Bioinformatics and evolutionary genomics of transposable elements in yeast

by

Jingxuan Chen

(Under the Direction of Casey Bergman)

Abstract

Transposable elements (TEs) are mobile repetitive DNA sequences that are ubiquitous in eukaryotic organisms. *Saccharomyces cerevisiae* is a leading model organism for studying TEs at the molecular level. However, the species-wide polymorphism and evolution of TE for *Saccharomyces* yeasts have not been fully revealed, partially due to the technical challenges in identifying repetitive DNA in high-throughput sequencing data. In this dissertation, I focused on developing new bioinformatics approaches and investigating the evolutionary history of TEs among *Saccharomyces* species.

First, I updated a meta-pipeline called McClintock that consists of 12 polymorphic TE detectors using short-read whole genome sequencing (WGS) data, and developed a reproducible simulation framework to comprehensively evaluate the performance of component TE detectors. Simulation analysis using *Saccharomyces* yeast as a paradigm identified four "best-in-class" methods that were further confirmed using empirical *S. cerevisiae* WGS datasets, which was also used to provide novel insights into TE biology of *S. cerevisiae*. This work presented a user-friendly pipeline and reproducible simulation evaluation framework for TE detection with short-read WGS data, which would facilitate future studies on TE polymorphism in different organisms.

Next, I sequenced and generated two high-quality genome assemblies of *Saccharomyces* strains that are useful for studies on TEs: the *S. paradoxus* strain DG1768 and the *S. uvarum* strain CBS 7001 combining the PacBio and/or Illumina sequencing data. I documented genetic alterations of the key lab strain – DG1768 – in molecular biological research about Ty1 mobility. This study also provided resources of an important reference genome for *S. uvarum* type strain CBS 7001 which should benefit future genomic studies on this species.

Finally, I investigated the evolutionary history of the Ty4 family among *Saccharomyces* species using both short-read and long-read WGS data. First, I implemented an integrated pipeline that compiled WGS datasets, reconstructed host species phylogeny, and estimated TE abundance. The results revealed species-wide ancestral states of Ty4/Tsu4 subfamilies across multiple *Saccharomyces* species. Next, I conducted genome-wide TE annotation with a RepeatMasker-based pipeline from public genome assemblies of *Saccharomyces* species, that cross-validated observations from short-read data and provided a rich dataset of full-length TE sequences for phylogenetic analysis. My results reported several new horizontal TE transfer (HTT) events in the Ty4 family of *Saccharomyces* species and further supported the significant role of HTT in shaping Ty content in *Saccharomyces* yeasts.

INDEX WORDS: Transposable element, *Saccharomyces*, Yeast, Evolution, Genomics, Dissertations, Theses (academic)

Bioinformatics and evolutionary genomics of transposable elements in yeast

by

Jingxuan Chen

B.S., Beijing Normal University

China, June, 2018

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2023

© 2023

Jingxuan Chen

All Rights Reserved

Bioinformatics and evolutionary genomics of transposable elements in yeast

by

Jingxuan Chen

Approved:

Major Professor: Casey Bergman

Committee: Justin Bahl
Travis Glenn
Zachary Lewis

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
December 2023

Acknowledgments

I would like to express my immense gratitude to my esteemed mentor and PI, Dr. Casey Bergman, for his exceptional mentorship and invaluable guidance throughout my doctoral journey. His vast knowledge, expertise, innovative mind and emphasis on maintaining the highest standard of scientific research deeply influenced my path in academic research. In addition, I want to express my appreciation to members of the Bergman lab that I have worked with over the past five years: Dr. Shunhua Han, Dr. Guilherme Dias, Preston Basting, and Oliver Gregory. I enjoyed the time working with them. I also want to thank my advisory committee members, Dr. Justin Bahl, Dr. Travis Glenn, Dr. Dave Hall, and Dr. Zachary Lewis, for their helpful comments on my research projects and dissertation. Finally, I want to thank members of the Garfinkel Lab, Dr. David Garfinkel, Dr. Sean Beckwith, and Adam Hannon-Hatfield, for all the inspiring discussions during joint lab meetings.

For Chapter 2, I want to thank members of the Bergman and Garfinkel Labs for helpful discussions throughout the project; the BioConda community for contributing software recipes for software dependencies used in this project; and the Georgia Advanced Computing Resource Center for computing resources and HPC support. This work was funded by a University of Georgia Research Education Award Traineeship (P.J.B.), the University of Georgia Research Foundation (C.M.B.), and NIH grant R01GM124216 (D.J.G and C.M.B.).

For Chapter 3, I want to thank thank Michael Ciriacy, who performed all genetic steps essential for the Ty1-155 knock-out experiment at the University of Düsseldorf and tragically died before seeing the fruition of his work. I also thank Kyria Boundy-Mills (University of California-Davis) and Jasper Rine (University of California-Berkeley) for assistance determining the provenance of strain 155. I thank Marissa Howard, Noah Workman, and Magdy Alabady at the University of Georgia Genomics and Bioinformatics Core for assistance

with PacBio library preparation and sequencing, Tyler Earp for assistance with Illumina library preparation, GeneWiz for Illumina sequencing, and the Georgia Advanced Computing Resource Center for computational resources. I thank Caiti Smukowski Heil (North Carolina State University) for helpful comments on the manuscript. This work was funded by the University of Georgia Research Foundation (CMB) and NIH grant R01GM124216 (DJG and CMB).

For Chapter 4, I want to thank members of the Bergman and Garfinkel Labs for inspiring discussion and comments along the project. I want to thank Marissa Howard, Noah Workman, and Magdy Alabady at the University of Georgia Genomics and Bioinformatics Core for assistance with PacBio library preparation and sequencing, and Georgia Advanced Computing Resource Center for technical supporting and computational resources. This work was funded by the University of Georgia Research Foundation (CMB) and NIH grant R01GM124216 (DJG and CMB).

Finally, I would like to express my deepest gratitude to my parents, Xiaoyan Li and Jiarong Chen, for their unwavering support, kind understanding and encouragement along my path of study and life. I also want to thank my friends especially the "best bioinformaticians group" who have made my life in Athens more colorful.

Table of Contents

Page

Acknowledgments iv

List of Figures viii

List of Tables xi

Chapter

1 Introduction 1

1.1 Transposable elements 1

1.2 Computational methods to annotate and detect TEs in genome sequences 4

1.3 *Saccharomyces* yeasts and population structure 8

1.4 Transposable elements in *Saccharomyces* yeasts 22

1.5 Dissertation Outline 29

2 Reproducible evaluation of transposable element detectors with McClintock 2 31

2.1 abstract 32

2.2 Introduction 33

2.3 Implementation 36

2.4 Materials and Methods 43

2.5 Results and Discussion 49

2.6 Conclusions 78

2.7 Data availability 81

2.8 Acknowledgements 81

3 High-quality genome assemblies of *S. paradoxus* and *S. uvarum* . . . 83

3.1 Abstract 84

3.2 Genome Assembly of the Ty1-Less *Saccharomyces