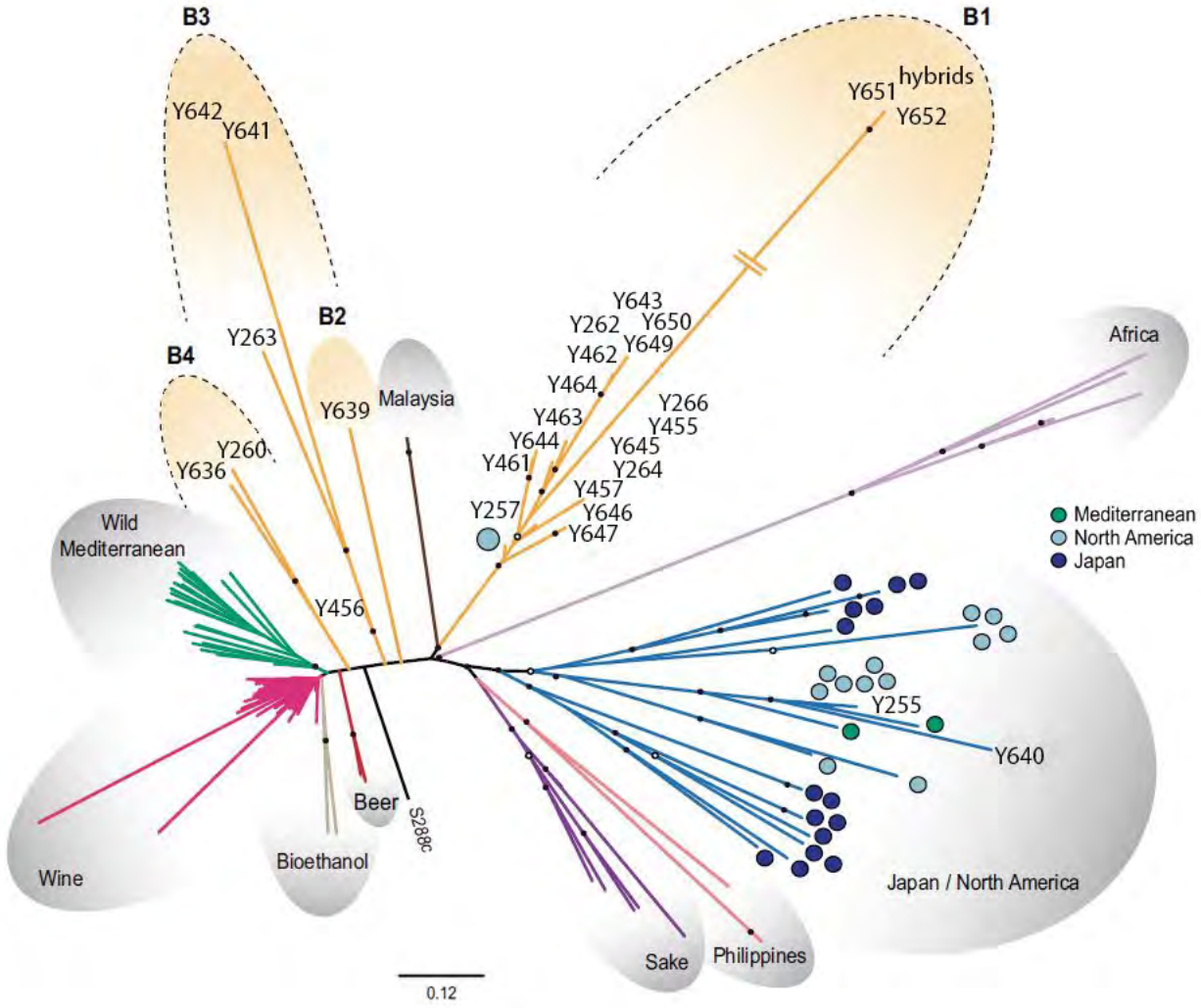

Figure 6.9: **Ancestry of the Brazilian strains of *S. cerevisiae*.** Genome phylogeny of 143 strains to establish the ancestry of the Brazilian *S. cerevisiae* strains in the context of representatives of the main populations currently sequenced. The phylogeny was determined from an alignment of 69,321 SNP positions. The newly sequenced population of Brazilian strains consists of clades B1-4 and two strains which fall within the Japan/North America cluster, presumably the result of migrations. Adapted from Barbosa *et al.* (2016).

Barbosa *et al.* (2016) reported evidence of introgression in 21 strains, no instances of which are fixed in the population. Strains typically contain a single foreign ORF, but larger regions containing up to five ORFs were also discovered. Although the configuration of introgressed region among the Brazilian strains varied, the authors noted two major patterns that corresponded to the phylogenetic groups (the split between B1 and B2-4). The authors assigned most introgressions to the North American population of *S. paradoxus*, determined by comparisons of orthologous genes, but also noted similarities with the sequenced strains of *S. cariocanus*. Nine ORFs are undetermined as to their precise origin but nevertheless belonged to an unknown *S. paradoxus* population that shows more sequence identity to the North American population than to the European or Far East. The authors also discovered that three ORFs commonly gained are *SNF5*, *BMH1* and *MET28*. The complete set of introgressed genes is significantly enriched for transmembrane transporters.

Barbosa *et al.* (2016) submitted raw sequencing reads to the European Nucleotide Archive (ENA) which were downloaded in order to construct the genomes. Construction was performed with SPAdes v.3.9 and organised into scaffolds with MeDuSa v.1.3. Finally, the genomes were screened with RepeatMasker v.4.0.7 using a custom library (Chapter 5 and Appendix P). The highly dispersed regions of introgression from *S. paradoxus* (section 7.1.3) caused issues with scaffolding of contigs, and therefore locations of insertions were not confidently assigned to a particular chromosome. Although the location in the *S. cerevisiae* reference genome could be approximated via flanking DNA, this would not take into account any chromosomal rearrangements that may have occurred during and post-hybridisation.

6.3.1 TE variation in Brazilian strains

All insertions were extracted from the strains, aligned and LTRs unique to the Brazilian strains extracted for nucleotide diversity (π) analysis with DnaSP v5.10 (Table 6.3). Insertions were divided into their respective subfamilies in order to improve alignments and therefore generate more accurate π values, which are indicative of familial age. Values would suggest that *Ty1*, followed by European *Ty4*, *Ty5* and *Ty3p* have inhabited the genomes far longer than the other four subfamilies. Alternatively, the insertions of these families are less deleterious to their hosts and therefore able to persist in the genomes and accumulate mutations.

TE and GC content, level of introgression and *Ty* copy numbers were collected for each Brazilian strain in Appendix Q. The TE content ranges from 0.96% in Y255 to 1.85% in hybrid strain

Y652, with a mean content of 1.31%. GC content is relatively stable, with the exception of North American migratory strain Y255 (Table Q.1, Q). Interestingly, this strain also possesses the lowest TE content and introgression level from *S. paradoxus*. Y263 is the only strain to experience extinction of all families.

Family	Subfamily	Unique LTRs			Nucleotide diversity (π)
		FLE	Solo	Total	
Ty1/2	*Ty1*	22	376	398	0.26001
	Ty1p	13	185	198	0.07727
	Ty2	4	16	20	0.03881
Ty3	*Ty3*	5	79	84	0.06384
	Ty3p	-	15	15	0.11113
Ty4	American	3	28	31	0.02481
	European	*	78	78	0.16459
Ty5	-	5	39	44	0.12982

Table 6.3: **Nucleotide diversity calculated for unique insertions in the Brazilian strains of *S. cerevisiae*.** *LTRs from European elements were not able to be extracted in full as they are disrupted by the ends of reads.

Rate of genomic recombination may be indicated by the number of contigs built by SPAdes and scaffolded by MeDuSa onto the reference genomes of *S. cerevisiae* and *S. paradoxus*. Contigs containing a recombination breakpoint would not be integrated into a scaffold if that breakpoint differed from the reference strains onto which it was being mapped. Therefore recombination may be estimated by number of contigs that failed to be built further due to differing breakpoints relative to the reference genomes. TE insertions are known to cause genome rearrangements as they serve as templates for ectopic recombination (Hoang *et al.*, 2010). As LTRs in particular are commonly present at the ends of contigs in the Brazilian genomes (>30%), the relationship between the number of contigs and genomic TE content was examined. Figure 6.10 displays the weak positive correlation observed between the two (R^2=0.3774). It is highly likely that *Ty* sequences may be at least partially responsible for the level of recombination observed in the genomes of the Brazilian strains. However, it should be considered that this may also be due to the fact that repetitive sequences cause assembly issues (Hoban *et al.*, 2016; Treangen and Salzberg, 2011) rather than being purely representative of recombination.

6.3.2 Species hybridisation and *Ty* elements

Barbosa *et al.* (2016) previously identified strains (Y651 and Y652 as *S. cerevisiae* x *S. paradoxus* hybrids by analysing sequence divergence. Here, the raw reads were mapped onto a reference file

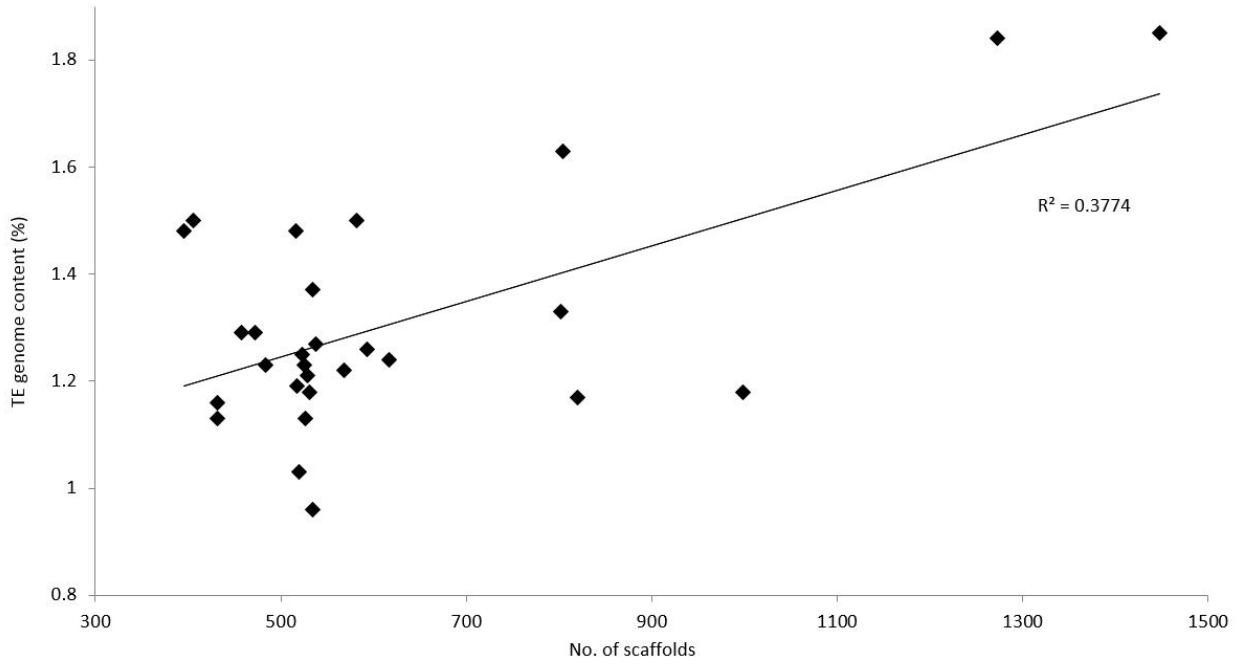

Figure 6.10: **Correlation between genomic TE content and no. of contigs in the Brazilian strains of *S. cerevisiae*.** Weak positive correlation is observed between genomic TE content % and no. of contigs (indicative of recombination breakpoints). It was ensured that no contigs contain regions of duplication before analysis.

of *S. cerevisiae* and *S. paradoxus* genomes (Borneman and Pretorius, 2015) in order to visualise hybrid genomes. A number of observations were unique to the two hybrid strains. Firstly, they possess the highest TE content percentage of all the isolated Brazilian strains (1.84 and 1.85% respectively). Secondly, they are the only strains to contain multiple copies of *Ty3* and *Ty4* FLEs. Copy number variation resulting from aneuploidy is a mechanism of adaptation to environmental changes (Dunham *et al.*, 2002; Kondrashov, 2012). Further evidence of aneuploidy is observed in only one other Brazilian strain (section 6.3.3). Y651 possesses an aneuploid genome (Figure 6.11) and Y652 is a complex diploid (Figure 6.12). Aneuploid Y651 contains an additional copy of chromosome V, but is otherwise haploid. Its chromosomes are shared between parental species, skewed slightly toward *S. cerevisiae*. Chromosomes I, II, VI, XII and XVI are entirely from *S. cerevisiae*, whereas chromosomes III, VIII, X and XV are of *S. paradoxus* origin. Chromosomes IV, VII, XI, XIII and XIV are non-homologous and have undergone recombination/large scale translocation events (Figure 6.11).

Y652 has a complex diploid genome (Figure 6.12). Interestingly, one copy of chromosome XII is entirely of *S. cerevisiae* origin, but the other is a combination of *S. cerevisiae* and *S. paradoxus*. Similar recombination events have occurred in chromosomes IV, VII, IX and XIV, detailed in Figure

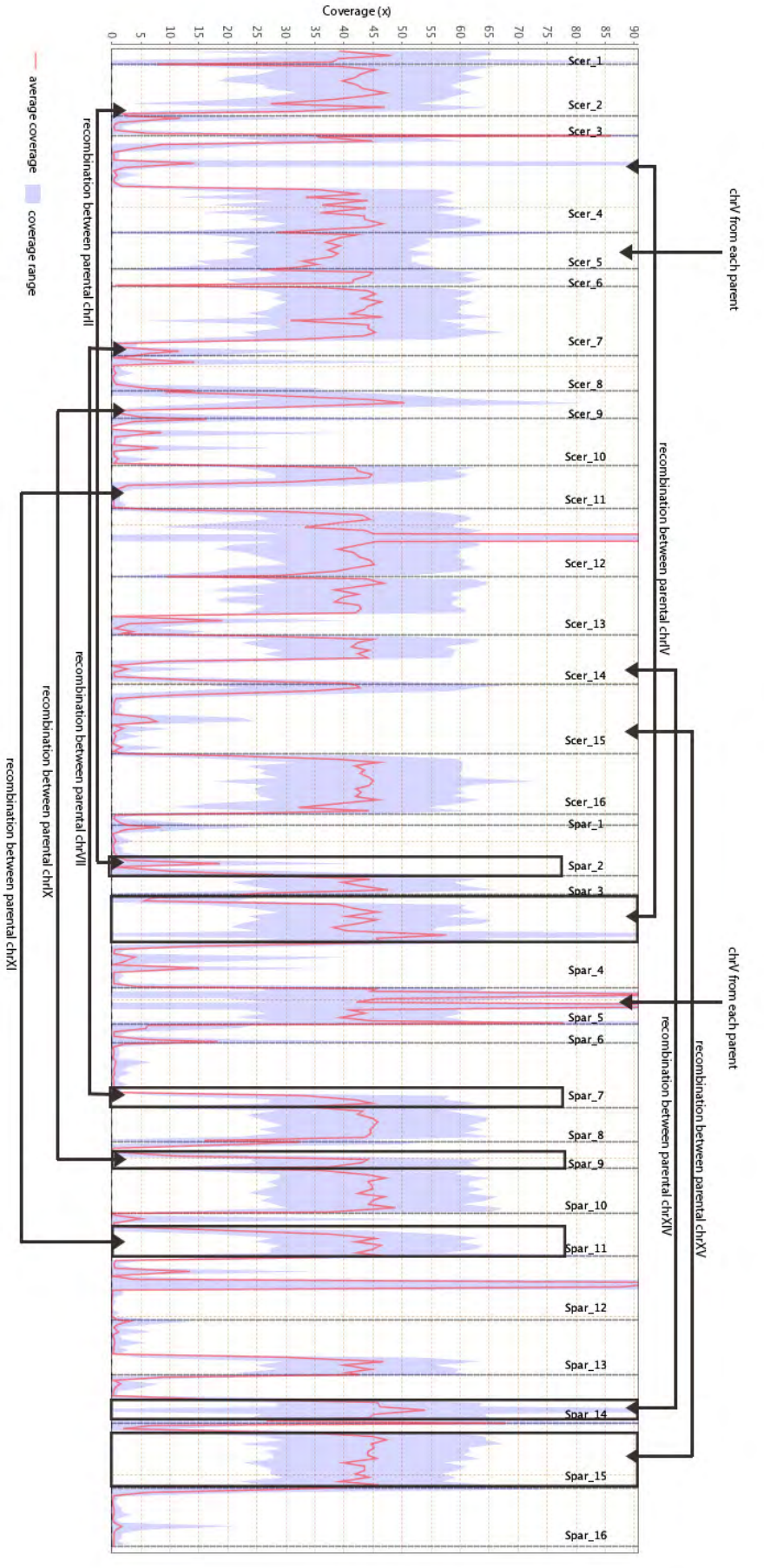

Figure 6.11: **Chromosomal organisation illustrated by genome coverage of aneuploid strain Y651 with two copies of chromosome V.** Recombination events (boxed regions) between chromosomes (x-axis) are detailed in the figure. Translocations between chromosomes could not be visualised due to the contigs ending at insertions or discrepancies with the reference genomes to which they were mapped. Legend indicates average and range of coverage of mapped reads onto *S. cerevisiae* and *S. paradoxus* reference genomes.

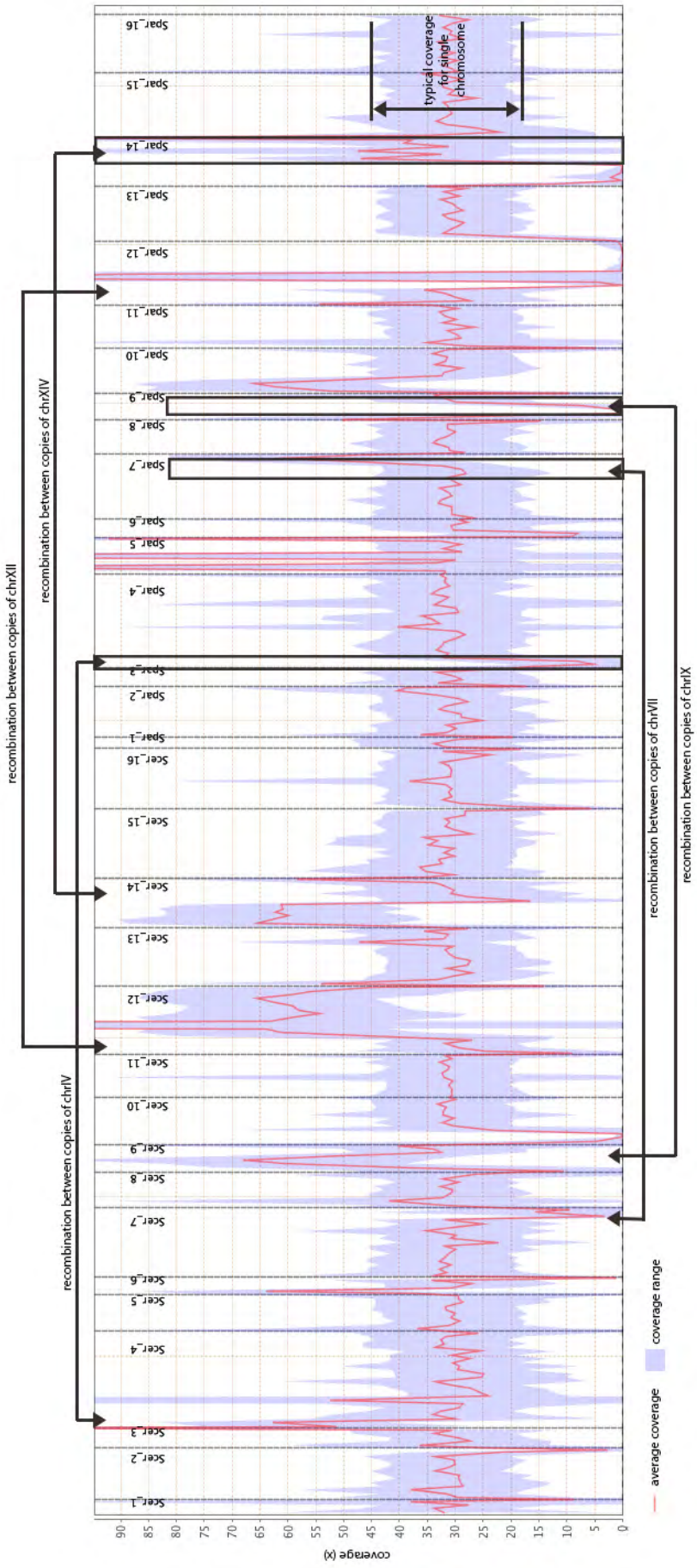

Figure 6.12: **Chromosomal organisation illustrated by genome coverage of diploid strain Y652.** Recombination events (boxed regions) between chromosomes are detailed in the figure. Translocations between chromosomes could not be visualised due to the contigs ending at insertions or discrepancies with the reference genomes to which they were mapped. Legend is the same as in Figure 6.11.

6.12.

6.3.3 Introgression is widespread in the genomes of Brazilian wild *S. cerevisiae*

By mapping reads onto a reference file of *S. cerevisiae* and *S. paradoxus* genomes, aneuploidy and introgressed regions were visualised such as in strain Y456 in Figure 6.13. S288c was used as a negative control for locating regions of introgression which identified a region on chromosome XII of *S. paradoxus* as a false positive (see B of Figure 6.13). Reads from S288c mapped to this region, so it was omitted from the results of the Brazilian stains. The percentage of the genome that was gained from *S. paradoxus* by introgression was calculated by aligning reads to the *S. cerevisiae* reference genome, then with a reference containing both *S. cerevisiae* and *S. paradoxus* genomes. The difference between these two percentages, taking into account the region on chromosome XII, provided a conservative estimate of the proportion of the genome gained from *S. paradoxus*. The false positive region on chromosome XII is highly conserved between *S. paradoxus* and *S. cerevisiae* only, and consists of a duplicated region containing copies of the *TAR1* gene and a number of putative ORFs within the region of $\sim$435-480kb, depending on the species. As no more than 2% of reads could be attributed to the false positive region of chromosome XII, this was used to establish the margin for false positives when determining the degree of introgression. This 2% was disregarded, which meant that introgression is observed in 23 (89%) of the 26 non-hybrid strains (Table Q.1 Q). Although regions of introgression in chromosomes V and VI are commonly observed across the strains, none were found to be fixed across the population. Regions that failed to map to either *S. cerevisiae* or *S. paradoxus* reference genomes were used as BLAST queries to identify other strains and species in which they may have originated. All are present in other strains of their parental species, but not in other species, reflecting similar results obtained by Barbosa *et al.* (2016). None contain *Ty* sequences.

Strains Y263 and Y266 possess $\sim$2-4% genomic introgression from *S. paradoxus*. However, only $\sim$80-89% of their genomes were mapped onto *S. cerevisiae*, leaving >8% unaccounted for. Their raw reads were each mapped to reference genomes of all *Saccharomyces* species and North American *S. paradoxus* to trace their possible origin. However, this did not reduce the percentage of unmapped reads. Using these regions as BLAST queries shows that they are of *Saccharomyces* origin, but that the population(s) in which they originated are yet to be identified and/or sequenced, as only partial hits were found.

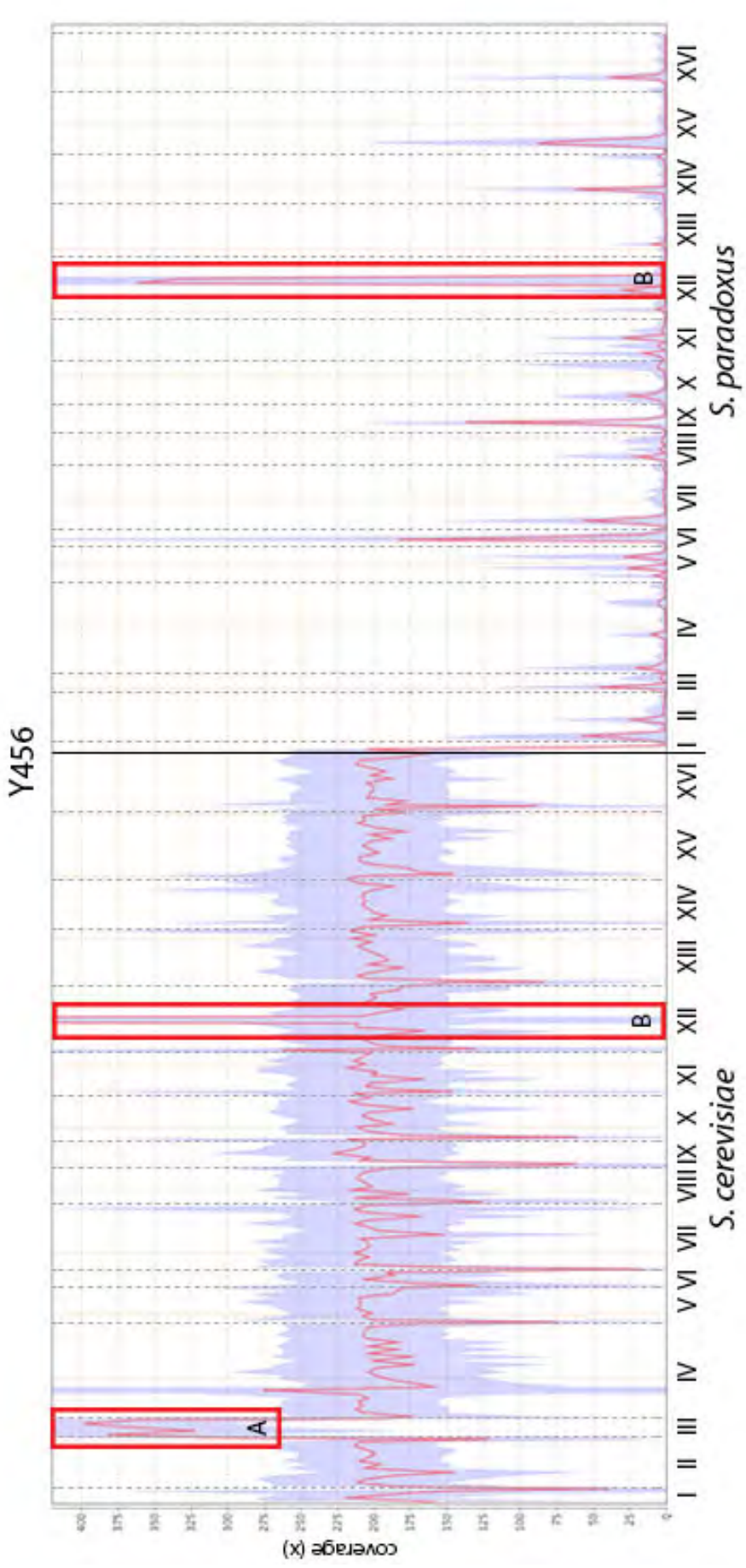

Figure 6.13: **Relative coverage of reads from strain Y456 onto the reference genomes of *S. cerevisiae* (left) and *S. paradoxus* (right).** Dotted vertical lines represent chromosome boundaries. Box A: aneuploidy in chromosome III of Y456 illustrated by approximately double the coverage of other chromosomes. Boxes B – region on chromosome XII with abnormally high coverage. This spike in coverage appeared on the genome maps regardless of the strain investigated (even S288c as a control possesses a single spike in this region) and so represented a falsely positive introgressed region due to conservation between the species.

With 40 protein coding genes identified as of *S. paradoxus* origin by Barbosa *et al.* (2016), strain Y263 was expected to contain the largest proportion of DNA gained from *S. paradoxus*. However, it is in fact strain Y640 that contains the largest non-*S. cerevisiae* proportion of the genome outside of the hybrid strains. Interestingly, Barbosa *et al.* (2016) did not recover any ORFs of *S. paradoxus* origin in this strain. Upon examination of its TE content, this strain contains a large number of *S. paradoxus*-like insertions, which may account for its apparent increased level of introgression.

6.3.4 Varying copy numbers of *Ty1/2* insertions

All 28 strains contain *S. paradoxus*-like *Ty1* insertions to varying degrees (*Ty1p*; Table 6.4), with full-length elements present in 19 strains (68%). *Ty1/2* coding regions of *S. cerevisiae* origin are found in all strains except Y263 which has undergone extinction of the entire superfamily (Table 6.4). *Ty2* elements share 99% similarity, and together with strain-specific *Ty1* (*n*=12) and *Ty2* (*n*=4) full-length insertions show that transposition of elements within these subfamilies has occurred relatively recently and/or since these strains share an ancestor. A small number of *Ty1* FLEs are present at loci as solo LTRs in other strains, having undergone intra-element recombination (*n*=4).

Both *Ty1* and *Ty2* are lost in strains Y641 and Y642, with only relics and solo LTRs to attest to their previous existence. *Ty2* is extinct in a further eight isolates, which correlates with strains falling within clades B2-4 in Figure 6.9, suggesting a common point of loss. *Ty1p* pseudoelements are present in all but four strains, with the complete loss of this subfamily in almost a third (*n*=9). No evidence of *Ty1*' FLEs was uncovered. Brazilian *Ty1p* LTRs typically share $\sim$93% identity with those of *S. paradoxus*.

All *Ty1/2* elements in the hybrids are confined to their original chromosomal background. Y651 and Y652 each contain only one copy of elements in the *S. cerevisiae*-like *Ty1* and *Ty2* subfamilies, which are present in *S. cerevisiae* DNA. Additionally, Y651 contains a single copy of *Ty1p*, whereas Y652 possesses three, all of which are present within *S. paradoxus* DNA, pointing to unlikely activity since the hybridisation event, unless transposition is confined to the *S. paradoxus* chromosomes.

Five strains (Y641, Y642, Y652, Y456 and Y263) each contain a single copy of a 1.4kb *Ty1*-like relic which shares 90% identity with *Tsk1* from *L. kluyveri*, than the endogenous *Ty1* relic commonly found in SGRP strains. However, only the IN domain remains intact in these five highly conserved relics (99% identity and similarity), and the rest of the element has degraded. The AA sequence

Strain	*S. paradoxus*-like	*S. cerevisiae*-like	
	Ty1p	*Ty1*	*Ty2*
Y260	0(1)	1(1)	0(2)
Y262	0(1)	1(1)	1(1)
Y264	0(1)	0(1)	1(1)
Y455	0(1)	0(1)	1(0)
Y461	1(1)	1(1)	1(1)
Y462	1(1)	2(1)	0(2)
Y464	1(1)	2(1)	1(0)
Y639	1(1)	1(1)	0(3)
Y640	1(1)	1(1)	0(2)
Y641	2(1)	0(1)	0(1)
Y642	2(1)	0(1)	0(1)
Y645	0(1)	1(1)	1(1)
Y646	2(1)	0(1)	1(1)
Y647	2(0)	0(1)	1(1)
Y649	1(1)	2(1)	1(1)
Y650	2(1)	1(1)	1(1)
Y651	2(1)	1(1)	0(1)
Y652	3(0)	1(1)	1(1)
Y255	0(0)	1(1)	0(2)
Y257	1(0)	1(0)	1(1)
Y263	0(1)	0(1)	0(1)
Y266	1(1)	0(1)	1(1)
Y456	0(1)	2(1)	1(1)
Y457	0(1)	1(1)	1(1)
Y463	1(1)	1(1)	1(1)
Y636	1(1)	1(1)	0(2)
Y643	1(1)	1(1)	1(1)
Y644	2(1)	1(1)	1(1)

Table 6.4: **Copy numbers of *Ty1*-like coding regions in the Brazilian strains of *S. cerevisiae*.** Element numbers: FLE(pseudo).

also shares 90% similarity with a relic copy in *S. cariocanus*, and the flanking DNA in the five strains is shared, likely originating from *S. paradoxus*. Possible functional constraint was identified by calculating Ka/Ks values with default settings (http://services.cbu.uib.no/tools/kaks), all of which are close to zero (0-0.004; Appendix R). This indicates that the positions are possibly evolving under purifying selection, and that the IN domain may provide function to the host strains. However, as too few changes have occurred since the strains diverged, the exact nature of selection at this locus remained unclear.

A total of 614 *Ty1/2* LTRs were recovered from the 28 strains. 398 (65%) sequences were identified as *S. cerevisiae Ty1* and subfamilies, 198 (32%) as *S. paradoxus Ty1p* and 20 (3%) as *Ty2*. The subfamilies could not be confidently determined, either due to divergence from canonical sequences or breakpoints indicating recombination. Excluding one *Ty1*-like relic (discussed above), no divergent LTRs are associated with coding regions.

Unique insertions represent independent activity of elements in their respective genomes, observed in 23 non-hybrid strains (Table 6.5). Three strains (Y257, Y266 and Y463) possess no unique activity in either subfamily, and are therefore unlikely to be active in these populations since the divergence of these strains. All three are early branching strains in the B1 clade of Figure 6.9, suggesting the activity level seen in the other strains began after the divergence of these from the ancestor.

Strain	*S. paradoxus*-like *Ty1p*	*S. cerevisiae*-like *Ty1*
Y260	12	58
Y262	16	9
Y264	11	20
Y455	3	18
Y461	3	11
Y462	9	6
Y464	9	13
Y639	4	26
Y640	10	20
Y641	36	9
Y642	13	5
Y645	2	6
Y646	2	6
Y647	1	3
Y649	7	5
Y650	4	1
Y255	1	2
Y257	0	0
Y263	1	4
Y266	0	0
Y456	2	3
Y457	2	1
Y463	0	0
Y636	3	0
Y643	1	2
Y644	1	2

Table 6.5: **Copy numbers of unique *Ty1* and *Ty1p* insertions in the non-hybrid Brazilian strains of *S. cerevisiae*.**

There is however no correlation between percentage of introgressed genome from *S. paradoxus* and *Ty1p* copy number (R^2= 0.0641; Figure 6.14), suggesting that few *Ty1p* insertions were gained during the introgression events, and that the activity of this subfamily occurred later (discussed in Section 6.3.5).

Furthermore, hybridisation between *Ty1/2* FLEs of these strains seems to be uncommon. Recombination events occurring within LTRs are not observed, and coding regions were identified

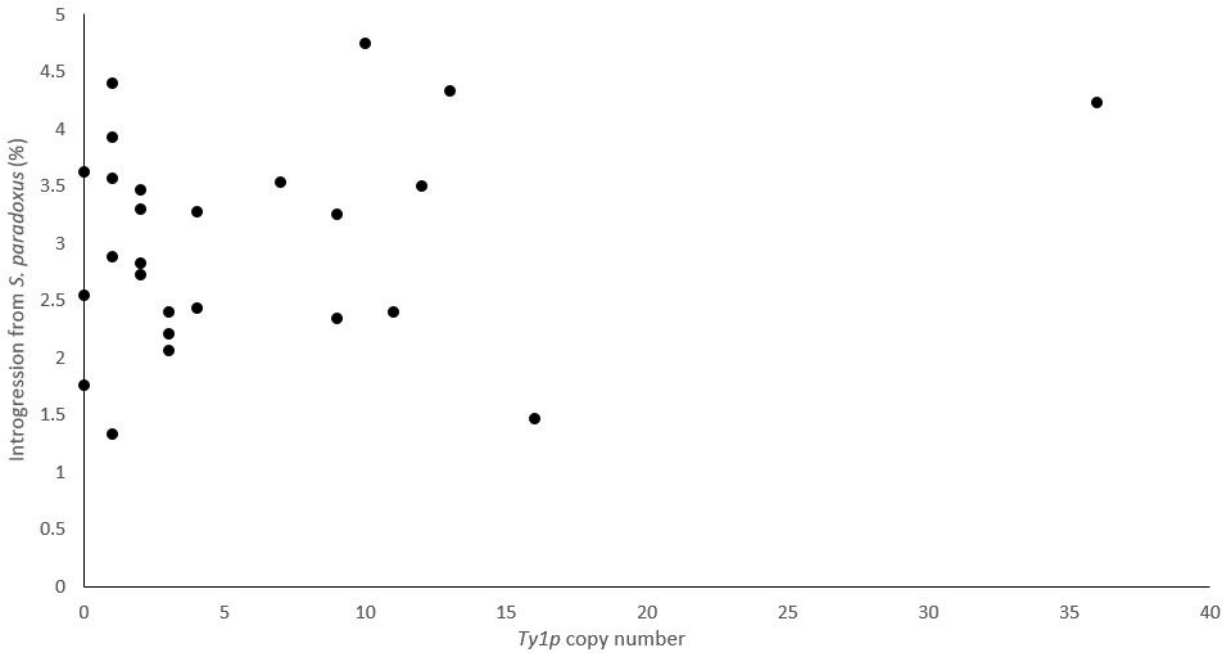

Figure 6.14: **Lack of correlation between the introgression and *Ty1p* copy number.** No correlation is observed between the percentage of genome gained from *S. paradoxus* and *Ty1p* copy number, suggesting that activity of this subfamily occurred post-hybridisation/introgression.

as distinctly *Ty1* or *Ty2* in their entirety. The exception to this is in nine strains which contain a degraded element whose identity is unclear. The 5' LTR belongs to that of *Ty1*, but the degraded *gag* region is more like that of *Ty2* (data not shown). However, the element is interrupted after ∼700bp by the host gene *ASE1*, perhaps due to an assembly error, so it cannot be confidently identified as a hybrid. The degraded insertion is however highly conserved in the three strains (95-99% identity).

6.3.5 A *S. paradoxus Ty1* subfamily was active post-introgression

To establish whether these are simply present in the introgressed regions gained from *S. paradoxus* or have been active in the time since introgression, strain Y641 was chosen for investigation as it possesses the highest frequency of unique *S. paradoxus*-like *Ty1* LTRs (*Ty1p*; Table Q.1). The strain contains significantly fewer *S. cerevisiae*-like *Ty1* LTRs (*n*=9), none of which are associated with FLEs. Additionally, Y641 contains both solo *Ty1p* LTRs (*n*=63 total; *n*=33 unique) as well as those associated with FLEs (*n*=3; the LTR from the remaining FLE was not able to be extracted in full due to fragmentation). Upon investigation of coding regions within this strain, both autonomous *Ty1*-like elements were found be that of *S. paradoxus* origin as opposed to *S. cerevisiae*. The FLEs (*n*=2) sre present on scaffolds where enough DNA flanking at least one end of each insertion is

present to extract 1kb to use as a BLAST query to ascertain its origin. A *S. paradoxus*-like element in *S. cerevisiae* flanking DNA would indicate transposition post-introgression. In both cases, the *Ty1p* elements are present in *S. cerevisiae* regions, indicating transposition of active elements has occurred after the introgression events.

Solo LTRs were also investigated in strain Y641 to ascertain previous activity levels of *Ty1* and *Ty1p*. The majority of *Ty1p* LTRs possess *S. cerevisiae* flanking DNA (*n*=60) and were therefore active post-introgression and have since undergone recombination. The remaining *Ty1p* solo LTRs are present in the *S. paradoxus* regions (*n*=3) and therefore likely present in this position when the regions were gained by introgression. It is unclear whether the *S. cerevisiae Ty1/2* insertions (*n*=28) have been active in the time since introgression, as their positions within endogenous genomic regions do not reveal their level of activity as with *Ty1p* insertions. None are found in the introgressed regions of Y641 or any other strain. Brazilian *Ty1p* LTRs and copies in *S. paradoxus* strains share ≥93% nucleotide identity, including solo LTRs, indicating the acquisition of this subfamily was relatively recent as few mutations have have time to accumulate.

The possible source of the *Ty1p* family in the Brazilian strains is unlikely to be a single shared insertion, as unsurprisingly, the introgression patterns are not fixed in the population. Most *Ty1p* insertions are low frequency and polymorphic. However, a number of solo insertions are of high frequency (*n*=4; $\sim$20 strains). No fixed insertions were discovered in all strains.

6.3.6 *Ty3* elements were gained from both parental species

Only the hybrid strains contain full-length copies of *Ty3p* from *S. paradoxus*. Strain Y457 contains partial *Ty3p* coding regions alongside the single endogenous *S. cerevisiae* copy. In the remaining 25 strains, all of the *S. paradoxus*-like sequences are present as solo LTRs. The unique solo *Ty3p* insertions (*n*=15) show almost twice the level of diversity than the endogenous *S. cerevisiae Ty3* sequences, which are a mix of solos (*n*=79) and those associated with FLEs (*n*=5). Copies in both subfamilies are present only in their respective regions, i.e. *Ty3p* in the introgressed regions and *Ty3* in the main *S. cerevisiae* genome, indicating that although *Ty3* has likely been recently active, it has not yet transposed into the *S. paradoxus* regions.

The endogenous full-length *Ty3* copy is highly conserved throughout the 25 strains π=0.0075). Stochastic loss of *Ty3* has occurred in three strains (Y641, Y642 and Y263), as no coding regions remain.

Additionally, strain Y640 contains a ∼1kb remnant of a *Ty3/gypsy*–like element that shares 81% identity with the RT-RH region of *Tlf3* in *L. fermentati*. However, no further domains are recognisable.

6.3.7 American and European *Ty4* insertions in the Brazilian strains

Ty4 is lost in seven of the strains, present as a single copy in the remaining strains and only the hybrids contain multiple copies (Table 6.6). Upon examination of the elements in hybrid Y651, the two proved to be of different origins: one *S. uvarum*-like (American) and one *S. cerevisiae*-like (European). The copies share 77% identity over the entire coding region. The LTRs, as seen previously in Chapter 4, share little identity and also differed in length (291bp and 370bp, respectively). Elements in hybrid Y652 unfortunately could not be compared as they are split over multiple reads, and even after scaffolding could not be manually assembled. The strain did however contain the flanking LTRs of three elements (two American and one European) but autonomy could not be ascertained due to the lack of contigs covering entire coding regions.

In the strains containing European *Ty4*, elements share 96% identity with those of their parental species, *S. cerevisiae* and *S. paradoxus*. They also share ∼72% identity with the American type over the majority of the element. No evidence of recombination between elements of the two types is not observed in coding regions or LTRs.

No full-length European insertions are unique to the Brazilian strains, as all are present in their parental species. However, the majority of solo LTRs are unique to individual strains, indicating that recent activity in this family has possibly occurred, but all elements have since undergone recombination. Unfortunately, most coding regions span entire contigs, beginning and ending in the LTRs, so TSDs and flanking DNA could usually not be ascertained. The short length of raw reads (70bp) prevented elements being confidently identified. Those LTRs with enough flanking DNA to be used as BLAST queries show that the insertions are still confined to their original loci. Notably, no European pseudo or partial elements are present, as only LTR-LTR recombination has caused the loss of this subfamily in around a third of the strains (n=9).

LTRs from the American subfamily of *Ty4* are present in five strains: the two hybrids, Y641 and Y642, plus a single solo LTR in strain Y456. Limited to these five strains, nucleotide diversity is far lower in American LTRs (π=0.02481) than European (π=0.16459). Two European-type solo LTRs are fixed in all 28 strains, and an additional insertion is fixed in 23 strains. The remaining insertions

Strain	*S. uvarum*-like American	*S. cerevisiae*-like European
Y260	0(0)	1(0)
Y262	0(0)	1(0)
Y264	0(0)	1(0)
Y455	0(0)	1(0)
Y461	0(0)	1(0)
Y462	0(0)	1(0)
Y464	0(0)	0(0)
Y639	0(0)	1(0)
Y640	0(0)	0(0)
Y641	1(1)	0(0)
Y642	1(1)	0(0)
Y645	0(0)	1(0)
Y646	0(0)	1(0)
Y647	0(0)	1(0)
Y649	0(0)	0(0)
Y650	0(0)	0(0)
Y651	1(1)	1(0)
Y652	2(1)	1(0)
Y255	0(0)	1(0)
Y257	0(0)	0(0)
Y263	0(0)	0(0)
Y266	0(0)	1(0)
Y456	0(0)	1(0)
Y457	0(0)	1(0)
Y463	0(0)	0(0)
Y636	0(0)	1(0)
Y643	0(0)	1(0)
Y644	0(0)	1(0)

Table 6.6: **Copy numbers of *Ty4* coding regions in the Brazilian strains of *S. cerevisiae*.** Element numbers: FLE(pseudo).

are low to moderate frequency (*n*=75; ≤10 strains) and typically polymorphic. The American type however, confined to five strains, are typically fixed in ≤4 strains (*n*=16). The remaining insertions (*n*=15) are polymorphic and present in single instances. Additionally, the subfamily shows possible evidence of recent strain-specific activity, as the solo LTRs and those associated FLEs in the hybrid strains are not found in the corresponding loci in *S. paradoxus*. However, this may also be due to the fact that the exact population of *S. paradoxus* that underwent hybridisation with these strains of *S. cerevisiae* is yet to be discovered.

The American *Ty4* elements in the hybrid strains possess divergent *gag* regions. The assumed partial copies are unfortunately present on short reads with their 5' LTRs. In Y652, this element could not be constructed in full as regions covering the PR and IN domains were not located in the reads by RepeatMasker, perhaps due to low identity, or simply low sequencing coverage. The

sequence shares 68% identity with the American element and none with the European copy. The FLE was manually constructed in Y651 using overlapping sequences of an alignment of all coding regions in this strain. The *gag* region shares higher identity with the American type (73%) than the European (66%). When used as a query in BLASTp, a Gag region from *S. eubayanus* is the closest hit, with 78% similarity. The *pol* region of this element shares 87% identity with the American copy, and far less with European ($\sim$30%). The LTRs are also shorter as in the American elements (291bp) and show little evidence of mutation. The entire Y651 element did however share 95% similarity with the element from *S. cariocanus* lacking RT (Chapter 4), hinting at a possible link between the elements of these populations. It was therefore concluded that this is a further subfamily of American *Ty4*, much like *Ty1*', elements of which differ from *Ty1* mainly in the *gag* region.

An alignment of translated Gag regions in all strains (Figure 6.15) shows that the European copies are all highly conserved and typically differed from the canonical *S. cerevisiae* Gag region at 16 sites. In contrast, American-type Gag sequences share far less similarity. The two American subtypes are clearly seen – those like *S. uvarum* (with one copy in *S. eubayanus* CRUB1971) and those like the remaining *S. eubayanus* elements. Variability in the Gag region is currently unreported outside the *Ty1/2* superfamily.

6.3.8 *Ty4* elements may contain extra domains

An NCBI Conserved Domain Search (CDS) with the *pol* regions of American and European elements revealed the presence of potential extra domains between IN and RT (Figure 6.16). The domain in the European elements shares similarity with the COG3942 superfamily, a surface antigen usually confined to bacterial species (e value 7.28^{-03}). In the American elements, an extra putative domain is located in the hybrid strain Y651 and determined as Utp14, one of the 40 protein components of the ribosomal small subunit (e value $9.85e^{-03}$; Dragon *et al.*, 2002). In the *S. uvarum Tsu4* element, this location is occupied by a zinc-finger domain (Chapter 4). No further American copies contain the extra domain.

Searching with the coding regions of all European elements in the Brazilian strains returned a further 10 copies with the functional domain. Upon alignment of the protein sequences, the presence of the putative antigen domain appears to coincide with an insertion of 17 residues (Figure 6.17). Elements from the reference strains of *S. cerevisiae* and *S. paradoxus* are missing this 17

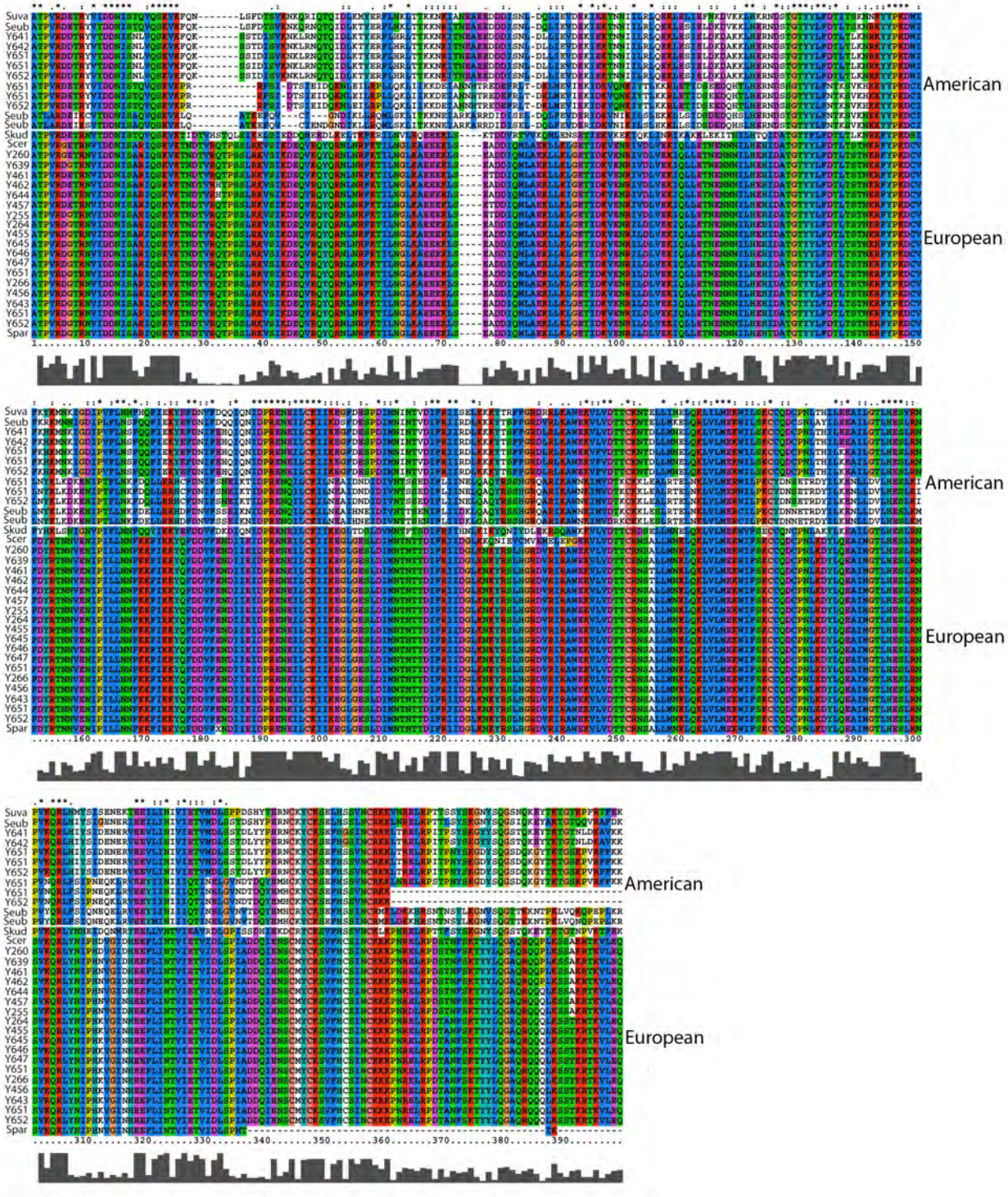

Figure 6.15: **Alignment of the divergent Gag regions in the types of *Ty4* elements in the Brazilian strains of *S. cerevisiae*.** Top section consists of the 'American' species and strains of Brazilian *S. cerevisiae*, including examples from *S. eubayanus* (Seub) and *S. uvarum* (Suva). Lower 'European' section contains *S. cerevisiae* (Scer), *S. paradoxus* (Spar), *S. kudriavzevii* (Skud) and sequences from the European elements of the Brazilian strains. Conservation indicators are as in Figure 6.3.

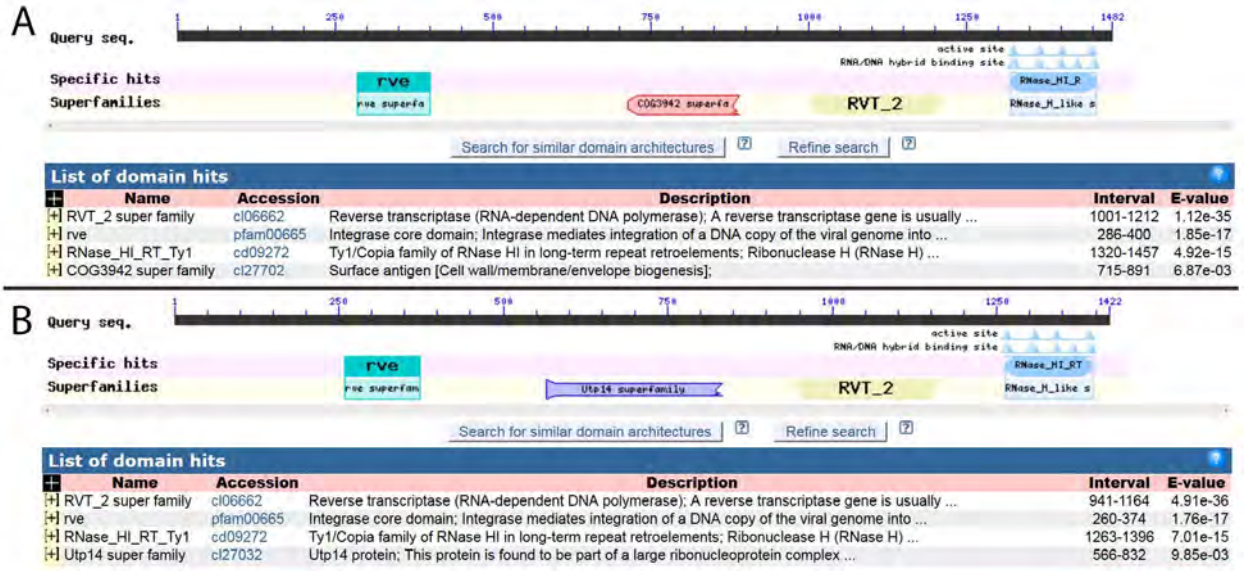

Figure 6.16: **NCBI CDS result of Pol regions in American and European type *Ty4* elements.** European elements (represented by the copy in strain Y266; A) and American elements (represented by a copy in strain Y651; B) contain putative extra domains, resembling a surface antigen and small ribosomal subunit. Both are positioned between IN and RT domains in their respective elements. Although present, the PR domains are not annotated.

residue insertion and lack the extra potential domain. However, a further four Brazilian elements may contain this insertion, but ∼13 changes to the surrounding domain do not result in the same BLAST hit, therefore a combination of the extra insertion and the changes may determine the presence of the extra domain. The element from the remaining strain (Y462) is fragmented over multiple contigs and could not be reliably constructed.

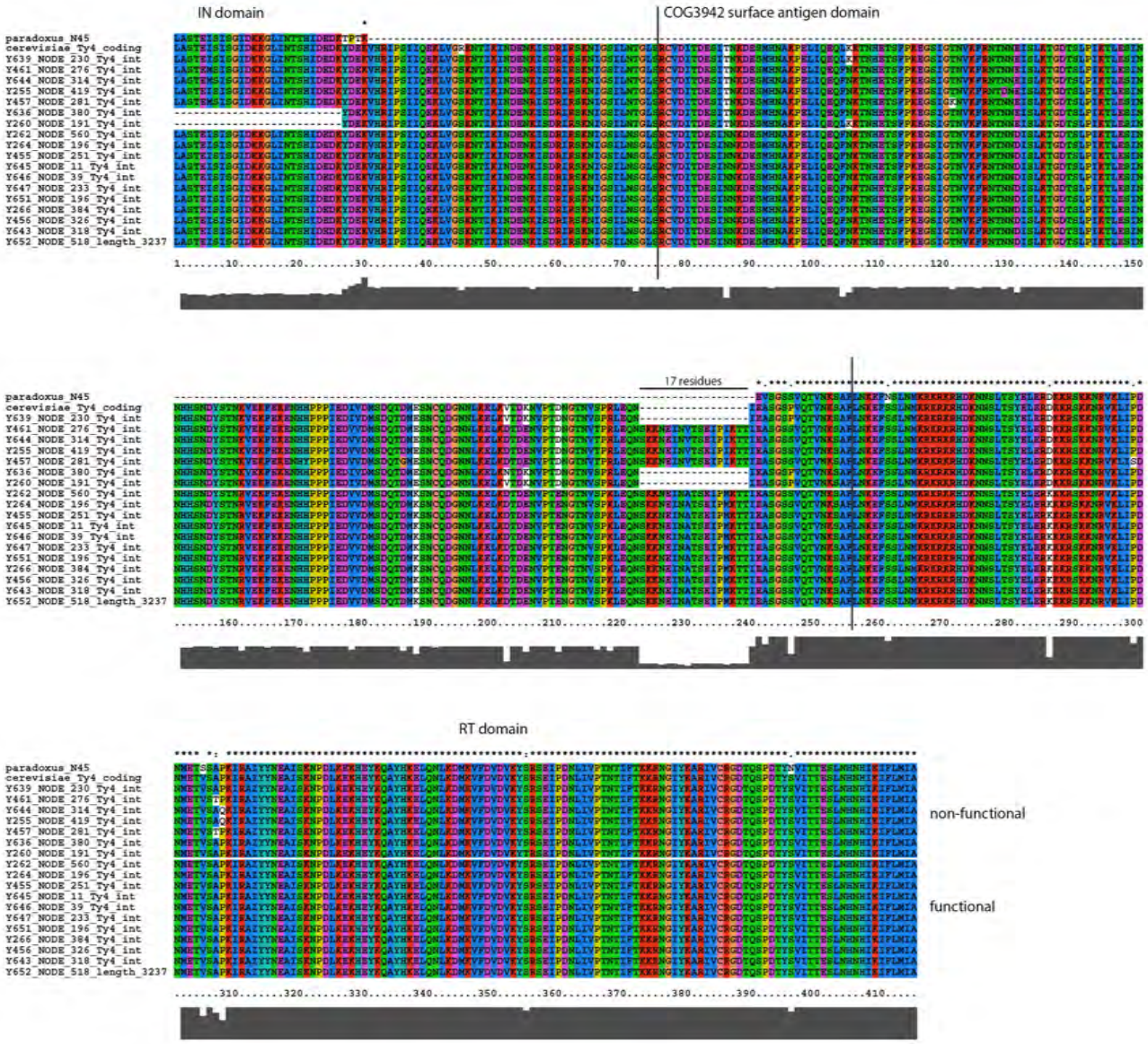

Figure 6.17: **Alignment of the putative surface antigen domain in European type *Ty4* elements.** Alignment of the surface antigen domain (spanned positions 75-255 according to NCBI CDS) and the coding regions surrounding the extra domain in European-type copies of the *Ty4* element. For BLAST to result in the putative extra domain, both the additional 17 residues and the changes as seen in the lower half of the alignment are required. The element in *S. paradoxus* N45 lacked 210 residues in the region between IN and RT observed in *S. cerevisiae* and Brazilian strains. Conservation indicators are as in Figure 6.3.

6.3.9 The *S. cerevisiae Ty5* relic is uncommon in the Brazilian strains

20 of the non-hybrid strains (71%) are entirely devoid of *Ty5* coding regions, possessing a solo LTR in position of the *S. cerevisiae* relic instead. The remaining six strains contain the *S. cerevisiae* relics, totalling seven elements. In two strains, Y462 and Y464, the *Ty5* relic has been interrupted by a newer *Ty1* element which has transposed within the existing *Ty5* coding region before undergoing recombination to result in a solo LTR. Surprisingly, neither hybrid contain intact *Ty5* elements. In strain Y651, the chromosome III is of *S. paradoxus* origin, and so does not possess the *S. cerevisiae Ty5* relic (Chapter 4). However, the relic is also absent in Y652, despite possessing a copy of *S. cerevisiae* chromosome III. The strain instead possesses a solo LTR at the locus, having undergone LTR-LTR recombination. Interestingly, no *S. paradoxus*-like copies of FLEs are found in any strains, only solo LTRs in the *S. paradoxus* regions.

Y260 is the only strain to contain multiple copies of *Ty5*, which possess different TSDs to the relic in the *S. cerevisiae* reference strain on chromosome III. When compared to the reference strain, it is clear that a rearrangement has occurred to result in the altered TSDs (Figure 6.18). The DNA flanking one end of each element originates from chromosome III (grey in Figure 6.18), whereas the opposite ends of the elements (both with TSDs of ATTTT) are flanked by DNA from chromosome VII (black in Figure 6.18). The simplest explanation for this rearrangement is that the strain once possessed two copies of *Ty5*, which underwent a recombination/conversion event to result in the translocation between chromosomes III and VII and the loss of regions PR and IN.

The coding regions and 3' LTRs are identical, with nucleotide changes confined to the 5' LTRs (n=10). Overall, the elements share 99% identity. The position of the stop codon between the *gag* and PR regions causing loss of function in these elements differs to that in the *Ty5* relic in other strains of *S. cerevisiae* (after the PR region), suggesting independent accumulation of stop codons.

As many as 20 solo LTRs are observed in a single strain (hybrid Y652), but no evidence of activity since the introgression events is observed as all insertions remained confined to their respective genomic origins. Solo LTRs unique to the Brazilian strains are observed however, indicating activity since the isolation of the population from its origin in Europe, or loss in the SGRP strains by drift or selection. Four *Ty5* solo insertions are fixed at high frequency, whereas the remaining 35 insertions varied in frequency (≤18 strains), few of which are fixed.

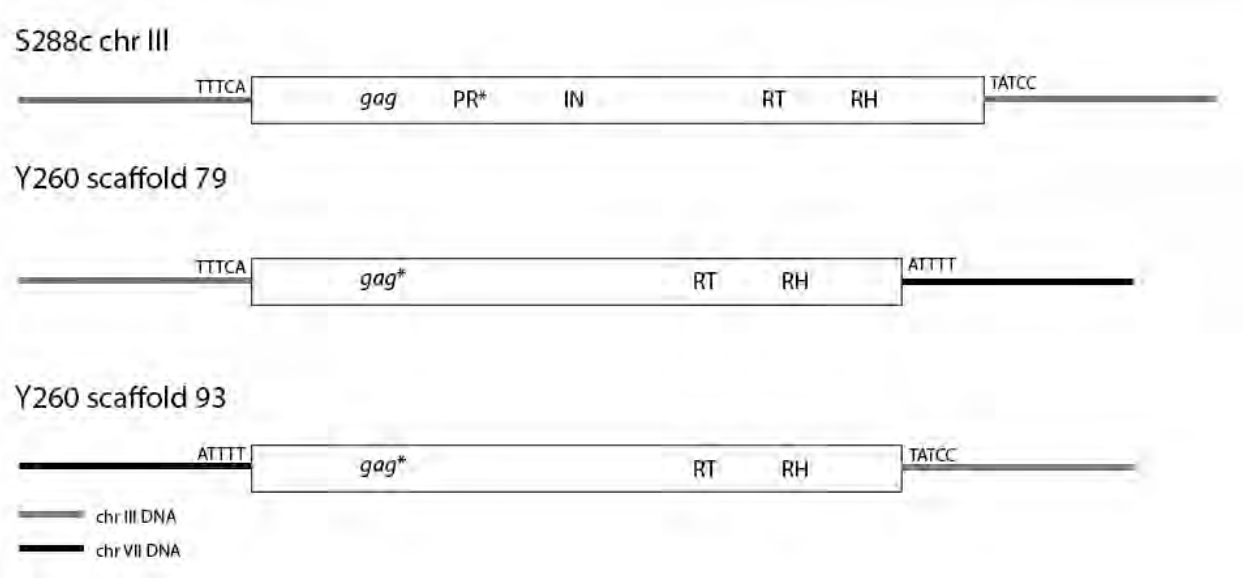

Figure 6.18: **Illustration of the potential recombination between two *Ty5* elements in strain Y260.** Recombination between multiple elements may have caused a chromosomal rearrangement in strain Y260. The recombination event most likely occurred in or around the PR-IN region, and this has resulted in shorter elements with these two domains lost or unrecognisable. *indicates the presence of a stop codon. TSDs are shown immediately 5' and 3' to the elements, respectively. LTRs are not annotated in the figure. Not drawn to scale.

6.4 Brazilian *S. cerevisiae* phylogenetics

In order to minimise phylogenetic noise caused by excessive numbers of sequences, the relationships between the LTR sequences in the Brazilian strains of *S. cerevisiae* with high rates of introgression from *S. paradoxus* were explored separately from the Peterhof sequences. Sequences from all 28 strains were included as both hybrids and introgression strains contain evidence of *Ty* families from both parental species. Duplicate insertions and poor quality sequences were removed before phylogenetic analysis.

6.4.1 *S. paradoxus*-like *Ty1* sequences dominate the Brazilian strains

Despite being predominately *S. cerevisiae* strains, the *Ty1/2* phylogeny (Figure 6.19) displays the majority of Brazilian sequences forming a sister group with *S. paradoxus*, or their own group of long-branched sequences separate from the bulk of SGRP *S. cerevisiae* sequences. Outside the SGRP *S. cerevisiae Ty1* and *Ty2* groups, three main Brazilian groups form (A-C). Group A (*n*=66) consists of long-branched *Ty1*-like LTRs all of which are solo LTRs. Group B (*n*=198) forms the distinct sister group to *S. paradoxus Ty1p*, 7% of which are from FLEs (*n*=14). Almost half of the sequences in this grouping originate in the two hybrid strains (*n*=99). Finally, group C (*n*=58) also

consists of long-branched sequences, and are all solo LTRs except one, which is associated with the degenerate element (section 6.3.4). A small number of Brazilian sequences fall within the main *S. paradoxus Ty1* grouping (*n*=3), indicating a more recent transfer.

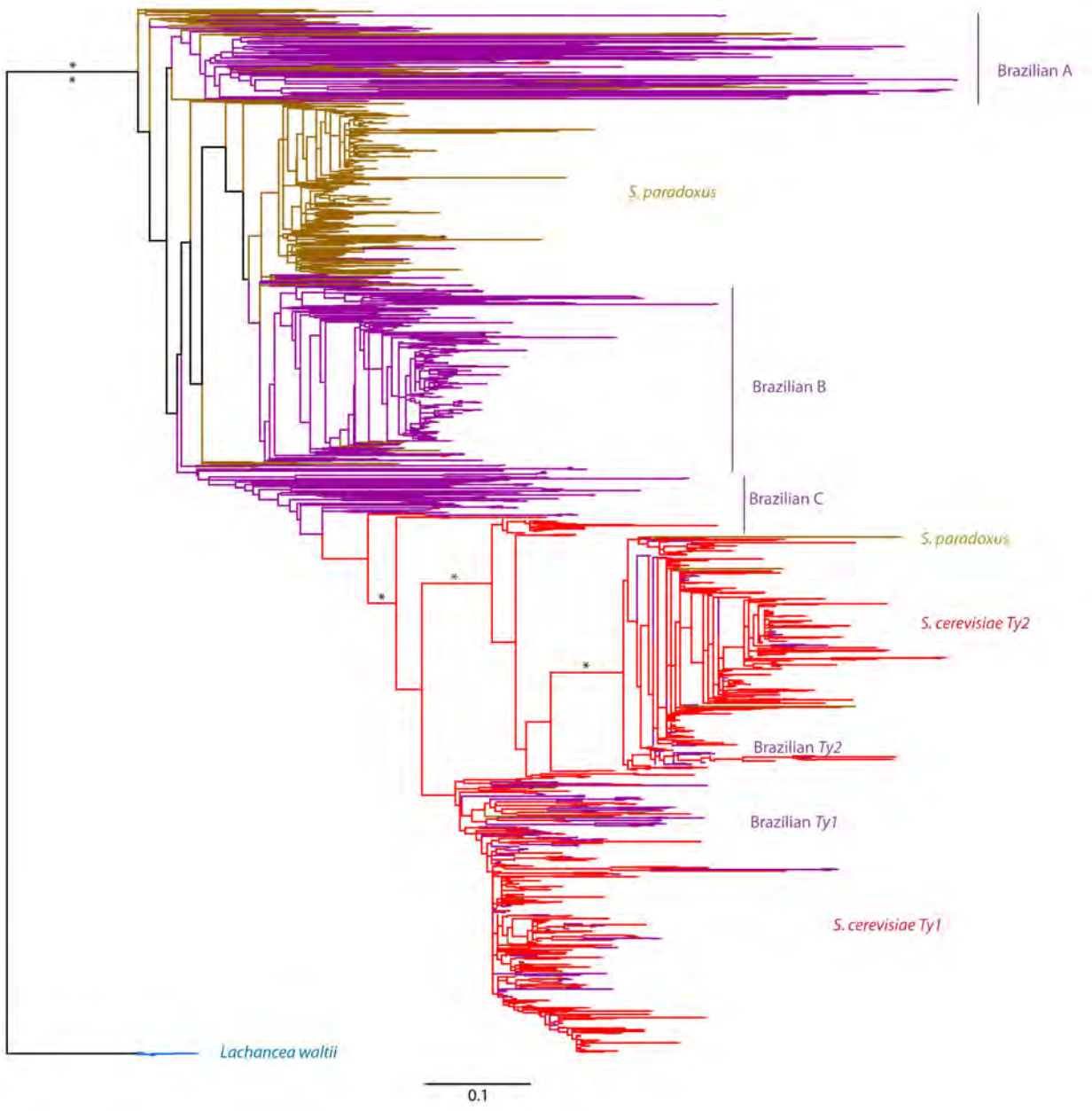

Figure 6.19: ***Ty1/2* LTR phylogeny of sequences from Brazilian *S. cerevisiae*.** Formatting is the same as in Figure 6.5, but based on an alignment of 333 sites and rooted with *Tlw2* sequences from *L. waltii*. The SGRP *S. cerevisiae* sequences falling within the Brazilian group C sequences, it is unlikely that this is the true relationship, and more likely a phylogenetic artefact resulting from the diversity seen in the long-branched sequences in group C of the Brazilian LTRs.

Unique *Ty2* LTRs (*n*=29) were collected, almost two thirds of which fall on short branches (*n*=19), consistent with recent activity or gain. Although degenerate *Ty2* coding regions were discovered in most Brazilian genomes, intact LTRs are present in less than half of the strains (*n*=11;

39%). A minority of LTRs (*n*=7; 24%) are associated with FLEs, but present in only five strains. *Ty2* is therefore extinct in 23 of the Brazilian strains, assuming the majority of Brazilian population has not been isolated from the source of *Ty2*. Far more *Ty1* sequences in the Brazilian strains are shared with SGRP *S. cerevisiae* strains (*n*=96), around a third of which are older, long-branched insertions (*n*=30).

6.4.2 Brazilian *Ty3* LTRs cluster only with parental species

Figure 6.20 displays the *Ty3* phylogeny of the Brazilian strains in relation to their parental species' sequences.

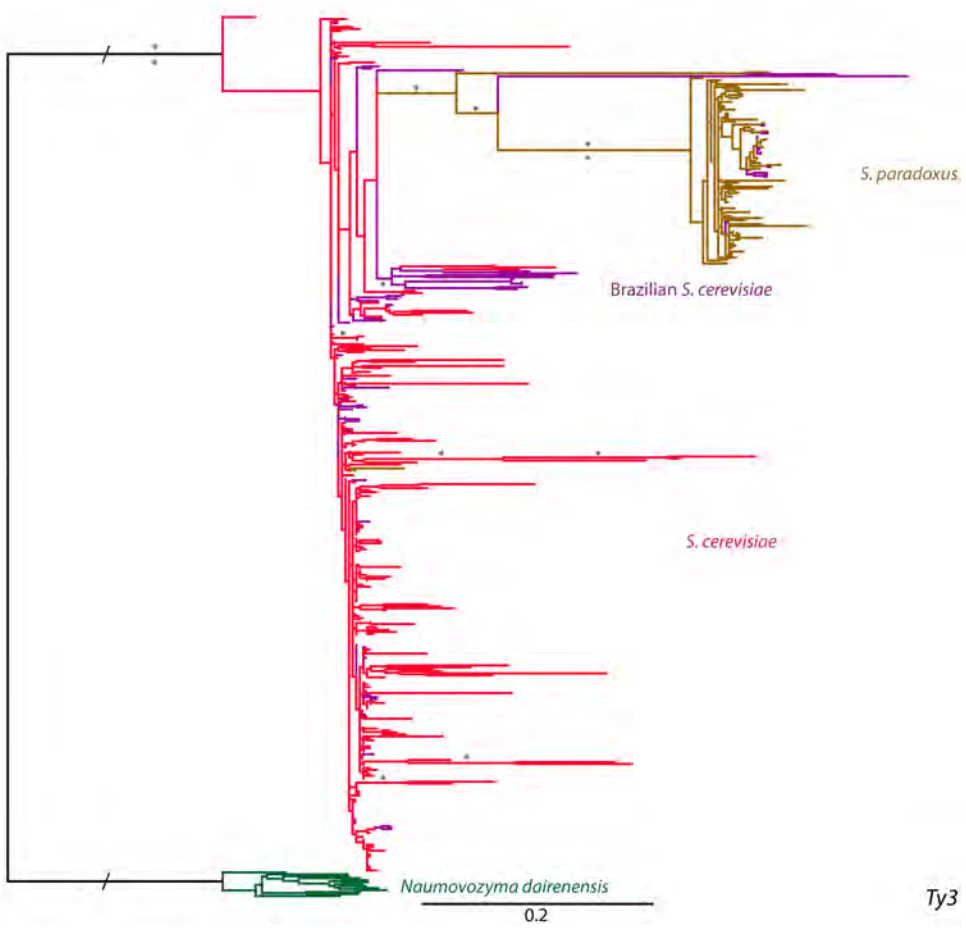

Figure 6.20: ***Ty3* LTR phylogeny of sequences from Brazilian *S. cerevisiae*.** Formatting is the same as in Figure 6.5, but based on an alignment of 353 sites and rooted with sequences from *N. dairenensis*. / indicates the root is arbitrarily shortened.

Although *S. paradoxus* sequences nest within those from *S. cerevisiae*, this is likely an artefact and as established in Chapter 5, a sister relationship is most likely the true connection. Only the root and the position of the *S. paradoxus* group are supported by both methods. Unlike in

the *Ty1/2* superfamily, sequences from the Brazilian strains do not form their own groupings but instead cluster with sequences of their parental species. The majority of sequences cluster with *S. cerevisiae* (*n*=78), whereas the remaining sequences fall with those of *S. paradoxus* (*n*=31). Furthermore, short-branched sequences (*n*=66) outweigh long-branched sequences (*n*=43) in the Brazilian strains, and also tend to cluster within groups of their parental species' sequences.

6.4.3 European and American *Ty4* inhabit the Brazilian population

Figure 6.21 displays the *Ty4* phylogeny of the Brazilian strains in relation to their parental species' sequences in the two main populations: American and European.

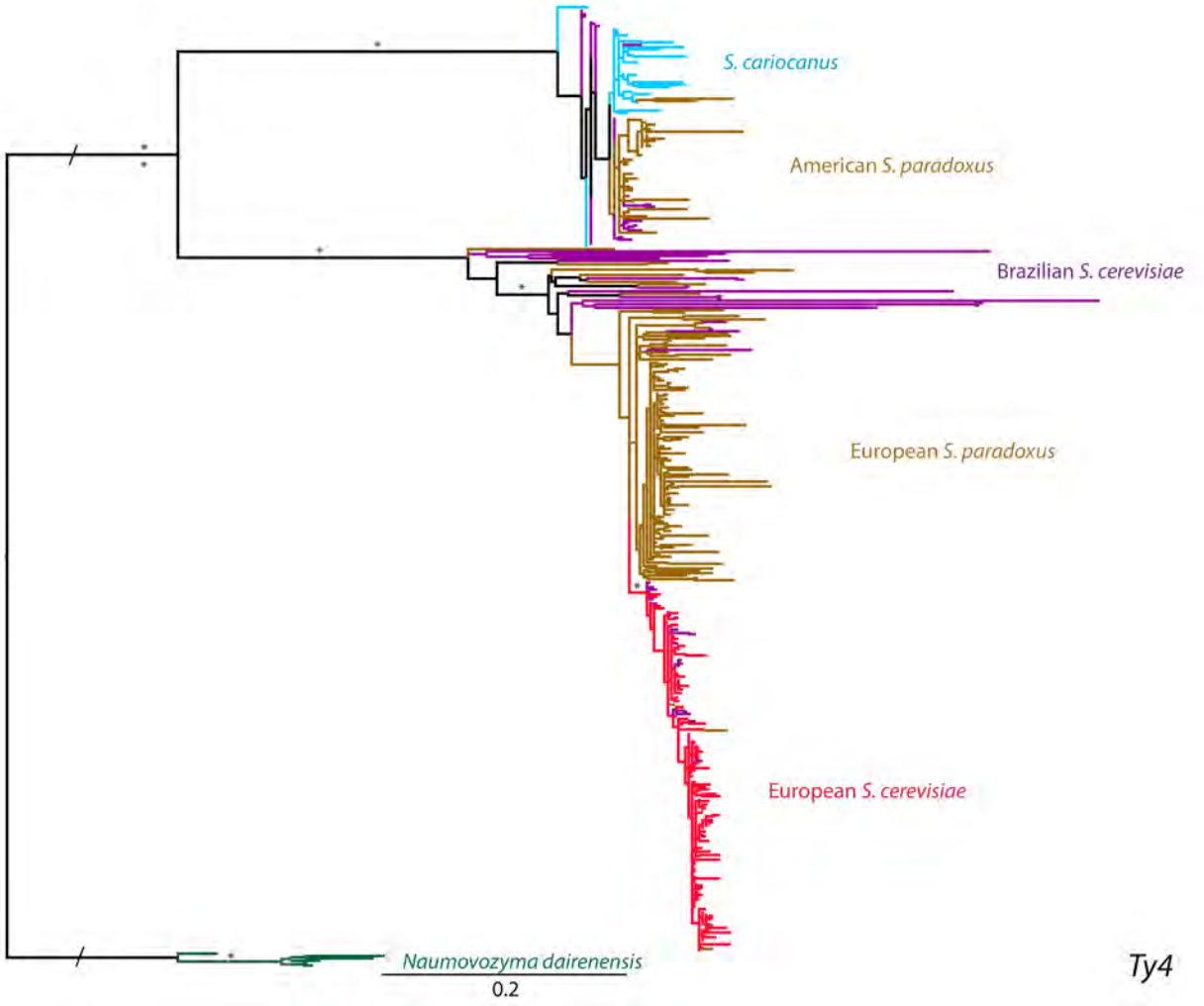

Figure 6.21: ***Ty4* LTR phylogeny of sequences from Brazilian *S. cerevisiae*.** Formatting is the same as in Figure 6.5, but based on an alignment of 383 sites and rooted with sequences from *N. dairenensis*. / indicates the root is arbitrarily shortened.

The split between sequences of the American and European populations is maximally supported by both methods. Long-branched Brazilian sequences mainly cluster with European *S.*

paradoxus LTRs (*n*=17), whereas short-branched sequences are prevalent within those of *S. cerevisiae* (*n*=29). Lower copy numbers of short-branched sequences are also present nested within American *S. paradoxus* (*n*=12) and *S. cariocanus* sequences (*n*=6). American sequences are confined to five Brazilian strains (section 6.3.7).

The long-branched sequences of *S. paradoxus* and Brazilian *S. cerevisiae* are most likely in this position due to diversity and age. Long-branched attraction, the phylogenetic artefact whereby rapidly mutating sequences are inferred as sharing a far closer relationship than in the true phylogeny (reviewed by Bergsten, 2005), and the fact that the population from which the Brazilian strains gained their *S. paradoxus* introgressions is yet to be discovered may also have influenced the phylogeny.

6.4.4 *Ty5* is extinct in the Brazilian population

Figure 6.22 on the following page displays the *Ty5* phylogeny of the Brazilian strains in relation to their parental species' sequences.

Both *Ty5* and *Ty5p* LTRs (*n*=44) are present in the Brazilian strains from each of their parental species, but only a small minority are associated with FLEs (*n*=5). As all coding regions are present as partial or non-autonomous elements, *Ty5* is therefore extinct in these strains. Long-branched sequences (*n*=5) nested within *S. paradoxus* suggested this particular *Ty5* lineage is far older than that of *S. cerevisiae*, as nucleotide changes have had time to accumulate. Sequences from the European strains of *S. paradoxus* (Figure 6.22) show evidence of independent activity, possibly since the introgression events resulting in the Brazilian strains, as none of these insertions are shared by the Brazilian strains.

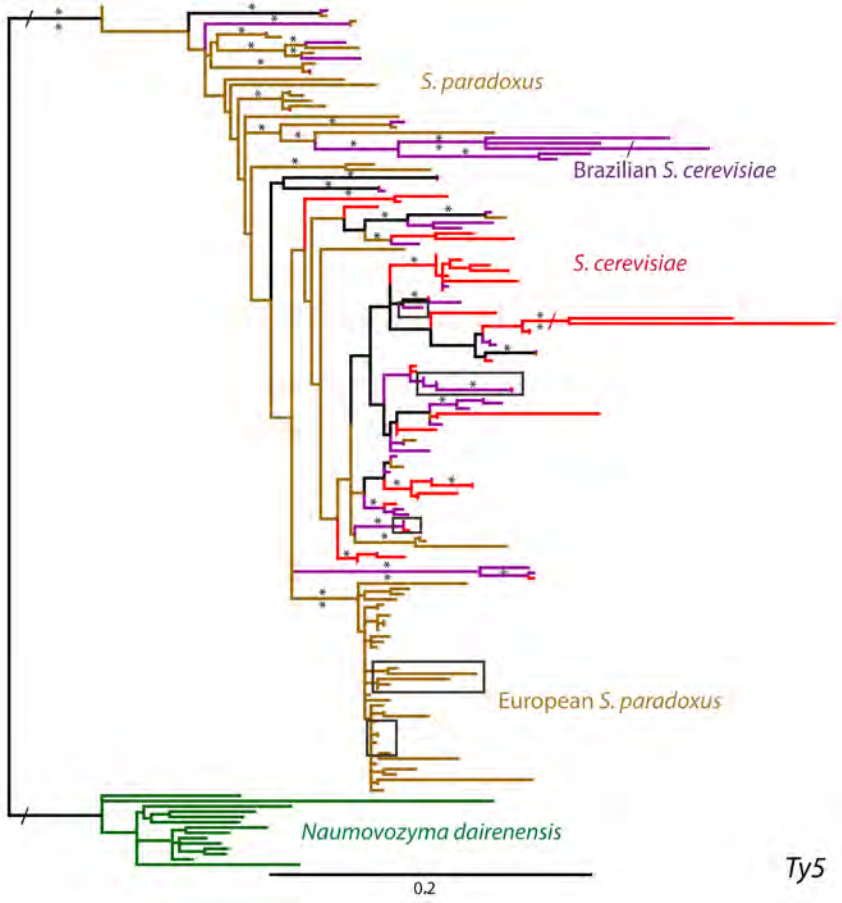

Figure 6.22: ***Ty5* LTR phylogeny of sequences from Brazilian *S. cerevisiae*.** Formatting is the same as in Figure 6.5, but based on an alignment of 251 sites and rooted with *N. dairenensis* sequences. / indicates the root is arbitrarily shortened. Boxed regions highlight the LTRs associated with FLEs.

6.5 Discussion

In this chapter, the *Ty* elements of two distinct populations of *Saccharomyces cerevisiae* were examined for common evolutionary characteristics and insertion profiles. The industrial Peterhof collection, distinct from common S288c-derived laboratory strains (Drozdova *et al.*, 2016), experienced minimal *Ty* activity in the time since this population was isolated from the SGRP strains. In contrast, the Brazilian population of wild strains which has experienced varying levels of hybridisation and introgression with *S. paradoxus* (Barbosa *et al.*, 2016), which allowed the spread of a *S. paradoxus*-like family, *Ty1p*, to propagate. Limited *Ty3-5* activity is observed in either population, and the introduction of a *Ty4* subfamily gained by the Brazilian strains adds to previously accumulated evidence (Chapters 4 and 5) that this subfamily may be confined to the American continent. Furthermore, *Ty2* is relatively unsuccessful in both populations, with little strain-specific activity and extinction in all PGC and a third of Brazilian strains. Since its acquisition from *S. mikatae* (Carr *et al.*, 2012), *Ty2* became widespread throughout *S. cerevisiae*, with evidence of this subfamily being reported in all sequenced strains to date (e.g. Ibeas and Jimenez, 1996; Dunn *et al.*, 2005; Liti *et al.*, 2005, 2009; Novo *et al.*, 2009; Akao *et al.*, 2011; Bleykasten-Grosshans *et al.*, 2013). Although the Peterhof population was clearly isolated in the time following the acquisition of *Ty2*, if the initial copy number in the ancestral population was low, the subfamily would have a poorer chance of survival due to the likelihood of recombination and negative selection acting upon the elements. The subfamily may have undergone a burst of activity after the HT event, but it was not enough to avoid extinction in the PGC population. Similarly in Brazilian *S. cerevisiae*, the introduction of a foreign subfamily such as *Ty1p* may also have introduced competition between the two (Abrusan and Krambeck, 2006; Le Rouzic and Capy, 2006). The presence of specifically *S. paradoxus* regions clearly did not cause the loss of *Ty2*, as Liti *et al.* (2005) noted *S. paradoxus* x *S. cerevisiae* hybrids containing the subfamily. However, as *Ty2* has yet to be found as a functional family in pure *S. paradoxus*, how well this would thrive in this species is unknown.

Widespread introgression from *S. paradoxus* into a Brazilian population of *S. cerevisiae*

Evidence of introgression from *S. paradoxus* in the shape of chimeric chromosomes is found in 23 non-hybrid strains to varying degrees, an increase upon the 21 strains identified by Barbosa *et al.*

(2016). This was likely a result of the differing identification methods employed, as the authors focussed on ORFs, whereas a whole genome mapping approach was used here.

Supporting the results of Barbosa *et al.* (2016) is the conclusion that introgression patterns are not fixed in the Brazilian population, and likely a result of multiple hybridisation and backcrossing events. In order to gain the elements observed in the strains, one hybridisation event with *S. paradoxus* occurred within the European populations in order to gain *Ty4*, as evidenced previously in Chapter 5. The migration to the Americas then occurred, where they underwent further hybridisation with American populations of *S. paradoxus* in order to gain *Ty1p* in all strains and the American *Ty4* in a minority of strains. This may have consisted of multiple hybridisation events and repeated backcrossing to cause the introgression patterns of the mosaic genomes as hypothesised by the authors. However, as seen in a study of *S. cerevisiae* and *S. uvarum* hybrids by Dunn *et al.* (2013), repeated backcrossing may not be required for introgression as the daughter cells can simply undergo independent recombination, losing regions of *S. uvarum* chromosomes to become genomically stable. The possibility that this occurred in the Brazilian strains in order to result in the complex patterns of introgressed *S. paradoxus* regions cannot be discounted. The elements and regions respectively gained by migratory strains Y255 and Y640 likely resulted from independent interactions with a *S. paradoxus* population (section 6.3.3).

Although there was no observed correlation between genomic proportion of introgression and TE content, the two hybrid strains possess the highest genome content, due to the presence of elements from the combined parentage and activity post-hybridisation. Liti *et al.* (2009) made a similar observation concerning strains with mosaic genomes caused by introgression. Additionally, a number of the introgressed regions and elements share identity with those of the *S. cariocanus* population, suggesting blurred lines between populations and species on this continent.

In this study, the number of assembled contigs, which is to an extent representative of frequency recombination breakpoints, shows a weak positive correlation with genomic TE content. In the ectopic exchange model, recombination rate is typically inversely correlated with insertion density (as opposed to genomic content), as selection is weaker in areas of low recombination, therefore allowing TEs to proliferate (Montgomery *et al.*, 1987; Langley *et al.*, 1988; Charlesworth *et al.*, 1992*a*,*b*; Dolgin and Charlesworth, 2008). Examined in Drosophila (Bartolome *et al.*, 2002; Rizzon *et al.*, 2002) and rice (Tian *et al.*, 2009), the negative correlation followed the model, whereas

no correlation was discovered in *Arabidopsis* (Wright *et al.*, 2003). In yeast, however, the relationship between insertions and recombination rate appeared to be more complicated. Although Ben-Aroya *et al.* (2004) reported a reduction in recombination due to full-length *Ty* insertions, Pan *et al.* (2011) noted that events resulting in recombination were not simple, and were more likely to occur between solo LTRs than regions containing FLEs. With this in mind, the apparent correlation between recombination rate and genomic TE content may also be explained in part by the range of complexity observed in the assembled genomes, in that the higher the insertion rate and potential rearrangements that occurred, the more difficult scaffolding was to perform. The number of contigs assembled for each genome, while indicative of recombination breakpoints due to the failure to build into scaffolds due to the differences in the reference genomes, may also simply be indicative of insertions that differ to that of the reference, and due to the highly repetitive nature of LTRs in particular, contigs were not able to be scaffolded further.

A new *Ty* family in Brazilian *S. cerevisiae*

The spread of *Ty1p* in the Brazilian strains of *S. cerevisiae* marks a successful transfer event into a new population from *S. paradoxus*. Although the regions of introgression are not fixed in the Brazilian population, all strains, including migratory strains Y255 and Y640, gained the new subfamily from *S. paradoxus*. This was likely gained by the ancestor population of the 26 non-migratory strains. The possibility that *Ty1p* was gained by Brazilian *S. cerevisiae* in multiple events cannot be discounted, as the introgressed regions contain evidence of the *Ty1p* family, usually as solo LTRs. If these regions are the source of the new family, recombination post-hybridisation with *S. paradoxus* caused these original elements to be lost in their new hosts, but not before they spread throughout the genomes. The possibility that the source for this family could have been a single copy gained by all strains also cannot be discounted, as the assembly process caused disruption of some solo insertions. It is however more likely that multiple elements were gained by the population(s) of Brazilian *S. cerevisiae* during hybridisation with *S. paradoxus*, as no insertion was found in common across all strains.

Full-length elements, solo LTRs, usually in high abundance, were discovered across the strains, together with low nucleotide diversity in LTRs, all of which indicated that this family was recently active since its acquisition, and perhaps continues to be so. LTR-LTR recombination caused the loss in some strains post-introgression, as the resulting solo LTRs are observed in the *S. cerevisiae*

regions of the genome. No unique activity in strains Y257, Y266 and Y463, all early branching strains in B1 of Figure 6.9 suggests a restriction of activity not observed in the other strains in this clade.

Activity of a new TE family introduced by hybridisation/introgression has previously been documented after the crossing of two *Drosophila* species (de Lucca Jr. *et al.*, 2007) and in plants, such as subspecies of rice (Liu and Wendel, 2000; Shen *et al.*, 2005), wheat (Liu *et al.*, 2015; Senerchia *et al.*, 2015) and sunflowers (Kawakami *et al.*, 2010). Activity of *Ty1p* may have been induced by the introgression events themselves, as seen in a variety of organisms including *Drosophila* (Guerreiro, 2014), rice (Han *et al.*, 2004), sugarcane (de Araujo *et al.*, 2005) and *Arabidopsis* (Josefsson *et al.*, 2006).

Phylogenetics of Peterhof and Brazilian populations of *S. cerevisiae*

All families share similar evolutionary genomics regardless of whether they originated in the Peterhof collection or Brazilian population, as evidenced by phylogenies for each family. Combining SGRP, Peterhof and Brazilian sequences caused the phylogenetic signals to be lost in LTR trees and provided very little insight into their evolutionary history unless smaller, representative datasets were created (data not shown). Therefore, splitting datasets into Peterhof and Brazilian allowed the clearest and most plausible topologies to be presented here.

When forming the phylogenies, it was hypothesised that LTR sequences from the Peterhof and Brazilian strains would primarily cluster with those of SGRP *S. cerevisiae*, which exist as a fair representation of insertions present within the species as a whole. However, in both populations, only a minority of insertions in each family are shared with the SGRP strains, and the majority, particularly in the *Ty1/2* superfamily, formed separate groupings. The Brazilian *Ty1/2* superfamily phylogeny illustrated the success with which *Ty1p* spread from donor *S. paradoxus* into the new *S. cerevisiae* population and has since evolved into a distinct subfamily alongside endogenous *S. paradoxus* subfamily.

In both populations, there is a prevalence of long-branched sequences within each familial phylogeny. Sequences in the PGC strains tended to assume long-branched groups in the basal position which is not observed for the Brazilian strains. They indicated ancient activity, often unique to the individual populations and not shared by the SGRP strains, potentially occurring post-isolation

of these two populations. Long-branched solo LTR sequences tend to become fixed in the populations as they are less likely to be deleterious and formed poor templates for ectopic recombination Hoang *et al.* (2010). With LTR phylogenies commonly displaying long-branch attraction (LBA) (Benachenhou *et al.*, 2013), there is a possibility that phylogenies may not reflect the true positions of the insertions due to the tendency for long branches to cluster together. Simultaneously, the clustering of short-branched insertions may have caused further alterations to the topologies. Short-branched insertions are indicative of relatively recent activity in most families. No evidence in the regions containing solo LTRs was discovered that might suggest duplication, which could also account for short-branched insertions, therefore it was concluded that recent transposition has occurred.

Population isolation and environmental effects on *Ty* families

The differences seen in activity levels and genomic TE contents within the two populations from SGRP strains may only be explained in part by their geographical differences, as the SGRP strains themselves were collected from a variety of locations. Population isolation occurred predominantly in the PGC strains, whereas those in Brazil are free to interact with other populations and species. This is starkly pronounced in their genomic TE content: without the repeated introduction of new *Ty* sources in the Peterhof population, their isolation caused the loss of most families through drift (Le Rouzic and Deceliere, 2005). A contributor to the poor success of families in the PGC strains may have been the stable environment in which the population is maintained, as various stressful conditions and environments have been reported to induce transposition (Lesage and Todeschini, 2005). In contrast, the Brazilian strains in their wild environment would be subjected to stressful conditions such as temperature changes, which would lead to increased transposition. Hosid *et al.* (2012) suggested that populations with increased TE activity are more likely to survive during environmental changes as there would be more genomic variety for natural selection to act upon.

Population history could be determined by *Ty* sequences: a future application

Interdelta sequencing is currently extensively used as a method of genotyping strains by amplifying randomly distributed regions between solo *Ty1/2* LTRs (Legras and Karst, 2003; Tristezza *et al.*, 2009; Franco-Duarte *et al.*, 2011; Xufre *et al.*, 2011; Sun, Guo, Liu and Liu, 2014; Sun *et al.*,

2017), but its reliability has recently been questioned (Pfliegler and Sipiczki, 2016). Identifying the presence of subfamilies may provide an alternative method in determining geographical origin, possible introgression/HT or at the very least provide a history of populations with which contact has occurred. A number of the Brazilian strains contain both types of *Ty4* and the *Ty1p* subfamily, the presence of which confirmed the interactions with multiple populations on multiple continents. Specifically designed PCR primers used to screen strains or species for e.g. both types of *Ty4*, potentially revealing migrations, interactions with overseas populations and inter-species hybridisation, thus identifying appealing candidates for genome sequencing and analysis. Additionally, LTR sequences could then be added to existing phylogenies in order to determine the possible relative age of the interaction(s).

6.6 Summary and conclusions

Two isolated populations of *S. cerevisiae* were investigated here for their *Ty* content. Consisting of a series of varyingly isolated populations throughout the world, *S. cerevisiae* is able to hybridise with other species and displays susceptibility to newly invading *Ty* families. The populations' contrasting environments promote low activity and extinction of families in the Peterhof collection, and new additions in the Brazilian strains through hybridisation and introgression.

Although the mosaic Peterhof strains were briefly exposed to a new source of elements during mating with S288c-derived strains, the insertions failed to thrive. There is little difference between those and the progenitor strains and they saw little change in their genomic landscapes from low levels of *Ty* activity. Confined to a stable environment, their elements are lost to genetic drift and LTR-LTR recombination.

The introduction of *Ty1p* elements into the Brazilian population of *S. cerevisiae* coincided with the adaptation to the climate by the acquisition of established *S. paradoxus* genomic regions. Few insertions are fixed and are unlikely to cause localised benefits, while others likely contribute to genomic diversity in the rearrangements TEs naturally promote. In all likelihood, *Ty1p* simply took advantage of a new population in which it could propagate.

Chapter 7

Discussion

The following is a brief discussion of the findings reported throughout this thesis, as more detailed discussions are included within each previous results chapter.

TEs are typically under negative selection in the genomes of their hosts due to deleterious effects (Charlesworth and Charlesworth, 1983; Biémont *et al.*, 1997; Charlesworth *et al.*, 1997). Hosts therefore employ successful methods of limiting TE effects and proliferation, such as RNAi, methylation and antisense RNAs (Kidwell and Lisch, 2000; Matsuda and Garfinkel, 2009). Despite this, TEs are able to persist in the genomes of most eukaryotic species. In this work, the genomes of more than 50 budding yeast species were investigated for evidence of TEs. All possess very different genomic TE contents, and few contain copy numbers or genomic fractions as high as *S. cerevisiae*. Some species, such as *Eremothecium* and *Ogataea*, are entirely devoid of elements. Due to the methods of identifying insertions, there is little doubt that these species - and perhaps many others - contain solo LTRs of families that cannot be located without coding sequences. Unlike genera such as *Drosophila* (McCullers and Steiniger, 2017) and most plants (Jiao *et al.*, 2017; Xia *et al.*, 2017; Stritt *et al.*, 2018), very few yeast contain large numbers of TEs, with the exceptions of *Candida albicans* (Goodwin and Poulter, 2000) and *Schizosaccharomyces japonicus* (Rhind *et al.*, 2011). These differences in genomic content suggests that the relationship between host and insertions is very much on a species-by-species basis. These yeasts also possess efficient TE control mechanisms, such as LTR-LTR recombination, necessary in small, compact genomes where methylation and RNAi are absent (Proffitt *et al.*, 1984; Wolfe *et al.*, 2015). Using the model species of budding yeast, the work presented here investigated two main ways in which TEs may avoid host defences and linger in genomes: by providing a benefit to their host and crossing species barriers via horizontal transfer of TE families.

The genomic population data of two yeast species, *Saccharomyces cerevisiae* and *S. paradoxus*, were examined for signatures of positive selection that may be acting upon insertions and

their adjacent genes. These insertions may provide benefits to the host genome, perhaps by altering expression of neighbouring genes. Candidate insertions for positive selection were found to be typically solo LTRs, and therefore no longer of use to the functionality of the *Ty* family. Insertions that may benefit their host's fitness are more likely to be older, having been well conserved and selected for, and present at high frequency or fixed across populations. Previous studies have reported the general benefits of *Ty* insertions (e.g. Wilke and Adams, 1992) or serendipitous, single findings (e.g. Paquin and Williamson, 1986; Brady *et al.*, 2008; Servant *et al.*, 2008), and often focus on full-length elements (FLEs) (e.g. Boeke and Sandmeyer, 1991; Roelants *et al.*, 1995, 1997; Lesage and Todeschini, 2005). Therefore, the work presented here is the first full genome screening of all insertions in *Saccharomyces*.

Of the candidates in *S. cerevisiae*, 12 were investigated for potentially significant differences in expression of adjacent genes in a stable laboratory environment. Candidate LTRs neighbouring *AVL9* and *SCS7* may cause significant differences in the expression of these host genes, therefore potentially improving host fitness. Although the reported findings suggest that positive selection may be acting on far more insertions than previously thought, they may only be representative of the *Ty* insertions of *Saccharomyces*. Hybridisation is relatively common in *Saccharomyces* (Peris *et al.*, 2017; reviewed by Morales and Dujon, 2012) and thought to be promoted by the stressful, industrial environments of many yeasts (Zeyl *et al.*, 1996; Matzke *et al.*, 1999; Tofalo *et al.*, 2013; Marti-Raga *et al.*, 2017). Hybridisation provides a simple mechanism whereby *Ty* elements and other genomic DNA can pass between species. Subsequent HT may not be as common in other species, particularly those which do not undergo asexual reproduction or have lower rates of recombination.

In comparison to unicellular species such as budding yeast, multicellular organisms employ germline separation which may protect them from horizontally transferred insertions being passed onto next generation, thereby reducing the apparent rate of HT occurrence. Effective population size may also play a role in the propagation of elements through HT, as species with higher effective population sizes are more efficient at removing TEs from the population (Charlesworth and Charlesworth, 1983; Brookfield and Badge, 1997; Groth and Blumenstiel, 2016). Furthermore, investigations into selection acting upon TEs in other species have revealed that insertions may have greater effects on a host when FLEs are under positive selection. The genomes of *Drosophila* species - in contrast to those of *Saccharomyces* - contain far greater numbers of FLEs

(McCullers and Steiniger, 2017). Reports of TEs resulting in benefits such as insecticide and viral resistance in *Drosophila* (Daborn *et al.*, 2002; Aminetzach *et al.*, 2005; Magwire *et al.*, 2011; Mateo *et al.*, 2014) are the result of autonomous full-length retrotransposons and DNA transposons, which may account for the more subtle effects observed here in *Saccharomyces*. Species within predator/prey, parasitic/host or symbiotic relationships may also experience atypical occurrences of HT, as *Drosophila* may undergo HT with its symbiotic bacteria *Wolbachia* (Brown and Lloyd, 2015; Ortiz *et al.*, 2015).

Reconstruction of the possible evolutionary relationships between the families of TE sequences in all available budding yeast genomes allowed the identification of potential horizontal transfer of families between species. Species phylogenies are typically created using the sequences of multiple highly conserved, slowly evolving host genes such as ribosomal 18S, 26S and internal transcribed spacers. These contrast with the phylogenetic analysis of TE sequences, which are expected to present long branches, reflecting their relatively fast evolution. Therefore, the short-branched *Ty* phylogenies reported here, where they share little congruence with those of their host species, strongly indicate transfer between populations and species. Short-branched sequences can also result from gene conversion, where branch lengths are indicative of the last conversion event rather than transposition. However, gene conversion occurs at relatively low frequency (Kupiec and Petes, 1988*b*), and the presence of ancient, long-branched insertions in all families would suggest that conversion is not particularly active in the *Ty* sequences of these species.

In order to recover maximum evidence of HT occurrence, both LTR and RT phylogenies were constructed. Although phylogenetic analyses of RH (Malik and Eickbush, 2001), IN (Malik and Eickbush, 1999); (Llorens and Marin, 2001) and PR (Llorens *et al.*, 2009) domains are congruent with those of the most conserved domain of RT (Eickbush and Jamburuthugoda, 2008), unresolved relationships may be determined in the future by increasing the number of domains included in phylogenies. TE trees are renowned for incongruence between LTR and internal coding regions and their poor support values, regardless of inference method (e.g. González *et al.*, 2008; Benachenhou *et al.*, 2009, 2013).

Despite these complications, the potential evolutionary histories of all *Ty* families presented here allowed for identification of in excess of 75 HT events in 19 species, with around half of these successful in the further propagation in recipient genomes. Furthermore, *Ty1p* of *S. paradoxus* spread throughout a Brazilian population of *S. cerevisiae* (Chapter 6). Although transfers of TE

families across greater phylogenetic distances have been reported (e.g.Tang *et al.*, 2015; Lin *et al.*, 2016; Peccoud *et al.*, 2017; Gao *et al.*, 2018), the events documented here concern only yeasts, suggesting a confinement of *Ty*-like elements to this particular species group. Evidence of HT was discovered in all *Ty* families, with the high rate of *Ty4* being the most surprising, having previously been assumed to be transcriptionally active at a very low level (Hug and Feldmann, 1996).

The ease with which families can be transferred across species is very much of interest, as the artificial transfer of *LINE* elements from *Candida albicans* into *S. cerevisiae* (Dong *et al.*, 2009) shows that 'foreign' elements are able to survive - and proliferate - in other species of budding yeast. However, as *Saccharomyces* do not typically possess DNA transposons and non-LTR-retrotransposons, it is possible that many cellular requirements for their activity have not survived in extant genomes. Furthermore, as hybridisation and introgression appear to be the main methods by which new TE families are gained in yeast, the types of elements gained are limited by the species which can undergo hybridisation. For example, the *S. cerevisiae* strain AWRI1631 contains degenerate copies of *Rover* DNA transposons (Borneman *et al.*, 2008), likely the result of hybridisation(s) with another yeast - perhaps of the *Lachancea* genus (Sarilar *et al.*, 2015) - which possesses this family.

During analysis, a further method of persistence was uncovered: sequence divergence of a TE family in differing populations, which has yet to be reported in *Saccharomyces* (*Ty4*; Chapters 4 and 5). The impact of insertions driving population divergence is documented in a variety of organisms (e.g. Begin and Schoen, 2007; Senerchia *et al.*, 2015; Oppold *et al.*, 2017), while divergence typically coincides with major evolutionary events such as speciation, as seen in the differing elements of *Saccharomyces* species (Chapter 4, Liti *et al.*, 2005) and *Drosophila*, such as the divergence of *roo* and *rooA* (de la Chaux and Wagner, 2009). It may be a method by which insertions are able escape deletion via inter-element recombination (Charlesworth and Langley, 1989; Petrov *et al.*, 2003) and small RNA targeting in eukaryotes (reviewed in Castillo and Moyle, 2012). A review of current literature revealed only two instances of divergence of TEs with isolation of host populations. *Arabidopsis* (Lockton *et al.*, 2008; Lockton and Gaut, 2010) and Mediterranean grass (Stritt *et al.*, 2018) show divergence in TE families, where polymorphisms are specific to populations. It is reasonable to think that such divergences can and do occur in the elements of other species, yet as the discovery and annotation of TEs is often performed as an entirely automated process, such divergences may be classified by the software as entirely different families. Without

manual investigations into the TE contents of genomes - particularly those of non-model species - to complement automated TE pipelines, many interesting and unique evolutionary histories may be misinterpreted or missed altogether.

The taxonomy of closely related species such as budding yeast has been a difficult and often controversial subject, particularly before NGS. Investigations into genomic TE content here led to the discovery of two wrongly designated species: *Vanderwaltozyma yarrowii* and *Naumovozyma dairenensis*. These were in fact strains of *Lachancea waltii* and *Candida albicans*, respectively, the latter of which was very recently independently reported by Stavrou *et al.* (2018). Furthermore, the accuracy of repeat rich regions of genomes is often questionable due to sequencing methods (Treangen and Salzberg, 2011; Hoban *et al.*, 2016). Genomes can be submitted to GenBank with their repetitive regions masked by programs such as RepeatMasker, or assembled in a fashion that severely under-represents these regions, thereby preventing the true TE content being established and hindering investigations into TEs and their host genomes (Choudhury *et al.*, 2017). The increasing availability of single-molecule techniques such as PacBio and Oxford Nanopore, vast improvements on previous Illumina-style sequencing methods, will hopefully improve studies of TEs (Feng *et al.*, 2015; Rhoads and Au, 2015; Salazar *et al.*, 2017). However, even with improved sequencing techniques, automated TE discovery and annotation methods should not be completely relied upon to accurately examine genomes, as seen here with *Sz. japonicus* (Rhind *et al.*, 2011) and species of *Eremothecium* (Dietrich *et al.*, 2004, 2013; Wendland and Walther, 2011).

The investigations reported here represent the first genomic screening of *Ty* insertions in *Saccharomyces* for signatures of positive selection, and an updated, comprehensive search for evidence of HT between species of budding yeast, which both may act as methods for TE families to persist in the genomes of their hosts. The results documented for these species are unlikely to characterise the relationship between host and TEs and rate of HT events in and between other species. Therefore, the work presented in this thesis will hopefully serve as inspiration for similar analyses of TEs in other species.

Appendix A

List of software used

The following is a list of software used throughout this work.

Software	Reference/link
UCSC *S. cerevisiae* genome browser	Link
UCSC *S. paradoxus* genome browser	Link
Wellcome Trust Sanger Institute *S. cerevisiae* BLAST	Link
Wellcome Trust Sanger Institute *S. paradoxus* BLAST	Link
NCBI BLAST	Altschul *et al.* (1990)
SGD *S. cerevisiae* WU-BLAST2	Link
SGD fungal genome WU-BLAST2	Link
SGD GO-Slim mapper	Link
PANTHER Classification System	Link
MAFFT v.7.205	Katoh and Standley (2013)
revseq	Link
Primer3	Untergasser *et al.* (2012)
CFX Manager	Link
qbase+	Link
GraphPad Prism v4.0	Link
CIPRES Science Gateway	Miller *et al.* (2013)
SIB ExPASy Translate	Link
UniProt	Link
FASTA sequence length sorter	Blankenberg *et al.* (2010)
Sequence extractor	Blankenberg *et al.* (2010)
DnaSP v.5.10	Rozas and Rozas (1999)

TOPALi v.2	Milne *et al.* (2009)
SplitsTree4 v.4.14.4	Huson and Bryant (2006)
ClustalX v.2.1	Larkin *et al.* (2007)
FigTree v.1.4.0	Link
RepeatMasker v.4.0.7	Smit *et al.* (2013)
RepBase v.20150807	Bao *et al.* (2015)
bwa c.0.6	Li and Durbin (2009)
SAMtools v1.3.1	Li *et al.* (2009a)
MeDuSa v.1.3	Bosi *et al.* (2015)
SPAdes v.3.6.2	Bankevich *et al.* (2012)
Mauve v.2.4.0	Darling *et al.* (2010)
SyMAP v.3.4	Soderlund *et al.* (2011)
RAxML v.8	Stamatakis (2014)
UGENE UniPro v.1.29	Okonechnikov *et al.* (2012)
Qualimap v.2	Okonechnikov *et al.* (2016)
Adobe Photoshop CS6 v.13	Link

Appendix B

Bayesian Inference and Maximum Likelihood parameters

Bayesian Inferences (Yang and Rannala, 1997) parameters used with MrBayes v.3.2.6 (Huelsenbeck and Crandall, 1997) on XSEDE, available on the CIPRES Science Gateway v.3.3 (Miller *et al.*, 2013). All other parameters were set to default.

Maximum hours	72
Data type	nt/AA (changed as required)
BEAGLE	disabled for nt phylogenies
Nst	6
Nt substitution model	4 x 4
Nt site variation	gamma
Protein model	Mixed or RT - allow MrBayes to select
MCMC generations	500,000
MCMC sample time	1000
MCMC burn in	0.25
Stop if convergence falls below stop	No
Sumt burn in	1250
Sumt consensus tree	All compatible groups
Show sumt probabilities	Yes
Sump burn in	1250

The following priors were determined by MrBayes during the initial configuration part of analysis.

transition/transversion ratio substitution rate coalescence parameter	DNA only
synonymous/nonsynonymous ratio	
frequencies of sites under purifying, neutral and positive selection	
state frequencies	
proportion of invariable sites	
autocorrelation for gamma distribution	
covarion switching rate	
stationary frequencies	
branch length probability distribution	

Maximum Likelihood (Huelsenbeck and Crandall, 1997) parameters used with RAxMLGUI v.8 (Stamatakis, 2014).

Bootstrap	ML+thorough
Runs	100
Repeats	1000
Model	GTRCAT (nt) PROTCAT (AA)
Matrix	RTREV (AA)

Appendix C

Candidate LTR presence/absence matrix

On the following page is a presence/absence matrix for candidate LTRs in the SGRP strains of *Saccharomyces cerevisiae*.

NCYC plate position	Strain	Family LTR name Adjacent gene Accession no. Use	*Ty1/2* YLRCdelta9 *ADY4* 1922305461	*Ty3* YDRCsigma3 *AMD2* 1750380219	*Ty3* YLRWsigma2 *AVL9* 1903887229	*Ty1/2* YMLWdelta4 *CAT2* 1892466728	*Ty1/2* YNRCdelta8 *HOL1* 1253419984	*Ty1/2* YORWdelta17 *MPC54* 423190716	*Ty1/2* YMLWdelta4 *RRN11* 1892466728	*Ty4* YMRCtau3 *SCS7* 423291242	*Ty1/2* YDRWdelta11 *SEC7* 1253549221	*Ty4* YERCtau2 *UBP9* 422853302	*Ty3* XII-681209 *YCS4* 1253558475	*Ty1/2* YERCdelta16 *MPD1* 1253554310
A1	DBVPG6765	wild	Present	Absent	Present	Present	Present	Absent	Present	Present	Present	Present	Present	Present
A2	BC187	fermentation	Present	Present	Absent	Absent	Absent	Present	Absent	Present	Absent	Absent	Absent	Present
A3	NCYC361	beer	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Absent	Absent	Absent
A4	Y9	sake	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent
A5	YJM981	clinical	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent
B1	SK1	laboratory	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Absent
B2	YPS606	wild	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent
B3	K11	sake	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent
B4	UWOPS03-461-4	wild	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent
B5	YJM975	clinical	Absent	Present	Present	Absent	Absent	Present	Absent	Absent	Present	Absent	Absent	Absent
C1	Y55	laboratory	Present	Present	Absent	Present	Absent	Absent	Present	Absent	Absent	Absent	Absent	Present
C2	L_1374	wine	Absent	Absent	Present	Absent	Absent	Absent	Absent	Present	Absent	Absent	Present	Absent
C3	YS4	baking	Absent	Absent	Present	Absent	Present	Absent	Absent	Absent	Present	Absent	Present	Absent
C4	UWOPS05_217_3	wild	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent
C5	NCYC110	beer	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent
D1	YPS128	wild	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent
D2	L-1528	wine	Absent	Present	Absent	Present	Present	Absent	Present	Absent	Present	Present	Absent	Present
D3	YS9	baking	Absent	Present	Present	Absent	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent
D4	S288c	laboratory	Absent	Absent	Absent	Absent	Present	Present	Absent	Present	Present	Absent	Absent	Present
D5	YS2	baking	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent
E1	DBVPG6044	wine	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent
E2	Y12	wine	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent
E3	322134S	clinical	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent
E4	W303	laboratory	Present	Absent	Present	Absent	Absent	Present	Absent	Absent	Present	Absent	Absent	Present
F1	DBVPG1788	wild	Present	Present	Present	Present	Absent	Absent	Present	Absent	Present	Present	Absent	Present
F2	DBVPG1106	wild	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent
F3	378604X	clinical	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Absent
F4	UWOPS05_227_2	wild	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent
G1	DBVPG1373	wild	Present	Absent	Present	Present	Absent	Present	Present	Present	Absent	Present	Absent	Absent
G2	UWOPS83_787_3	wild	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent
G3	273614N	clinical	Absent	Present	Absent	Absent	Absent	Present	Absent	Present	Absent	Absent	Present	Absent
G4	DBVPG6040	wild	Absent	Present	Present	Absent	Absent	Absent	Absent	Absent	Present	Absent	Absent	Absent
H1	DBVPG1853	beer	Absent	Present	Absent	Absent	Absent	Absent	Absent	Present	Present	Present	Absent	Absent
H2	UWOPS87_2421	wild	Present	Absent	Absent	Present	Absent	Absent	Present	Absent	Absent	Absent	Absent	Absent
H3	YJM978	clinical	Absent	Present	Present	Absent	Present	Present	Absent	Absent	Present	Absent	Present	Absent
H4	YIIc17-E5	wine	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Absent

Table C.1: **Candidate LTR presence/absence matrix.** Strain DBVPG6765 failed to culture and was therefore discounted from qPCR.

Appendix D

RNA extraction quality

The following table is supporting data for Chapter 3.

Strain code	Concentration (ng/µl)	A260 (Abs)	A280 (Abs)	260/280	260/230
A2	273	6.83	3.23	2.12	2.12
A3	185	4.64	2.17	2.14	2.17
A4	251	6.27	2.94	2.13	1.95
A5	143	3.56	1.72	2.08	1.99
B1	360	9.01	4.20	2.14	2.10
B2	258	6.46	3.23	2.00	1.32
B3	308	7.70	3.59	2.14	2.07
B4	204	5.11	2.48	2.06	1.71
B5	177	4.42	2.14	2.07	1.44
C1	163	4.07	2.05	1.99	1.35
C2	246	6.15	2.89	2.13	2.22
C3	202	5.04	2.40	2.10	2.09
C4	217	5.42	2.58	2.10	2.11
C5	106	2.64	1.28	2.06	1.79
D1	140	3.50	1.88	1.86	0.97
D2	271	6.77	3.16	2.14	2.14
D3	212	5.31	2.52	2.11	2.03
D4	141	3.52	1.70	2.07	1.80
D5	316	7.89	4.52	1.74	2.04
E1	148	3.70	1.77	2.09	1.95
E2	199	4.97	2.36	2.10	2.13
E3	215	5.37	2.56	2.09	1.93
E4	289	7.22	3.41	2.11	2.01
F1	150	3.75	1.80	2.09	1.96
F2	447	11.16	5.67	1.97	1.54
F3	437	10.92	5.20	2.10	2.04
F4	391	9.78	4.70	2.08	1.82
G1	102	2.55	1.26	2.03	1.48
G2	204	5.11	2.42	2.11	2.10
G3	207	5.18	2.43	2.13	2.06
G4	133	3.33	1.61	2.07	1.70
H1	433	10.82	5.11	2.12	2.05
H2	305	7.62	3.66	2.08	1.67
H3	127	3.18	1.54	2.07	1.76
H4	156	3.89	1.89	2.06	1.79

Table D.1: **RNA extraction qualities.** RNA was extracted from cultured *S. cerevisiae* SGRP strains using the method of RNASwift (Nwokeoji *et al.*, 2016). The quality and concentration of genomic RNA was evaluated with the NanoDrop system (Fisher Scientific).

Appendix E

qPCR primer details

No.	Primer name	Gene	Direction	Type	Sequence	Volume H_2O added (µl)*	Tm	GC content (%)
1	refALG9-qPCR-F	ALG9	F	Housekeeping	TAAGGGAATGAATAACAAG	324	48.0	31.6
2	refALG9-qPCR-R		R		TCAGATGTAGAGGGTTGA	292	51.4	44.4
3	refTAF10-qPCR-F	TAF10	F	Housekeeping	GCGGTATCTAATGCTAAC	420	51.4	44.4
4	refTAF10-qPCR-R		R		CTTGCTGTAGTCTTCTCATT	450	53.2	40
5	refUCB6-qPCR-F	UCB6	F	Housekeeping	GATTACCACCCTGATACTT	688	52.4	42.1
6	refUCB6-qPCR-R		R		ATGTCTTCTTCTGATGGTC	682	52.4	42.1
7	ADY4-qPCR-F	ADY4	F	GOI	TTGGTAGGCAATAATCTAA	534	48.0	31.6
8	ADY4-qPCR-R		R		CTCAATAACATCTACCTCGT	542	53.2	40
9	AMD2-qPCR-F	AMD2	F	GOI	CATTGATTGTCAGTTTTCT	716	48.0	31.6
10	AMD2-qPCR-R		R		TTGTGTATCTGTAGCCATC	682	52.4	42.1
11	AVL9-qPCR-F	AVL9	F	GOI	GGTCACTAAAGATAAGGATG	414	53.2	40
12	AVL9-qPCR-R		R		TAAGTTTGTCCTTGATTTG	590	48.0	31.6
13	CAT2-qPCR-F	CAT2	F	GOI	TGCTAAAATCTAATGATGAC	484	49.1	30
14	CAT2-qPCR-R		R		AGATAGTTGAGATGTGGAGA	336	53.2	40
15	HOL1-qPCR-F	HOL1	F	GOI	CATTAGGTCTCTTTGGTG	550	51.4	44.4
16	HOL1-qPCR-R		R		AGCACAGGACTCACTAATAC	476	55.3	45
17	MPC54-qPCR-F	MPC54	F	GOI	AAGAGTTTATGGAACAAAAG	422	49.1	30
18	MPC54-qPCR-R		R		AGTGTTTTCATACTGCTTTC	642	51.1	35
19	RRN11-qPCR-F	RRN11	F	GOI	ACAAGGAAGATAGTGATGAG	288	53.2	40
20	RRN11-qPCR-R		R		GTTTATTTTGTGAGGACACT	574	51.1	35
21	SCS7-qPCR-F	SCS7	F	GOI	TATTTGATTGGTTACTTGG	500	48.0	31.6
22	SCS7-qPCR-R		R		CAAAAGTAGTGGAGTCAAA	358	50.2	36.8
23	SEC7-qPCR-F	SEC7	F	GOI	CAGAAACAAAAGAAACAAG	170	48.0	31.6
24	SEC7-qPCR-R		R		CTTCATCTTCATCTTCATCT	556	51.1	35
25	UBP9-qPCR-F	UBP9	F	GOI	CAAATGTTATGGTTTACAAG	512	49.1	30
26	UBP9-qPCR-R		R		TCCTCTGAATACTTGAATCT	624	51.1	35
27	YCS4-qPCR-F	YC54	F	GOI	AGAGAAAAGACAATGATGAC	254	51.1	35
28	YCS4-qPCR-R		R		ATTGACTGAGAGAAACAACT	320	51.1	35
29	YER134C-qPCR-F	YERC134	F	GOI	ACAAAGAAGTGGAGAAATAC	358	51.1	35
30	YER134C-qPCR-R		R		TCCACTCAGGTAATCTTG	506	51.4	44.4

Table E.1: **Primer details used in qPCR for Chapter 3.** *Volume of H_2O added for primer concentration of 50µmol/µl.

Appendix F

qPCR plate layout

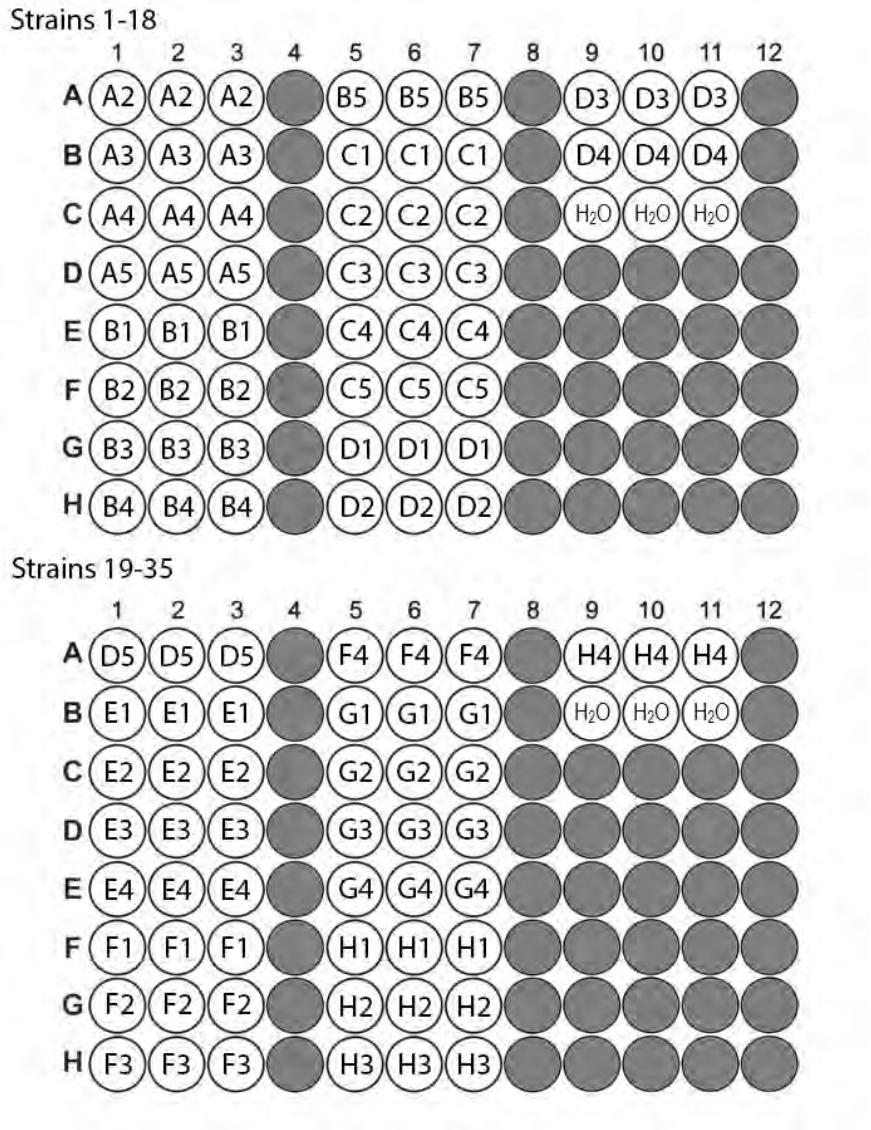

Figure F.1: **qPCR plate layout** used for qPCR in Chapter 3. Strain codes correspond to those used in the presence/absence matrix of Appendix C.

Appendix G

Tajima's *D* results for all insertions in *S. cerevisiae*

The following tables are supporting results for Chapter 3, which cover Tajima's *D* tests completed for all insertions of all families in *S. cerevisiae*.

In all tables, N/A indicates a lack of polymorphisms which prevented the test from being completed. Those insertions with significant results are listed in Results Chapter 3.

Ty1/2 superfamily

Sequence	No. of strains	Tajima's *D*	*P* value	π value	θ value (per site)
423095111	14	-1.24392	>0.10	0.01359	0.01935
423274700	30	-1.37590	>0.10	0.00425	0.00751
423284107	5	-0.68448	>0.10	0.02229	0.02458
452457071	13	-1.46463	>0.10	0.00416	0.00677
1253241397	15	-1.27331	>0.10	0.02262	0.03232
1253243015	6	-1.24578	>0.10	0.01370	0.01720
1253289677	10	-1.64861	0.05-0.10	0.01984	0.03027
1253319819	4	-0.78012	>0.10	0.00606	0.00661
1253343310	5	N/A	N/A	N/A	N/A
1253416948	6	-1.33942	>0.10	0.01998	0.02545
1253434488	12	-1.29119	>0.10	0.00323	0.00497
1253460430	15	0.67276	>0.10	0.01087	0.00924
1253465121	7	0.27225	>0.10	0.01558	0.01484
1253467398	17	-1.33013	>0.10	0.02170	0.03207
1253537608	4	N/A	N/A	N/A	N/A
1253538259	20	-1.33679	>0.10	0.01332	0.02038
1253555414	29	-1.61533	0.05-0.10	0.00763	0.01428
1253555750	9	-1.19734	>0.10	0.01515	0.02007
1253585482	6	-0.67613	>0.10	0.00463	0.00529
1253601153	4	-0.84729	>0.10	0.02252	0.02457
1253612939	13	0.29407	>0.10	0.02146	0.02008
1253635781	5	1.11678	>0.10	0.02840	0.02465
1253651207	23	-0.64597	>0.10	0.01393	0.01688
1253790050	4	N/A	N/A	N/A	N/A
1253791920	5	0.31504	>0.10	0 .08605	0.08261
1253829756	10	0.17997	>0.10	0.03755	0.03620
1253848883	11	-0.51784	>0.10	0.01278	0.01448
1750365951	12	N/A	N/A	N/A	N/A
1750371567	20	-1.34880	>0.10	0.01325	0.02038

1750546426	25	0.45318	>0.10	0.01308	0.01158
1750589650	17	-0.97450	>0.10	0.01136	0.01515
1750592606	4	-0.75445	>0.10	0.00445	0.00486
1750603848	10	0.81980	>0.10	0.00141	0.00106
1750813820	13	0.52008	>0.10	0.01538	0.01367
1750817734	4	-0.70990	>0.10	0.00300	0.00328
1750842981	12	-1.12253	>0.10	0.00346	0.00497
1893510956	15	-0.77356	>0.10	0.00280	0.00369
1906203936	5	-0.41017	>0.10	0.00541	0.00577
1945049940	4	-0.81734	>0.10	0.01064	0.01161
1945050283	4	-0.75445	>0.10	0.00452	0.00493
1948757264	5	-0.81650	>0.10	0.00118	0.00142
1975691073	10	-0.76222	>0.10	0.02242	0.02670
1975719999	5	-0.17475	>0.10	0.00420	0.00432
1994531502	15	-1.00042	>0.10	0.00968	0.01292
2017501236	9	0.04541	>0.10	0.02138	0.02118
2063021721	4	-0.38921	>0.10	0.01104	0.01150
2063028992	4	0.16766	>0.10	0.00502	0.00493
2063030395	6	0.02803	>0.10	0.01325	0.01319
2064339580	7	-1.51388	0.05-0.10	0.03715	0.05041
2064463329	7	-1.00623	>0.10	0.00086	0.00123
2064890711	9	-1.53917	>0.10	0.01537	0.02237
2068868355	4	-0.75445	>0.10	0.00455	0.00496
2068928477	26	-0.95463	>0.10	0.01689	0.02275
2070746256	8	-1.45938	>0.10	0.01017	0.01433
2293172043	19	-0.98513	>0.10	0.01551	0.02075
2293267940	6	1.12414	>0.10	0.00482	0.00396
1253601963	18	-0.98092	>0.10	0.00765	0.01041
2017445866	8	-0.26382	>0.10	0.03044	0.03204
1975692398	8	-0.28622	>0.10	0.03022	0.03195
1750949753	7	-1.23716	>0.10	0.00174	0.00248

1963750385	19	-1.33865	>0.10	0.01541	0.02334
1752141171	21	-1.45758	>0.10	0.01513	0.02428
2293171133	7	1.10686	>0.10	0.00460	0.00370
1945055670	12	-0.42854	>0.10	0.00260	0.00299
1904702417	9	-0.46495	>0.10	0.02157	0.02381
1948027081	7	-1.55311	0.05-0.10	0.00606	0.00866
1253298553	21	-1.25552	>0.10	0.01519	0.02254
1901770950	8	-0.27692	>0.10	0.02428	0.02563
2017463375	15	-0.59679	>0.10	0.03459	0.04020
1253860549	5	-0.74682	>0.10	0.00912	0.01021
1922304422	4	N/A	N/A	N/A	N/A
1750390760	18	-0.93985	>0.10	0.01602	0.02102
1253888926	14	-0.48484	>0.10	0.00745	0.00852
1253788471	12	-0.90201	>0.10	0.00298	0.00400
1253305308	11	-1.34124	>0.10	0.00473	0.00707
2064877131	8	-1.45938	>0.10	0.01017	0.01433
1253546086	5	-1.04849	>0.10	0.00361	0.00434
1948027290	18	1.32989	>0.10	0.02004	0.01489
1253642065	9	N/A	N/A	N/A	N/A
1253895868	16	0.63878	>0.10	0.01380	0.01184
1750367443	8	-1.23716	>0.10	0.00176	0.00251
2017465203	8	-1.64262	0.05-0.10	0.02310	0.03349
1750549190	10	-0.79994	>0.10	0.01767	0.02129
1751352537	19	1.17988	>0.10	0.02148	0.01642
1253841100	10	-0.61494	>0.10	0.00923	0.01071
1253643613	16	-1.12617	>0.10	0.02319	0.03177
1906197045	21	-1.17904	>0.10	0.00376	0.00591
1253626575	16	-1.51429	>0.10	0.01139	0.01826
1904694409	23	-0.83369	>0.10	0.00622	0.00826
1750547686	27	-0.31615	>0.10	0.00340	0.00383
1750427858	22	-0.98526	>0.10	0.01017	0.01392

1253642825	11	-1.37058	>0.10	0.01295	0.01862
1253842613	16	-0.83602	>0.10	0.00269	0.00368
1750415694	10	-1.16934	>0.10	0.01531	0.02037

Table G.1: **Tajima's *D* results for *Ty1/2* insertions.**

Ty3

Sequence	No. of strains	Tajima's *D*	*P* value	π value	θ value (per site)
1963300263	4	0.59158	>0.10	0.00343	0.00321
1893505639	15	-0.47812	>0.10	0.01364	0.01547
1752137827	8	-1.57586	0.05-0.10	0.00946	0.01377
1750609673	4	-0.78012	>0.10	0.00587	0.00640
1750579190	16	-1.58514	0.05-0.10	0.00455	0.00800
1750424241	6	-1.33698	>0.10	0.00489	0.00642
1750414164	6	-1.29503	>0.10	0.00396	0.00520
1750397574	4	-0.70990	>0.10	0.00294	0.00321
1750391442	4	-0.78012	>0.10	0.00588	0.00642
1253845381	13	-1.38139	>0.10	0.00686	0.01043
1253845040	4	1.66214	>0.10	0.01124	0.00960
1253563606	4	-0.61237	>0.10	0.00147	0.00160
1253542549	5	-1.14554	>0.10	0.00708	0.00850
1253421223	9	-1.27944	>0.10	0.02302	0.03094
1253333458	6	0.88776	>0.10	0.01045	0.00907
1253291586	10	0.52594	>0.10	0.00243	0.00209
423299768	14	-1.49064	>0.10	0.01443	0.02226
423119869	9	N/A	N/A	N/A	N/A
423092408	10	-1.03446	>0.10	0.00222	0.00312
1750853728	5	-0.97256	>0.10	0.00236	0.00283
1750951189	9	1.42885	>0.10	0.01557	0.01194
1752139269	9	-1.39844	>0.10	0.00441	0.00649
1892472183	10	-1.03446	>0.10	0.00222	0.00312
422848513	13	-1.38139	>0.10	0.00686	0.01043
423108613	8	-0.56068	>0.10	0.00597	0.00679
423104489	11	-1.42961	>0.10	0.00107	0.00200
1253576906	4	-0.81734	>0.10	0.01036	0.01130
2056181479	6	0.61082	>0.10	0.01144	0.01037
1253612664	5	-1.14554	>0.10	0.00708	0.00850

1253564256	15	-1.53857	>0.10	0.02717	0.04236
1253290099	14	-0.79031	>0.10	0.00281	0.00370
1892463544	13	1.47542	>0.10	0.00158	0.00095
1922462233	21	-0.89661	>0.10	0.02212	0.02870
422856666	21	-1.18019	>0.10	0.02004	0.02870
1750790573	4	-0.70990	>0.10	0.00298	0.00325
1253542355	6	-0.36201	>0.10	0.03138	0.03329
423324250	14	-0.95219	>0.10	0.01063	0.01387
1253234179	11	-1.62322	0.05-0.10	0.01402	0.02177
1253296927	14	-1.14505	>0.10	0.00653	0.00925
1253426982	13	-1.64267	0.05-0.10	0.00558	0.00948
1253541193	4	-0.82407	>0.10	0.01183	0.01291
1253624022	4	-0.70990	>0.10	0.00294	0.00321
1253802574	20	-1.16547	>0.10	0.00378	0.00585
1253808383	6	-1.36732	>0.10	0.00587	0.00771
1253812633	6	-1.35927	>0.10	0.02124	0.02713
1253840950	7	-0.75333	>0.10	0.01575	0.01821
1750361205	15	-1.11524	>0.10	0.01328	0.01825
1750603135	4	-0.61237	>0.10	0.00133	0.00145
1750614606	21	-1.38070	>0.10	0.00734	0.01181
1750621922	15	-0.31699	>0.10	0.00892	0.00972
1750847635	9	-0.73969	>0.10	0.00541	0.00651
1751355977	13	0.10070	>0.10	0.00880	0.00858
1893519769	14	-1.03805	>0.10	0.01112	0.01489
1945055758	25	-1.43329	>0.10	0.01005	0.01660
1962582664	13	-1.75462	0.05-0.10	0.00652	0.01144
1962594677	6	-1.33698	>0.10	0.00492	0.00646
1963600102	10	-1.58602	0.05-0.10	0.00739	0.01144
1975691243	19	-0.95616	>0.10	0.01019	0.01367
2017502064	12	-1.17901	>0.10	0.00187	0.00291
2017505814	15	-1.04163	>0.10	0.01003	0.01357

2056181752	8	-1.14142	>0.10	0.00695	0.00910
2063030282	6	-1.33698	>0.10	0.00493	0.00648
2064458468	19	-0.05144	>0.10	0.00249	0.00254
2064642908	6	-1.29503	>0.10	0.00378	0.00496

Table G.2: **Tajima's *D* results for *Ty3* insertions.**

Ty4

Sequence	No. of strains	Tajima's *D*	*P* value	π value	θ value (per site)
1253545921	5	-0.81650	>0.10	0.00108	0.00129
1253547397	4	N/A	N/A	N/A	N/A
1253550240	4	2.08033	>0.10	0.00719	0.00588
1253610629	6	-0.73166	>0.10	0.02096	0.02374
1253808752	5	-0.54556	>0.10	0.02168	0.02341
1253858106	7	-1.02394	>0.10	0.01662	0.02035
253859846	6	-0.73166	>0.10	0.02096	0.02374
1750567884	6	-0.98980	>0.10	0.01913	0.02274
1750571613	4	1.21400	>0.10	0.01993	0.01779
1750582399	4	0.03892	>0.10	0.01033	0.01029
1922468385	4	-0.87425	>0.10	0.28940	0.31571
1922470484	4	-0.06867	>0.10	0.01168	0.01176
2017463612	4	-0.75445	>0.10	0.00407	0.00443
2063016019	4	-0.38921	>0.10	0.00988	0.01029
2064869783	5	0.69900	>0.10	0.00431	0.00388
2293152340	20	-1.72331	0.05-0.10	0.00076	0.00215
1750546887	4	-0.82407	>0.10	0.01081	0.01179
1750363477	6	-1.44477	0.05-0.10	0.00991	0.01302
1253610629	6	-0.73166	>0.10	0.02096	0.02374
1253833151	4	-0.86098	>0.10	0.03580	0.03906
1750601028	4	2.08033	>0.10	0.00658	0.00539
1750621338	4	-0.61237	>0.10	0.00135	0.00147

Table G.3: **Tajima's *D* results for *Ty4* insertions.**

Ty5

Sequence	No. of strains	Tajima's *D*	*P* value	π value	θ value (per site)
422846906	22	-1.44848	>0.10	0.04279	0.06740
1253885689	5	-0.33192	>0.10	0.01275	0.01339
1253607370	6	N/A	N/A	N/A	N/A
1253905252	4	-0.70990	>0.10	0.00402	0.00438
1750830484	8	-0.18166	<0.10	0.03259	0.03377
1752115423	4	-0.38921	>0.10	0.01455	0.01515
1893503712	9	-1.28067	>0.10	0.01173	0.01612
1253468709	4	-0.38921	>0.10	0.01455	0.01515
1253541860	9	-1.28067	>0.10	0.01173	0.01612
1253605367	4	-0.61237	< 0.10	0.00199	0.00217
1253540480	21	-0.59736	>0.10	0.04271	0.05032
1253449340	11	-0.97277	>0.10	0.02702	0.03428
2068870413	7	N/A	N/A	N/A	N/A
1752124229	7	1.34164	>0.10	0.00239	0.00171
1253622465	16	1.11147	>0.10	0.02523	0.01962
1750393967	17	-0.39214	>0.10	0.03921	0.04335

Table G.4: **Tajima's *D* results for *Ty5* insertions.**

Appendix H

Tajima's *D* results for all insertions in *S. paradoxus*

The following tables are supporting results for Chapter 3, which cover Tajima's *D* tests completed for all insertions of all families in *S. paradoxus*.

In all tables, N/A indicates a lack of polymorphisms which prevented the test from being completed. Those insertions with significant results are listed in Results Chapter 3.

Ty1/2 superfamily

Sequence	No. of strains	Tajima's *D*	*P* value	π value	θ value (per site)
203316420	5	-1.14554	>0.10	0.00757	0.00909
1253978543	12	0.53727	>0.10	0.05031	0.04501
1254000917	7	-1.43414	>0.10	0.00342	0.00489
1254003216	11	-0.16811	>0.10	0.00720	0.00752
1254015836	7	-1.33626	>0.10	0.01851	0.02431
1254042274	7	-1.44309	>0.10	0.03090	0.04130
1254082394	15	N/A	N/A	N/A	N/A
1254106805	10	-0.77544	>0.10	0.02086	0.02494
1254109789	12	0.56310	>0.10	0.00773	0.00676
1254140487	15	-1.45596	>0.10	0.01354	0.02097
1254157663	10	0.51245	>0.10	0.01363	0.01223
1254171365	10	-1.38473	>0.10	0.01996	0.02814
1254296184	16	-0.32139	>0.10	0.01891	0.02055
1254299738	9	-1.69754	0.05-0.10	0.00891	0.01388
1254380201	7	-1.00623	>0.10	0.00090	0.00129
1254385225	4	-0.79684	>0.10	0.00786	0.00858
1254391570	5	-1.24176	>0.10	0.03041	0.03649
1254407467	7	N/A	N/A	N/A	N/A
1254417005	8	1.76414	0.05-0.10	0.00804	0.00579
1254417560	7	-1.00623	>0.10	0.00090	0.00128
1254466181	14	-0.71962	>0.10	0.02914	0.03494
1254475535	7	0.25402	>0.10	0.00809	0.00770
1254487069	4	N/A	N/A	N/A	N/A
1254568436	7	0.93086	>0.10	0.02590	0.02220
1254568461	7	-1.43414	>0.10	0.00359	0.00513
1254570171	5	N/A	N/A	N/A	N/A
1254581381	5	-0.41017	>0.10	0.00568	0.00606
1750933499	4	-0.81734	>0.10	0.01108	0.01208
1750978940	14	-0.60752	>0.10	0.00946	0.01116

1750987481	8	-1.02972	>0.10	0.00368	0.00482
1750988685	10	-1.66706	0.05-0.10	0.00252	0.00445
1751000800	6	-0.44736	>0.10	0.00376	0.00412
1751002336	4	-0.78012	>0.10	0.00621	0.00678
1751004647	4	-0.75445	>0.10	0.00472	0.00515
1751224491	12	-1.67054	0.05-0.10	0.02704	0.04263
1751228721	15	-0.67092	>0.10	0.00770	0.00935
1751271947	5	1.32768	>0.10	0.01212	0.01018
205950307	10	-0.10094	>0.10	0.02875	0.02937
205954333	8	0.23069	>0.10	0.00783	0.00746
1250539987	5	-0.80734	>0.10	0.01069	0.01208
1253940161	4	0.16766	>0.10	0.00526	0.00516
1253944513	4	-0.85430	>0.10	0.03021	0.03296
1253965680	6	-0.44736	>0.10	0.00379	0.00414
1253968255	6	-1.13197	>0.10	0.00212	0.00279
1253971669	6	-1.46056	0.05-0.10	0.01313	0.01725
1254005374	7	-1.33021	>0.10	0.02367	0.03092
1254011475	9	-1.08823	>0.10	0.00069	0.00115
1254087361	7	1.24854	>0.10	0.02402	0.01961
1254124084	5	-0.81650	>0.10	0.00121	0.00145
1254145435	9	0.40121	>0.10	0.02297	0.02123
1254159332	9	1.45715	>0.10	0.04074	0.03154
1254169217	5	0.98145	>0.10	0.01796	0.01581
1254194682	13	0.01076	>0.10	0.01473	0.01469
1254242779	11	-1.28247	>0.10	0.01239	0.01740
1254247165	6	-0.73110	>0.10	0.03851	0.04352
1254251068	15	-0.85107	>0.10	0.01698	0.02137
1254252169	4	-0.82943	>0.10	0.01420	0.01549
1254253248	12	-1.29282	>0.10	0.02385	0.03343
1254256455	5	-1.17432	>0.10	0.00991	0.01189
1254262592	7	-0.65405	>0.10	0.00315	0.00368

1254263237	9	-1.58469	0.05-0.10	0.01005	0.01504
1254273457	8	-1.05482	>0.10	0.00078	0.00120
1254280322	9	1.76414	0.05-0.10	0.00804	0.00579
1254290491	10	-0.10539	>0.10	0.01381	0.01414
1254373522	12	1.09997	>0.10	0.01199	0.00943
1254373626	12	-0.90560	>0.10	0.02330	0.02919
1254379648	14	0.29262	>0.10	0.02839	0.02657
1254385016	10	-0.69098	>0.10	0.00175	0.00222
1254388801	13	-0.56342	>0.10	0.01045	0.01212
1254394238	7	N/A	N/A	N/A	N/A
1254397255	6	1.03370	>0.10	0.01614	0.01377
1254404140	11	-1.49107	>0.10	0.01738	0.02577
1254408721	9	-0.06382	>0.10	0.00227	0.00231
1254409542	10	1.54637	>0.10	0.01076	0.00788
1254413515	7	0.20619	>0.10	0.00270	0.00258
1254425300	13	-1.65136	0.05-0.10	0.01830	0.02939
1254459340	5	1.22474	>0.10	0.00185	0.00148
1254461638	10	-0.85103	>0.10	0.02335	0.02841
1254470966	17	0.91388	>0.10	0.05138	0.04213
1254488181	11	-1.46460	>0.10	0.00332	0.00537
1254494629	5	1.57274	>0.10	0.00566	0.00453
1254563162	4	-0.61237	>0.10	0.00151	0.00164
1254564766	6	-1.36732	>0.10	0.00604	0.00794
1254569124	4	-0.61237	>0.10	0.00157	0.00172
1254578732	11	-1.40055	>0.10	0.00906	0.01333
1750931608	10	-1.01586	>0.10	0.03012	0.03815
1750936593	7	-0.33869	>0.10	0.00710	0.00761
1750987759	6	-1.13197	>0.10	0.00218	0.00286
1750991037	6	-0.93302	>0.10	0.00105	0.00138
1750994717	11	-0.81068	>0.10	0.01852	0.02255
1750994807	11	-0.16811	>0.10	0.00720	0.00752

1750998205	10	N/A	N/A	N/A	N/A
1750999991	8	-1.45671	>0.10	0.01130	0.01587
1751000090	9	-1.16790	>0.10	0.01776	0.02329
1751004305	6	-1.47739	>0.10	0.01656	0.02176
1751004609	9	-1.14944	>0.10	0.00329	0.00458
1751225035	8	1.16650	>0.10	0.00169	0.00122
1751230612	6	1.29710	>0.10	0.01509	0.01239
1751234179	16	0.30922	>0.10	0.01688	0.01566
1751251098	12	-0.38175	>0.10	0.00187	0.00215
1751253950	7	0.55902	>0.10	0.00150	0.00128
1751258912	9	-1.59839	0.05-0.10	0.01690	0.02501
1751259342	8	-1.42130	>0.10	0.00689	0.00976
1751262858	11	N/A	N/A	N/A	N/A
1751263383	10	-1.11173	>0.10	0.00061	0.00107
1751265408	11	-1.67815	0.05-0.10	0.00725	0.01188
1751266573	5	-0.81650	>0.10	0.00128	0.00153
1751269079	14	-0.41027	>0.10	0.00847	0.00947

Table H.1: **Tajima's *D* results for *Ty1/2* insertions.**

Ty3

Sequence	No. of strains	Tajima's *D*	*P* value	π value	θ value (per site)
204019754	10	0.22171	>0.10	0.00220	0.00206
1253958237	8	0.16843	>0.10	0.00583	0.00562
1253979580	10	-0.10174	>0.10	0.01073	0.01098
1253987674	4	N/A	N/A	N/A	N/A
1253990620	8	-1.05482	>0.10	0.00073	0.00112
1254020327	7	0.40249	>0.10	0.00389	0.00357
1254122758	9	-0.20713	>0.10	0.01193	0.01247
1254141673	11	0.31974	>0.10	0.01654	0.01543
1254237181	9	0.24844	>0.10	0.00669	0.00633
1254245650	9	-1.02653	>0.10	0.00643	0.00831
1254264772	8	-0.16751	>0.10	0.00523	0.00543
1254371088	5	1.45884	>0.10	0.00350	0.00280
1254371870	12	-1.74606	0.05-0.10	0.00736	0.01251
1254418141	5	-1.18441	>0.10	0.01050	0.01259
1254448868	4	0.03892	>0.10	0.01071	0.01067
1254489464	11	-1.17529	>0.10	0.02008	0.02693
1254493036	4	-0.83741	>0.10	0.01603	0.01749
1751223678	7	0.20619	>0.10	0.00250	0.00238
1751255781	7	-1.55311	0.05-0.10	0.00583	0.00833
1751260386	6	-1.40833	0.05-0.10	0.00777	0.01021
1751261699	8	-1.44684	>0.10	0.01541	0.02136
1751267326	8	0.08124	>0.10	0.00445	0.00437

Table H.2: **Tajima's *D* results for *Ty3* insertions.**

Ty4

Sequence	No. of strains	Tajima's *D*	*P* value	π value	θ value (per site)
1254245884	9	0.92073	>0.10	0.03038	0.02565
1254102474	7	-1.35841	>0.10	0.00223	0.00318
1254461557	12	-0.33218	>0.10	0.00583	0.00637
1254246245	9	0.74175	>0.10	0.02370	0.02060
1254012963	7	-0.04378	>0.10	0.00857	0.00864
1254009871	6	0.85057	>0.10	0.00144	0.00118
1253969717	8	-1.05482	>0.10	0.00067	0.00104
1751255985	7	-0.27492	>0.10	0.00197	0.00211
1254242465	9	0.24844	>0.10	0.00629	0.00595
1254146633	12	0.73262	>0.10	0.01254	0.01068
1751257415	11	-0.21051	>0.10	0.02047	0.02146

Table H.3: **Tajima's *D* results for *Ty4* insertions.**

Ty5

Sequence	No. of strains	Tajima's *D*	*P* value	π value	θ value (per site)
205759631	4	-0.70990	>0.10	0.00398	0.00435
1254083747	6	-0.39875	>0.10	0.01301	0.01396
1254102451	9	-0.46495	>0.10	0.03245	0.03582
1254411251	5	-1.09380	>0.10	0.00637	0.00765
1750994560	5	-1.14554	>0.10	0.01062	0.01274
1751006167	10	-1.66706	0.05-0.10	0.00321	0.00568
1751219545	9	-0.57015	>0.10	0.02997	0.03389
1751247193	9	0.15647	>0.10	0.00155	0.00147
1751274729	4	-0.22234	>0.10	0.02186	0.02235
1751276348	5	-0.97256	>0.10	0.00349	0.00419

Table H.4: **Tajima's *D* results for *Ty5* insertions.**

Appendix I

GO process categories of genes adjacent to candidate LTRs

The table on the following page lists the GO categories of adjacent genes and is supporting data for Chapter 3.

SGD GO-Slim Biological Process	No. of associated genes in species	
	S. cerevisiae	*S. paradoxus*
Amino acid transport	1	
Carbohydrate metabolic process	1	2
Cell budding and morphogenesis	2	
Cellular amino acid metabolic process	1	
Cellular ion homeostasis	2	1
Cellular organisation	5	6
Chromatin organisation	3	1
Chromosome segregation	2	1
Cofactor metabolic process	1	2
Conjugation	1	
Cytokinesis	1	
Cytoplasmic translation	2	
DNA replication and transcription	4	
Endosomal transport	1	1
Generation of precursor metabolites	1	1
Golgi vesicle transport	4	2
Histone modification	1	
Ion transport	3	4
Lipid metabolic process	2	2
Meiotic cell cycle	4	3
Membrane fusion	1	2
Mitotic cell cycle	2	3
Monocarboxylic acid metabolic process		2
Nuclear transport and organisation	1	5
Nucleobase transport and metabolism	3	4
Organelle assembly, fission and fusion	4	4
Protein biogenesis and modification	6	5
Proteolysis	1	2
Regulation of cell cycle	6	4
Response to stimuli and stress	8	3
Ribosomal biogenesis and export	1	3
RNA processing	3	3
Signalling	1	2
Sporulation	4	2
Transcription from RNA pol promoters	3	3
Translational elongation	1	
Vitamin metabolic process	1	
Other	2	3
Unknown	20	10

Table I.1: **GO categories of genes adjacent to candidate LTRs.** Breakdown of functions of genes adjacent to candidate LTRs in *S. cerevisiae* and *S. paradoxus* using the SGD GO-Slim Process Mapper. Genes are present in multiple process categories.

Appendix J

tRNA-candidate LTR associations

The following tables list those candidates associated with tRNAs and are supporting data for Chapter 4.

Family	No.	SGD ID	Chr.	Associated tRNA
Ty1/2	1	YERWdelta22	V	Glutamate
	2	YMLCdelta2	XIII	Tyrosine
	3	YERCdelta16	V	Lysine
	4	YORWdelta17	XV	n/a
	5	YHRCdelta10	VIII	Proline
	6	YPLWdelta5	XVI	n/a
	7	YELWdelta6	V	Arginine
	8	YDRWdelta11	IV	Glutamine
	9	YCRCdelta6	III	Proline
	11	YGRWdelta32	VII	n/a
	12	YLRCdelta9	XII	Leucine
	13	YNRCdelta8	XIV	Leucine
	14	YMLWdelta4	XIII	n/a
	15	YPLWdelta3	XVI	Tryptophan
	16	YGRWdelta23	VII	Glycine
	17	YLRCdelta21	XII	Arginine
	18	VII-77305	VII	Valine
	19	YJRWdelta18	X	n/a
	20	YOLCdelta3	XV	Threonine
	21	YGRCdelta12	VII	n/a
Ty3	2	YDRCsigma3	IV	Tyrosine
	3	XIII-760251	XIII	Alanine
	4	XII-681209	XII	n/a
	5	YORWsigma3	XV	Alanine
	7	V-487854	V	Glutamate
	8	YGLCsigma1	VII	Histidine
	9	YILWsigma3	IX	Aspartate
	10	YHRCsigma2	VIII	Alanine
	11	VI-212340	VI	n/a
	12	YLRWsigma2	XII	Arginine
	13	III-168387	III	Glutamine
	15	YBLWsigma1	II	Isoleucine
	16	YNLWsigma3	XIV	Proline
	17	YERCsigma3	V	Histidine
	18	YILWsigma2	IX	Isoleucine
Ty4	1	YERCtau2	V	Glutamate
	2	YMRCtau3	XIII	Glutamine
	4	YMRWtau2	XIII	n/a
Ty5	1	YGLWomega1	VII	n/a
	2	YCRWomega3	III	n/a
	3	YHLComega1	VIII	n/a

Table J.1: **Candidate LTRs associated with tRNAs in *S. cerevisiae*.**

Family	No.	Designation	Associated tRNA
Ty1/2	1	IV-364275	Lysine
	2	XI-472807	n/a
	3	X-513745	Arginine
	4	IV-445330	Serine
	5	VII-326627	Glutamate
	6	XVI-856470	Alanine
	7	II-236002	n/a
	8	VIII-112510	Serine
	9	XV-911530	Methionine
	10	II-786219	n/a
	11	XV-100955	Glycine
	12	XV-7897	n/a
	13	XIV-704237	n/a
	14	VII-318823	Histidine
	15	XV-610717	Valine
Ty3	1	XV-285166	Threonine
Ty4	1	VIII-442749	Threonine
	2	VIII-445249	n/a
Ty5	1	VII-642530	n/a

Table J.2: **Candidate LTRs associated with tRNAs in *S. paradoxus*.**

Appendix K

Details of gene adjacent to candidate LTRs in *S. cerevisiae*

The following table split over two pages details the genes adjacent to candidate LTR insertions in *S. cerevisiae* and are supporting data for Chapter 3.

Family	No.	SGD I.D.[a]	Chr.	Strain[b]	LTR co-ordinates	Adjacent gene(s)	Position[c]	Function of adjacent gene(s)
Ty1/2	1	YERWdelta22	V		487835-488165	*COG3*	Upstream	Component of oligomeric Golgi complex
	2	YMLCdelta2	XIII		168350-168681	YML053C	Upstream	Unknown
	3	YERCdelta16	V		435947-436277	*MDP-1* (YER134C)	Upstream	Magnesium-dependent acid phosphatase
	4	YORWdelta17	XV		664814-665145	*MPC54*	Downstream	Component of the meiotic outer plaque
	5	YHRCdelta10	VIII		389178-389509	*DCD1*	Downstream	Deoxycytidine monophosphate (dCMP) deaminase required for biosynthesis
	6	YPLWdelta5	XVI	CLIB215[d]	62390-62720[b]	YPL257W	Downstream	Unknown, interacts with a heat shock protein
	7	YELWdelta6	V		138222-138553	*GCN4*	Downstream	Basic leucine zipper (bZIP) transcriptional activator used in amino acid starvation
	8	YDRWdelta11	IV		802906-803235	*SEC7*	Upstream	Guanine nucleotide exchange factor (GEF) associated with Golgi body
	9	YCRCdelta6	III		124135-124465	YCR007C	Upstream	Integral membrane protein of DUP240 family
	11	YGRWdelta32	VII		931689-932019	*CRM1*	Downstream	Major karyopherin involved in export of proteins, RNAs and ribosomal subunits from nucleus
	12	YLRCdelta9	XII		592707-593035	*ADY4*	Downstream	Structural component of meiotic outer plaque
	13	YNRCdelta8	XIV		726615-726945	*HOL1*	Upstream	Transporter in the major facilitator superfamily (DHA1 family) of cation and histidinol
	14	YMLWdelta4	XIII	CBS7960[d]	189753-190083	*RRN11* *CAT2*	Upstream Downstream	Component of CF rDNA transcription factor complex Carnitine acetyl-CoA transferase
	15	YPLWdelta3	XVI	CBS7960[d]	56453-56788	*THI21*	Upstream	Hydroxymethylpyrimidine phosphate (HMP) kinase
	16	YGRWdelta23	VII		778785-779112	ARS729	Upstream	Replication origin
	17	YLRCdelta21	XII		818074-818403	*GAS2* RPL26A	Downstream Upstream	1,3-beta-glucanosyltransferase Ribosomal 60S subunit protein L26A
	18	VII-77305	VII	YJM789	77305-77638	*VRG4*	Downstream	Golgi GDP-mannose transporter
	20	YOLCdelta3	XV		113295-113626	ARS1509 YOL107W	Upstream Downstream	Replication origin Unknown
	21	YGRCdelta12	VII		547654-547956	*MSP1* RPS25A	Upstream Downstream	ATPase in mitochondrial outer membrane Component of 40S ribosomal subunit
Ty3	2	YDRCsigma3	IV		945956-946295	*AMD2*	Upstream	Amidase (putative)
	3	XIII-760251	XIII	Y55	760251-760591	*RKR1*	Downstream	Ubiquitin ligase associated with chromatin
	4	XII-681209	XII	SK1	681209-681548	*YCS4*	Upstream	Condensin complex subunit

Ty3 cont.	5	YORWsigma3	XV		854276-854614	YOR289W	Upstream	Unknown
	7	V-487854	V	RM11-1A	487854-488198	YER158C	Downstream	Unknown
	8	YGLCsigma1	VII		319424-319762	*USE1*	Downstream	SNARE protein required for targeting of Golgi-derived retrograde transport vesicles
	9	YILWsigma3	IX		324392-324734	*BAR1*	Upstream	Aspartyl protease that cleaves α-factor
	10	YHRCsigma2	VIII		146332-146671	YHR020W	Downstream	tRNA synthetase
	11	VI-212340	VI	RM11-1A	212340-212679	*MET10*	Downstream	Subunit of assimilatory sulfite reductase
	12	YLRWsigma2	XII		374000-374338	*AVL9* ARS1213	Downstream Upstream	Involved in exocytic transport from Golgi body Replication origin
	13	III-168387	III	DBVPG1853	168387-169064	*RHB1*	Upstream	Rheb-related GTPase (putative)
	15	YBLWsigma1	II		197715-198054	ARS231	Downstream	Replication origin
	16	YNLWsigma3	XIV		546739-547077	YNL042W-B BOP3	Downstream	Unknown Involved in methylmercury resistance
	17	YERCsigma3	V		434632-434971	*GLC7*	Downstream	Protein phosphatase catalytic subunit involved in glycogen metabolism
	18	YILWsigma2	IX	M22[d]	210309-210647	*AIR1*	Downstream	Zinc knuckle protein involved in RNA processing; component of the TRAMP complex
Ty4	1	YERCtau2	V		354361-354730	*UBP9*	Upstream	Ubiquitin protease
	2	YMRCtau3	XIII		808538-808906	*SCS7*	Upstream	Sphingolipid alpha-hydroxylase
	4	YMRWtau2	XIII		503772-504141	ARS1319	Upstream	Replication origin
Ty5	1	YGLWomega1	VII		839-1079	TEL07L	Upstream	Telomeric region
	2	YCRWomega3	III		291923-292167	ARS317 HMR	Downstream	Replication origin Silenced mating-type cassette
	3	YHLComega1	VIII		7995-8225	*ARN2* ARS802	Upstream Downstream	Transport of siderophore triacestylfusarinine C Replication origin

Table K.1: **Details of neighbouring genes in *S. cerevisiae*.** (Table continued from previous page) Data collected from the UCSC Genome Browser and SGD, and for LTRs not present or not annotated in the reference genome, supplemented with data from the Sanger Institute. Data on adjacent genes were only collected for candidate insertions. All genes that possessed a more significantly negative value of *D* or were too distant were excluded here, but the numbering system is kept consistent throughout. [a]I.D. as determined by SGD. If not present in the reference, insertions were allocated a name based on the chromosome and 5' coordinate. [b]All sequences are found within the reference strain S288c unless the original strain is specified, in which case, position within specified strain is given. [c]relative to LTR. [d]although present in the reference strain as one of the paired LTRs flanking FLEs, these insertions were present as solo LTRs in this and the multiple other strains used for Tajima's *D*.

Appendix L

Details of gene adjacent to candidate LTRs in *S. paradoxus*

The following table split over two pages details the genes adjacent to candidate LTR insertions in *S. paradoxus* and are supporting data for Chapter 3.

Family	No.	Designation	Chr.	Strain	LTR co-ordinates	Adjacent gene(s)	Position[a]	Function of adjacent gene(s)
Ty1/2	1	IV-364275	IV	N-17	364275-364593	*MCH1*	Downstream	Monocarboxylate permease
	2	XI-472807	XI	N-44	472807-473118	*FOX2*	Downstream	Enzyme of peroxisomal fatty acid β–oxidation pathway
						TOF2	Upstream	Protein required for rDNA silencing
	3	X-513745	X	Q95.3	513745-514063	*KCH1*	Downstream	Potassium transporter
						HIT1	Upstream	Box C/D snoRNP assembly factor
	4	IV-445330	IV	Z1.1	445330-445647	YDL007C-A	Downstream	Unknown
						RPT2	Upstream	ATPase
	5	VII-326627	VII	Z1.1	326627-326882	*TOS8*	Upstream	Putative transcription factor
						VPS45	Downstream	Involved in membrane traffic control between the Golgi and the vacuole
	6	XVI-856470	XVI	CBS432	856470-856795	*CUR1*	Downstream	Sorting factor involved in protein quality control and the destabilisation of URE3 prions
						KRE6	Upstream	Integral membrane protein with β-glucan synthase activity
	7	II-236002	II	CBS432	236002-236319	*IPP1*	Downstream	Inorganic pyrophosphatase
						YBR012C	Upstream	Unknown
	8	VIII-112510	VIII	CBS432	112510-112810	*SPO13*	Downstream	Required for segmentation during meiosis I
						MIP6	Upstream	Putative RNA-binding protein involved in nuclear mRNA exportation
	9	XV-911530	XV	CBS432	911530-911844	*TYE7*	Upstream	Transcription factor for *Ty1*-mediated gene expression
	10	II-786219	II	KPN3828	786219-786552	*PAU24*	Upstream	Cell wall mannoprotein of seripauperin multigene family
						ARR1	Downstream	Transcriptional activator of the bZIP family, involved in resistance to arsenic compounds
	11	XV-100955	XV	KPN3829	100955-101284	*ZEO1*	Upstream	Plasma membrane protein
						INO4	Downstream	Transcription factor required for derepression of phospholipid synthesis genes
	12	XV-7897	XV	KPN3829	7897-8204	*ENB1*	Downstream	Endosomal ferric enterobactin transporter
						CSS3	Upstream	Unknown; may suppress *Ty1* retrotransposition
	13	XIV-704237	XIV	Q32.3	704237-704603	*HOL1*	Upstream	Cation and histidinol transporter, member of DHA1 family
	14	VII-318823	VII	Q32.3	318823-319141	*USE1*	Upstream	SNARE protein required for targeting of Golgi-derived retrograde transport vesicles
						SRM1	Downstream	Nucleotide exchange factor for *Gsp1p*

Ty1/2 cont.	15	XV-610717	XV	Q95.3	610717-611028	*HEM15*	Upstream	Ferrochelatase membrane protein, catalyses the final step in haem biosynthetic pathway
						MPC54	Downstream	Component of the meiotic outer plaque which assembles during meiosis II
	16[e]	XV-285166	XV	CBS432	285166-285492	YOL014W	Upstream	Unknown
						HRD1	Downstream	Ubiquitin-protein ligase in endoplasmic reticulum
	17[e]	VIII-442749	VIII	KPN3829	442749-443078	*OYE3*	Downstream	NADPH oxidoreductase
Ty3	1	VIII-445249	VIII	CBS432	445249-445590	*PFS1*	Upstream	Sporulation protein
						KOG1	Downstream	Subunit of TORC1 complex involved in growth control
Ty4	1	VII-642530	VII	CBS5829	642530-642903	*NNF2*	Downstream	Involved in chromosome segregation; interacts with RNA pol subunits
						UTP22	Upstream	Component of the U3 snoRNA-associated protein complex
	2	III-276699	III	Q95.3	276699-277065	OCA4	Upstream	Cytoplasmic protein
						HMLα2	Downstream	Silenced copy of mating type protein α2 at HML (in *S. cerevisiae*)
Ty5	1	XVI-24300	XVI	CBS5829	24300-24550	YHL042W	Upstream	Unknown
						RMD6	Downstream	Required for sporulation

Table L.1: **Details of neighbouring genes in *S. paradoxus*** (Table continued from previous page) Insertions are not annotated in the *S. paradoxus* genome browser, unlike *S. cerevisiae*, therefore names were designated according to chromosome and 5' co-ordinate. [α]relative to LTR.

Appendix M

Paralogues associated with *Ty* insertions in *S. cerevisiae* and *S. paradoxus*

Species	GOI	Paralogue	Function of paralogue product	Adjacent LTR(s)
S. cerevisiae	YER158C	*AFR1*	α-factor receptor regulator	n/a
	UBP9	*UBP13*	ubiquitin protease	n/a
	ATO2	*ADY2*	acetate transporter	n/a
	CLB6	*CLB5*	B-type cyclin	*Ty1* upstream, opposite orientation
	AIR1	*AIR2*	RNA-binding subunit of TRAMP complex	n/a
	RPL26A	RPL26B	ribosomal 60S subunit	n/a
	VRG4	HVG1	unknown	n/a
S. paradoxus	*TOF2*	*NET1*	subunit of RENT complex	n/a
	MIP6	*PES4*	poly(A) binding protein	Partial *Ty3p* and *Ty4*E downstream, same orientation
	KRE6	*SKN1*	lipid biosynthesis membrane protein	*Ty1* upstream; same orientation; Partial *Ty1* downstream; opposite orientation
	CUR1	*BTN2*	SNARE-binding protein	*Ty1* downstream; same orientation
	TOS8	*CUP9*	transcriptional repressor	*Ty1* downstream; opposite orientation
	KHC1	*PRM6*	potassium transporter	n/a

Table M.1: **Paralogues associated with *Ty* insertions in *Saccharomyces*.** Many genes adjacent to potential candidate LTRs documented in Chapter 3 have paralogues which arose in the WGD event in an ancestor of *Saccharomyces* species. The function of the paralogue is also documented, as many differ from the original copy of the gene. *Ty4*E=European.

Appendix N

Supporting *Saccharomyces* LTR alignments

The following figures are supporting alignments for Chapter 4.

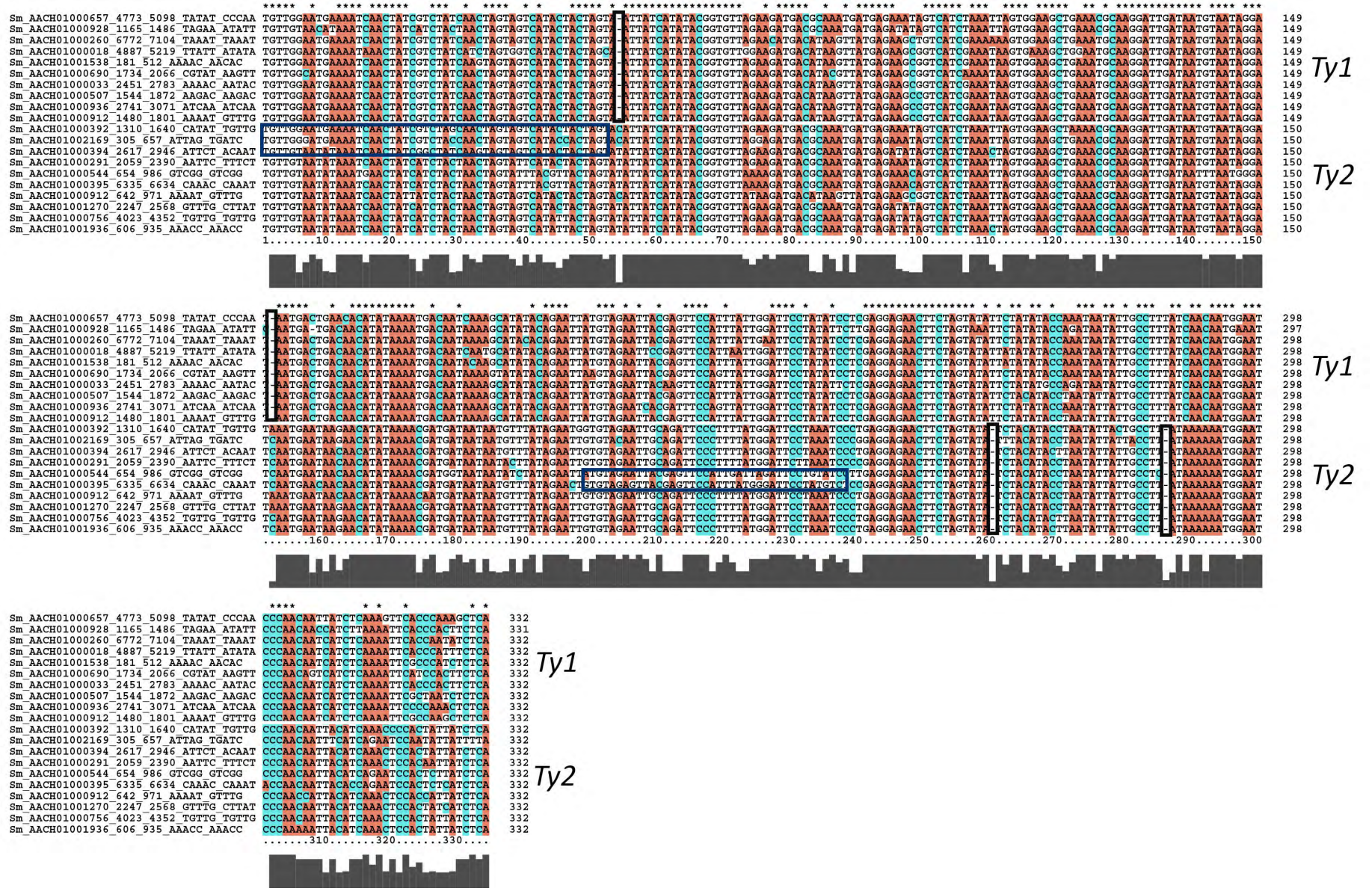

Figure N.1: ***Ty1/2* LTR alignment allows the differentiation of the families of *S. mikatae*.** Sequences from the *Ty1/2* families in *S. mikatae* possess four typical deletions (black boxes), two in each family – at sites 54 and 152 in *Ty1*, and 260 and 288 in *Ty2*, respectively. Additionally, two regions highlighted in blue represented potential recombination breakpoints between the two families, as these regions of *Ty2* LTRs were more like those of *Ty1*.

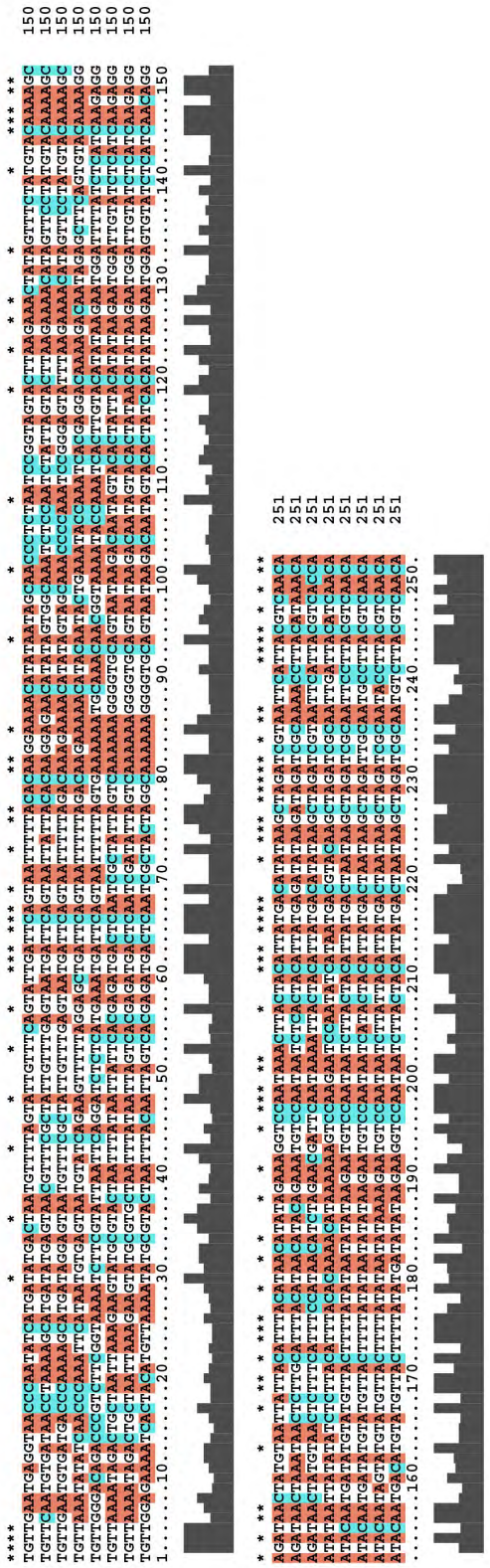

Figure N.2: **Alignment of *Ty5* LTRs in *S. mikatae*** Although SplitsTree found evidence for recombination breakpoints (P=0.001), this was likely a result of the age of the LTRs in this family.

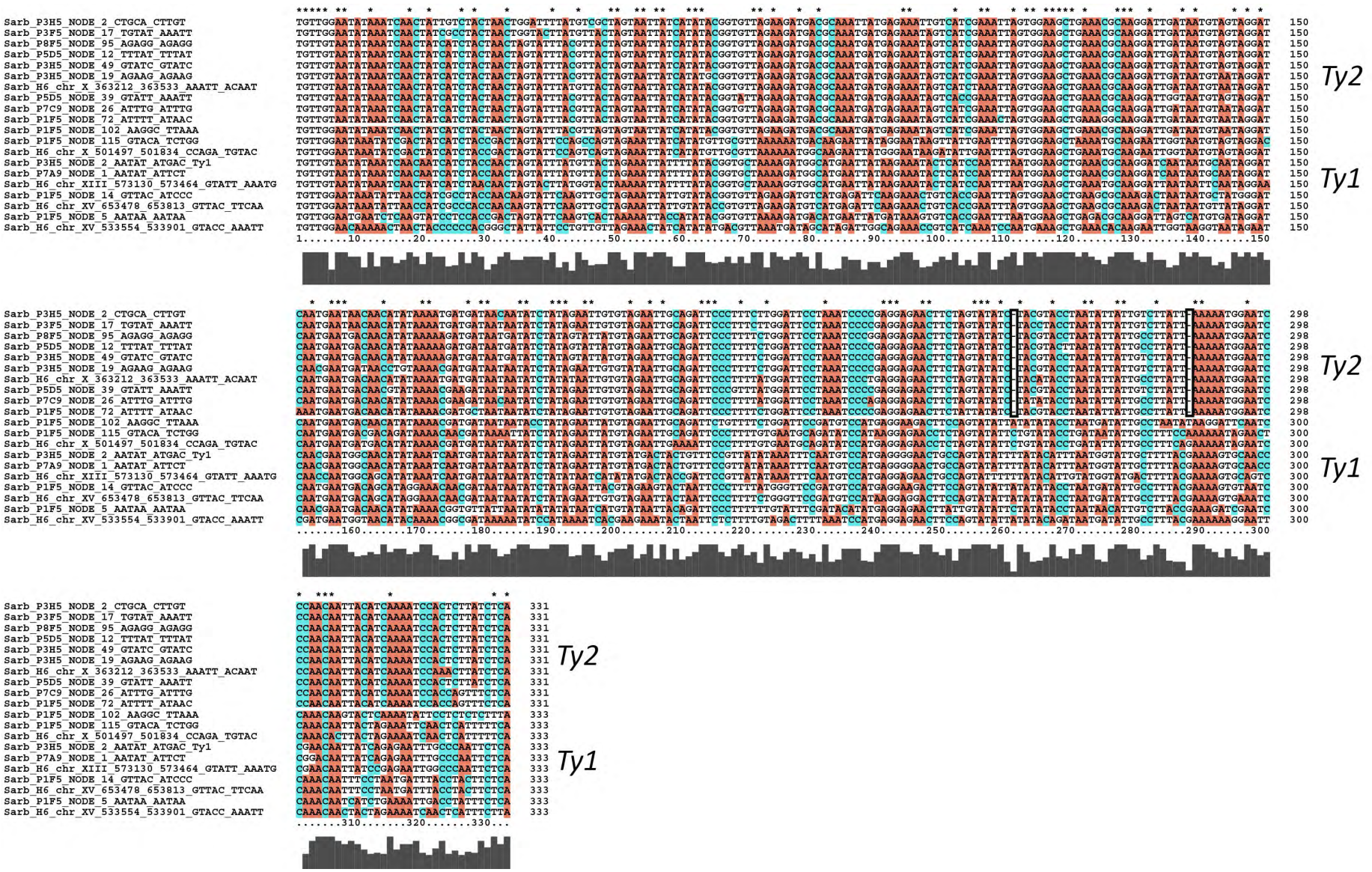

Figure N.3: **Alignment of *Ty1/2* LTRs in *S. arboricola***. The characteristic deletions of *Ty2* at positions 261 and 289 allowed families to be distinguished.

Appendix O

Summaries of genome contents of surveyed species

The following table contains supporting data for Chapter 5.

Species		Strain	Genome content %
Kazachstania (Clade 2)	*K. africana*	CBS2517	0.18
	*K. exigua**	CBS379	-
	K. naganishii	CBS8797	0.58
	K. saulgeensis	CLIB1764	0.18
	K. servazzii	SRCM102023	0.32
Naumovozyma (Clade 3)	*N. castellii*	CBS4309	0.53
	N. dairenensis	CBS421	1.89
Nakaseomyces (Clade 4)	*Nk. bacillisporus*	CBS7720	0.46
	Nk. bracarensis	CBS10154	0.03
	Nk. castellii	CBS4332	0.03
	Nk. delphensis	CBS2170	0.09
	Nk. glabrata	DSY562	0.70
	Nk. nivarensis	CBS9983	0.04
Tetrapisispora (Clade 5)	*T. blattae*	CBS6284	1.88
	T. phaffii	CBS4417	0.41
Vanderwaltozyma (Clade 6)	*V. polyspora*	DSM 70294	1.61
Zygosaccharomyces (Clade 7)	*Z. baillii*	ISA1307	0.05
	Z. parabaillii	ATCC 60483	0.05
	Z. rouxii	CBS732	0.08
Torulaspora (Clade 9)	*T. delbrueckii*	CBS1146	0.46
	T. microellipsoides	CBS427	0.16

Species		Strain	Genome content %
Lachancea (Clade 10)	*L. cidri*	CBS2950	0.04
	L. dasiensis	CBS10888	0.69
	L. fantastica	CBS6924	0.12
	L. fermentati	CBS6772	0.16
	L. kluyveri	NRRL Y-12651	1.31
	L. lanzarotensis	CBS12615	0.24
	L. meyersii	CBS8951	0.04
	L. mirantina	CBS11717	0.08
	L. nothofagi	CBS11611	0.16
	L. quebecensis	LL2012_068	0.54
	L. thermotolerans	CBS6340	0.41
	L. waltii	NCYC 2644	1.79
Kluyeromyces (Clade 11)	*K. aestuarii*	ATCC 18862	0.33
	K. dobzhanskii	CBS2104	0.15
	K. lactis	NRRL Y-1140	0.25
	K. marxianus	CMKU3-1042	0.77
	K. wickerhamii	UCD 54-210	0.05
Eremothecium (Clade 12)	*E. aceri*	FD-2008	0.26
	E. coryli	CBS749	0.52
	E. cymbalariae	DBVPG7215	0.40
	E. gossypii	ATCC 10895	0.08
	E. sinecaudum	ATCC 58844	0.27
Schizosaccharomyces	*Sz. cryophilus*	OY26	0.42
	Sz. japonicus	yFS275	3.75
	Sz. kambucha	SPK1820	0.56
	Sz. octosporus	yFS286	0.03
	Sz. pombe	972h	1.08
Ogataea	*O. angusta**	CBS4732	-
	O. arabinofermentans	NRRL Y-2248	0.30
	O. boidinii	JCM 9604	0.37

Species		Strain	Genome content %
	O. methanolica	JCM 10240	0.91
Ogataea	*O. parapolymorpha*	DL-1	0.30
(cont.)	*O. polymorpha*	NCYC 495	0.34
	O. succiphila	JCM 9445	0.01

Table O.1: **Summary contents of *sensu lato*, *Schizosaccharomyces* and *Ogataea* species.** Representatives for *Zygotorulaspora* (clade 8) were not available. *The genomes of *K. exigua* and *O. angusta* have not been fully sequenced, therefore the presence of RNAi pathway proteins and genome content % could not be ascertained at this stage. The reference genome for each species was used where previously designated in the literature. P - *Argonaute* and/or *Dicer* present; A - absent.

Appendix P

Characteristics of LTR-retrotransposon families

The following sections contain supporting data for Chapter 5 and detail the *Ty*-like families of *Saccharomyces sensu lato, Schizosaccharomyces* and *Ogataea* species. Family names are assigned with respect to host species names as suggested by Neuvéglise *et al.* (2002).

Tajima's *D* values in bold indicate significant results (Chapter 5).

Kazachstania (Clade 2)

The genome of *K. exigua* has yet to be sequenced in full, and only genomic survey sequencing (GSS) reads were available, representing approximately 20% of the genome (Bon *et al.*, 2000). Annotated elements, named *Tse1-5* for this species' previous designation of *Saccharomyces exigua*, were available on the NCBI database (Neuvéglise *et al.*, 2002). Therefore, at this stage the copy numbers of elements and solo LTRs could not be accurately determined in this species.

Family	*Ty1/2*-like					*Ty4-like*		*Ty5*-like			
	Tka1	*Tse1*	*Tkn1*	*Tksa1*	*Tkse1*	*Tksa4*	*Tkse4*	*Tse5*	*Tkn5*	*Tksa5*	*Tkse5*
FLE(pseudo)	0(1)	≥3(1)	0(2)	1(1)	0(1)	0(1)	1(1)	≥2(0)	1(1)	0(1)	6(1)
solo LTRs	0	≥4	42	18	0	1	9	≥1	8	0	7
LTR length (bp)	-	424	365	426	-	287	278	370	240	-	229
LTR diversity (π)	-	0.00788	0.16844	0.08396	-	-	0.41306	0.08184	0.03993	-	0.04556
Tajima's *D*	-	-1.43477	-0.50296	-1.67762	-	-	-0.76227	0.23771	-1.49628	-	**-2.20653**

Table P.1: **Characteristics of *Ty1/copia* families within *Kazachstania* species.** Copy numbers are likely to be an underestimate in *K. exigua* due to its incomplete genome sequencing. –indicates no evidence of LTRs in this family was found and/or too few sequences were available to complete the nucleotide diversity and Tajima's *D* tests. Copy numbers are: FLE(partial or pseudo); solo LTRs.

Family	*gypsy*				
	Tka3	*Tse3*	*Tkn3*	*Tksa3*	*Tkse3*
FLE(pseudo)	0(1)	≥2(1)	0(1)	1(1)	0(4)
solo LTRs	40	≥5	39	9	0
LTR length (bp)	467	947	360	702	656
LTR diversity (π)	0.13006	0.00546	0.14187	0.03391	-
Tajima's *D*	-1.51484	-0.61195	-1.19995	-1.42930	-

Table P.2: **Characteristics of *Ty3/gypsy* families within *Kazachstania* species.**

Naumovozyma (Clade 3)

Previously members of the *Saccharomyces* clade, *Naumovozyma castellii* and *N. dairenensis* were transferred to a new genus (Kurtzman and Robnett, 2003), now including *N. baii* in the basal position of the species phylogeny (Liu *et al.*, 2012).

While screening strains of *Naumovozyma* for TEs, strain 763_NDAI, sequenced as part of a large project of clinical isolates (Roach *et al.*, 2015), was wrongly identified as *N. dairenensis*. Less than 1% of sequencing reads mapped to the reference strain of *N. dairenensis*, as the strain was in fact an isolate of *Candida albicans* (93% of reads mapped to the reference strain). This finding was independently documented by Stavrou *et al.* (2018).

Family	***Ty1/2*-like**					***Ty4-like***	***Ty5*-like**
	Tnc1	*Ty1*	*Tnd2*-like*	*Tnd1*	*Tnd2*	*Tnd4*	*Tnd5*
FLE(pseudo)	0(3)	0(1)	0(0)	1(1)	1(0)	1(1)	1(0)
solo LTRs	5	3	3	15	15	23	19
LTR length (bp)	352	323	414	563	432	310	309
LTR diversity (π)	0.06379	0.11607	0.01449	0.06806	0.01257	0.08862	0.12156
Tajima's *D*	-0.88634	-	-	-1.20761	**-2.01629**	-2.18220	-1.53150

Table P.3: **Characteristics of *Ty1/copia* families within *Naumovozyma* species.** Tajima's *D* test required ≥4 sequences therefore was not completed for two *Ty1/copia* families of *N. castellii*. **Tnd2*-like in *N. castellii*.

Family	***gypsy***				
	Tnc3	*Tnc3s*	*Tnd3*	*Tnd3s*	*Tnc3-like**
FLE(pseudo)	3(8)	0(1)	1(0)	1(1)	0(2)
solo LTRs	37	4	21	14	10
LTR length (bp)	297	232	971	252	308
LTR diversity (π)	0.03335	0.09646	0.08234	0.07446	0.12605
Tajima's *D*	**-2.21807**	-0.60945	-1.40725	-1.14384	-1.56182

Table P.4: **Characteristics of *Ty3/gypsy* families within *Naumovozyma* species.** **Tnc3*-like family in *N. dairenensis*.

Nakaseomyces (Clade 4)

Family	*gypsy*			*Ty4-like*	*Ty5*-like			
	Tnkc3	*Tnkd3*	*Tnkg3*	*Tnkb4*	*Tnkb5*	*Tnkc5*	*Tnkg5*	*Tnkg6*
FLE(pseudo)	0(1)	1(1)	1(1)	1(2)	0(3)	0(1)	3*(0)	5(0)
solo LTRs	0	?	5	56	0	0	3	9
LTR length (bp)	-	?	966	338	-	-	395	485
LTR diversity (π)	-	-	0.03475	0.12623	-	-	0.06527	0.08076
Tajima's *D*	-	-	-2.03181	**-1.94715**	-	-	1.82421	1.20149

Table P.5: **Characteristics of families within *Nakaseomyces* species.** *tandem insertions. The copy number and LTR length of the *Ty3/gypsy* family in *Nk. delphensis* were not able to be determined.

Tetrapisispora (Clade 5)

Family	*Ty1/2*-like		*gypsy*		*Ty4*-like		*Ty5*-like		*Tca2*-like
	Ttb1	*Ttp1*	*Ttb3*	*Ttp3*	*Ttb4*	*Ttp4*	*Ttb5*	*Ttp5*	*Ttb2*
FLE(pseudo)	0(2)	0(3)	1(4)	0(1)	1(0)	0(7)	0(4)	0(1)	0(1)
solo LTRs	125	102	81	0	20	78	35	0	0
LTR length (bp)	216	406	415	592	366	408	425	~650	-
LTR diversity (π)	0.06181	0.08395	0.05994	-	0.08537	0.08817	0.06319	-	-
Tajima's *D*	**-2.53199**	**-2.19733**	**-2.45781**	-	**-1.99580**	-1.96707	**-2.13189**	-	-

Table P.6: **Characteristics of families within *Tetrapisispora* species.**

Vanderwaltozyma (Clade 6)

An NCBI Trace Archive strain (listed as *K. yarrowii*) was incorrectly designated as a species of *Vanderwaltozyma*. <2% of reads mapped to the genome of sister species *V. polyspora* whereas 95% onto the reference genome of *Lachancea waltii*.

Family	*Ty1/2*-like	*gypsy*			*Ty4*-like	*Ty5*-like
	Tkp1	*Tkp2*	*Tkp3*	*Ty3*-like	*Tkp4*	*Tkp5*
FLE(pseudo)	1(2)	2(1)	1(1)	1(4)	1(5)	1(5)
solo LTRs	12	3	21	2	25	31
LTR length (bp)	552	319	321	286	321	412
LTR diversity (π)	0.08925	0.13433	0.12860	0.08627	0.02837	0.07560
Tajima's *D*	-1.02076	-1.13296	-0.73454	-0.24490	-1.16391	0.32245

Table P.7: **Characteristics of families within *Vanderwaltozyma polyspora.***

Zygosaccharomyces and *Zygotorulaspora* (Clades 7-8)

Zygosaccharomyces bailii, *Z. parabailii* and *Z. rouxii* all contained a single *Ty3/gypsy* pseudoelement with pseudogenes RT and RH. Additionally, no LTRs could be recovered from any species

in this clade, Therefore, the sequences of these species were not included in the phylogenies of Chapter 5. Furthermore, no nuclear genomes of *Zygotorulaspora* species have currently been sequenced, therefore the TE content remains unknown.

Torulaspora (Clade 9)

Family	*Ty1/2*-like			*gypsy*	
	Ttd1	*Ttd2*	*Ttm1*	*Ttd3*	*Ttm3*
FLE(pseudo)	6*(3)	3(1)	0(2)	3(2)	0(2)
solo LTRs	4	1	0	2	0
LTR length (bp)	324	517	-	141	-
LTR diversity (π)	0.01120	0.04123	-	0.00504	-
Tajima's *D*	-0.37277	**-1.97698**	-	0.01889	-

Table P.8: **Characteristics of families within *Torulaspora* species.** *indicates presence of tandem formations.

Lachancea (Clade 10)

The phylogenetic analysis of Kurtzman and Robnett (2003) allowed the reconciliation of several species previously spread throughout the *sensu lato* clades into the new genus of *Lachancea*. *L. fantastica* was previously designated as a strain of *L. thermotolerans*, but renamed by Vakirlis *et al.* (2016).

Family	*Ty1/2*-like												
	Tld1	*Tlfa1*	*Tlfe1*	*Tsk1*	*Tll1*	*Tlme1*	*Tlmi1*	*Tln1*	*Tlq1*	*Tlt1*	*Tlw1*	*Tlw2*	*Ty1p*
FLE(pseudo)	2(1)	0(2)	6(0)	8(1)	0(1)	0(1)	0(3)	0(2)	1(1)	1(3)	0(4)	15(1)	1(0)
solo LTRs	75	0	62	173	0	0	0	0	29	44	46	29	15
LTR length (bp)	456	-	333	321	-	-	-	-	416	418	395	399	317
LTR diversity (π)	0.09557	-	0.03754	0.07832	-	-	-	-	0.19346	0.24300	0.22052	0.02010	0.10390
Tajima's *D*	**-2.30971**	-	**-1.91668**	**-2.28109**	-	-	-	-	-0.85849	-1.39510	-1.57070	**-2.27988**	**-1.85843**

Table P.9: **Characteristics of *Ty1/copia* families within *Lachancea* species.** No *Ty1/copia* elements beyond *Ty1/2* like were discovered in these species, with the exception of *Ty4*-like solo LTRs in one strain of *L. waltii* (Table P.10).

Family	*Ty1/2*-like			*Ty4*-like
	Tlw1	*Tlw2*	*Ty1p*	*Ty4p**
FLE(pseudo)	0(4)	15(1)	1(0)	0(0)
solo LTRs	46	29	25	3
LTR length (bp)	395	399	317	373
LTR diversity (π)	0.22052	0.02010	0.10390	0.06110
Tajima's *D*	-1.57070	**-2.27988**	**-1.85843**	-

Table P.10: **Characteristics of *Ty1/copia* families within *L. waltii*.** *trace archive strain only.

Family	*gypsy*								
	Tld3	*Tlfa3*	*Tlfe3*	*Tlk3*	*Tll3*	*Tlme3*	*Tlq3*	*Tlt3*	*Tlw3*
FLE(pseudo)	1(1)	0(3)	2(0)	1(0)	0(2)	0(1)	1(1)	0(0)	3(0)
solo LTRs	5	0	11	3	4	0	24	18	2
LTR length (bp)	349	-	200	364	357	-	261	262	174
LTR diversity (π)	0.14483	-	0.06623	0.20092	0.13039	-	0.17354	0.08334	0.03941
Tajima's *D*	-1.35529	-	-1.75933	-0.37059	-0.57398	-	-1.78122	**-1.91813**	-1.66658

Table P.11: **Characteristics of *Ty3/gypsy* families within *Lachancea* species.**

Kluyveromyces (Clade 11)

Family	*Ty1/2*-like					*gypsy*
	Tkd1	*Tkl1*	*Tkm1*	*Tkm2*	*Tkw1*	*Tka3*
FLE(pseudo)	0(1)	2(1)	8(7)	1(3)	0(1)	2(1)
solo LTRs	6	44	12	1	2	9
LTR length (bp)	393	395	385	815	393	441
LTR diversity (π)	0.27157	0.19200	0.08661	0.07504	-	0.15186
Tajima's *D*	-1.15811	-1.54141	-1.84854	-1.67451	-	-1.59484

Table P.12: **Characteristics of families within *Kluyveromyces* species.**

Eremothecium (Clade 12)

Family	*gypsy*					
	Tea3	*Teco3*	*Tecy3-like**	*Tecy3*	*Teg3*	*Tes3*
FLE(pseudo)	1(1)	1(4)	0(1)	1(1)	0(3)	1(1)
solo LTRs	7	40	16	30	0	6
LTR length (bp)	397	372	412	419	-	385
LTR diversity (π)	0.09155	0.20033	0.39887	0.04105	-	0.14705
Tajima's *D*	-1.26257	-1.10024	-1.25400	-1.96784	-	-0.17494

Table P.13: **Characteristics of *Ty3/gypsy* families within *Eremothecium* species.** **E. cymbalariae*-like family in *E. coryli*.

Schizosaccharomyces

Sz. kambucha and *Sz. octosporus* did not contain any full-length LTRs.

Family	*Tcry1*	*Tcry2*	*Tcry3*	*Tcry4*
Copy number	2(0) 4	1(1) 9	0(1) 1	0(1) 7
LTR length	416	469	440	464
LTR diversity (π)	0.12191	0.03961	0.05327	0.13757
Tajima's *D*	-0.42159	-1.64947	-	-1.21736

Table P.14: **Characteristics of *Ty3/gypsy* families within *Sz. cryophilus* species.**

Family	*Tj2*	*Tj3*	*Tj4*	*Tj5*	*Tj6*	*Tj7*	*Tj9*	*Tj10*	*Tj14*[+]
Copy number	1(1) 8	3(4) 13	2*(0) 3	2(0) 6	2(0) 12	1(0) 3	8*(0) 3	2(1) 2	2*(0) 0
LTR length	339	268	411	247	239	239	456	309	267
LTR diversity (π)	0.05238	0.02568	0.00438	0.02456	0.06628	0.06844	0.03712	0.05409	0.00499
Tajima's *D*	-1.58382	**-1.93778**	0.14908	-0.66695	-1.36138	-1.58637	-0.71975	-0.84843	-

Table P.15: **Characteristics of *Ty3/gypsy* families within *Sz. japonicus* species.** *tandem elements. +Newly identified family. The remaining families (*Tj1, Tj8*) could not be reliably identified given the co-ordinates of Rhind et al. (2011).

Family	972h		NCYC132	
	Tf1	*Tf2*	*Tf1*	*Tf2*
Copy number	0(0) 24	15(1) 41	1(1) 19	0(1) 2
LTR length	360	349	360	349
LTR diversity (π)	0.08237	0.13908	0.44593	-
Tajima's *D*	-1.69010	**-1.99130**	-1.08784	-

Table P.16: **Characteristics of *Ty3/gypsy* families within *Sz. pombe* species.**

Ogataea

Due to only recently resolved taxonomies, *Ogataea* is synonymous with *Candida* and *Pichia* (Kurtzman *et al.*, 2011).

	Ty3/gypsy			
	Toa3	*Tob3*	*Tom3*	*Tom4*
Copy number	0(1) no LTRs	2(0) 3	2(4) 4	5(2) 7
LTR length (bp)	-	535	348	373
LTR diversity (π)	-	0.04377	0.12191	0.20549
Tajima's *D*	-	-1.11622	1.40937	-0.80141

Table P.17: **Characteristics of *Ty3/gypsy* families within *Ogataea* species.**

	Ty5-like							
	Toan5	*Toar5*	*Tob5*	*Tom5*	*Top5*	*Top6*	*Toan5**	*Top6**
Copy number	≥1	1(2) no LTRs	0(1) 0	0(2) no LTRs	2(0) 0	2(0) 24	4(1) 11	0(1) 4
LTR length (bp)	322	-	331	-	264	282	320	281
LTR diversity (π)	-	-	-	-	+	0.08917	0.01216	0.23086
Tajima's *D*	-	-	-	-	+	-1.65426	-0.76959	-0.51720

Table P.18: **Characteristics of *Ty5*-like families within *Ogataea* species.** *families in *O. polymorpha*. +nucleotide polymorphism and Tajima's tests could not be completed for identical sequences.

Appendix Q

Genomic contents of Brazilian strains of *S. cerevisiae*

The following table contains supporting data for Chapter 6.

ENA Accession Number	Strain	TE content (%)	GC content (%)	No. of scaffolds	Introgression		No. of full-length elements in family (solo LTRs)			
					No. of coding genes	*S. paradoxus* genome %	*Ty1/2*	*Ty3*	*Ty4*	*Ty5*
ERR1111486	Y260	1.50	38.17	406	-	3.50	2^{2} (73)	1 (19)	1 (9)	2 (3)
ERR1111487	Y262	1.18	38.28	998	19	1.47	2^{T} (46)	1 (8)	1 (13)	0 (10)
ERR1111488	Y264	1.13	38.17	431	8	2.40	1^{T} (65)	1 (14)	1 (13)	0 (12)
ERR1111489	Y455	1.13	38.19	526	8	2.40	1 (67)	1 (14)	1 (10)	0 (11)
ERR1111490	Y461	1.26	38.19	593	3	2.21	3^{T} (56)	1 (10)	1 (10)	0 (12)
ERR1111491	Y462	1.27	38.21	538	4	2.34	3^{2} (72)	1 (15)	1 (8)	0*(10)
ERR1111492	Y464	1.29	38.15	458	12	3.25	4^{2} (82)	1 (15)	0 (13)	0*(13)
ERR1111493	Y639	1.63	38.26	804	-	2.43	4^{2} (75)	1 (15)	1 (7)	1 (6)
ERR1111494	Y640	1.22	38.15	568	-	4.75	3 (84)	1^{T} (15)	0 (9)	1 (9)
ERR1111495	Y641	1.48	38.11	516	24	4.23	2^{T} (88)	0 (16)	1 (20)	0 (5)
ERR1111496	Y642	1.50	38.15	582	24	4.33	2^{T} (88)	0 (11)	1 (18)	0 (6)
ERR1111497	Y645	1.23	38.27	525	8	2.73	2 (69)	1 (11)	1 (13)	1 (14)
ERR1111498	Y646	1.21	38.17	529	11	2.83	3^{2} (60)	1 (12)	1 (7)	0 (14)
ERR1111499	Y647	1.23	38.18	483	9	2.88	3 (65)	1 (10)	1 (13)	0 (14)
ERR1111500	Y649	1.24	38.23	617	21	3.53	4^{2} (61)	1 (9)	0 (10)	0 (12)
ERR1111501	Y650	1.25	38.20	523	20	3.28	4^{2} (53)	1 (8)	0 (14)	0 (14)
ERR1111502	Y651	1.84	38.35	1273	^{H}	27.54	3 (104)	2 (11)	3 (20)	0 (10)
ERR1111503	Y652	1.85	38.32	1448	^{H}	28.66	$5^{T\,2}$ (170)	2 (21)	2 (38)	0 (20)
ERR1111504	Y255	0.96	37.16	534	-	1.33	2 (33)	1 (4)	1 (4)	1 (6)
ERR1111505	Y257	1.03	38.09	520	1^{R}	1.76	3 (44)	1 (5)	0 (3)	0 (7)
ERR1111506	Y263	1.33	38.13	802	40	3.93	0^{2} (44)	0 (1)	0 (7)	0 (4)
ERR1111507	Y266	1.16	38.06	432	8	2.55	2 (48)	1 (4)	1 (4)	0 (6)
ERR1111508	Y456	1.48	38.04	395	27	3.47	3^{T} (52)	1 (2)	1 (5)	0 (10)
ERR1111509	Y457	1.17	38.10	820	8	3.30	2 (50)	1 (7)	1 (3)	0 (8)
ERR1111510	Y463	1.19	38.13	517	21	3.62	3 (40)	1 (5)	0 (4)	0 (7)
ERR1111511	Y636	1.37	38.12	534	-	2.06	3 (37)	1 (4)	1 (2)	1 (3)
ERR1111512	Y643	1.29	38.09	472	19	3.57	3 (52)	1 (2)	1 (5)	0 (14)
ERR1111513	Y644	1.18	38.09	531	10	4.40	4 (50)	1 (4)	1 (3)	0 (9)
	Average	1.31	38.13		15	3.03	3 (65)	1 (10)	1 (10)	0 (10)

Table Q.1: **Statistics of the 28 Brazilian strains of *S. cerevisiae*.** **Ty5* coding regions disrupted by *Ty1* element in these strains; ^{R}recombinant gene; ^{H}hybrid genomes; introgressed coding genes determined by Barbosa *et al.* (2016). **hybrids were not counted towards the average. Conservative frequencies as partial sequences and those fragmented by the ends of contig reads were not included. ^{T}strains contained a *Tsk1* or *Tlf3* relic, described in the text. ^{2}contained a *Ty1/2* hybrid element described in the text.

Appendix R

Test of functional constraint on a *Ty1* relic in Brazilian strains of *S. cerevisiae*

The following table and figure contain supporting data for Chapter 6.

Node	Ka/Ks Branch1	Ka Branch1	Ks Branch1	Ka/Ks Branch2	Ka Branch2	Ks Branch2
1	0.4718	0.2746	0.5819	0.0000	0.0000	0.0025
2	0.1441	0.0005	0.0038	0.5236	0.0019	0.0038
3	0.2703	0.0002	0.0006	0.4345	0.0005	0.0012
4	0.0000	0.0000	$1e^{-10}$	1.4286	0.0014	$1e^{-10}$
5	0.9115	0.0013	0.0015	1.0075	0.0012	0.0012

Table R.1: **Output of Ka/Ks calculation on the potential functionally constrained IN region of *Ty1* relics in Brazilian strains of *S. cerevisiae*.**

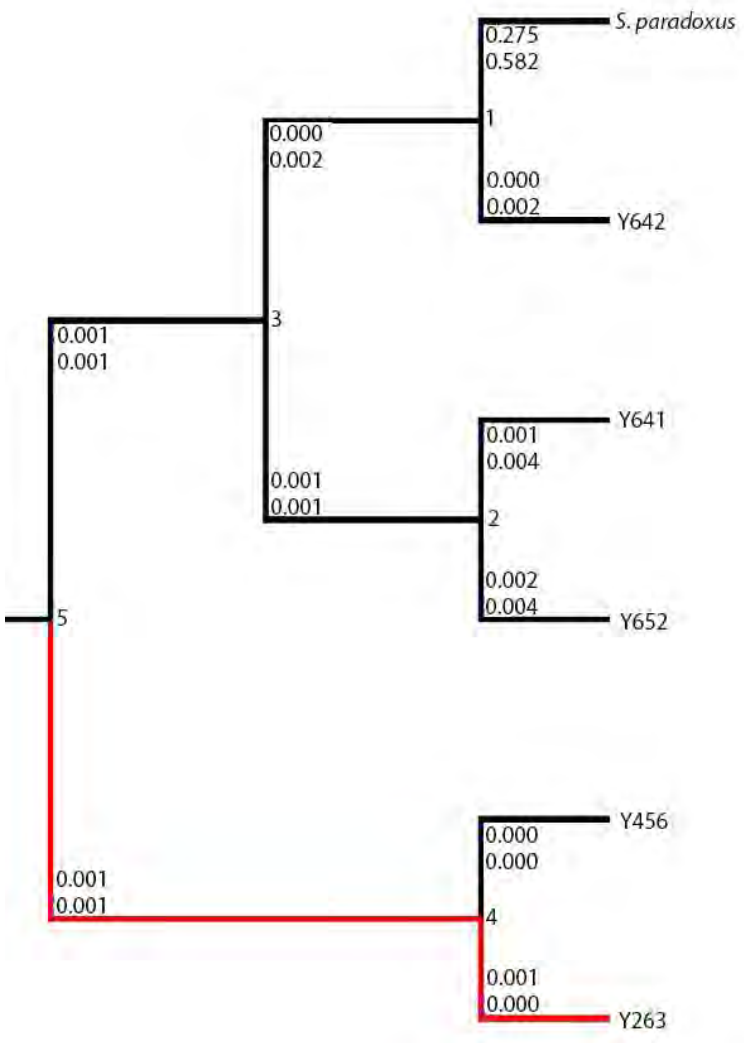

Figure R.1: **Output of Ka/Ks calculation on the potential functionally constrained IN region of *Ty1* relics in Brazilian strains of *S. cerevisiae*.**

References

Aa, E., Townsend, J. P., Adams, R. I., Nielsen, K. M. and Taylor, J. W. (2006), 'Population structure and gene evolution in *Saccharomyces cerevisiae*', *FEMS Yeast Res* **6**(5), 702–715.

Abrusan, G. and Krambeck, H. J. (2006), 'Competition may determine the diversity of transposable elements', *Theor Popul Biol* **70**(3), 364–375.

Adams, J. and Oeller, P. (1986), 'Structure of evolving populations of *Saccharomyces cerevisiae*: Adaptive changes are frequently associated with sequence alterations involving mobile elements belonging to the *Ty* family', *PNAS* **83**(18), 7124–7127.

Ahn, H. W., Tucker, J. M., Arribere, J. A. and Garfinkel, D. J. (2017), 'Ribosome biogenesis modulates *Ty1* copy number control in *Saccharomyces cerevisiae*', *Genetics* **207**(4), 1441–1456.

Akao, T., Yashiro, I., Hosoyama, A., Kitagaki, H., Horikawa, H., Watanabe, D., Akada, R., Ando, Y., Harashima, S., Inoue, T., Inoue, Y., Kajiwara, S., Kitamoto, K., Kitamoto, N., Kobayashi, O., Kuhara, S., Masubuchi, T., Mizoguchi, H., Nakao, Y., Nakazato, A., Namise, M., Oba, T., Ogata, T., Ohta, A., Sato, M., Shibasaki, S., Takatsume, Y., Tanimoto, S., Tsuboi, H., Nishimura, A., Yoda, K., Ishikawa, T., Iwashita, K., Fujita, N. and Shimoi, H. (2011), 'Whole-genome sequencing of sake yeast *Saccharomyces cerevisiae* Kyokai no. 7', *Res* **18**(6), 423–434.

Almeida, P., Barbosa, R., Bensasson, D., Gonçalves, P. and Sampaio, J. P. (2017), 'Adaptive divergence in wine yeasts and their wild relatives suggests a prominent role for introgressions and rapid evolution at noncoding sites', *Mol Ecol* **26**(7), 2167–2182.

Almeida, P., Barbosa, R., Zalar, P., Imanishi, Y., Shimizu, K., Turchetti, B., Legras, J. L., Serra, M., Dequin, S., Couloux, A., Guy, J., Bensasson, D., Gonçalves, P. and Sampaio, J. P. (2015), 'A population genomics insight into the mediterranean origins of wine yeast domestication', *Mol Ecol* **24**(21), 5412–5427.

Almeida, P., Gonçalves, C., Teixeira, S., Libkind, D., Bontrager, M., Masneuf-Pomarede, I., Albertin, W., Durrens, P., Sherman, D. J., Marullo, P., Hittinger, C. T., Gonçalves, P. and Sampaio,

J. P. (2014), 'A Gondwanan imprint on global diversity and domestication of wine and cider yeast *Saccharomyces uvarum*', *Nat Commun* **5**(4044), e.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990), 'Basic local alignment search tool', *J Mol Biol* **215**(3), 403–410.

Amberg, D., Burke, D. and Strathern, J. (2005), *Methods in Yeast Genetics: A Cold Spring Harbor Laboratory Course Manual*, 2005 edn, Cold Spring Harbor Laboratory Press, USA.

Aminetzach, Y. T., Macpherson, J. M. and Petrov, D. A. (2005), 'Pesticide resistance via transposition-mediated adaptive gene truncation in *Drosophila*', *Science* **309**(5735), 764–767.

Anderson, F. E. and Swofford, D. L. (2004), 'Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA', *Mol Phylogenet Evol* **33**(2), 440–51.

Andersson, J. (2012), 'Phylogenomic approaches underestimate eukaryotic gene transfer', *Mob Genet Elements* **2**(1), 59–62.

Andersson, J. O. (2009), 'Horizontal gene transfer between microbial eukaryotes', *Methods Mol Biol* **532**, 473–487.

Ansari, A. and Hampsey, M. (2005), 'A role for the CPF 3'-end processing machinery in RNAP II-dependent gene looping', *Genes Dev* **19**(24), 2969–78.

Anxolabehere, D., Kidwell, M. G. and Periquet, G. (1988), 'Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of *Drosophila melanogaster* by mobile *P* elements', *Mol Biol Evol* **5**(3), 252–269.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Gelpke, M. D., Roach, J., Oh, T., Ho, I. Y., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S. F., Clark, M. S., Edwards, Y. J., Doggett, N., Zharkikh, A., Tavtigian, S. V., Pruss, D., Barnstead, M., Evans, C., Baden, H., Powell, J., Glusman, G., Rowen, L., Hood, L., Tan, Y. H., Elgar, G., Hawkins, T., Venkatesh, B., Rokhsar, D. and Brenner, S. (2002), 'Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*', *Science* **297**(5585), 1301–1310.

Argueso, J. L., Carazzolle, M. F., Mieczkowski, P. A., Duarte, F. M., Netto, O. V., Missawa, S. K., Galzerani, F., Costa, G. G., Vidal, R. O., Noronha, M. F., Dominska, M., Andrietta, M. G., Andrietta, S. R., Cunha, A. F., Gomes, L. H., Tavares, F. C., Alcarde, A. R., Dietrich, F. S., McCusker, J. H., Petes, T. D. and Pereira, G. A. (2009), 'Genome structure of a *Saccharomyces cerevisiae* strain widely used in bioethanol production', *Genome Res* **19**(12), 2258–2270.

Argueso, J. L., Westmoreland, J., Mieczkowski, P. A., Gawel, M., Petes, T. D. and Resnick, M. A. (2008), 'Double-strand breaks associated with repetitive can reshape the genome', *Proc Natl Acad Sci U S A* **105**(33), 11845–11850.

Arkhipova, I. R. and Morrison, H. G. (2001), 'Three retrotransposon families in the genome of *Giardia lamblia*: two telomeric, one dead', *Proc Natl Acad Sci U S A* **98**(25), 14497–502.

Arndt, K. and Fink, G. (1986), 'Gcn4 protein, a positive transcription factor in yeast, binds general control promoters at all 5' tgactc 3' sequences', *Proc Natl Acad Sci U S A* **83**(22), 8516–20.

Artieri, C. G. and Fraser, H. B. (2014), 'Evolution at two levels of gene expression in yeast', *Genome Res* **24**(3), 411–421.

Aye, M., Dildine, S. L., Claypool, J. A., Jourdain, S. and Sandmeyer, S. B. (2001), 'A truncation mutant of the 95-kilodalton subunit of transcription factor IIIC reveals asymmetry in *Ty3* integration', *Mol Cell Biol* **21**(22), 7839–7851.

Aye, M., Irwin, B., Beliakova-Bethell, N., Chen, E., Garrus, J. and Sandmeyer, S. (2004), 'Host factors that affect *Ty3* retrotransposition in *Saccharomyces cerevisiae*', *Genetics* **168**(3), 1159–1176.

Aye, M. and Sandmeyer, S. B. (2003), '*Ty3* requires yeast la homologous protein for wild-type frequencies of transposition', *Molecular Microbiology* **49**(2), 501–515.

Bachhawat, N., Ouyang, Q. and Henry, S. A. (1995), 'Functional characterization of an inositol-sensitive upstream activation sequence in yeast. a cis-regulatory element responsible for inositol-choline mediated regulation of phospholipid biosynthesis', *J Biol Chem* **270**(42), 25087–95.

Bachman, N., Eby, Y. and Boeke, J. D. (2004), 'Local definition of *Ty1* target preference by long terminal repeats and clustered tRNA genes', *Genome Res* **14**(7), 1232–1247.

Baker, E., Wang, B., Bellora, N., Peris, D., Hulfachor, A. B., Koshalek, J. A., Adams, M., Libkind, D. and Hittinger, C. T. (2015), 'The genome sequence of *Saccharomyces eubayanus* and the domestication of lager-brewing yeasts', *Mol Biol Evol* **32**(11), 2818–2831.

Baller, J. A., Gao, J., Stamenova, R., Curcio, M. J. and Voytas, D. F. (2012), 'A nucleosomal surface defines an integration hotspot for the *Saccharomyces cerevisiae Ty1* retrotransposon', *Genome Res* **22**(4), 704–713.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A. and Pevzner, P. A. (2012), 'SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing', *J Comput Biol* **19**(5), 455–477.

Bao, W., Kojima, K. K. and Kohany, O. (2015), 'Repbase update, a database of repetitive elements in eukaryotic genomes', *Mob DNA* **6**(11).

Barberis, A., Pearlberg, J., Simkovich, N., Farrell, S., Reinagel, P., Bamdad, C., Sigal, G. and Ptashne, M. (1995), 'Contact with a component of the polymerase II holoenzyme suffices for gene activation', *Cell* **81**(3), 359–368.

Barbosa, R., Almeida, P., Safar, S. V., Santos, R. O., Morais, P. B., Nielly-Thibault, L., Leducq, J. B., Landry, C. R., Gonçalves, P., Rosa, C. A. and Sampaio, J. P. (2016), 'Evidence of natural hybridization in brazilian wild lineages of *Saccharomyces cerevisiae*', *Genome Biol Evol* **8**(2), 317–329.

Barreiro, L. B. and Quintana-Murci, L. (2010), 'From evolutionary genetics to human immunology: how selection shapes host defence genes', *Nat Rev Genet* **11**(1), 17–30.

Barros Lopes, M., Bellon, J., Shirley, N. and Ganter, P. (2002), 'Evidence for multiple interspecific hybridization in *Saccharomyces sensu stricto* species', *FEMS Yeast Res* **1**.

Bartolome, C., Bello, X. and Maside, X. (2009), 'Widespread evidence for horizontal transfer of transposable elements across *Drosophila* genomes', *Genome Biol* **10**(2), R22.

Bartolome, C., Maside, X. and Charlesworth, B. (2002), 'On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*', *Mol Biol Evol* **19**(6), 926–37.

Barton, N. H. (2000), 'Genetic hitchhiking', *Philos Trans R Soc Lond B Biol Sci* **355**(1403), 1553–1562.

Bass, D., Howe, A., Brown, N., Barton, H., Demidova, M., Michelle, H., Li, L., Sanders, H., Watkinson, S. C., Willcock, S. and Richards, T. A. (2007), 'Yeast forms dominate fungal diversity in the deep oceans', *Proc Biol Sci* **274**(1629), 3069–77.

Bass, D. and Richards, T. A. (2011), 'Three reasons to re-evaluate fungal diversity 'on Earth and in the ocean'', *Fungal Biology Reviews* **25**(4), 159–164.

Beauregard, A., Curcio, M. J. and Belfort, M. (2008), 'The take and give between retrotransposable elements and their hosts', *Annu Rev Genet* **42**, 587–617.

Beggs, J. D. (1978), 'Transformation of yeast by a replicating hybrid plasmid', *Nature* **275**(5676), 104–9.

Begin, M. and Schoen, D. J. (2007), 'Transposable elements, mutational correlations, and population divergence in *Caenorhabditis elegans*', *Evolution* **61**(5), 1062–70.

Belloch, C., Perez-Torrado, R., González, S. S., Perez-Ortin, J. E., Garcia-Martinez, J., Querol, A. and Barrio, E. (2009), 'Chimeric genomes of natural hybrids of *Saccharomyces cerevisiae* and *Saccharomyces kudriavzevii*', *Appl Environ Microbiol* **75**(8), 2534–2544.

Bellon, J. R., Schmid, F., Capone, D. L., Dunn, B. L. and Chambers, P. J. (2013), 'Introducing a new breed of wine yeast: interspecific hybridisation between a commercial *Saccharomyces cerevisiae* wine yeast and *Saccharomyces mikatae*', *PLoS One* **8**(4), e62053.

Belyayev, A. (2014), 'Bursts of transposable elements as an evolutionary driving force', *J Evol Biol* **27**(12), 2573–2584.

Ben-Aroya, S., Mieczkowski, P. A., Petes, T. D. and Kupiec, M. (2004), 'The compact chromatin structure of a *Ty* repeated sequence suppresses recombination hotspot activity in *Saccharomyces cerevisiae*', *Mol Cell* **15**(2), 221–231.

Benachenhou, F., Jern, P., Oja, M., Sperber, G., Blikstad, V., Somervuo, P., Kaski, S. and Blomberg, J. (2009), 'Evolutionary conservation of orthoretroviral long terminal repeats (LTRs) and *ab initio* detection of single LTRs in genomic data', *PLoS One* **4**(4), e5179.

Benachenhou, F., Sperber, G. O., Bongcam-Rudloff, E., Andersson, G., Boeke, J. D. and Blomberg, J. (2013), 'Conserved structure and inferred evolutionary history of long terminal repeats (LTRs)', *Mob DNA* **4**(1), 5.

Bensasson, D. (2011), 'Evidence for a high mutation rate at rapidly evolving yeast centromeres', *BMC Evol Biol* **11**, 211.

Bensasson, D., Zarowiecki, M., Burt, A. and Koufopanou, V. (2008), 'Rapid evolution of yeast centromeres in the absence of drive', *Genetics* **178**(4), 2161–2167.

Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Sayers, E. W. (2015), 'Genbank', *Nucleic Acids Res* **43**(Database issue), D30–35.

Bergman, C. M. (2018), 'Horizontal transfer and proliferation of *Tsu4* in *Saccharomyces paradoxus*', *In press* .

Bergsten, J. (2005), 'A review of long-branch attraction', *Cladistics* **21**, 163–193.

Bilanchone, V., Claypool, J. A., Kinsey, P. and Sandmeyer, S. (1993), 'Positive and negative regulatory elements control expression of the yeast retrotransposon *Ty3*', *Genetics* **134**(3), 685–700.

Biémont, C. (1994), 'Dynamic equilibrium between insertion and excision of *P* elements in highly inbred lines from an M' strain of *Drosophila melanogaster*', *J Mol Evol* **39**, 466–472.

Biémont, C., Lemeunier, F., Garcia Guerreiro, M. P., Brookfield, J. F., Gautier, C., Aulard, S. and Pasyukova, E. G. (1994), 'Population dynamics of the *copia, mdg1, mdg3, gypsy*, and *P* transposable elements in a natural population of *Drosophila melanogaster*', *Genet Res* **63**(3), 197–212.

Biémont, C., Tsitrone, A., Vieira, C. and Hoogland, C. (1997), 'Transposable element distribution in *Drosophila*', *Genetics* **147**(4), 1997–9.

Bing, J., Han, P. J., Liu, W. Q., Wang, Q. M. and Bai, F. Y. (2014), 'Evidence for a far east asian origin of lager beer yeast', *Curr Biol* **24**(10), 380–381.

Blackwood, E. M. and Kadonaga, J. (1998), 'Going the distance: A current view of enhancer action', *Science* **281**(5373), 60–63.

Blanc, V. and Adams, J. (2003), 'Evolution in *Saccharomyces*: identification of mutations increasing fitness in laboratory populations', *Genetics* **165**, 975–983.

Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J., Nekrutenko, A. and Galaxy, T. (2010), 'Manipulation of FASTQ data with Galaxy', *Bioinformatics* **26**(14), 1783–1785.

Bleykasten-Grosshans, C., Friedrich, A. and Schacherer, J. (2013), 'Genome wide analysis of intraspecific transposon diversity in yeast', *BMC Genomics* **14**(399).

Bleykasten-Grosshans, C., Jung, P. P., Fritsch, E. S., Potier, S., de Montigny, J. and Souciet, J. L. (2011), 'The *Ty1* LTR-retrotransposon population in *Saccharomyces cerevisiae* genome: dynamics and sequence variations during mobility', *FEMS Yeast Res* **11**(4), 334–344.

Bleykasten-Grosshans, C. and Neuvéglise, C. (2011), 'Transposable elements in yeasts', *C R Biol* **334**(8-9), 679–686.

Boeke, J. D. and Corces, V. G. (1989), 'Transcription and reverse transcription of retrotransposons', *Annu Rev Microbiol* **43**, 403–434.

Boeke, J. D. and Devine, S. E. (1998), 'Yeast retrotransposons: Finding a nice quiet neighborhood', *Cell* **83**.

Boeke, J. D. and Sandmeyer, S. (1991), *Yeast Transposable Elements*, Vol. 1, Cold Spring Harbor Laboratory Press, New York, book section 4, pp. 193–261.

Bolton, E. C. and Boeke, J. D. (2003), 'Transcriptional interactions between yeast tRNA genes, flanking genes and *Ty* elements: a genomic point of view', *Genome Res* **13**(2), 254–263.

Bon, E., Neuvéglise, C., Lepingle, A., Wincker, P., Artiguenave, F., Gaillardin, C. and Casaregola, S. (2000), 'Genomic exploration of the hemiascomycetous yeasts: 6. *Saccharomyces exiguus*', *FEBS Lett* **487**, 42–46.

Bonchev, G. and Parisod, C. (2013), 'Transposable elements and microevolutionary changes in natural populations', *Mol Ecol Resour* **13**(5), 765–775.

Borneman, A. R., Desany, B. A., Riches, D., Affourtit, J. P., Forgan, A. H., Pretorius, I. S., Egholm, M. and Chambers, P. J. (2011), 'Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of *Saccharomyces cerevisiae*', *PLoS Genet* **7**(2), e1001287.

Borneman, A. R., Desany, B. A., Riches, D., Affourtit, J. P., Forgan, A. H., Pretorius, I. S., Egholm, M. and Chambers, P. J. (2012), 'The genome sequence of the wine yeast VIN7 reveals an allotriploid hybrid genome with *Saccharomyces cerevisiae* and *Saccharomyces kudriavzevii* origins', *FEMS Yeast Res* **12**(1), 88–96.

Borneman, A. R., Forgan, A. H., Pretorius, I. S. and Chambers, P. J. (2008), 'Comparative genome analysis of a *Saccharomyces cerevisiae* wine strain', *FEMS Yeast Res* **8**(7), 1185–1195.

Borneman, A. R. and Pretorius, I. S. (2015), 'Genomic insights into the *Saccharomyces sensu stricto* complex', *Genetics* **199**(2), 281–291.

Bosi, E., Donati, B., Galardini, M., Brunetti, S., Sagot, M. F., Lio, P., Crescenzi, P., Fani, R. and Fondi, M. (2015), 'MeDuSa: a multi-draft based scaffolder', *Bioinformatics* **31**(15), 2443–51.

Bowen, N. J. and McDonald, J. F. (2001), '*Drosophila* euchromatic LTR retrotransposons are much younger than the host species in which they reside', *Genome Res* **11**(9), 1527–1540.

Bowen, N., Jordan, I., Epstein, J., Wood, V. and Levin, H. L. (2003), 'Retrotransposons and their recognition of pol ii promoters: A comprehensive survey of the transposable elements from the complete genome sequence of *Schizosaccharomyces pombe*', *Genome Res* **13**(1).

Bradshaw, V. A. and McEntee, K. (1989), 'Dna damage activates transcription and transposition of yeast *Ty* retrotransposons', *Mol Gen Genet* **218**(3), 465–474.

Brady, T. L., Fuerst, P. G., Dick, R. A., Schmidt, C. and Voytas, D. F. (2008), 'Retrotransposon target site selection by imitation of a cellular protein', *Mol Cell Biol* **28**(4), 1230–1239.

Brand, A. H., Micklem, G. and Nasmyth, K. (1987), 'A yeast silencer contains sequences that can promote autonomous plasmid replication and transcriptional activation', *Cell* **51**(5), 709–719.

Braverman, J., Hudson, R., Kaplan, N., Langley, C. H. and Stephan, W. (1995), 'The hitchhiking effect on site frequency spectrum of polymorphisms', *Genetics* **140**(2), 783–796.

Breslauer, K. J., Frank, R., Blocker, H. and Marky, L. A. (1986), 'Predicting duplex stability from the base sequence', *Proc Natl Acad Sci U S A* **83**(11), 3746–3750.

Bridier-Nahmias, A., Tchalikian-Cosson, A., Baller, J. A., Menouni, R., Fayol, H., Flores, A., Saib, A., Werner, M., Voytas, D. F. and Lesage, P. (2015), 'Retrotransposons. an RNA polymerase III subunit determines sites of retrotransposon integration', *Science* **348**(6234), 585–588.

Brisco, P. and Kohlhaw, G. (1990), 'Regulation of yeast LEU2', *J Biol Chem* **265**(20).

Brookfield, J. and Badge, R. (1997), 'Population genetics models of transposable elements', *Genetica* **100**.

Brookfield, J. F. Y. (2005), 'Evolutionary forces generating sequence homogeneity and heterogeneity within retrotransposon families', *Cyto Gen Res* **110**.

Brown, A. N. and Lloyd, V. K. (2015), 'Evidence for horizontal transfer of *Wolbachia* by a *Drosophila* mite', *Exp Appl Acarol* **66**(3), 301–311.

Burns, K. H. and Boeke, J. D. (2012), 'Human transposon tectonics', *Cell* **149**(4), 740–752.

Bushman, F. (2003), 'Targeting survival: integration site selection by retroviruses and LTR retrotransposons', *Cell* **115**.

Cameron, J. R., Loh, E. Y. and Davis, R. W. (1979), 'Evidence for transposition of dispersed repetitive families in yeast', *Cell* **16**(4), 739–751.

Carr, M., Bensasson, D. and Bergman, C. M. (2012), 'Evolutionary genomics of transposable elements in *Saccharomyces cerevisiae*', *PLoS One* **7**(11), e50978.

Carr, M., Soloway, J. R., Robinson, T. E. and Brookfield, J. F. (2002), 'Mechanisms regulating the copy numbers of six LTR retrotransposons in the genome of *Drosophila melanogaster*', *Chromosoma* **110**(8), 511–518.

Carr, M. and Suga, H. (2014), 'The holozoan *Capsaspora owczarzaki* possesses a diverse complement of active transposable element families', *Genome Biol Evol* **6**(4), 949–963.

Carreto, L., Eiriz, M. F., Gomes, A. C., Pereira, P. M., Schuller, D. and Santos, M. A. (2008), 'Comparative genomics of wild type yeast strains unveils important genome diversity', *BMC Genomics* **9**(524).

Casacuberta, E. and González, J. (2013), 'The impact of transposable elements in environmental adaptation', *Mol Ecol* **22**(6), 1503–1517.

Casarégola, S., Neuvéglise, C., Bon, E. and Gaillardin, C. (2002), 'Ylli, a non-ltr retrotransposon l1 family in the dimorphic yeast yarrowia lipolytica', *Mol Biol Evol* **19**(5), 664–77.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/11961100*

Casey, G. P. and Pedersen, M. B. (1988), 'Sequence polymorphisms in the genus *Saccharomyces*. V. cloning and characterization of a LEU2 gene from *S. carlsbergensis*', *Carlsberg Res Commun* **53**(3), 209–219.

Casola, C., Hucks, D. and Feschotte, C. (2008), 'Convergent domestication of *pogo*-like transposases into centromere-binding proteins in fission yeast and mammals', *Mol Biol Evol* **25**(1), 29–41.

Castanera, R., Lopez-Varas, L., Borgognone, A., LaButti, K., Lapidus, A., Schmutz, J., Grimwood, J., Perez, G., Pisabarro, A. G., Grigoriev, I. V., Stajich, J. E. and Ramirez, L. (2016), 'Transposable elements versus the fungal genome: Impact on whole-genome architecture and transcriptional profiles', *PLoS Genet* **12**(6), e1006108.

Castillo, D. M. and Moyle, L. C. (2012), 'Evolutionary Implications of Mechanistic Models of TE-Mediated Hybrid Incompatibility', *Int J Evol Biol* **2012**, 698198.

Cavrak, V. V., Lettner, N., Jamge, S., Kosarewicz, A., Bayer, L. M. and Mittelsten Scheid, O. (2014), 'How a retrotransposon exploits the plant's heat stress response for its activation', *PLoS Genet* **10**(1), e1004115.

Chaisson, M. J., Wilson, R. K. and Eichler, E. E. (2015), 'Genetic variation and the *de novo* assembly of human genomes', *Nat Rev Genet* **16**(11), 627–40.

Chaleff, D. T. and Fink, G. R. (1980), 'Genetic events associated with an insertion mutation in yeast', *Cell* **21**(1), 227–37.

Chan, J. E. and Kolodner, R. D. (2011), 'A genetic and structural study of genome rearrangements mediated by high copy repeat *Ty1* elements', *PLoS Genet* **7**(5), e1002089.

Chapman, K., Bystrom, A. and Boeke, J. D. (1992), 'Initiator methionine tRNA is essential for *Ty1* transposition in yeast', *PNAS* **89**.

Charlesworth, B. and Charlesworth, D. (1983), 'The population dynamics of transposable elements', *Genet Res (Camb)* **42**.

Charlesworth, B. and Langley, C. H. (1989), 'The population genetics of *Drosophila* transposable elements', *Annu Rev Genet* **23**, 251–87.

Charlesworth, B., Langley, C. H. and Sniegowski, P. D. (1997), 'Transposable element distributions in *Drosophila*', *Genetics* **147**(4), 1993–1995.

Charlesworth, B., Lapid, A. and Canada, D. (1992*a*), 'The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster*. I: element frequencies and distribution in *Drosophila*', *Genet Res (Camb)* **60**(2), 103–114.

Charlesworth, B., Lapid, A. and Canada, D. (1992*b*), 'The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster*. II: inferences on the nature of selection against elements', *Genet Res (Camb)* **60**(2), 115–130.

Charlesworth, B., Morgan, M. and Charlesworth, D. (1993), 'The effect of deleterious mutations on neutral molecular variation', *Genetics* **134**(4), 1289–1303.

Charlesworth, B., Sniegowski, P. and Stephan, W. (1994), 'The evolutionary dynamics of repetitive in eukaryotes', *Nature* **371**(6494), 215–220.

Cheeseman, K., Ropars, J., Renault, P., Dupont, J., Gouzy, J., Branca, A., Abraham, A. L., Ceppi, M., Conseiller, E., Debuchy, R., Malagnac, F., Goarin, A., Silar, P., Lacoste, S., Sallet, E., Bensimon, A., Giraud, T. and Brygoo, Y. (2014), 'Multiple recent horizontal transfers of a large genomic region in cheese making fungi', *Nat Commun* **5**(2876).

Chen, S. and Corces, V. G. (2001), 'The *gypsy* insulator of *Drosophila* affects chromatin structure in a directional manner', *Genetics* **159**(4), 1649–1658.

Chenais, B., Caruso, A., Hiard, S. and Casse, N. (2012), 'The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments', *Gene* **509**(1), 7–15.

Cheng, X., Zhang, D., Cheng, Z., Keller, B. and Ling, H. Q. (2009), 'A new family of *Ty1-copia*-like retrotransposons originated in the tomato genome by a recent horizontal transfer event', *Genetics* **181**(4), 1183–1193.

Choudhury, R. R., Neuhaus, J. M. and Parisod, C. (2017), 'Resolving fine-grained dynamics of retrotransposons: comparative analysis of inferential methods and genomic resources', *Plant J*

.

Chung, N., Jenkins, G., Hannun, Y. A., Heitman, J. and Obeid, L. M. (2000), 'Sphingolipids signal heat stress-induced ubiquitin-dependent proteolysis', *J Biol Chem* **275**(23), 17229–32.

Chuong, E. B., Elde, N. C. and Feschotte, C. (2016), 'Regulatory evolution of innate immunity through co-option of endogenous retroviruses', *Science* **351**(6277), 1083–7.

Chuong, E. B., Elde, N. C. and Feschotte, C. (2017), 'Regulatory activities of transposable elements: from conflicts to benefits', *Nat Rev Genet* **18**(2), 71–86.

Ciriacy, M. and Breilmann, D. (1982), 'δ sequences mediate rearrangements in *Saccharomyces cerevisiae*', *Curr Genet* **6**(1), 55–61.

Ciriacy, M. and Williamson, V. M. (1981), 'Analysis of mutations affecting *Ty*-mediated gene expression in *Saccharomyces cerevisiae*', *Mol Gen Genet* **182**(1), 159–163.

Clark, A. G., Bilanchone, V., Haywood, L., Dildine, S. L. and Sandmeyer, S. (1988), 'A yeast sigma composite element, *Ty3*, has properties of a retrotransposon', *J Biol Chem* **263**(3).

Cliften, P. F., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B. and Johnston, M. (2003), 'Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting', *Science* **301**(71), 71–76.

Clowers, K. J., Heilberger, J., Piotrowski, J. S., Will, J. L. and Gasch, A. P. (2015), 'Ecological and genetic barriers differentiate natural populations of *Saccharomyces cerevisiae*', *Mol Biol Evol* **32**(9), 2317–2327.

Combina, M., Perez-Torrado, R., Tronchoni, J., Belloch, C. and Querol, A. (2012), 'Genome-wide gene expression of a natural hybrid between *Saccharomyces cerevisiae* and *S. kudriavzevii* under enological conditions', *Int J Food Microbiol* **157**(3), 340–345.

Company, M. and Errede, B. (1987), 'Cell-type-dependent gene activation by yeast transposon *Ty1* involves multiple regulatory determinants', *Mol Cell Biol* **7**(9).

Coney, L. and Roeder, G. (1988), 'Control of yeast gene expression by transposable elements: Maximum expression requires a functional *Ty* activator sequence and a defective *Ty* promoter', *Mol Cell Biol* **8**(10).

Cordaux, R. and Batzer, M. A. (2009), 'The impact of retrotransposons on human genome evolution', *Nat Rev Genet* **10**(10), 691–703.

Cridland, J., Thornton, K. R. and Long, A. D. (2015), 'Gene expression variation in *Drosophila melanogaster* due to rare transposable element insertion alleles of large effect', *Genetics* **199**, 85–93.

Cubillos, F. A., Billi, E., Zorgo, E., Parts, L., Fargier, P., Omholt, S., Blomberg, A., Warringer, J., Louis, E. J. and Liti, G. (2011), 'Assessing the complex architecture of polygenic traits in diverged yeast populations', *Mol Ecol* **20**(7), 1401–1413.

Cullen, B. R., Raymond, K. and Ju, G. (1985), 'Functional analysis of the transcriptional control region located within the avian retroviral long terminal repeat', *Mol Cell Biol* **5**(3).

Curcio, M. J., Lutz, J. and Lesage, P. (2015), 'The *Ty1* LTR-retrotransposon of budding yeast, *Saccharomyces cerevisiae*', *Microbiology Spectrum* **3**(2).

Daborn, P. J., Yen, J. L., Bogwitz, M. R., Le Goff, G., Feil, E., Jeffers, S., Tijet, N., Perry, T., Heckel, D., Batterham, P., Feyereisen, R., Wilson, T. G. and ffrench Constant, R. H. (2002), 'A single p450 allele associated with insecticide resistance in *Drosophila*', *Science* **297**(5590), 2253–2256.

Dai, J., Xie, W., Brady, T. L., Gao, J. and Voytas, D. F. (2007), 'Phosphorylation regulates integration of the yeast *Ty5* retrotransposon into heterochromatin', *Mol Cell* **27**(2), 289–299.

Dangel, A. W., Baker, B. J., Mendoza, A. R. and Yu, C. Y. (1995), 'Complement component c4 gene intron 9 as a phylogenetic marker for primates: long terminal repeats of the endogenous retrovirus erv-k(c4) are a molecular clock of evolution', *Immunogenetics* **42**(1), 41–52.

Daniels, S. B., Peterson, K. R., Strausbaugh, L. D., Kidwell, M. G. and Chovnick, A. (1990), 'Evidence for horizontal transmission of the *P* transposable element between *Drosophila* species', *Genetics* **124**(2), 339–355.

Darling, A. E., Mau, B. and Perna, N. T. (2010), 'progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement', *PLoS One* **5**(6), e11147.

de Araujo, P. G., Rossi, M., de Jesus, E. M., Saccaro, N. L., J., Kajihara, D., Massa, R., de Felix, J. M., Drummond, R. D., Falco, M. C., Chabregas, S. M., Ulian, E. C., Menossi, M. and Van Sluys, M. A. (2005), 'Transcriptionally active transposable elements in recent hybrid sugarcane', *Plant J* **44**(5), 707–17.

de Boer, J. G., Yazawa, R., Davidson, W. S. and Koop, B. F. (2007), 'Bursts and horizontal evolution of transposons in the speciation of pseudotetraploid salmonids', *BMC Genomics* **8**(422).

de la Chaux, N. and Wagner, A. (2009), 'Evolutionary dynamics of the LTR retrotransposons *roo* and *rooA* inferred from twelve complete *Drosophila* genomes', *BMC Evol Biol* **9**(205).

de Lucca Jr., M., Carareto, C. M. A. and Ceron, C. R. (2007), 'Distribution of the Bari-I transposable element in stable hybrid strains between *Drosophila melanogaster* and *Drosophila simulans* and in Brazillian populations of these species', *Genetics and Molecular Biology* **30**(3), 676–680.

de Setta, N., Van Sluys, M. A., Capy, P. and Carareto, C. M. (2009), 'Multiple invasions of *Gypsy* and *Micropia* retroelements in genus *Zaprionus* and *melanogaster* subgroup of the genus *Drosophila*', *BMC Evol Biol* **9**(279).

de Setta, N., Van Sluys, M. A., Capy, P. and Carareto, C. M. (2011), '*Copia* retrotransposon in the *Zaprionus* genus: another case of transposable element sharing with the *Drosophila melanogaster* subgroup', *J Mol Evol* **72**(3), 326–338.

Delneri, D., Colson, I., Grammenoudi, S., Roberts, I. N., Louis, E. J. and Oliver, S. G. (2003), 'Engineering evolution to study speciation in yeasts', *Nature* **422**(6927), 68–72.

Diao, X., Freeling, M. and Lisch, D. (2006), 'Horizontal transfer of a plant transposon', *PLoS Biol* **4**(1), e5.

Dietrich, F. S., Voegeli, S., Brachat, S., Lerch, A., Gates, K., Steiner, S., Mohr, C., Pohlmann, R., Luedi, P., Choi, S., Wing, R. A., Flavier, A., Gaffney, T. D. and Philippsen, P. (2004), 'The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome', *Science* **304**(5668), 304–307.

Dietrich, F. S., Voegeli, S., Kuo, S. and Philippsen, P. (2013), 'Genomes of *Ashbya* fungi isolated from insects reveal four mating-type loci, numerous translocations, lack of transposons, and distinct gene duplications', *G3 (Bethesda)* **3**(8), 1225–1239.

Dolgin, E. S. and Charlesworth, B. (2008), 'The effects of recombination rate on the distribution and abundance of transposable elements', *Genetics* **178**(4), 2169–2177.

Dong, C., Poulter, R. T. and Han, J. S. (2009), '*LINE*-like retrotransposition in *Saccharomyces cerevisiae*', *Genetics* **181**(1), 301–311.

Donnart, T., Piednoel, M., Higuet, D. and Bonnivard, E. (2017), 'Filamentous ascomycete genomes provide insights into *Copia* retrotransposon diversity in fungi', *BMC Genomics* **18**(1), 410.

Doolittle, W. F. and Sapienza, C. (1980), 'Selfish genes, the phenotype paradigm and genome evolution', *Nature* **284**(5757), 601–3.

Douady, C. J., Delsuc, F., Boucher, Y., Doolittle, W. F. and Douzery, E. J. (2003), 'Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability', *Mol Biol Evol* **20**(2), 248–54.

Downs, K. M., Brennan, G. and Liebman, S. W. (1985), 'Deletions extending from a single *Ty1* element in *Saccharomyces cerevisiae*', *Mol Cell Biol* **5**(12), 3451–7.

Dragon, F., Gallagher, J. E., Compagnone-Post, P. A., Mitchell, B. M., Porwancher, K. A., Wehner, K. A., Wormsley, S., Settlage, R. E., Shabanowitz, J., Osheim, Y., Beyer, A. L., Hunt, D. F. and Baserga, S. J. (2002), 'A large nucleolar U3 ribonucleoprotein required for 18S ribosomal rna biogenesis', *Nature* **417**(6892), 967–70.

Drinnenberg, I. A., Weinberg, D. E., Xie, K. T., Mower, J. P., Wolfe, K. H., Fink, G. R. and Bartel, D. P. (2009), 'RNAi in budding yeast', *Science* **326**(5952), 544–550.

Drozdova, P. B., Tarasov, O. V., Matveenko, A. G., Radchenko, E. A., Sopova, J. V., Polev, D. E., Inge-Vechtomov, S. G. and Dobrynin, P. V. (2016), 'Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strains of the Peterhof Genetic Collection', *PLoS One* **11**(5), e0154722.

Du, Z. and Li, L. (2014), 'Investigating the interactions of yeast prions: [SWI+], [PSI+], and [PIN+]', *Genetics* **197**(2), 685–700.

Dudley, A. M., Gansheroff, L. J. and Winston, F. (1999), 'Specific components of the SAGA complex are required for Gcn4- and Gcr1-mediated activation of the his4-912delta promoter in *Saccharomyces cerevisiae*', *Genetics* **151**(4), 1365–78.

Dujon, B. (2006), 'Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution', *Trends Genet* **22**(7), 375–387.

Dujon, B. A. and Louis, E. J. (2017), 'Genome diversity and evolution in the budding yeasts (Saccharomycotina)', *Genetics* **206**(2), 717–750.

Dunham, M. J., Badrane, H., Ferea, T., Adams, J., Brown, P. O., Rosenzweig, F. and Botstein, D. (2002), 'Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*', *Proc Natl Acad Sci U S A* **99**(25), 16144–16149.

Dunn, B., Levine, R. P. and Sherlock, G. (2005), 'Microarray karyotyping of commercial wine yeast strains reveals shared, as well as unique, genomic signatures', *BMC Genomics* **6**(53).

Dunn, B., Paulish, T., Stanbery, A., Piotrowski, J., Koniges, G., Kroll, E., Louis, E. J., Liti, G., Sherlock, G. and Rosenzweig, F. (2013), 'Recurrent rearrangement during adaptive evolution in an interspecific yeast hybrid suggests a model for rapid introgression', *PLoS Genet* **9**(3), e1003366.

Dunn, B., Richter, C., Kvitek, D. J., Pugh, T. and Sherlock, G. (2012), 'Analysis of the *Saccharomyces cerevisiae* pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments', *Genome Res* **22**(5), 908–924.

Dunn, T. M., Haak, D., Monaghan, E. and Beeler, T. J. (1998), 'Synthesis of monohydroxylated inositolphosphorylceramide (IPC-C) in *Saccharomyces cerevisiae* requires Scs7p, a protein with both a cytochrome b5-like domain and a hydroxylase/desaturase domain', *Yeast* **14**(4), 311–321.

Dunthorn, M., Kauserud, H., Bass, D., Mayor, J. and Mahe, F. (2017), 'Yeasts dominate soil fungal communities in three lowland Neotropical rainforests', *Environ Microbiol Rep* **9**(5), 668–675.

Dupeyron, M., Leclercq, S., Cerveau, N., Bouchon, D. and Gilbert, C. (2014), 'Horizontal transfer of transposons between and within crustaceans and insects', *Mob DNA* **5**(1), 4.

Dupuy, C., Periquet, G., Serbielle, C., Bezier, A., Louis, F. and Drezen, J. M. (2011), 'Transfer of a chromosomal *Maverick* to endogenous bracovirus in a parasitoid wasp', *Genetica* **139**(4), 489–96.

Eichten, S. R., Ellis, N. A., Makarevitch, I., Yeh, C. T., Gent, J. I., Guo, L., McGinnis, K. M., Zhang, X., Schnable, P. S., Vaughn, M. W., Dawe, R. K. and Springer, N. M. (2012), 'Spreading of heterochromatin is limited to specific families of maize retrotransposons', *PLoS Genet* **8**(12), e1003127.

Eickbush, T. H. and Jamburuthugoda, V. K. (2008), 'The diversity of retrotransposons and the properties of their reverse transcriptases', *Virus Res* **134**(1-2), 221–234.

El Baidouri, M., Carpentier, M. C., Cooke, R., Gao, D., Lasserre, E., Llauro, C., Mirouze, M., Picault, N., Jackson, S. A. and Panaud, O. (2014), 'Widespread and frequent horizontal transfers of transposable elements in plants', *Genome Res* **24**(5), 831–838.

Elbarbary, R. A., Lucas, B. A. and Maquat, L. E. (2016), 'Retrotransposons as regulators of gene expression', *Science* **351**(6274), aac7247.

Elder, R. T., Loh, E. Y. and Davis, R. W. (1983), 'RNA from the yeast transposable element *Ty1* has both ends in the direct repeats, a structure similar to retrovirus RNA', *Proc Natl Acad Sci U S A* **80**(9), 2432–6.

Elion, E. A. and Warner, J. R. (1984), 'The major promoter element of rRNA transcription in yeast lies 2 kb upstream', *Cell* **39**(3 Pt 2), 663–673.

Engel, S. R., Dietrich, F. S., Fisk, D. G., Binkley, G., Balakrishnan, R., Constanzo, M., Dwight, S. S., Hitz, B. C., Karra, K., Nash, R. S., Weng, S., Wong, E. D., Lloyd, P., Skrzypek, M. S., Miyasato, S. R., Simison, M. and Cherry, J. M. (2014), 'The reference genome sequence of *Saccharomyces cerevisiae*: Then and now', *G3 (Bethesda)* **4**(3), 389–398.

Errede, B., Cardillo, T. S., Sherman, F., Dubois, E., Deschamps, J. and Wiame, J. M. (1980), 'Mating signals control expression of mutations resulting from insertion of a transposable repetitive element adjacent to diverse yeast genes', *Cell* **22**(2 Pt 2), 427–436.

Errede, B., Company, M., Ferchak, J. D., Hutchison, C. A., r. and Yarnell, W. S. (1985), 'Activation regions in a yeast transposon have homology to mating type control sequences and to mammalian enhancers', *Proc Natl Acad Sci U S A* **82**(16), 5423–5427.

Errede, B., Company, M. and Hutchinson III, C. (1987), '*Ty1* sequence with enhancer and mating-type-dependent regulatory activities', *Mol Cell Biol* **7**(1), 258–265.

Esberg, A., Muller, L. A. and McCusker, J. H. (2011), 'Genomic structure of and genome-wide recombination in the *Saccharomyces cerevisiae* s288c progenitor isolate em93', *PLoS One* **6**(9), e25211.

Escher, D., Bodmer-Glavas, M., Barberis, A. and Schaffner, W. (2000), 'Conservation of glutamine-rich transactivation function between yeast and humans', *Mol Cell Biol* **20**(8), 2774–2782.

Farabaugh, P. J. and Fink, G. R. (1980), 'Insertion of the eukaryotic transposable element *Ty1* creates a 5-base pair duplication', *Nature* **286**(5771), 352–356.

Fay, J. C. and Benavides, J. A. (2005), 'Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*', *PLoS Genet* **1**(1), 66–71.

Fay, J. C., McCullough, H., Sniegowski, P. D. and Eisen, M. (2004), 'Population genetic variation in gene expression is associated with phenotypic variation in *Saccharomyces cerevisiae*', *Genome Biology* **5**(26).

Felsenstein, J. (1985), 'Confidence limits on phylogenies: An approach using the bootstrap', *Evolution* **39**(4), 783–791.

Feng, G., Leem, Y. E. and Levin, H. L. (2013), 'Transposon integration enhances expression of stress response genes', *Nucleic Acids Res* **41**(2), 775–789.

Feng, Y., Zhang, Y., Ying, C., Wang, D. and Du, C. (2015), 'Nanopore-based fourth-generation sequencing technology', *Genomics Proteomics Bioinformatics* **13**(1), 4–16.

Ferguson-Yankey, S. R., Skrzypek, M. S., Lester, R. L. and Dickson, R. C. (2002), 'Mutant analysis reveals complex regulation of sphingolipid long chain base phosphates and long chain bases during heat stress in yeast', *Yeast* **19**(7), 573–86.

Fietto, J. L., Araujo, R. S., Valadao, F. N., Fietto, L. G., Brandao, R. L., Neves, M. J., Gomes, F. C., Nicoli, J. R. and Castro, I. M. (2004), 'Molecular and physiological comparisons between *Saccharomyces cerevisiae* and *Saccharomyces boulardii*', *Can J Microbiol* **50**(8), 615–621.

Finatto, T., de Oliveira, A. C., Chaparro, C., da Maia, L. C., Farias, D. R., Woyann, L. G., Mistura, C. C., Soares-Bresolin, A. P., Llauro, C., Panaud, O. and Picault, N. (2015), 'Abiotic stress and genome dynamics: specific genes and transposable elements response to iron excess in rice', *Rice (N Y)* **8**(13).

Fingerman, E. G., Dombrowski, P. G., Francis, C. A. and Sniegowski, P. D. (2003), 'Distribution and sequence analysis of a novel *Ty3*-like element in natural *Saccharomyces paradoxus* isolates', *Yeast* **20**(9), 761–770.

Fink, G. R., Boeke, J. D. and Garfinkel, D. J. (1986), 'The mechanism and consequences of retrotransposition', *Trends Genet* **2**, 118–123.

Fitzpatrick, D. A. (2012), 'Horizontal gene transfer in fungi', *FEMS Microbiol Lett* **329**(1), 1–8.

Fortune, P. M., Roulin, A. and Panaud, O. (2008), 'Horizontal transfer of transposable elements in plants', *Commun Integr Biol* **1**(1), 74–77.

Frame, I. G., Cutfield, J. F. and Poulter, R. T. (2001), 'New BEL-like LTR-retrotransposons in *Fugu rubripes, Caenorhabditis elegans*, and *Drosophila melanogaster*', *Gene* **263**(1-2), 219–230.

Franco-Duarte, R., Bigey, F., Carreto, L., Mendes, I., Dequin, S., Santos, M. A., Pais, C. and Schuller, D. (2015), 'Intrastrain genomic and phenotypic variability of the commercial *Saccharomyces cerevisiae* strain Zymaflore VL1 reveals microevolutionary adaptation to vineyard environments', *FEMS Yeast Res* **15**(6).

Franco-Duarte, R., Mendes, I., Gomes, A. C., Santos, M. A., de Sousa, B. and Schuller, D. (2011), 'Genotyping of *Saccharomyces cerevisiae* strains by interdelta sequence typing using automated microfluidics', *Electrophoresis* **32**(12), 1447–1455.

Fraser, H. B., Moses, A. M. and Schadt, E. E. (2010), 'Evidence for widespread adaptive evolution of gene expression in budding yeast', *Proc Natl Acad Sci U S A* **107**(7), 2977–2982.

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. and Dubchak, I. (2004), 'VISTA: computational tools for comparative genomics', *Nucleic Acids Res* **32**(Web Server issue), W273–9.

Fritsch, E. S., Schacherer, J., Bleykasten-Grosshans, C., Souciet, J. L., Potier, S. and de Montigny, J. (2009), 'Influence of genetic background on the occurrence of chromosomal rearrangements in *Saccharomyces cerevisiae*', *BMC Genomics* **10**, 99.

Frost, L. S., Leplae, R., Summers, A. O. and Toussaint, A. (2005), 'Mobile genetic elements: the agents of open source evolution', *Nat Rev Microbiol* **3**(9), 722–732.

Fu, Y. and Li, W. (1993), 'Statistical tests of neutrality of mutations', *Genetics* **133**(3), 693–709.

Fulton, A., Rathjen, P., Kingman, S. and Kingsman, A. (1988), 'Upstream and downstream transcriptional control signals in the yeast retrotransposon, *Ty*', *Nucleic Acids Res* **16**(12).

Gabriel, A. and Mules, E. (1999), 'Fidelity of retrotransposon replication', *Annals New York Academy of Sciences* **870**, 108–118.

Gabriel, A., Willems, M., Mules, E. and Boeke, J. D. (1996), 'Replication infidelity during a single cycle of *Ty1* retrotransposition', *PNAS* **93**.

Gafner, J. and Philippsen, P. (1980), 'The yeast transposon *Ty1* generates duplications of target on insertion', *Nature* **286**(5771), 414–418.

Gai, X. and Voytas, D. F. (1998), 'A single amino acid change in the yeast retrotransposon *Ty5* abolishes targeting to silent chromatin', *Mol Cell* **1**(7), 1051–5.

Galeote, V., Bigey, F., Beyne, E., Novo, M., Legras, J. L., Casaregola, S. and Dequin, S. (2011), 'Amplification of a *Zygosaccharomyces bailii* segment in wine yeast genomes by extrachromosomal circular formation', *PLoS One* **6**(3), e17872.

Gao, D., Chu, Y., Xia, H., Xu, C., Heyduk, K., Abernathy, B., Ozias-Akins, P., Leebens-Mack, J. H. and Jackson, S. A. (2018), 'Horizontal Transfer of Non-LTR Retrotransposons from Arthropods to Flowering Plants', *Mol Biol Evol* **35**(2), 354–364.

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., Paulsen, I. T., James, K., Eisen, J. A., Rutherford, K., Salzberg, S. L., Craig, A., Kyes, S., Chan, M. S., Nene, V., Shallom, S. J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M. W., Vaidya, A. B., Martin, D. M., Fairlamb, A. H., Fraunholz, M. J., Roos, D. S., Ralph, S. A., McFadden, G. I., Cummings, L. M., Subramanian, G. M., Mungall, C., Venter, J. C., Carucci, D. J., Hoffman, S. L., Newbold, C., Davis, R. W., Fraser, C. M. and Barrell, B. (2002), 'Genome sequence of the human malaria parasite *Plasmodium falciparum*', *Nature* **419**(6906), 498–511.

Garfinkel, D. J., Hedge, A., Youngren, S. and Copeland, T. (1991), 'Proteolytic processing of *pol-TYB* proteins from the yeast retrotransposon *Ty1*', *J Virol* **65**(9).

Garfinkel, D. J., Mastrangelo, M., Sanders, N., Shafer, B. and Strathern, J. (1988), 'Transposon tagging using *Ty* elements in yeast', *Genetics* **120**.

Garfinkel, D. J., Nyswaner, K. M., Stefanisko, K. M., Chang, C. and Moore, S. P. (2005), '*Ty1* copy number dynamics in *Saccharomyces*', *Genetics* **169**(4), 1845–57.

Garfinkel, D. J., Nyswaner, K. M., Wang, J. and Cho, J. (2003), 'Post-transcriptional cosuppression of *Ty1* retrotransposition', *Genetics* **165**.

Garfinkel, D. J., Stefanisko, K. M., Nyswaner, K. M., Moore, S. P., Oh, J. and Hughes, S. H. (2006), 'Retrotransposon suicide: formation of *Ty1* circles and autointegration via a central flap', *J Virol* **80**(24), 11920–34.

Garud, N. R., Messer, P. W., Buzbas, E. O. and Petrov, D. A. (2015), 'Recent selective sweeps in north american *Drosophila melanogaster* show signatures of soft sweeps', *PLoS Genet* **11**(2), e1005004.

Gayevskiy, V. and Goddard, M. R. (2016), '*Saccharomyces eubayanus* and *Saccharomyces arboricola* reside in North Island native New Zealand forests', *Environ Microbiol* **18**(4), 1137–47.

Gibson, B., Krogerus, K., Ekberg, J., Monroux, A., Mattinen, L., Rautio, J. and Vidgren, V. (2015), 'Variation in alpha-acetolactate production within the hybrid lager yeast group *Saccharomyces pastorianus* and affirmation of the central role of the ILV6 gene', *Yeast* **32**(1), 301–16.

Gilbert, C., Chateigner, A., Ernenwein, L., Barbe, V., Bezier, A., Herniou, E. A. and Cordaux, R. (2014), 'Population genomics supports baculoviruses as vectors of horizontal transfer of insect transposons', *Nat Commun* **5**, 3348.

Gilbert, C. and Cordaux, R. (2013), 'Horizontal transfer and evolution of prokaryote transposable elements in eukaryotes', *Genome Biol Evol* **5**(5), 822–32.

Gilbert, C., Hernandez, S. S., Flores-Benabib, J., Smith, E. N. and Feschotte, C. (2012), 'Rampant horizontal transfer of *SPIN* transposons in squamate reptiles', *Mol Biol Evol* **29**(2), 503–15.

Gilbert, C., Schaack, S., Pace, J. K., n., Brindley, P. J. and Feschotte, C. (2010), 'A role for host-parasite interactions in the horizontal transfer of transposons across phyla', *Nature* **464**(7293), 1347–50.

Gilbert, C., Waters, P., Feschotte, C. and Schaack, S. (2013), 'Horizontal transfer of *OC1* transposons in the Tasmanian devil', *BMC Genomics* **14**, 134.

Gillespie, J. H. (2000), 'Genetic drift in an infinite population. the pseudohitchhiking model', *Genetics* **155**(2), 909–19.

Ginzburg, L. R., Bingham, P. M. and Yoo, S. (1984), 'On the theory of speciation induced by transposable elements', *Genetics* **107**(2), 331–41.

Göke, J. and Ng, H. H. (2016), 'CTRL+INSERT: retrotransposons and their contribution to regulation and innovation of the transcriptome', *EMBO Rep* **17**(8), 1131–44.

Glushakova, A. M., Ivannikova Iu, V., Naumova, E. S., Chernov, I. and Naumov, G. I. (2007), 'Massive isolation and identification of *Saccharomyces paradoxus* yeasts from plant phyllosphere [in Russian]', *Mikrobiologiia* **76**(2), 236–42.

Goddard, M. R., Anfang, N., Tang, R., Gardner, R. C. and Jun, C. (2010), 'A distinct population of *Saccharomyces cerevisiae* in New Zealand: evidence for local dispersal by insects and human-aided global dispersal in oak barrels', *Environ Microbiol* **12**(1), 63–73.

Goddard, M. R. and Greig, D. (2015), '*Saccharomyces cerevisiae*: a nomadic yeast with no niche?', *FEMS Yeast Res* **15**(3).

Goel, A. and Pearlman, R. (1988), 'Transposable element mediated enhancement of gene expression in *Saccharomyces* involves sequence specific binding of a trans acting factor', *Mol Cell Biol* **8**(6), 2572–2580.

Goeman, J. J. and Solari, A. (2014), 'Multiple hypothesis testing in genomics', *Stat Med* **33**(11), 1946–78.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S. G. (1996), 'Life with 6000 genes', *Science* **274**(5287), 546, 563–7.

Gonçalves, J. W., Valiati, V. H., Delprat, A., Valente, V. L. and Ruiz, A. (2014), 'Structural and sequence diversity of the transposon *Galileo* in the *Drosophila willistoni* genome', *BMC Genomics* **15**, 792.

González, J., Karasov, T. L., Messer, P. W. and Petrov, D. A. (2010), 'Genome-wide patterns of adaptation to temperate environments associated with transposable elements in *Drosophila*', *PLoS Genet* **6**(4), e1000905.

González, J., Lenkov, K., Lipatov, M., Macpherson, J. M. and Petrov, D. A. (2008), 'High rate of recent transposable element-induced adaptation in *Drosophila melanogaster*', *PLoS Biol* **6**(10), e251.

González, P. and Lessios, H. A. (1999), 'Evolution of sea urchin retroviral-like (*SURL*) elements: evidence from 40 echinoid species', *Mol Biol Evol* **16**(7), 938–52.

González, S. S., Barrio, E., Gafner, J. and Querol, A. (2006), 'Natural hybrids from *Saccharomyces cerevisiae, Saccharomyces bayanus* and *Saccharomyces kudriavzevii* in wine fermentations', *FEMS Yeast Res* **6**(8), 1221–34.

Goodwin, T. J., Dalle Nogare, D. E., Butler, M. I. and Poulter, R. T. (2003), '*Ty3/gypsy*-like retrotransposons in *Candida albicans* and *Candida dubliniensis*: *Tca3* and *Tcd3*', *Yeast* **20**(6), 493–508.

Goodwin, T. J., Ormandy, J. E. and Poulter, R. T. (2001), 'L1-like non-LTR retrotransposons in the yeast *Candida albicans*', *Curr Genet* **39**(2), 83–91.

Goodwin, T. J. and Poulter, R. T. (2000), 'Multiple LTR-retrotransposon families in the asexual yeast *Candida albicans*', *Genome Res* **10**(2), 174–91.

Goto, K., Iwase, T., Kichise, K., Kitano, K., Totuka, A., Obata, T. and Hara, S. (1990), 'Isolation and properties of a chromosome-dependent KHR killer toxin in *Saccharomyces cerevisiae*', *Agric Biol Chem* **54**(2), 505–9.

Grandaubert, J., Lowe, R. G., Soyer, J. L., Schoch, C. L., Van de Wouw, A. P., Fudal, I., Robbertse, B., Lapalu, N., Links, M. G., Ollivier, B., Linglin, J., Barbe, V., Mangenot, S., Cruaud, C., Borhan, H., Howlett, B. J., Balesdent, M. H. and Rouxel, T. (2014), 'Transposable element-assisted evolution and adaptation to host plant within the *Leptosphaeria maculans-Leptosphaeria biglobosa* species complex of fungal pathogens', *BMC Genomics* **15**, 891.

Grandbastien, M. A., Audeon, C., Bonnivard, E., Casacuberta, J. M., Chalhoub, B., Costa, A. P., Le, Q. H., Melayah, D., Petit, M., Poncet, C., Tam, S. M., Van Sluys, M. A. and Mhiri, C. (2005), 'Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae', *Cytogenet Genome Res* **110**(1-4), 229–41.

Greig, D. (2009), 'Reproductive isolation in *Saccharomyces*', *Heredity (Edinb)* **102**(1), 39–44.

Greig, D., Borts, R. H., Louis, E. J. and Travisano, M. (2002), 'Epistasis and hybrid sterility in *Saccharomyces*', *Proc Biol Sci* **269**(1496), 1167–71.

Groth, C., Hansen, J. and Piskur, J. (1999), 'A natural chimeric yeast containing genetic material from three species', *Int J Syst Bacteriol* **49 Pt 4**, 1933–8.

Groth, S. B. and Blumenstiel, J. P. (2016), 'Horizontal transfer can drive a greater transposable element load in large populations', *J Hered* **108**(1), 36–44.

Guerreiro, M. P. (2014), 'Interspecific hybridization as a genomic stressor inducing mobilization of transposable elements in *Drosophila*', *Mob Genet Elements* **4**, e34394.

Guio, L., Barron, M. G. and Gonzalez, J. (2014), 'The transposable element *Bari-Jheh* mediates oxidative stress response in *Drosophila*', *Mol Ecol* **23**(8), 2020–30.

Han, F. P., Liu, Z. L., Tan, M., Hao, S., Fedak, G. and Liu, B. (2004), 'Mobilized retrotransposon Tos17 of rice by alien DNA introgression transposes into genes and causes structural and methylation alterations of a flanking genomic region', *Hereditas* **141**(3), 243–51.

Hani, J. and Feldmann, H. (1998), 'tRNA genes and retroelements in the yeast genome', *Nucleic Acids Res* **26**(3), 689–696.

Hansen, J. and Kielland-Brandt, M. C. (1994), '*Saccharomyces carlsbergensis* contains two functional MET2 alleles similar to homologues from *S. cerevisiae* and *S. monacensis*', *Gene* **140**(1), 33–40.

Harsay, E. and Schekman, R. (2007), 'Avl9p, a member of a novel protein superfamily, functions in the late secretory pathway', *Mol Biol Cell* **18**(4), 1203–19.

Hartl, D. L. and Clark, A. (2007), *Principles of Population Genetics*, 4th edn, Sinauer, Massachusetts.

Hartl, D. L., Lohe, A. R. and Lozovskaya, E. R. (1997), 'Modern thoughts on an ancyent marinere: function, evolution, regulation', *Annu Rev Genet* **31**, 337–58.

Hartwell, L. H., McLaughlin, C. S. and Warner, J. R. (1970), 'Identification of ten genes that control ribosome formation in yeast', *Mol Gen Genet* **109**(1), 42–56.

Havecker, E. R., Gao, G. and Voytas, D. F. (2004), 'The diversity of LTR retrotransposons', *Genome Biol* **5**(6).

Hazzouri, K. M., Mohajer, A., Dejak, S. I., Otto, S. P. and Wright, S. I. (2008), 'Contrasting patterns of transposable-element insertion polymorphism and nucleotide diversity in autotetraploid and allotetraploid *Arabidopsis* species', *Genetics* **179**(1), 581–92.

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), 522–9.

Herskowitz, I. (1988), 'Life cycle of the budding yeast *Saccharomyces cerevisiae*', *Microbiol Rev* **52**(4), 536–53.

Hickey, D. A. (1982), 'Selfish DNA: a sexually-transmitted nuclear parasite', *Genetics* **101**(3-4), 519–31.

Hill, D. E., Hope, I. A., Macke, J. P. and Struhl, K. (1986), 'Saturation mutagenesis of the yeast his3 regulatory site: requirements for transcriptional induction and for binding by *GCN4* activator protein', *Science* **234**(4775), 451–7.

Hillis, D. M. (1996), 'Inferring complex phylogenies', *Nature* **383**(6596), 130–1.

Hillis, D. M. (1998), 'Taxonomic sampling, phylogenetic accuracy, and investigator bias', *Syst Biol* **47**(1), 3–8.

Hillis, D. M. and Bull, J. J. (1993), 'An empirical test of bootstrapping as a method for assessing confidence in phylogentic analaysis', *Syst Biol* **42**(2), 182–192.

Hirsch, C. D. and Springer, N. M. (2017), 'Transposable element influences on gene expression in plants', *Biochim Biophys Acta* **1860**(1), 157–165.

Hittinger, C. T. (2013), '*Saccharomyces* diversity and evolution: a budding model genus', *Trends Genet* **29**(5), 309–17.

Hittinger, C. T., Gonçalves, P., Sampaio, J. P., Dover, J., Johnston, M. and Rokas, A. (2010), 'Remarkably ancient balanced polymorphisms in a multi-locus gene network', *Nature* **464**(7285), 54–8.

Ho, K. L., Ma, L., Cheung, S., Manhas, S., Fang, N., Wang, K., Young, B., Loewen, C., Mayor, T. and Measday, V. (2015), 'A role for the budding yeast separase, esp1, in *Ty1* element retrotransposition', *PLoS Genet* **11**(3), e1005109.

Hoang, M. L., Tan, F. J., Lai, D. C., Celniker, S. E., Hoskins, R. A., Dunham, M. J., Zheng, Y. and Koshland, D. (2010), 'Competitive repair by naturally dispersed repetitive DNA during non-allelic homologous recombination', *PLoS Genet* **6**(12), e1001228.

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., Poss, M. L., Reed, L. K., Storfer, A. and Whitlock, M. C. (2016), 'Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions', *Am Nat* **188**(4), 379–97.

Horn, A. V. and Han, J. S. (2016), 'Line retrotransposition assays in *Saccharomyces cerevisiae*', *Methods Mol Biol* **1400**, 131–7.

Hosid, E., Brodsky, L., Kalendar, R., Raskina, O. and Belyayev, A. (2012), 'Diversity of long terminal repeat retrotransposon genome distribution in natural populations of the wild diploid wheat *Aegilops speltoides*', *Genetics* **190**(1), 263–74.

Huelsenbeck, J. P. and Crandall, K. A. (1997), 'Phylogeny estimation and hypothesis testing using maximum likelihood', *Annual Review of Ecology and Systematics* **28**, 437–466.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R. and Bollback, J. P. (2001), 'Bayesian inference of phylogeny and its impact on evolutionary biology', *Science* **294**(5550), 2310–4.

Hug, A. and Feldmann, H. (1996), 'Yeast retrotransposon *Ty4*: the majority of the rare transcripts lack a u3-r sequence', *Nucleic Acids Res* **24**(12).

Hurst, G. D. D. and Schilthuizen, M. (1998), 'Selfish genetic elements and speciation', *Heredity (Edinb)* pp. 2–8.

Huson, D. H. and Bryant, D. (2006), 'Application of phylogenetic networks in evolutionary studies', *Mol Biol Evol* **23**(2), 254–67.

Hwang, G. W., Furuoya, Y., Hiroshima, A., Furuchi, T. and Naganuma, A. (2005), 'Overexpression of bop3 confers resistance to methylmercury in *Saccharomyces cerevisiae* through interaction with other proteins such as fkh1, rts1, and msn2', *Biochem Biophys Res Commun* **330**(2), 378–85.

Ibeas, J. I. and Jimenez, J. (1996), 'Genomic complexity and chromosomal rearrangements in wine-laboratory yeast hybrids', *Curr Genet* **30**(5), 410–6.

Istace, B., Friedrich, A., d'Agata, L., Faye, S., Payen, E., Beluche, O., Caradec, C., Davidas, S., Cruaud, C., Liti, G., Lemainque, A., Engelen, S., Wincker, P., Schacherer, J. and Aury, J. M. (2017), 'de novo assembly and population genomic survey of natural yeast isolates with the oxford nanopore minion sequencer', *Gigascience* **6**(2), 1–13.

Ivancevic, A. M., Walsh, A. M., Kortschak, R. D. and Adelson, D. L. (2013), 'Jumping the fine line between species: horizontal transfer of transposable elements in animals catalyses genome evolution', *Bioessays* **35**(12), 1071–82.

James, S. A., Cai, J., Roberts, I. N. and Collins, M. (1997), 'A phylogenetic analysis of the genus *Saccharomyces* based on 18s rRNA gene sequences: Description of *Saccharomyces kunashirensis sp. nov.* and *Saccharomyces martiniae sp. nov*.', *Int J Sys Bact* **47**(2).

Janetzky, B. and Lehle, L. (1992), '*Ty4*, a new retrotransposon from *Saccharomyces cerevisiae*, flanked by tau-elements', *J Biol Chem* **267**(28).

Jelinsky, S. A., Estep, P., Church, G. M. and Samson, L. D. (2000), 'Regulatory networks revealed by transcriptional profiling of damaged *Saccharomyces cerevisiae* cells: Rpn4 links base excision repair with proteasomes', *Mol Cell Biol* **20**(21), 8157–67.

Jenkins, G. M., Richards, A., Wahl, T., Mao, C., Obeid, L. and Hannun, Y. (1997), 'Involvement of yeast sphingolipids in the heat stress response of *Saccharomyces cerevisiae*', *J Biol Chem* **272**(51), 32566–72.

Jiang, Y. W. (2008), 'An essential role of tap42-associated pp2a and 2a-like phosphatases in *Ty1* transcriptional silencing of *S. cerevisiae*', *Yeast* **25**(10), 755–64.

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Wang, B., Campbell, M. S., Stein, J. C., Wei, X., Chin, C. S., Guill, K., Regulski, M., Kumari, S., Olson, A., Gent, J., Schneider, K. L., Wolfgruber, T. K., May, M. R., Springer, N. M., Antoniou, E., McCombie, W. R., Presting, G. G., McMullen, M., Ross-Ibarra, J., Dawe, R. K., Hastie, A., Rank, D. R. and Ware, D. (2017), 'Improved maize reference genome with single-molecule technologies', *Nature* **546**(7659), 524–527.

Johnson, L. J., Giraud, T., Anderson, R. and Hood, M. E. (2010), 'The impact of genome defense on mobile elements in *Microbotryum*', *Genetica* **138**(3), 313–9.

Johnson, L., Koufopanou, V., Goddard, M., Hetherington, R., Schafer, S. and Burt, A. (2003), 'Population genetics of the wild yeast *Saccharomyces paradoxus*', *Genetics* **160**, 43–52.

Johnson, S. P. and Warner, J. R. (1989), 'Unusual enhancer function in yeast rRNA transcription', *Molecular and Cellular Biology* **9**(11), 4986–4993.

Jordan, I. K. and McDonald, J. F. (1999*a*), 'Phylogenetic perspective reveals abundant *Ty1/Ty2* hybrid elements in the *Saccharomyces cerevisiae* genome', *Mol Biol Evol* **16**(3), 419–22.

Jordan, I. K., Rogozin, I. B., Glazko, G. V. and Koonin, E. V. (2003), 'Origin of a substantial fraction of human regulatory sequences from transposable elements', *Trends Genet* **19**(2), 68–72.

Jordan, I. and McDonald, J. F. (1998), 'Evidence for the role of recombination in the regulatory evolution of *Ty* elements', *J Mol Evol* **47**(1), 14–20.

Jordan, I. and McDonald, J. F. (1999*b*), 'Comparative genomics and evolutionary dynamics of *Saccharomyces cerevisiae Ty* elements', *Genetica* **107**(1-3), 3–13.

Jordan, I. and McDonald, J. F. (1999*c*), 'Tempo and mode of *Ty* element evolution in *Saccharomyces cerevisiae*', *Genetics* **151**.

Josefsson, C., Dilkes, B. and Comai, L. (2006), 'Parent-dependent loss of gene silencing during interspecies hybridization', *Curr Biol* **16**(13), 1322–8.

Kaneko, Y. and Banno, I. (1991), 'Re-examination of *Saccharomyces bayanus* strains by DNA-hybridization and electrophoretic karyotyping', *IFO Res Comm* **15**, 30–41.

Kapitonov, V. V. and Jurka, J. (2007), 'Helitrons on a roll: eukaryotic rolling-circle transposons', *Trends Genet* **23**(10), 521–9.

Kashkush, K., Feldman, M. and Levy, A. A. (2003), 'Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat', *Nat Genet* **33**(1), 102–6.

Katoh, K. and Standley, D. M. (2013), 'MAFFT multiple sequence alignment software version 7: improvements in performance and usability', *Mol Biol Evol* **30**(4), 772–80.

Kawakami, T., Strakosh, S. C., Zhen, Y. and Ungerer, M. C. (2010), 'Different scales of *Ty1/copia*-like retrotransposon proliferation in the genomes of three diploid hybrid sunflower species', *Heredity (Edinb)* **104**(4), 341–50.

Kazazian, H. H., J. (2004), 'Mobile elements: drivers of genome evolution', *Science* **303**(5664), 1626–32.

Kück, P. and Wägele, J. W. (2015), 'Plesiomorphic character states cause systematic errors in molecular phylogenetic analyses: a simulation study', *Cladistics* **32**, 461–478.

Ke, N., Irwin, P. and Voytas, D. F. (1997), 'The pheromone response pathway activates transcription of *Ty5* retrotransposons located within silent chromatin of *Saccharomyces cerevisia*e', *EMBO Journal* **16**(20), 6272–6280.

Ke, N. and Voytas, D. F. (1997), 'High frequency crecombination of the *Saccharomyces* retrotransposon *Ty5*: The LTR mediates formation of tandem elements', *Genetics* **147**, 545–556.

Keeling, P. J. and Palmer, J. D. (2008), 'Horizontal gene transfer in eukaryotic evolution', *Nat Rev Genet* **9**(8), 605–18.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B. and Lander, E. (2003), 'Sequencing and comparison of yeast species to identify genes and regulatory elements', *Nature* **429**.

Khatri, I., Tomar, R., Ganesan, K., Prasad, G. S. and Subramanian, S. (2017), 'Complete genome sequence and comparative genomics of the probiotic yeast *Saccharomyces boulardii*', *Sci Rep* **7**(1), 371.

Khoury, G. and Gruss, P. (1983), 'Enhancer elements', *Cell* **33**(2), 313–4.

Kidwell, M. G. and Lisch, D. (1997), 'Transposable elements as sources of variation in animals and plants', *Proc Natl Acad Sci U S A* **94**(15), 7704–11.

Kidwell, M. G. and Lisch, D. R. (2000), 'Transposable elements and host genome evolution', *Trends Ecol Evol* **15**(3), 95–99.

Kijima, T. and Innan, H. (2010), 'On the estimation of the insertion time of LTR retrotransposable elements', *Mol Biol Evol* **27**, 896–904.

Kim, A., Terzian, C., Santamaria, P., Pelisson, A., Purd'homme, N. and Bucheton, A. (1994), 'Retroviruses in invertebrates: the *gypsy* retrotransposon is apparently an infectious retrovirus of *Drosophila melanogaster*', *Proc Natl Acad Sci U S A* **91**(4), 1285–9.

Kim, J., Vanguri, S., Boeke, J. D., Gabriel, A. and Voytas, D. F. (1998), 'Transposable elements and genome organization: A comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence', *Genome Res* **8**(5), 464–478.

Kim, Y. and Stephan, W. (2002), 'Detecting a local signature of genetic hitchhiking along a recombining chromosome', *Genetics* **160**, 765–777.

Kimura, K. and Kidwell, M. G. (1994), 'Differences in *P* element population dynamics between the sibling species *Drosophila melanogaster* and *Drosophila simulans*', *Genet Res* **63**(1), 27–38.

Kimura, M. (1968), 'Evolutionary rate at the molecular level', *Nature* **217**(5129), 624–6.

Kinsey, P. and Sandmeyer, S. (1991), 'Adjacent pol II and pol III promoters: transcription of the yeast retrotransposon *Ty3* and a target tRNA gene', *Nucleic Acids Res* **19**(6).

Kirchner, J. and Sandmeyer, S. (1993), 'Proteolytic processing of *Ty3* proteins is required for transposition', *J Virol* **67**(1), 19–28.

Kirkland, T. N., Muszewska, A. and Stajich, J. E. (2018), 'Analysis of transposable elements in coccidioides species', *J Fungi (Basel)* **4**(1).

Klein, H. and Petes, T. D. (1984), 'Genetic mapping of *Ty* elements in *Saccharomyces cerevisiae*', *Mol Cell Biol* **4**(2).

Knight, S. A., Labbe, S., Kwon, L. F., Kosman, D. J. and Thiele, D. J. (1996), 'A widespread transposable element masks expression of a yeast copper transport gene', *Genes and Development* **10**(15), 1917–1929.

Kofler, R., Betancourt, A. J. and Schlotterer, C. (2012), 'Sequencing of pooled samples (pool-seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*', *PLoS Genet* **8**(1), e1002487.

Kondrashov, F. A. (2012), 'Gene duplication as a mechanism of genomic adaptation to a changing environment', *Proc Biol Sci* **279**(1749), 5048–57.

Kornitzer, D., Raboy, B., Kulka, R. G. and Fink, G. R. (1994), 'Regulated degradation of the transcription factor gcn4', *EMBO Journal* **13**(24).

Kreitman, M. (2000), 'Methods to detect selection in populations with applications to the human', *Annu Rev Genomics Hum Genet* **1**, 539–59.

Kuhn, A., Ong, Y. M., Cheng, C. Y., Wong, T. Y., Quake, S. R. and Burkholder, W. F. (2014), 'Linkage disequilibrium and signatures of positive selection around *LINE*-1 retrotransposons in the human genome', *Proc Natl Acad Sci U S A* **111**(22), 8131–6.

Kupiec, M. and Petes, T. D. (1988*a*), 'Allelic and ectopic recombination between *Ty* elements in yeast', *Genetics* **119**.

Kupiec, M. and Petes, T. D. (1988*b*), 'Meiotic recombination between repeated transposable elements in *Saccharomyces cerevisiae*', *Mol Cell Biol* **8**(7).

Kurtzman, C. (2003), 'Phylogenetic circumscription of *Saccharomyces, Kluyveromyces* and other members of the Saccharomycetaceae, and the proposal of the new genera *Lachancea, Nakaseomyces, Naumovia, Vanderwaltozyma* and *Zygotorulaspora*.', *FEMS Yeast Research* **4**(3), 233–245.

Kurtzman, C., Fell, J. W. and Boekhout, T. (2011), *The Yeasts: A Taxonomic Study*, Elsevier Science, St. Louis.

Kurtzman, C. and Robnett, C. (2003), 'Phylogenetic relationships among yeasts of the *Saccharomyces* complex determined from multigene sequence analyses', *FEMS Yeast Research* **3**(4), 417–432.

Lalo, D., Steffan, J. S., Dodd, J. A. and Nomura, M. (1996), '*RRN11* encodes the third subunit of the complex containing Rrn6p and Rrn7p that is essential for the initiation of rDNA transcription by yeast RNA polymerase I', *J Biol Chem* **271**(35), 21062–7.

Landry, C. R., Townsend, J. P., Hartl, D. L. and Cavalieri, D. (2006), 'Ecological and evolutionary genomics of *Saccharomyces cerevisiae*', *Mol Ecol* **15**(3), 575–91.

Landry, J. R., Rouhi, A., Medstrand, P. and Mager, D. L. (2002), 'The Opitz syndrome gene mid1 is transcribed from a human endogenous retroviral promoter', *Mol Biol Evol* **19**(11), 1934–42.

Langley, C. H., Montgomery, E., Hudson, R., Kaplan, N. and Charlesworth, B. (1988), 'On the role of unequal exchange in the containment of transposable element copy number', *Genet Res* **52**(3), 223–35.

Laricchia, K. M., Zdraljevic, S., Cook, D. E. and Andersen, E. C. (2017), 'Natural variation in the distribution and abundance of transposable elements across the *Caenorhabditis elegans* species', *Mol Biol Evol* .

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. and Higgins, D. G. (2007), 'Clustal W and Clustal X version 2.0', *Bioinformatics* **23**(21), 2947–8.

Laten, H. M., Majumdar, A. and Gaucher, E. A. (1998), '*SIRE-1*, a *copia/Ty1*-like retroelement from soybean, encodes a retroviral envelope-like protein', *Proc Natl Acad Sci U S A* **95**(12), 6897–902.

Lauermann, V. and Boeke, J. D. (1997), 'Plus-strand strong-stop transfer in yeast *Ty* retrotransposons', *EMBO J* **16**(21), 6603–12.

Lawler, J. F., Haeusser, D. P., Dull, A., Boeke, J. D. and Keeney, J. B. (2002), '*Ty1* defect in proteolysis at high temperature', *Journal of Virology* **76**(9), 4233–4240.

Le Rouzic, A., Boutin, T. S. and Capy, P. (2007), 'Long-term evolution of transposable elements', *Proc Natl Acad Sci U S A* **104**(49), 19375–80.

Le Rouzic, A. and Capy, P. (2005), 'The first steps of transposable elements invasion: parasitic strategy vs. genetic drift', *Genetics* **169**(2), 1033–43.

Le Rouzic, A. and Capy, P. (2006), 'Population genetics models of competition between transposable element subfamilies', *Genetics* **174**(2), 785–93.

Le Rouzic, A. and Deceliere, G. (2005), 'Models of the population genetics of transposable elements', *Genet Res* **85**(3), 171–81.

Le, T. N., Miyazaki, Y., Takuno, S. and Saze, H. (2015), 'Epigenetic regulation of intragenic transposable elements impacts gene transcription in *Arabidopsis thaliana*', *Nucleic Acids Res* **43**(8), 3911–21.

League, G. P., Slot, J. C. and Rokas, A. (2012), 'The asp3 locus in *Saccharomyces cerevisiae* originated by horizontal gene transfer from *Wickerhamomyces*', *FEMS Yeast Res* **12**(7), 859–63.

Leducq, J. B., Charron, G., Samani, P., Dube, A. K., Sylvester, K., James, B., Almeida, P., Sampaio, J. P., Hittinger, C. T., Bell, G. and Landry, C. R. (2014), 'Local climatic adaptation in a widespread microorganism', *Proc Biol Sci* **281**(1777), 20132472.

Leem, Y. E., Ripmaster, T. L., Kelly, F. D., Ebina, H., Heincelman, M. E., Zhang, K., Grewal, S. I., Hoffman, C. S. and Levin, H. L. (2008), 'Retrotransposon *Tf1* is targeted to Pol II promoters by transcription activators', *Mol Cell* **30**(1), 98–107.

Legras, J. L., Erny, C. and Charpentier, C. (2014), 'Population structure and comparative genome hybridization of European flor yeast reveal a unique group of *Saccharomyces cerevisiae* strains with few gene duplications in their genome', *PLoS One* **9**(10), e108089.

Legras, J. L., Galeote, V., Bigey, F., Camarasa, C., Marsit, S., Nidelet, T., Sanchez, I., Couloux, A., Guy, J., Franco-Duarte, R., Marina, M. H., Gabaldon, T., Schuller, D., Sampaio, J. P. and Dequin, S. (2018), 'Adaptation of *S. cerevisiae* to fermented food environments reveals remarkable genome plasticity and the footprints of domestication', *Mol Biol Evol* .

Legras, J. L. and Karst, F. (2003), 'Optimisation of interdelta analysis for *Saccharomyces cerevisiae* strain characterisation', *FEMS Microbiol Lett* **221**(2), 249–55.

Lemoine, F. J., Degtyareva, N. P., Lobachev, K. and Petes, T. D. (2005), 'Chromosomal translocations in yeast induced by low levels of polymerase a model for chromosome fragile sites', *Cell* **120**(5), 587–98.

Lerat, E., Capy, P. and Biémont, C. (2002), 'Codon usage by transposable elements and their host genes in five species', *J Mol Evol* **54**(5), 625–37.

Lerat, E., Rizzon, C. and Biemont, C. (2003), 'Sequence divergence within transposable element families in the *Drosophila melanogaster* genome', *Genome Res* **13**(8), 1889–96.

Lesage, P. and Todeschini, A. L. (2005), 'Happy together: the life and times of *Ty* retrotransposons and their hosts', *Cytogenet Genome Res.* **110**(1-4), 70–90.

Li, H. and Durbin, R. (2009), 'Fast and accurate short read alignment with burrows-wheeler transform', *Bioinformatics* **25**(14), 1754–60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009a), 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics* **25**(16), 2078–9.

Li, Y. D., Liang, H., Gu, Z., Lin, Z., Guan, W., Zhou, L., Li, Y. Q. and Li, W. H. (2009b), 'Detecting positive selection in the budding yeast genome', *J Evol Biol* **22**(12), 2430–7.

Liao, X., Clare, J. and Farabaugh, P. J. (1987), 'The upstream activation site of a *Ty2* element of yeast is necessary but not sufficient to promote maximal transcription of the element', *PNAS* **84**(23), 8520–4.

Libkind, D., Hittinger, C. T., Valerio, E., Gonçalves, C., Dover, J., Johnston, M., Gonçalves, P. and Sampaio, J. P. (2011), 'Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast', *Proc Natl Acad Sci U S A* **108**(35), 14539–44.

Liebman, S., Shalit, P. and Picologlou, S. (1981), '*Ty* elements are involved in the formation of deletions in del1 strains of *Saccharomyces cerevisiae*', *Cell* **26**(3 Pt 1), 401–9.

Lin, X., Faridi, N. and Casola, C. (2016), 'An ancient transkingdom horizontal transfer of *Penelope*-like retroelements from Arthropods to Conifers', *Genome Biol Evol* **8**(4), 1252–66.

Lindegren, C. C. (1949), *The Yeast Cell, Its Genetics And Cytology*, 1st edn, Educational Publishers Inc., Saint Louis.

Liti, G., Barton, D. B. and Louis, E. J. (2006), 'Sequence diversity, reproductive isolation and species concepts in *Saccharomyces*', *Genetics* **174**(2), 839–50.

Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., Davey, R. P., Roberts, I. N., Burt, A., Koufopanou, V., Tsai, I. J., Bergman, C. M., Bensasson, D., O'Kelly, M. J., van Oudenaarden, A., Barton, D. B., Bailes, E., Nguyen, A. N., Jones, M., Quail, M. A., Goodhead, I., Sims, S., Smith, F., Blomberg, A., Durbin, R. and Louis, E. J. (2009), 'Population genomics of domestic and wild yeasts', *Nature* **458**(7236), 337–41.

Liti, G. and Louis, E. J. (2005), 'Yeast evolution and comparative genomics', *Annu Rev Microbiol* **59**, 135–53.

Liti, G., Nguyen Ba, A. N., Blythe, M., Muller, C. A., Bergstrom, A., Cubillos, F. A., Dafhnis-Calas, F., Khoshraftar, S., Malla, S., Mehta, N., Slow, C., Warringer, J., Moses, A., Louis, E. J. and Nieduszynski, C. (2013), 'High quality de novo sequencing and assembly of the *Saccharomyces arboricolus* genome', *BMC Genomics* **14**(69).

Liti, G., Peruffo, A., James, S. A., Roberts, I. N. and Louis, E. J. (2005), 'Inferences of evolutionary relationships from a population survey of LTR-retrotransposons and telomeric-associated sequences in the *Saccharomyces sensu stricto* complex', *Yeast* **22**(3), 177–92.

Liti, G. and Schacherer, J. (2011), 'The rise of yeast population genomics', *C R Biol* **334**(8-9), 612–9.

Liu, B. and Wendel, J. F. (2000), 'Retrotransposon activation followed by rapid repression in introgressed rice plants', *Genome* **43**(5), 874–80.

Liu, Q., Wang, H., Hu, H. and Zhang, H. (2015), 'Genome-wide identification and evolutionary analysis of positively selected miRNA genes in domesticated rice', *Mol Genet Genomics* **290**(2), 593–602.

Liu, W. Q., Han, P. J., Qiu, J. Z. and Wang, Q. M. (2012), '*Naumovozyma baii sp. nov.*, an ascomycetous yeast species isolated from rotten wood in a tropical forest', *Int J Syst Evol Microbiol* **62**(12), 3095–8.

Llorens, C. and Marin, I. (2001), 'A mammalian gene evolved from the integrase domain of an LTR retrotransposon', *Mol Biol Evol* **18**(8).

Llorens, C., Munoz-Pomer, A., Bernad, L., Botella, H. and Moya, A. (2009), 'Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees', *Biol Direct* **4**, 41.

Llorens, J. V., Clark, J. B., Martinez-Garay, I., Soriano, S., de Frutos, R. and Martinez-Sebastian, M. J. (2008), '*Gypsy* endogenous retrovirus maintains potential infectivity in several species of drosophilids', *BMC Evol Biol* **8**, 302.

Lockton, S. and Gaut, B. S. (2010), 'The evolution of transposable elements in natural populations of self-fertilizing *Arabidopsis thaliana* and its outcrossing relative *Arabidopsis lyrata*', *BMC Evol Biol* **10**, 10.

Lockton, S., Ross-Ibarra, J. and Gaut, B. S. (2008), 'Demography and weak selection drive patterns of transposable element diversity in natural populations of *Arabidopsis lyrata*', *Proc Natl Acad Sci U S A* **105**(37), 13965–70.

Long, Q., Bengra, C., Li, C., Kutlar, F. and Tuan, D. (1998), 'A long terminal repeat of the human endogenous retrovirus ERV-9 is located in the 5' boundary area of the human beta-globin locus control region', *Genomics* **54**(3), 542–55.

Lopandic, K., Gangl, H., Wallner, E., Tscheik, G., Leitner, G., Querol, A., Borth, N., Breitenbach, M., Prillinger, H. and Tiefenbrunner, W. (2007), 'Genetically different wine yeasts isolated from austrian vine-growing regions influence wine aroma differently and contain putative hybrids between *Saccharomyces cerevisiae* and *Saccharomyces kudriavzevii*', *FEMS Yeast Res* **7**(6), 953–65.

Lopes, M. R., Morais, C. G., Kominek, J., Cadete, R. M., Soares, M. A., Uetanabaro, A. P., Fonseca, C., Lachance, M. A., Hittinger, C. T. and Rosa, C. A. (2016), 'Genomic analysis and d-xylose fermentation of three novel *Spathaspora* species: *Spathaspora girioi sp. nov., Spathaspora hagerdaliae f. a., sp. nov*. and *Spathaspora gorwiae f. a., sp. nov*', *FEMS Yeast Res* **16**(4).

Loreto, E. L., Carareto, C. M. and Capy, P. (2008), 'Revisiting horizontal transfer of transposable elements in *Drosophila*', *Heredity (Edinb)* **100**(6), 545–54.

Louis, E. J. (2011), 'Population genomics and speciation in yeasts', *Fungal Biology Reviews* **25**(3), 136–142.

MacKenzie, D. A., Defernez, M., Dunn, W. B., Brown, M., Fuller, L. J., de Herrera, S. R., Gunther, A., James, S. A., Eagles, J., Philo, M., Goodacre, R. and Roberts, I. N. (2008), 'Relatedness of medically important strains of *Saccharomyces cerevisiae* as revealed by phylogenetics and metabolomics', *Yeast* **25**(7), 501–12.

Macpherson, J. M., Gonzalez, J., Witten, D. M., Davis, J. C., Rosenberg, N. A., Hirsh, A. E. and Petrov, D. A. (2008), 'Nonadaptive explanations for signatures of partial selective sweeps in *Drosophila*', *Mol Biol Evol* **25**(6), 1025–42.

Magwire, M. M., Bayer, F., Webster, C. L., Cao, C. and Jiggins, F. M. (2011), 'Successive increases in the resistance of *Drosophila* to viral infection through a transposon insertion followed by a duplication', *PLoS Genet* **7**(10), e1002337.

Maheshwari, S. and Barbash, D. A. (2011), 'The genetics of hybrid incompatibilities', *Annu Rev Genet* **45**, 331–55.

Malik, H. S. and Eickbush, T. H. (1999), 'Modular evolution of the integrase domain in the *Ty3/Gypsy* class of LTR retrotransposons', *J Virol* **73**(6).

Malik, H. S. and Eickbush, T. H. (2001), 'Phylogenetic analysis of ribonuclease h domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses', *Genome Res* **11**(1).

Marcet-Houben, M. and Gabaldon, T. (2015), 'Beyond the whole-genome duplication: Phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage', *PLoS Biol* **13**(8), e1002220.

Marinoni, G., Manuel, M., Petersen, R. F., Hvidtfeldt, J., Sulo, P. and Piskur, J. (1999), 'Horizontal transfer of genetic material among *Saccharomyces* yeasts', *J Bacteriol* **181**(20), 6488–96.

Markova, D. N. and Mason-Gamer, R. J. (2015), 'The role of vertical and horizontal transfer in the evolutionary dynamics of *PIF*-like transposable elements in triticeae', *PLoS One* **10**(9), e0137648.

Marsit, S., Mena, A., Bigey, F., Sauvage, F. X., Couloux, A., Guy, J., Legras, J. L., Barrio, E., Dequin, S. and Galeote, V. (2015), 'Evolutionary advantage conferred by an eukaryote-to-eukaryote gene transfer event in wine yeasts', *Mol Biol Evol* **32**(7), 1695–707.

Marti-Raga, M., Peltier, E., Mas, A., Beltran, G. and Marullo, P. (2017), 'Genetic causes of phenotypic adaptation to the second fermentation of sparkling wines in *Saccharomyces cerevisiae*', *G3 (Bethesda)* **7**(2), 399–412.

Maside, X., Bartolome, C. and Charlesworth, B. (2003), 'Inferences on the evolutionary history of the *S*-element family of *Drosophila melanogaster*', *Mol Biol Evol* **20**(8), 1183–7.

Mateo, L., Ullastres, A. and Gonzalez, J. (2014), 'A transposable element insertion confers xenobiotic resistance in *Drosophila*', *PLoS Genet* **10**(8), e1004560.

Matheson, K., Parsons, L. and Gammie, A. (2017), 'Whole-genome sequence and variant analysis of w303, a widely-used strain of *Saccharomyces cerevisiae*', *G3 (Bethesda)* **7**(7), 2219–2226.

Matsuda, E. and Garfinkel, D. J. (2009), 'Posttranslational interference of *Ty1* retrotransposition by antisense RNAs', *Proc Natl Acad Sci U S A* **106**(37), 15657–62.

Matsunaga, W., Ohama, N., Tanabe, N., Masuta, Y., Masuda, S., Mitani, N., Yamaguchi-Shinozaki, K., Ma, J. F., Kato, A. and Ito, H. (2015), 'A small rna mediated regulation of a stress-activated retrotransposon and the tissue specific transposition during the reproductive period in *Arabidopsis*', *Front Plant Sci* **6**, 48.

Matzke, M. A., Mittelsten Scheid, O. and Matzke, A. J. (1999), 'Rapid structural and epigenetic changes in polyploid and aneuploid genomes', *Bioessays* **21**(9), 761–7.

Maxwell, P. H. and Curcio, M. J. (2007), 'Host factors that control long terminal repeat retrotransposons in *Saccharomyces cerevisiae*: implications for regulation of mammalian retroviruses', *Eukaryot Cell* **6**(7), 1069–80.

Maynard Smith, J. and Haigh, J. (1974), 'The hitch-hiking effect of a favourable gene', *genet Res (Camb)* **23**.

McClintock, B. (1944), 'The relation of homozygous deficiencies to mutations and allelic series in maize', *Genetics* **29**(5), 478–502.

McCullers, T. J. and Steiniger, M. (2017), 'Transposable elements in *Drosophila*', *Mob Genet Elements* **7**(3), 1–18.

Medstrand, P., Landry, J. R. and Mager, D. L. (2001), 'Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans', *J Biol Chem* **276**(3), 1896–903.

Merenciano, M., Ullastres, A., de Cara, M. A., Barron, M. G. and González, J. (2016), 'Multiple independent retroelement insertions in the promoter of a stress response gene have variable molecular and functional effects in *Drosophila*', *PLoS Genet* **12**(8), e1006249.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D. and Thomas, P. D. (2017), 'PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements', *Nucleic Acids Res* **45**(D1), D183–D189.

Miller, M. A., Schwartz, T. and Pfeiffer, W. (2013), 'Embedding CIPRES science gateway capabilities in phylogenetics software environments', *XSEDE* p. 1.

Milne, I., Lindner, D., Bayer, M., Husmeier, D., McGuire, G., Marshall, D. F. and Wright, F. (2009), 'TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on hpc clusters and multi-core desktops', *Bioinformatics* **25**(1), 126–7.

Miousse, I. R., Chalbot, M. C., Lumen, A., Ferguson, A., Kavouras, I. G. and Koturbash, I. (2015), 'Response of transposable elements to environmental stressors', *Mutat Res Rev Mutat Res* **765**, 19–39.

Miyao, A., Tanaka, K., Murata, K., Sawaki, H., Takeda, S., Abe, K., Shinozuka, Y., Onosato, K. and Hirochika, H. (2003), 'Target site specificity of the *Tos17* retrotransposon shows a preference

for insertion within genes and against insertion in retrotransposon-rich regions of the genome', *Plant Cell* **15**(8), 1771–80.

Mizuguchi, T., Barrowman, J. and Grewal, S. I. (2015), 'Chromosome domain architecture and dynamic organization of the fission yeast genome', *FEBS Lett* **589**(20 Pt A), 2975–86.

Møller, H. D., Larsen, C. E., Parsons, L., Hansen, A. J., Regenberg, B. and Mourier, T. (2016), 'Formation of extrachromosomal circular from long terminal repeats of retrotransposons in *Saccharomyces cerevisiae*', *G3 (Bethesda)* **6**(2), 453–62.

Møller, H. D., Parsons, L., Jørgensen, T., Botstein, D. and Regenberg, B. (2015), 'Extrachromosomal circular is common in yeast', *PNAS* .

Montchamp-Moreau, C. (1990), 'Dynamics of *P-M* hybrid dysgenesis in *P*-transformed lines of *Drosophila simulans*', *Evolution* **44**(1), 194–203.

Montgomery, E., Charlesworth, B. and Langley, C. H. (1987), 'A test for the role of natural selection in the stabilization of transposable element copy number in a population of *Drosophila melanogaster*', *Genet Res* **49**(1), 31–41.

Moore, S. P., Liti, G., Stefanisko, K. M., Nyswaner, K. M., Chang, C., Louis, E. J. and Garfinkel, D. J. (2004), 'Analysis of a *Ty1*-less variant of *Saccharomyces paradoxus*: the gain and loss of *Ty1* elements', *Yeast* **21**(8), 649–60.

Morais, P. B., Hagler, A. N., Rosa, C. A., Mendonca-Hagler, L. C. and Klaczko, L. B. (1992), 'Yeasts associated with *Drosophila* in tropical forests of rio de janeiro, brazil', *Can J Microbiol* **38**(11), 1150–5.

Morales, L. and Dujon, B. (2012), 'Evolutionary role of interspecies hybridization and genetic exchanges in yeasts', *Microbiol Mol Biol Rev* **76**(4), 721–39.

Morillon, A., Benard, L., Springer, M. and Lesage, P. (2002), 'Differential effects of chromatin and gcn4 on the 50-fold range of expression among individual yeast *Ty1* retrotransposons', *Molecular and Cellular Biology* **22**(7), 2078–2088.

Morillon, A., Springer, M. and Lesage, P. (2000), 'Activation of the kss1 invasive-filamentous growth pathway induces *Ty1* transcription and retrotransposition in *Saccharomyces cerevisiae*', *Mol Cell Biol* **20**(15).

Mortimer, R. K. and Johnston, J. R. (1986), 'Genealogy of principal strains of the yeast genetic stock center', *Genetics* **113**(1), 35–43.

Mount, S. M. and Rubin, G. M. (1985), 'Complete nucleotide sequence of the *Drosophila* transposable element copia: homology between *copia* and retroviral proteins', *Mol Cell Biol* **5**(7), 1630–8.

Muller, L. A. and McCusker, J. H. (2009), 'A multispecies-based taxonomic microarray reveals interspecies hybridization and introgression in *Saccharomyces cerevisiae*', *FEMS Yeast Res* **9**(1), 143–52.

Muszewska, A., Hoffman-Sommer, M. and Grynberg, M. (2011), 'LTR retrotransposons in fungi', *PLoS One* **6**(12), e29425.

Nakanishi, H., Higuchi, Y., Kawakami, S., Yamashita, F. and Hashida, M. (2010), '*piggyBac* transposon-mediated long-term gene expression in mice', *Mol Ther* **18**(4), 707–14.

Naseeb, S., James, S. A., Alsammar, H., Michaels, C. J., Gini, B., Nueno-Palop, C., Bond, C. J., McGhie, H., Roberts, I. N. and Delneri, D. (2017), '*Saccharomyces jurei sp. nov.*, isolation and genetic identification of a novel yeast species from *Quercus robur*', *Int J Syst Evol Microbiol* **67**(6), 2046–2052.

Natarajan, K., Meyer, M. R., Jackson, B. M., Slade, D., Roberts, C., Hinnebusch, A. G. and Marton, M. J. (2001), 'Transcriptional profiling shows that gcn4p is a master regulator of gene expression during amino acid starvation in yeast', *Mol Cell Biol* **21**(13), 4347–68.

Natsoulis, G., Thomas, W. K., Roghmann, M., Winston, F. and Boeke, J. D. (1989), '*Ty1* transposition in *Saccharomyces cerevisiae* is nonrandom', *Genetics* **123**.

Naumov, G. I. (1996), 'Genetic identification of biological species in the *Saccharomyces sensu stricto* complex', *J Ind Microbiol* **17**(3), 295–302.

Naumov, G. I., James, S. A., Naumova, E. S., Louis, E. J. and Roberts, I. N. (2000), 'Three new species in the *Saccharomyces sensu stricto* complex: *Saccharomyces cariocanus, Saccharomyces kudriavzevii* and *Saccharomyces mikatae*', *Int J Sys Evol Micro* **50**, 1931–1942.

Naumov, G. I., Lee, C. F. and Naumova, E. S. (2013), 'Molecular genetic diversity of the *Saccharomyces* yeasts in taiwan: *Saccharomyces arboricola, Saccharomyces cerevisiae* and *Saccharomyces kudriavzevii*', *Antonie Van Leeuwenhoek* **103**(1), 217–28.

Naumova, E. S., Naumov, G. I., Masneuf-Pomarede, I., Aigle, M. and Dubourdieu, D. (2005), 'Molecular genetic study of introgression between *Saccharomyces bayanus* and *S. cerevisiae*', *Yeast* **22**(14), 1099–115.

Naumova, E. S., Naumov, G. I., Michailova, Y. V., Martynenko, N. N. and Masneuf-Pomarede, I. (2011), 'Genetic diversity study of the yeast *Saccharomyces bayanus var. uvarum* reveals introgressed subtelomeric *Saccharomyces cerevisiae* genes', *Res Microbiol* **162**(2), 204–13.

NBRC (2010), 'Catalogue of biological resources: Microorganisms, microorganism-related resources, human-related resources'.

Nei, M. and Li, W. H. (1979), 'Mathematical model for studying genetic variation in terms of restriction endonucleases', *Proc Natl Acad Sci U S A* **76**(10), 5269–73.

Neuvéglise, C., Chalvet, F., Wincker, P., Gaillardin, C. and Casaregola, S. (2005), 'Mutator-like element in the yeast *Yarrowia lipolytica* displays multiple alternative splicings', *Eukaryot Cell* **4**(3), 615–24.

Neuvéglise, C., Feldmann, H., Bon, E., Gaillardin, C. and Casaregola, S. (2002), 'Genomic evolution of the long terminal repeat retrotransposons in hemiascomycetous yeasts', *Genome Res* **12**(6), 930–943.

Nguyen, H. V. and Gaillardin, C. (2005), 'Evolutionary relationships between the former species *Saccharomyces uvarum* and the hybrids *Saccharomyces bayanus* and *Saccharomyces pastorianus*; reinstatement of *Saccharomyces uvarum* (beijerinck) as a distinct species', *FEMS Yeast Res* **5**(4-5), 471–83.

Nguyen, H. V., Legras, J. L., Neuvéglise, C. and Gaillardin, C. (2011), 'Deciphering the hybridisation history leading to the lager lineage based on the mosaic genomes of *Saccharomyces bayanus* strains nbrc1948 and cbs380', *PLoS One* **6**(10), e25821.

Nieduszynski, C. A., Knox, Y. and Donaldson, A. D. (2006), 'Genome-wide identification of replication origins in yeast by comparative genomics', *Genes Dev* **20**(14), 1874–9.

Novo, M., Bigey, F., Beyne, E., Galeote, V., Gavory, F., Mallet, S., Cambon, B., Legras, J. L., Wincker, P., Casaregola, S. and Dequin, S. (2009), 'Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast *Saccharomyces cerevisiae* ec1118', *Proc Natl Acad Sci U S A* **106**(38), 16333–8.

Nuzhdin, S. V. (1999), 'Sure facts, speculations, and open questions about the evolution of transposable element copy number', *Genetica* **107**(1-3), 129–37.

Nwokeoji, A. O., Kilby, P. M., Portwood, D. E. and Dickman, M. J. (2016), 'Rnaswift: A rapid, versatile RNA extraction method free from phenol and chloroform', *Anal Biochem* **512**, 36–46.

O'Brochta, D. A., Stosic, C. D., Pilitt, K., Subramanian, R. A., Hice, R. H. and Atkinson, P. W. (2009), 'Transpositionally active episomal *hAT* elements', *BMC Mol Biol* **10**, 108.

Ogden, R. C., Beckman, J. S., Abelson, J., Kang, H. S., Soll, D. and Schmidt, O. (1979), 'In vitro transcription and processing of a yeast tRNA gene containing an intervening sequence', *Cell* **17**(2), 399–406.

Okonechnikov, K., Conesa, A. and Garcia-Alcalde, F. (2016), 'Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data', *Bioinformatics* **32**(2), 292–4.

Okonechnikov, K., Golosova, O. and Fursov, M. (2012), 'Unipro UGENE: a unified bioinformatics toolkit', *Bioinformatics* **28**(8), 1166–7.

Oliphant, A., Brandl, C. J. and Struhl, K. (1989), 'Defining the sequence specificity of dna-binding proteins by selecting binding sites from random-sequence oligonucleotides: Analysis of yeast *GCN4* protein', *Mol Cell Biol* **9**(7), 2944–2949.

Oliver, K. R. and Greene, W. K. (2009), 'Transposable elements: powerful facilitators of evolution', *Bioessays* **31**(7), 703–14.

Oliver, K. R. and Greene, W. K. (2012), 'Transposable elements and viruses as factors in adaptation and evolution: an expansion and strengthening of the te-thrust hypothesis', *Ecol Evol* **2**(11), 2912–33.

Oppold, A. M., Schmidt, H., Rose, M., Hellmann, S. L., Dolze, F., Ripp, F., Weich, B., Schmidt-Ott, U., Schmidt, E., Kofler, R., Hankeln, T. and Pfenninger, M. (2017), '*Chironomus riparius* (diptera) genome sequencing reveals the impact of minisatellite transposable elements on population divergence', *Mol Ecol* **26**(12), 3256–3275.

Orgel, L. E. and Crick, F. H. (1980), 'Selfish DNA: the ultimate parasite', *Nature* **284**(5757), 604–7.

Ortiz, M. F., Wallau, G. L., Graichen, D. A. and Loreto, E. L. (2015), 'An evaluation of the ecological relationship between *Drosophila* species and their parasitoid wasps as an opportunity for horizontal transposon transfer', *Mol Genet Genomics* **290**(1), 67–78.

O'Sullivan, J. M., Tan-Wong, S. M., Morillon, A., Lee, B., Coles, J., Mellor, J. and Proudfoot, N. J. (2004), 'Gene loops juxtapose promoters and terminators in yeast', *Nat Genet* **36**(9), 1014–8.

Pace, J. K., n., Gilbert, C., Clark, M. S. and Feschotte, C. (2008), 'Repeated horizontal transfer of a transposon in mammals and other tetrapods', *Proc Natl Acad Sci U S A* **105**(44), 17023–8.

Page, R. D. M. and Holmes, E. C. (1998), *Molecular Evolution: A Phylogenetic Approach*, Blackwell Science, London.

Pan, J., Sasaki, M., Kniewel, R., Murakami, H., Blitzblau, H. G., Tischfield, S. E., Zhu, X., Neale, M. J., Jasin, M., Socci, N. D., Hochwagen, A. and Keeney, S. (2011), 'A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation', *Cell* **144**(5), 719–31.

Paques, F. and Haber, J. E. (1999), 'Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*', *Microbiol Mol Biol Rev* **63**(2), 349–404.

Paquin, C. E. and Adams, J. (1983), 'Relative fitness can decrease in evolving asexual populations of s. cerevisiae', *Nature* **306**(5941), 368–70.

Paquin, C. and Williamson, V. (1986), '*Ty* insertions at two loci account for most of the spontaneous antimycin a resistance mutations during growth at 15°c of *Saccharomyces cerevisiae* strains lacking adhi', *Mol Cell Biol* **6**(1).

Paris, M. and Despres, L. (2012), 'Identifying insecticide resistance genes in mosquito by combining AFLP genome scans and 454 pyrosequencing', *Mol Ecol* **21**(7), 1672–86.

Parsons, A. B., Brost, R. L., Ding, H., Li, Z., Zhang, C., Sheikh, B., Brown, G. W., Kane, P. M., Hughes, T. R. and Boone, C. (2004), 'Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways', *Nat Biotechnol* **22**(1), 62–9.

Pasyukova, E. G., Nuzhdin, S. V., Morozova, T. V. and Mackay, T. F. (2004), 'Accumulation of transposable elements in the genome of *Drosophila melanogaster* is associated with a decrease in fitness', *J Hered* **95**(4), 284–90.

Payen, C., Fischer, G., Marck, C., Proux, C., Sherman, D. J., Coppee, J. Y., Johnston, M., Dujon, B. and Neuvéglise, C. (2009), 'Unusual composition of a yeast chromosome arm is associated with its delayed replication', *Genome Res* **19**(10), 1710–21.

Peccoud, J., Loiseau, V., Cordaux, R. and Gilbert, C. (2017), 'Massive horizontal transfer of transposable elements in insects', *Proc Natl Acad Sci U S A* **114**(18), 4721–4726.

Pelechano, V., Garcia-Martinez, J. and Perez-Ortin, J. E. (2006), 'A genomic study of the inter-ORF distances in *Saccharomyces cerevisiae*', *Yeast* **23**(9), 689–99.

Pennaneach, V. and Kolodner, R. D. (2009), 'Stabilization of dicentric translocations through secondary rearrangements mediated by multiple mechanisms in *S. cerevisiae*', *PLoS One* **4**(7), e6389.

Pereira, V., Enard, D. and Eyre-Walker, A. (2009), 'The effect of transposable element insertions on gene expression evolution in rodents', *PLoS One* **4**(2), e4321.

Perez-Torrado, R., González, S. S., Combina, M., Barrio, E. and Querol, A. (2015), 'Molecular and enological characterization of a natural *Saccharomyces uvarum* and *Saccharomyces cerevisiae* hybrid', *Int J Food Microbiol* **204**, 101–10.

Perez-Traves, L., Lopes, C. A., Querol, A. and Barrio, E. (2014), 'On the complexity of the *Saccharomyces bayanus* taxon: hybridization and potential hybrid speciation', *PLoS One* **9**(4), e93729.

Peris, D., Langdon, Q. K., Moriarty, R. V., Sylvester, K., Bontrager, M., Charron, G., Leducq, J. B., Landry, C. R., Libkind, D. and Hittinger, C. T. (2016), 'Complex ancestries of lager-brewing hybrids were shaped by standing variation in the wild yeast *Saccharomyces eubayanus*', *PLoS Genet* **12**(7), e1006155.

Peris, D., Lopes, C., Belloch, C., Querol, A. and Barrio, E. (2012), 'Comparative genomics among *Saccharomyces cerevisiae* × *Saccharomyces kudriavzevii* natural hybrid strains isolated from wine and beer reveals different origins', *BMC Genomics* **13**(407).

Peris, D., Moriarty, R. V., Alexander, W. G., Baker, E., Sylvester, K., Sardi, M., Langdon, Q. K., Libkind, D., Wang, Q. M., Bai, F. Y., Leducq, J. B., Charron, G., Landry, C. R., Sampaio, J. P., Gonçalves, P., Hyma, K. E., Fay, J. C., Sato, T. K. and Hittinger, C. T. (2017), 'Hybridization and adaptive evolution of diverse *Saccharomyces* species for cellulosic biofuel production', *Biotechnol Biofuels* **10**, 78.

Peris, D., Sylvester, K., Libkind, D., Gonçalves, P., Sampaio, J. P., Alexander, W. G. and Hittinger, C. T. (2014), 'Population structure and reticulate evolution of *Saccharomyces eubayanus* and its lager-brewing hybrids', *Mol Ecol* **23**(8), 2031–45.

Perlman, P. S. and Boeke, J. D. (2004), 'Molecular biology. ring around the retroelement', *Science* **303**(5655), 182–4.

Pesheva, M., Krastanova, O., Staleva, L., Dentcheva, V., Hadzhitodorov, M. and Venkov, P. (2005), 'The *Ty1* transposition assay: a new short-term test for detection of carcinogens', *J Microbiol Methods* **61**(1), 1–8.

Pesheva, M., Krastanova, O., Stamenova, R., Kantardjiev, D. and Venkov, P. (2008), 'The response of *Ty1* test to genotoxins', *Arch Toxicol* **82**(10), 779–85.

Peter, J., De Chiara, M., Friedrich, A., Yue, J. X., Pflieger, D., Bergstrom, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., Cruaud, C., Labadie, K., Aury, J. M., Istace, B., Lebrigand, K., Barbry, P., Engelen, S., Lemainque, A., Wincker, P., Liti, G. and Schacherer, J. (2018), 'Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates', *Nature* .

Petrascheck, M., Escher, D., Mahmoudi, T., Verrijzer, C. P., Schaffner, W. and Barberis, A. (2005), 'DNA looping induced by a transcriptional enhancer in vivo', *Nucleic Acids Res* **33**(12), 3743–50.

Petrov, D. A., Aminetzach, Y. T., Davis, J. C., Bensasson, D. and Hirsh, A. E. (2003), 'Size matters: non-LTR retrotransposable elements and ectopic recombination in *Drosophila*', *Mol Biol Evol* **20**(6), 880–92.

Pfliegler, W. P. and Sipiczki, M. (2016), 'Does fingerprinting truly represent the diversity of wine yeasts? a case study with interdelta genotyping of *Saccharomyces cerevisiae* strains', *Lett Appl Microbiol* **63**(6), 406–411.

Piskurek, O. and Okada, N. (2007), 'Poxviruses as possible vectors for horizontal transfer of retroposons from reptiles to mammals', *Proc Natl Acad Sci U S A* **104**(29), 12046–51.

Poulter, R. T. and Goodwin, T. J. (2005), '*DIRS*-1 and the other tyrosine recombinase retrotransposons', *Cytogenet Genome Res* **110**(1-4), 575–88.

Prakash, L. and Higgins, D. (1982), 'Role of DNA repair in ethyl methanesulfonate-induced mutagenesis in *Saccharomyces cerevisiae*', *Carcinogenesis* **3**(4), 439–44.

Pretorius, I. S. (2000), 'Tailoring wine yeast for the new millennium: novel approaches to the ancient art of winemaking', *Yeast* **16**(8), 675–729.

Pritham, E. J. (2009), 'Transposable elements and factors influencing their success in eukaryotes', *J Hered* **100**(5), 648–55.

Proffitt, J. H., Davie, J. R., Swinton, D. and Hattman, S. (1984), '5-methylcytosine is not detectable in *Saccharomyces cerevisiae* DNA', *Mol Cell Biol* **4**(5), 985–8.

Pujol, C., Daniels, K. J., Lockhart, S. R., Srikantha, T., Radke, J. B., Geiger, J. and Soll, D. R. (2004), 'The closely related species *Candida albicans* and *Candida dubliniensis* can mate', *Eukaryot Cell* **3**(4), 1015–27.

Pulvirenti, A., Nguyen, H., Caggia, C., Giudici, P., Rainieri, S. and Zambonelli, C. (2000), '*Saccharomyces uvarum*, a proper species within *Saccharomyces sensu stricto*', *FEMS Microbiol Lett* **192**(2), 191–6.

Rabitsch, K. P., Toth, A., Galova, M., Schleiffer, A., Schaffner, G., Aigner, E., Rupp, C., Penkner, A. M., Moreno-Borchart, A. C., Primig, M., Esposito, R. E., Klein, F., Knop, M. and Nasmyth, K. (2001), 'A screen for genes required for meiosis and spore formation based on whole-genome expression', *Curr Biol* **11**(13), 1001–9.

Rai, S. K., Sangesland, M., Lee, M., J., Esnault, C., Cui, Y., Chatterjee, A. G. and Levin, H. L. (2017), 'Host factors that promote retrotransposon integration are similar in distantly related eukaryotes', *PLoS Genet* **13**(12), e1006775.

Rainieri, S., Zambonelli, C., Hallsworth, J. E., Pulvirenti, A. and Giudici, P. (1999), '*Saccharomyces uvarum*, a distinct group within *Saccharomyces sensu stricto*', *FEMS Microbiol Lett* **177**(1), 177–85.

Ramírez-Soriano, A., Ramos-Onsins, S. E., Rozas, J., Calafell, F. and Navarro, A. (2008), 'Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination', *Genetics* **179**(1), 555–67.

Rannala, B. and Yang, Z. (1996), 'Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference', *J Mol Evol* **43**(3), 304–11.

Rödelsperger, C. and Sommer, R. J. (2011), 'Computational archaeology of the *Pristionchus pacificus* genome reveals evidence of horizontal gene transfers from insects', *BMC Evol Biol* **11**, 239.

Replansky, T., Koufopanou, V., Greig, D. and Bell, G. (2008), '*Saccharomyces sensu stricto* as a model system for evolution and ecology', *Trends Ecol Evol* **23**(9), 494–501.

Rey, O., Danchin, E., Mirouze, M., Loot, C. and Blanchet, S. (2016), 'Adaptation to global change: A transposable element-epigenetics perspective', *Trends Ecol Evol* **31**(7), 514–26.

Rhind, N., Chen, Z., Yassour, M., Thompson, D. A., Haas, B. J., Habib, N., Wapinski, I., Roy, S., Lin, M. F., Heiman, D. I., Young, S. K., Furuya, K., Guo, Y., Pidoux, A., Chen, H. M., Robbertse, B., Goldberg, J. M., Aoki, K., Bayne, E. H., Berlin, A. M., Desjardins, C. A., Dobbs, E., Dukaj, L., Fan, L., FitzGerald, M. G., French, C., Gujja, S., Hansen, K., Keifenheim, D., Levin, J. Z., Mosher, R. A., Muller, C. A., Pfiffner, J., Priest, M., Russ, C., Smialowska, A., Swoboda, P., Sykes, S. M., Vaughn, M., Vengrova, S., Yoder, R., Zeng, Q., Allshire, R., Baulcombe, D., Birren, B. W., Brown, W., Ekwall, K., Kellis, M., Leatherwood, J., Levin, H., Margalit, H., Martienssen, R., Nieduszynski, C. A., Spatafora, J. W., Friedman, N., Dalgaard, J. Z., Baumann, P., Niki, H., Regev, A. and Nusbaum, C. (2011), 'Comparative functional genomics of the fission yeasts', *Science* **332**(6032), 930–6.

Rhoads, A. and Au, K. F. (2015), 'Pacbio sequencing and its applications', *Genomics Proteomics Bioinformatics* **13**(5), 278–89.

Ribeiro-dos Santos, G., Schenberg, A. C. G., Gardner, D. and Oliver, S. G. (1997), 'Enhancement of *Ty* transposition at the *ADH4* and *ADH2* loci in meiotic yeast cells', *Mol Genet Genomics* **254**(5), 555–61.

Richards, T. A. (2011), 'Genome evolution: horizontal movements in the fungi', *Curr Biol* **21**(4), R166–8.

Riley, R., Haridas, S., Wolfe, K. H., Lopes, M. R., Hittinger, C. T., Goker, M., Salamov, A. A., Wisecaver, J. H., Long, T. M., Calvey, C. H., Aerts, A. L., Barry, K. W., Choi, C., Clum, A., Coughlan, A. Y., Deshpande, S., Douglass, A. P., Hanson, S. J., Klenk, H. P., LaButti, K. M., Lapidus, A., Lindquist, E. A., Lipzen, A. M., Meier-Kolthoff, J. P., Ohm, R. A., Otillar, R. P., Pangilinan, J. L., Peng, Y., Rokas, A., Rosa, C. A., Scheuner, C., Sibirny, A. A., Slot, J. C., Stielow, J. B.,

Sun, H., Kurtzman, C. P., Blackwell, M., Grigoriev, I. V. and Jeffries, T. W. (2016), 'Comparative genomics of biotechnologically important yeasts', *Proc Natl Acad Sci U S A* **113**(35), 9882–7.

Risler, J., Kenny, A. E., Palumbo, R., Gamache, E. R. and Curcio, M. J. (2012), 'Host co-factors of the retrovirus like transposon *Ty1*', *Mob DNA* **3**(12).

Rizzon, C., Marais, G., Gouy, M. and Biémont, C. (2002), 'Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome', *Genome Res* **12**(3), 400–7.

Roach, D. J., Burton, J. N., Lee, C., Stackhouse, B., Butler-Wu, S. M., Cookson, B. T., Shendure, J. and Salipante, S. J. (2015), 'A year of infection in the intensive care unit: Prospective whole genome sequencing of bacterial clinical isolates reveals cryptic transmissions and novel microbiota', *PLoS Genet* **11**(7), e1005413.

Roeder, G., Farabaugh, P. J., Chaleff, D. and Fink, G. R. (1980), 'The origins of gene instability in yeast', *Science* **209**(4463), 1375–80.

Roeder, G. and Fink, G. R. (1982), 'Movement of yeast transposable elements by gene conversion', *PNAS* **79**, 5621–5625.

Roeder, G., Rose, A. and Pearlman, R. (1985), 'Transposable element sequences involved in the enhancement of gene expression', *Proc Natl Acad Sci U S A* **82**.

Roeder, G. S., Coney, L. R., Pearlman, R. E. and Rose, A. B. (1986), 'Control of yeast gene expression by transposable elements', *Basic Life Sci* **40**, 545–55.

Roeder, G., Smith, M. and Lambie, E. (1984), 'Intrachromosomal movement of genetically marked *Saccharomyces cerevisiae* transposons by gene conversion', *Mol Cell Biol* **4**(4).

Roelants, F., Potier, S., Souciet, J. L. and de Montigny, J. (1995), 'Reactivation of the ATCase domain of the URA2 gene complex: a positive selection method for *Ty* insertions and chromosomal rearrangements in *Saccharomyces cerevisiae*', *Mol Gen Genet* **246**(6), 767–73.

Roelants, F., Potier, S., Souciet, J. L. and de Montigny, J. (1997), 'δ sequence of *Ty1* transposon can initiate transcription of the distal part of the URA2 gene complex', *FEMS Microbiology Letters* **148**, 69–74.

Rolfe, M., Spanos, A. and Banks, G. (1986), 'Induction of yeast *Ty* element transcription by ultraviolet light', *Nature* **319**, 339–340.

Rolland, T., Neuvéglise, C., Sacerdot, C. and Dujon, B. (2009), 'Insertion of horizontally transferred genes within conserved syntenic regions of yeast genomes', *PLoS One* **4**(8), e6515.

Romanish, M. T., Lock, W. M., van de Lagemaat, L. N., Dunn, C. A. and Mager, D. L. (2007), 'Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian evolution', *PLoS Genet* **3**(1), e10.

Ronquist, F. and Huelsenbeck, J. P. (2003), 'MrBayes 3: Bayesian phylogenetic inference under mixed models', *Bioinformatics* **19**(12), 1572–4.

Roulin, A., Piegu, B., Fortune, P. M., Sabot, F., D'Hont, A., Manicacci, D. and Panaud, O. (2009), 'Whole genome surveys of rice, maize and sorghum reveal multiple horizontal transfers of the LTR-retrotransposon *Route66* in *Poaceae*', *BMC Evol Biol* **9**, 58.

Roulin, A., Piegu, B., Wing, R. A. and Panaud, O. (2008), 'Evidence of multiple horizontal transfers of the long terminal repeat retrotransposon *RIRE1* within the genus *Oryza*', *Plant J* **53**(6), 950–9.

Rozas, J. (2009), *Sequence Polymorphism Analysis Using DnaSP*, Methods in Molecular Biology, Humana Press, USA, pp. 337–350.

Rozas, J. and Rozas, R. (1999), 'DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis', *Bioinformatics* **15**(2), 174–5.

Rozpędowska, E., Piškur, J. and Wolfe, K. H. (2011), *Genome Sequences of Saccharomycotina: Resources and Applications in Phylogenomics*, 5th edn, Elsevier Science, Saint Louis, MO, USA, book section 11, pp. 145–157.

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**(9), 1077–81.

Ruggiero, R. P., Bourgeois, Y. and Boissinot, S. (2017), 'LINE insertion polymorphisms are abundant but at low frequencies across populations of *Anolis carolinensis*', *Front Genet* **8**, 44.

Rytka, J. (1975), 'Positive selection of general amino acid permease mutants in *Saccharomyces cerevisiae*', *J Bacteriol* **121**(2), 562–70.

Sacerdot, C., Mercier, G., Todeschini, A. L., Dutreix, M., Springer, M. and Lesage, P. (2005), 'Impact of ionizing radiation on the life cycle of *Saccharomyces cerevisiae Ty1* retrotransposon', *Yeast* **22**(6), 441–55.

Salazar, A. N., Gorter de Vries, A. R., van den Broek, M., Wijsman, M., de la Torre Cortes, P., Brickwedde, A., Brouwers, N., Daran, J. G. and Abeel, T. (2017), 'Nanopore sequencing enables near-complete *de novo* assembly of *Saccharomyces cerevisiae* reference strain CEN.PK113-7D', *FEMS Yeast Res* **17**(7).

Samani, P. and Bell, G. (2010), 'Adaptation of experimental yeast populations to stressful conditions in relation to population size', *J Evol Biol* **23**(4), 791–6.

Sampaio, J. P. and Gonçalves, P. (2008), 'Natural populations of *Saccharomyces kudriavzevii* in Portugal are associated with oak bark and are sympatric with *S. cerevisiae* and *S. paradoxus*', *Appl Environ Microbiol* **74**(7), 2144–52.

Sandmeyer, S. (1998), 'Targeting transposition: at home in the genome', *Genome Res* **8**(5), 416–8.

Sandmeyer, S. (2003), 'Integration by design', *Proc Natl Acad Sci U S A* **100**(10), 5586–8.

Sandmeyer, S. B. and Menees, T. M. (1996), 'Morphogenesis at the retrotransposon-retrovirus interface: *gypsy* and *copia* families in yeast and *Drosophila*', *Curr Top Microbiol Immunol* **214**, 261–96.

SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y. and Bennetzen, J. L. (1998), 'The paleontology of intergene retrotransposons of maize', *Nat Genet* **20**(1), 43–5.

SanMiguel, P., Tikhonov, A., Jin, Y. K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P. S., Edwards, K. J., Lee, M., Avramova, Z. and Bennetzen, J. L. (1996), 'Nested retrotransposons in the intergenic regions of the maize genome', *Science* **274**(5288), 765–8.

Santiago, T. C. and Mamoun, C. B. (2003), 'Genome expression analysis in yeast reveals novel transcriptional regulation by inositol and choline and new regulatory functions for opi1p, ino2p, and ino4p', *J Biol Chem* **278**(40), 38723–30.

Sarilar, V., Bleykasten-Grosshans, C. and Neuvéglise, C. (2015), 'Evolutionary dynamics of *hAT* transposon families in Saccharomycetaceae', *Genome Biol Evol* **7**(1), 172–90.

Sasaki, M., Tischfield, S. E., van Overbeek, M. and Keeney, S. (2013), 'Meiotic recombination initiation in and around retrotransposable elements in *Saccharomyces cerevisiae*', *PLoS Genet* **9**(8), e1003732.

Sawyer, S. L. and Malik, H. S. (2006), 'Positive selection of yeast nonhomologous end-joining genes and a retrotransposon conflict hypothesis', *Proc Natl Acad Sci U S A* **103**(47), 17614–9.

Scannell, D. R., Byrne, K. P., Gordon, J. L., Wong, S. and Wolfe, K. H. (2006), 'Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts', *Nature* **440**(7082), 341–5.

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 U S A* **104**(20), 8397–402.

Scannell, D., Zill, O. A., Rokas, A., Payen, C., Dunham, M. J., Eisen, M., Rine, J., Johnston, M. and Hittinger, C. T. (2011), 'The awesome power of yeast evolutionary genetics: New genome sequences and strain resources for the *Saccharomyces sensu stricto* genus', *G3 (Bethesda)* **11**.

Schaack, S., Gilbert, C. and Feschotte, C. (2010), 'Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution', *Trends Ecol Evol* **25**(9), 537–46.

Schacherer, J., Shapiro, J. A., Ruderfer, D. M. and Kruglyak, L. (2009), 'Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*', *Nature* **458**(7236), 342–5.

Schaffner, W. (2015), 'Enhancers, enhancers - from their discovery to today's universe of transcription enhancers', *Biol Chem* **396**(4), 311–27.

Schlenke, T. A. and Begun, D. J. (2004), 'Strong selective sweep associated with a transposon insertion in *Drosophila simulans*', *Proc Natl Acad Sci U S A* **101**(6), 1626–31.

Schröder, M. S., Martinez de San Vicente, K., Prandini, T. H., Hammel, S., Higgins, D. G., Bagagli, E., Wolfe, K. H. and Butler, G. (2016), 'Multiple origins of the pathogenic yeast *Candida orthopsilosis* by separate hybridizations between two parental species', *PLoS Genet* **12**(11), e1006404.

Sehgal, A., Lee, C. Y. and Espenshade, P. J. (2007), 'SREBP controls oxygen-dependent mobilization of retrotransposons in fission yeast', *PLoS Genet* **3**(8), e131.

Selker, E. U. (2002), 'Repeat-induced gene silencing in fungi', *Adv Genet* **46**, 439–50.

Senerchia, N., Felber, F. and Parisod, C. (2015), 'Genome reorganization in F1 hybrids uncovers the role of retrotransposons in reproductive isolation', *Proc Biol Sci* **282**(1804), 20142874.

Serfling, E., Lubbe, A., Dorsch-Hasler, K. and Schaffner, W. (1985), 'Metal-dependent SV40 viruses containing inducible enhancers from the upstream region of metallothionein genes', *EMBO J* **4**(13B), 3851–9.

Servant, G., Pennetier, C. and Lesage, P. (2008), 'Remodeling yeast gene transcription by activating the *Ty1* long terminal repeat retrotransposon under severe adenine deficiency', *Mol Cell Biol* **28**(17), 5543–54.

Servant, G., Pinson, B., Tchalikian-Cosson, A., Coulpier, F., Lemoine, S., Pennetier, C., Bridier-Nahmias, A., Todeschini, A. L., Fayol, H., Daignan-Fornier, B. and Lesage, P. (2012), 'Tye7 regulates yeast *Ty1* retrotransposon sense and antisense transcription in response to adenylic nucleotides stress', *Nucleic Acids Res* **40**(12), 5271–82.

Sharma, A. and Presting, G. G. (2014), 'Evolution of centromeric retrotransposons in grasses', *Genome Biol Evol* **6**(6), 1335–52.

Shen, Y., Lin, X.-Y., Shan, X.-H., Lin, C.-J., Han, F.-P., Pang, J.-S. and Liu, B. (2005), 'Genomic rearrangement in endogenous long terminal repeat retrotransposons of rice lines introgressed by wild rice (*Zizania latifolia Griseb*.)', *Journal of Integrative Plant Biology* **47**(8), 998–1008.

Shen, Z., Denton, M., Mutti, N., Pappan, K., Kanost, M. R., Reese, J. C. and Reeck, G. R. (2003), 'Polygalacturonase from *Sitophilus oryzae*: possible horizontal transfer of a pectinase gene from fungi to weevils', *J Insect Sci* **3**, 24.

Shibata, Y., Malhotra, A., Bekiranov, S. and Dutta, A. (2009), 'Yeast genome analysis identifies chromosomal translocation, gene conversion events and several sites of *Ty* element insertion', *Nucleic Acids Res* **37**(19), 6454–65.

Shimizu, K. K., Shimizu-Inatsugi, R., Tsuchimatsu, T. and Purugganan, M. D. (2008), 'Independent origins of self-compatibility in *Arabidopsis thaliana*', *Mol Ecol* **17**(2), 704–14.

Sicard, D. and Legras, J. L. (2011), 'Bread, beer and wine: yeast domestication in the *Saccharomyces sensu stricto* complex', *C R Biol* **334**(3), 229–36.

Silva, J. C. and Kidwell, M. G. (2000), 'Horizontal transfer and selection in the evolution of *P* elements', *Mol Biol Evol* **17**(10), 1542–57.

Silva, J., Loreto, E. L. and Clark, J. B. (2004), 'Factors that affect the horizontal transfer of transposable elements', *Curr Biol* **6**, 57–72.

Silverman, S. and Fink, G. R. (1984), 'Effects of *Ty* insertions on HIS4 transcription in *Saccharomyces cerevisiae*', *Mol Cell Biol* **4**(7).

Siow, C. C., Nieduszynska, S. R., Muller, C. A. and Nieduszynski, C. A. (2012), 'Oridb, the dna replication origin database updated and extended', *Nucleic Acids Res* **40**(Database issue), D682–6.

Sipiczki, M. (2008), 'Interspecies hybridization and recombination in *Saccharomyces* wine yeasts', *FEMS Yeast Res* **8**(7), 996–1007.

Skelly, D. A., Merrihew, G. E., Riffle, M., Connelly, C. F., Kerr, E. O., Johansson, M., Jaschob, D., Graczyk, B., Shulman, N. J., Wakefield, J., Cooper, S. J., Fields, S., Noble, W. S., Muller, E. G., Davis, T. N., Dunham, M. J., Maccoss, M. J. and Akey, J. M. (2013), 'Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast', *Genome Res* **23**(9), 1496–504.

Slotkin, R. K. and Martienssen, R. (2007), 'Transposable elements and the epigenetic regulation of the genome', *Nat Rev Genet* **8**(4), 272–85.

Smit, A. F. A., Hubley, R. and Green, P. (2013), 'Repeatmasker open-4.0'.
URL: *http://www.repeatmasker.org/*

Smukowski Heil, C., Burton, J. N., Liachko, I., Friedrich, A., Hanson, N. A., Morris, C. L., Schacherer, J., Shendure, J., Thomas, J. H. and Dunham, M. J. (2017), 'Identification of a novel interspecific hybrid yeast from a metagenomic spontaneously inoculated beer sample using hi-c', *Yeast* .

Sánchez-Gracia, A., Maside, X. and Charlesworth, B. (2005), 'High rate of horizontal transfer of transposable elements in *Drosophila*', *Trends Genet* **21**(4), 200–3.

Sniegowski, P., Dombrowski, P. G. and Fingerman, E. G. (2002), '*Saccharomyces cerevisiae* and *Saccharomyces paradoxus* coexist in a natural woodland site in North America and display different levels of reproductive isolation from European conspecifics', *FEMS Yeast Research* **1**, 299–306.

Soderlund, C., Bomhoff, M. and Nelson, W. M. (2011), 'SyMAP v3.4: a turnkey synteny system with application to plant genomes', *Nucleic Acids Res* **39**(10), e68.

Song, S. U., Gerasimova, T., Kurkulos, M., Boeke, J. D. and Corces, V. G. (1994), 'An *env*-like protein encoded by a *Drosophila* retroelement: evidence that *gypsy* is an infectious retrovirus', *Genes Dev* **8**(17), 2046–57.

Souciet, J., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., Bon, E., Brottier, P., Casaregola, S., de Montigny, J., Dujon, B., Durrens, P., Gaillardin, C., Lepingle, A., Llorente, B., Malpertuy, A., Neuvéglise, C., Ozier-Kalogeropoulos, O., Potier, S., Saurin, W., Tekaia, F., Toffano-Nioche, C., Wesolowski-Louvel, M., Wincker, P. and Weissenbach, J. (2000), 'Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies', *FEBS Lett* **487**(1), 3–12.

Spradling, A. C., Bellen, H. J. and Hoskins, R. A. (2011), '*Drosophila P* elements preferentially transpose to replication origins', *Proc Natl Acad Sci U S A* **108**(38), 15948–53.

Stamatakis, A. (2014), 'RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies', *Bioinformatics* **30**(9), 1312–3.

Stamenova, R., Dimitrov, M., Stoycheva, T., Pesheva, M., Venkov, P. and Tsvetkov, T. S. (2008), 'Transposition of *Saccharomyces cerevisiae Ty1* retrotransposon is activated by improper cryopreservation', *Cryobiology* **56**(3), 241–7.

Stanley, D., Fraser, S., Stanley, G. A. and Chambers, P. J. (2010), 'Retrotransposon expression in ethanol-stressed *Saccharomyces cerevisiae*', *Appl Microbiol Biotechnol* **87**(4), 1447–54.

Startek, M., Le Rouzic, A., Capy, P., Grzebelus, D. and Gambin, A. (2013), 'Genomic parasites or symbionts? Modeling the effects of environmental pressure on transposition activity in asexual populations', *Theor Popul Biol* **90**, 145–51.

Stavrou, A. A., Mixao, V., Boekhout, T. and Gabaldon, T. (2018), 'Misidentification of genome assemblies in public databases: the case of *Naumovozyma dairenensis* and proposal of a protocol to correct misidentifications', *Yeast* **[epub ahead of print]**.

Stefanini, I., Dapporto, L., Berna, L., Polsinelli, M., Turillazzi, S. and Cavalieri, D. (2016), 'Social wasps are a *Saccharomyces* mating nest', *Proc Natl Acad Sci U S A* **113**(8), 2247–51.

Stefanini, I., Dapporto, L., Legras, J. L., Calabretta, A., Di Paola, M., De Filippo, C., Viola, R., Capretti, P., Polsinelli, M., Turillazzi, S. and Cavalieri, D. (2012), 'Role of social wasps in *Saccharomyces cerevisiae* ecology and evolution', *Proc Natl Acad Sci U S A* **109**(33), 13398–403.

Stoycheva, T., Massardo, D. R., Pesheva, M., Venkov, P., Wolf, K., Del Giudice, L. and Pontieri, P. (2007), '*Ty1* transposition induced by carcinogens in *Saccharomyces cerevisiae* yeast depends on mitochondrial function', *Gene* **389**(2), 212–8.

Stoycheva, T., Pesheva, M. and Venkov, P. (2010), 'The role of reactive oxygen species in the induction of *Ty1* retrotransposition in *Saccharomyces cerevisiae*', *Yeast* **27**(5), 259–67.

Stritt, C., Gordon, S. P., Wicker, T., Vogel, J. P. and Roulin, A. C. (2018), 'Recent activity in expanding populations and purifying selection have shaped transposable element landscapes across natural accessions of the mediterranean grass *Brachypodium distachyon*', *Genome Biol Evol* **10**(1), 304–318.

Strope, P. K., Skelly, D. A., Kozmin, S. G., Mahadevan, G., Stone, E. A., Magwene, P. M., Dietrich, F. S. and McCusker, J. H. (2015), 'The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen', *Genome Res* **25**(5), 762–74.

Stucka, R., Lochmuller, H. and Feldmann, H. (1989), '*Ty4*, a novel low-copy number element in *Saccharomyces cerevisiae*: one copy is located in a cluster of *Ty* elements and tRNA genes', *Nucleic Acids Res* **17**(13).

Stucka, R., Schwarzlose, C., Lochmuller, H., Hacker, U. and Feldmann, H. (1992), 'Molecular analysis of the yeast *Ty4* element: homology with *Ty1*, *copia*, and plant retrotransposons', *Gene* **122**(1), 119–28.

Subramanian, R. A., Arensburger, P., Atkinson, P. W. and O'Brochta, D. A. (2007), 'Transposable element dynamics of the hAT element *Herves* in the human malaria vector *Anopheles gambiae s.s.*', *Genetics* **176**(4), 2477–87.

Sun, W., Shen, Y. H., Han, M. J., Cao, Y. F. and Zhang, Z. (2014), 'An adaptive transposable element insertion in the regulatory region of the EO gene in the domesticated silkworm, *Bombyx mori*', *Mol Biol Evol* **31**(12), 3302–13.

Sun, Y., Guo, J., Liu, F. and Liu, Y. (2014), 'Identification of indigenous yeast flora isolated from the five winegrape varieties harvested in Xiangning, China', *Antonie Van Leeuwenhoek* **105**(3), 533–40.

Sun, Y., Qin, Y., Pei, Y., Wang, G., Joseph, C., Bisson, L. and Liu, Y. (2017), 'Evaluation of Chinese *Saccharomyces cerevisiae* wine strains from different geographical origins', *Am J Enol Vitic* **68**(1), 73–80.

Sylvester, K., Wang, Q. M., James, B., Mendez, R., Hulfachor, A. B. and Hittinger, C. T. (2015), 'Temperature and host preferences drive the diversification of *Saccharomyces* and other yeasts: a survey and the discovery of eight new yeast species', *FEMS Yeast Res* **15**(3).

Syomin, B. V., Leonova, T. Y. and Ilyin, Y. V. (2002), 'Evidence for horizontal transfer of the LTR retrotransposon *mdg3*, which lacks an *env* gene', *Mol Genet Genomics* **267**(3), pp. 418–23.

Tajima, F. (1989), 'Statistical method for testing the neutral hypothesis by polymorphism', *Genetics* **123**(3), pp.585–595.

Tanaka, A., Tanizawa, H., Sriswasdi, S., Iwasaki, O., Chatterjee, A. G., Speicher, D. W., Levin, H. L., Noguchi, E. and Noma, K. (2012), 'Epigenetic regulation of condensin-mediated genome organization during the cell cycle and upon dna damage through histone H3 lysine 56 acetylation', *Mol Cell* **48**(4), 532–46.

Tang, Z., Zhang, H. H., Huang, K., Zhang, X. G., Han, M. J. and Zhang, Z. (2015), 'Repeated horizontal transfers of four DNA transposons in invertebrates and bats', *Mob DNA* **6**(1), 3.

Taylor, J. W. and Berbee, M. L. (2006), 'Dating divergences in the fungal tree of life: review and new analyses', *Mycologia* **98**(6), 838–49.

Taylor, J. W., Turner, E., Townsend, J. P., Dettman, J. R. and Jacobson, D. (2006), 'Eukaryotic microbes, species recognition and the geographic limits of species: examples from the kingdom Fungi', *Philos Trans R Soc Lond B Biol Sci* **361**(1475), 1947–63.

Temin, H. M. (1991), 'Sex and recombination in retroviruses', *Trends Genet* **7**(3), 71–4.

Teste, M. A., Duquenne, M., Francois, J. M. and Parrou, J. L. (2009), 'Validation of reference genes for quantitative expression analysis by real-time RT-PCR in *Saccharomyces cerevisiae*', *BMC Mol Biol* **10**(99).

Thomas, J., Schaack, S. and Pritham, E. J. (2010), 'Pervasive horizontal transfer of rolling-circle transposons among animals', *Genome Biol Evol* **2**, pp. 656–664.

Thompson, P. J., Macfarlan, T. S. and Lorincz, M. C. (2016), 'Long terminal repeats: From parasitic elements to building blocks of the transcriptional regulatory repertoire', *Mol Cell* **62**(5), 766–76.

Tian, Z., Rizzon, C., Du, J., Zhu, L., Bennetzen, J. L., Jackson, S. A., Gaut, B. S. and Ma, J. (2009), 'Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons?', *Genome Res* **19**(12), 2221–30.

Tkach, J. M., Yimit, A., Lee, A. Y., Riffle, M., Costanzo, M., Jaschob, D., Hendry, J. A., Ou, J., Moffat, J., Boone, C., Davis, T. N., Nislow, C. and Brown, G. W. (2012), 'Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress', *Nat Cell Biol* **14**(9), 966–76.

Todeschini, A. L., Morillon, A., Springer, M. and Lesage, P. (2005), 'Severe adenine starvation activates *Ty1* transcription and retrotransposition in *Saccharomyces cerevisiae*', *Mol Cell Biol* **25**(17), 7459–72.

Tofalo, R., Perpetuini, G., Schirone, M., Fasoli, G., Aguzzi, I., Corsetti, A. and Suzzi, G. (2013), 'Biogeographical characterization of *Saccharomyces cerevisiae* wine yeast by molecular methods', *Front Microbiol* **4**, 166.

Treangen, T. J. and Salzberg, S. L. (2011), 'Repetitive and next-generation sequencing: computational challenges and solutions', *Nat Rev Genet* **13**(1), 36–46.

Tristezza, M., Gerardi, C., Logrieco, A. and Grieco, F. (2009), 'An optimized protocol for the production of interdelta markers in *Saccharomyces cerevisiae* by using capillary electrophoresis', *J Microbiol Methods* **78**(3), 286–91.

Trizzino, M., Park, Y., Holsbach-Beltrame, M., Aracena, K., Mika, K., Caliskan, M., Perry, G. H., Lynch, V. J. and Brown, C. D. (2017), 'Transposable elements are the primary source of novelty in primate gene regulation', *Genome Res* **27**(10), 1623–1633.

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**(12), 4957–62.

Tucker, J. M., Larango, M. E., Wachsmuth, L. P., Kannan, N. and Garfinkel, D. J. (2015), 'The *Ty1* retrotransposon restriction factor p22 targets *Gag*', *PLoS Genet* **11**(10), e1005571.

Ullastres, A., Petit, N. and Gonzalez, J. (2015), 'Exploring the phenotypic space and the evolutionary history of a natural mutation in *Drosophila melanogaster*', *Mol Biol Evol* **32**(7), 1800–14.

Umezu, K., Hiraoka, M., Mori, M. and Maki, H. (2002), 'Structural analysis of aberrant chromosomes that occur spontaneously in diploid *Saccharomyces cerevisiae*: Retrotransposon *Ty1* plays a crucial role in chromosomal rearrangements', *Genetics* **160**.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M. and Rozen, S. G. (2012), 'Primer3–new capabilities and interfaces', *Nucleic Acids Res* **40**(15), e115.

Vakirlis, N., Sarilar, V., Drillon, G., Fleiss, A., Agier, N., Meyniel, J. P., Blanpain, L., Carbone, A., Devillers, H., Dubois, K., Gillet-Markowska, A., Graziani, S., Huu-Vang, N., Poirel, M., Reisser, C., Schott, J., Schacherer, J., Lafontaine, I., Llorente, B., Neuvéglise, C. and Fischer, G. (2016), 'Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus', *Genome Res* **26**(7), 918–32.

van Arsdell, S., Stetler, G. L. and Thorner, J. (1987), 'The yeast repeated element sigma contains a hormone-inducible promoter', *Mol Cell Biol* **7**(2), 749–59.

van der Aa Kuhle, A. and Jespersen, L. (2003), 'The taxonomic position of *Saccharomyces boulardii* as evaluated by sequence analysis of the D1/D2 domain of 26S rDNA, the ITS1-5.8S rDNA-ITS2 region and the mitochondrial cytochrome-c oxidase II gene', *Syst Appl Microbiol* **26**(4), 564–71.

van Houten, J. V. and Newlon, C. S. (1990), 'Mutational analysis of the consensus sequence of a replication origin from yeast chromosome iii', *Mol Cell Biol* **10**(8), 3917–25.

VanHoute, D. and Maxwell, P. H. (2014), 'Extension of *Saccharomyces paradoxus* chronological lifespan by retrotransposons in certain media conditions is associated with changes in reactive oxygen species', *Genetics* **198**(2), 531–45.

VanHulle, K., Lemoine, F. J., Narayanan, V., Downing, B., Hull, K., McCullough, C., Bellinger, M., Lobachev, K., Petes, T. D. and Malkova, A. (2007), 'Inverted repeats channel repair of distant double-strand breaks into chromatid fusions and chromosomal rearrangements', *Mol Cell Biol* **27**(7), 2601–14.

Varmus, H. (1988), 'Retroviruses', *Science* **240**(4858), 1427–35.

Varoquaux, N., Liachko, I., Ay, F., Burton, J. N., Shendure, J., Dunham, M. J., Vert, J. P. and Noble, W. S. (2015), 'Accurate identification of centromere locations in yeast genomes using Hi-C', *Nucleic Acids Res* **43**(11), 5331–9.

Vaughan-Martini, A. (1989), '*Saccharomyces paradoxus comb nov*., a newly separated species of the *Saccharomyces sensu stricto* complex based upon nDNA/nhomologies', *Systematic and Applied Microbiology* **12**, 179–182.

Vaughan-Martini, A. and Kurtzman, C. (1985), 'Deoxyribonucleic acid relatedness among species of the genus *Saccharomyces sensu stricto*', *Int J Sys Bact* **35**, 508–511.

Villanueva-Cañas, J., Rech, G., de Cara, M. and González, J. (2017), 'Beyond SNPs: how to detect selection on transposable element insertions', *Methods in Ecology and Evolution* **8**(6), 728–737.

Voytas, D. F. and Boeke, J. D. (1992), 'Yeast retrotransposon revealed', *Nature* **358**(6389), 717.

Voytas, D. F. and Boeke, J. D. (2002), *Ty1 and Ty5 of Saccharomyces cerevisiae*, 2nd edn, ASM Press, Washington, D.C., book section 26, pp. 631–662.

Wallau, G. L., Capy, P., Loreto, E., Le Rouzic, A. and Hua-Van, A. (2016), 'VHICA, a new method to discriminate between vertical and horizontal transposon transfer: Application to the mariner family within *Drosophila*', *Mol Biol Evol* **33**(4), 1094–109.

Wallau, G. L., Ortiz, M. F. and Loreto, E. L. (2012), 'Horizontal transposon transfer in eukarya: detection, bias, and perspectives', *Genome Biol Evol* **4**(8), 689–99.

Wang, Q. M., Liu, W. Q., Liti, G., Wang, S. A. and Bai, F. Y. (2012), 'Surprisingly diverged populations of *Saccharomyces cerevisiae* in natural environments remote from human activity', *Mol Ecol* **21**(22), 5404–17.

Wang, S. A. and Bai, F. Y. (2008), '*Saccharomyces arboricolus* sp. nov., a yeast species from tree bark', *Int J Syst Evol Microbiol* **58**(Pt 2), 510–4.

Wang, X., Weigel, D. and Smith, L. M. (2013), 'Transposon variants and their effects on gene expression in arabidopsis', *PLoS Genet* **9**(2), e1003255.

Warmington, J., Anwar, R., Waring, R., Davied, R., Indge, K. and Oliver, S. (1986), 'A 'hot-spot' for *Ty* transposition on the left arm of yeast chromosome III', *Nucleic Acids Res* **14**(8).

Warmington, J., Green, R., Newlon, C. and Oliver, S. (1987), 'Polymorphisms on the right arm of yeast chromosome iii associated with *Ty* transposition and recombination events', *Nucleic Acids Res* **15**(21).

Warnefors, M., Pereira, V. and Eyre-Walker, A. (2010), 'Transposable elements: insertion pattern and impact on gene expression evolution in hominids', *Mol Biol Evol* **27**(8), 1955–62.

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**(21).

Wei, W., McCusker, J. H., Hyman, R. W., Jones, T., Ning, Y., Cao, Z., Gu, Z., Bruno, D., Miranda, M., Nguyen, M., Wilhelmy, J., Komp, C., Tamse, R., Wang, X., Jia, P., Luedi, P., Oefner, P. J., David, L., Dietrich, F. S., Li, Y., Davis, R. W. and Steinmetz, L. M. (2007), 'Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM978', *Proc Natl Acad Sci U S A* **104**(31), 12825–30.

Wendland, J. and Walther, A. (2011), 'Genome evolution in the *Eremothecium* clade of the *Saccharomyces* complex revealed by comparative genomics', *G3 (Bethesda)* **1**(7), 539–48.

Wheelan, S. J., Scheifele, L. Z., Martinez-Murillo, F., Irizarry, R. A. and Boeke, J. D. (2006), 'Transposon insertion site profiling chip (tip-chip)', *Proc Natl Acad Sci U S A* **103**(47), 17632–7.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P. and Schulman, A. H. (2007), 'A unified classification system for eukaryotic transposable elements', *Nat Rev Genet* **8**(12), 973–82.

Wilcox, T. P., Zwickl, D. J., Heath, T. A. and Hillis, D. M. (2002), 'Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support', *Mol Phylogenet Evol* **25**(2), 361–71.

Wiley, E. and Lieberman, B. (2011), *Parametric Phylogenetics*, 2nd edn, John Wiley and Sons, Inc, book section 7, pp. 203–228.

Wilhelm, M., Boutabout, M., Heyman, T. and Wilhelm, F. X. (1999), 'Reverse transcription of the yeast *Ty1* retrotransposon: the mode of first strand transfer is either intermolecular or intramolecular', *J Mol Biol* **288**(4), 505–10.

Wilhelm, M. and Wilhelm, F. (2001), 'Reverse transcription of retroviruses and retrotransposons', *Cell Mol Life Sci* **58**(9), 1246–1262.

Wilke, C. and Adams, J. (1992), 'Fitness effects of *Ty* transposition in *Saccharomyces*', *Genetics* **131**(1), 31–42.

Wilke, C., Heidler, S., Brown, N. and Liebman, S. (1989), 'Analysis of yeast retrotransposon *Ty* insertions at the *CAN1* locus', *Genetics* **123**.

Williamson, V. M. (1983), 'Transposable elements in yeast', *Int Rev Cytol* **83**, 1–25.

Williamson, V. M., Cox, D., Young, E. T., Russell, D. W. and Smith, M. (1983), 'Characterization of transposable element-associated mutations that alter yeast alcohol dehydrogenase II expression', *Mol Cell Biol* **3**(1), 20–31.

Winckler, T., Dingermann, T. and Glockner, G. (2002), '*Dictyostelium* mobile elements: strategies to amplify in a compact genome', *Cell Mol Life Sci* **59**(12), 2097–111.

Winston, F., Chaleff, D., Valent, B. and Fink, G. R. (1984), 'Mutations affecting *Ty*-mediated espression of the HIS4 gene of *Saccharomyces cerevisiae*', *Genetics* **107**, 179–197.

Wloch, D. M., Szafraniec, K., Borts, R. H. and Korona, R. (2001), 'Direct estimate of the mutation rate and the distribution of fitness effects in the yeast *Saccharomyces cerevisiae*', *Genetics* **159**(2), 441–52.

Wolfe, K. H., Armisen, D., Proux-Wera, E., OhEigeartaigh, S. S., Azam, H., Gordon, J. L. and Byrne, K. P. (2015), 'Clade- and species-specific features of genome evolution in the Saccharomycetaceae', *FEMS Yeast Res* **15**(5), fov035.

Wolfe, K. H. and Shields, D. C. (1997), 'Molecular evidence for an ancient duplication of the entire yeast genome', *Nature* **387**(6634), 708–13.

Wong, G. K., Passey, D. A., Huang, Y., Yang, Z. and Yu, J. (2000), 'Is "junk" mostly intron DNA?', *Genome Res* **10**(11), 1672–8.

Wood, H. M., Grahame, J. W., Humphray, S., Rogers, J. and Butlin, R. K. (2008), 'Sequence differentiation in regions identified by a genome scan for local adaptation', *Mol Ecol* **17**(13), 3123–35.

Wrent, P., Rivas, E. M., Peinado, J. M. and de Siloniz, M. I. (2017), '*Zygosaccharomyces rouxii* strains CECT 11923 and *Z. rouxii* CECT 10425: Two new putative hybrids?', *Int J Food Microbiol* **241**, 7–14.

Wright, D. A. and Voytas, D. F. (1998), 'Potential retroviruses in plants: *Tat1* is related to a group of *Arabidopsis thaliana Ty3/gypsy* retrotransposons that encode envelope-like proteins', *Genetics* **149**(2), 703–15.

Wright, S. I., Agrawal, N. and Bureau, T. E. (2003), 'Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*', *Genome Res* **13**(8), 1897–903.

Xia, E. H., Zhang, H. B., Sheng, J., Li, K., Zhang, Q. J., Kim, C., Zhang, Y., Liu, Y., Zhu, T., Li, W., Huang, H., Tong, Y., Nan, H., Shi, C., Shi, C., Jiang, J. J., Mao, S. Y., Jiao, J. Y., Zhang, D., Zhao, Y., Zhao, Y. J., Zhang, L. P., Liu, Y. L., Liu, B. Y., Yu, Y., Shao, S. F., Ni, D. J., Eichler, E. E. and Gao, L. Z. (2017), 'The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis', *Mol Plant* .

Xie, W., Gai, X., Zhu, Y., Zappulla, D. C., Sternglanz, R. and Voytas, D. F. (2001), 'Targeting of the yeast *Ty5* retrotransposon to silent chromatin is mediated by interactions between integrase and *Sir4p*', *Molecular and Cellular Biology* **21**(19), 6606–6614.

Xiong, Y. and Eickbush, T. (1988), 'Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns', *Mol Biol Evol* **5**(6).

Xiong, Y. and Eickbush, T. H. (1990), 'Origin and evolution of retroelements based upon their reverse transcriptase sequences', *EMBO J* **9**(10), 3353–62.

Xu, H. and Boeke, J. D. (1987), 'High-frequency deletion between homologous sequences during retrotransposition of *Ty* elements in *Saccharomyces cerevisiae*', *PNAS* **84**.

Xufre, A., Albergaria, H., Girio, F. and Spencer-Martins, I. (2011), 'Use of interdelta polymorphisms of *Saccharomyces cerevisiae* strains to monitor population evolution during wine fermentation', *J Ind Microbiol Biotechnol* **38**(1), 127–32.

Yamada, Y., Mikata, K. and Banno, I. (1993), 'Reidentification of 121 strains of the genus *Saccharomyces*', *Bull JFCC* **9**, 95–119.

Yang, Z. and Rannala, B. (1997), 'Bayesian phylogenetic inference using sequences: a Markov Chain Monte Carlo Method', *Mol Biol Evol* **14**(7), 717–24.

Yieh, L., Kassavetis, G., Geiduschek, E. P. and Sandmeyer, S. B. (2000), 'The Brf and TATA-binding protein subunits of the RNA polymerase III transcription factor IIIB mediate position-specific integration of the *gypsy*-like element, *Ty3*', *J Biol Chem* **275**(38), 29800–7.

Yofe, I., Weill, U., Meurer, M., Chuartzman, S., Zalckvar, E., Goldman, O., Ben-Dor, S., Schutze, C., Wiedemann, N., Knop, M., Khmelinskii, A. and Schuldiner, M. (2016), 'One library to make them all: streamlining the creation of yeast libraries via a SWAp-Tag strategy', *Nat Methods* **13**(4), 371–378.

Yoshikawa, K., Tanaka, T., Ida, Y., Furusawa, C., Hirasawa, T. and Shimizu, H. (2011), 'Comprehensive phenotypic analysis of single-gene deletion and overexpression strains of *Saccharomyces cerevisiae*', *Yeast* **28**(5), 349–61.

Yu, K. and Elder, R. (1989), 'A region internal to the coding sequences is essential for transcription of the yeast *Ty*-D15 element', *Mol Cell Biol* **9**(9), 3667–3678.

Yue, J. X., Li, J., Aigrain, L., Hallin, J., Persson, K., Oliver, K., Bergstrom, 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**(6), 913–924.

Zeyl, C. (2000), 'Budding yeast as a model organism for population genetics', *Yeast* **16**.

Zeyl, C., Bell, G. and Green, D. (1996), 'Sex and the spread of retrotransposon *Ty3* in experimental populations', *Genetics* **143**.

Zhang, L., Yan, L., Jiang, J., Wang, Y., Jiang, Y., Yan, T. and Cao, Y. (2014), 'The structure and retrotransposition mechanism of LTR-retrotransposons in the asexual yeast *Candida albicans*', *Virulence* **5**(6), 655–64.

Zhang, Q. J. and Gao, L. Z. (2017), 'Rapid and recent evolution of LTR retrotransposons drives rice genome evolution during the speciation of aa- genome *Oryza* species', *G3 (Bethesda)* .

Zhang, Z., Qian, W. and Zhang, J. (2009), 'Positive selection for elevated gene expression noise in yeast', *Mol Syst Biol* **5**, 299.

Zhou, T., Gu, W. and Wilke, C. O. (2010), 'Detecting positive and purifying selection at synonymous sites in yeast and worm', *Mol Biol Evol* **27**(8), 1912–22.

Zhouravleva, G., Frolova, L., Le Goff, X., Le Guellec, R., Inge-Vechtomov, S., Kisselev, L. and Philippe, M. (1995), 'Termination of translation in eukaryotes is governed by two interacting polypeptide chain release factors, eRF1 and eRF3', *EMBO J* **14**(16), 4065–72.

Zhu, Y., Dai, J., Fuerst, P. G. and Voytas, D. F. (2003), 'Controlling integration specificity of a yeast retrotransposon', *Proc Natl Acad Sci U S A* **100**(10), 5891–5.

Zhu, Y., Zou, S., Wright, D. A. and Voytas, D. F. (1999), 'Tagging chromatin with retrotransposons: target specificity of the *Saccharomyces Ty5* retrotransposon changes with the chromosomal localization of *Sir3p* and *Sir4p*', *Genes Devel* **13**.

Zou, S., Kim, J. and Voytas, D. F. (1996*a*), 'The *Saccharomyces* retrotransposon *Ty5* influences the organisation of chromosome ends', *Nucleic Acids Res* **24**(23).

Zou, S., Kim, J. and Voytas, D. F. (1996*b*), 'The *Saccharomyces* retrotransposon *Ty5* integrates preferentially into regions of silent chromatin at the telomeres and mating loci', *Genes Dev* **10**, 634–645.

Zou, S. and Voytas, D. F. (1997), 'Silent chromatin determines target preference of the *Saccharomyces* retrotransposon *Ty5*', *PNAS* **94**, 7412–7416.

Zou, S., Wright, D. A. and Voytas, D. F. (1995), 'The *Saccharomyces Ty5* retrotransposon family is associated with origins of DNA replication at the telomeres and the silent mating locus HMR', *Proc Natl Acad Sci U S A* **92**(3), 920–4.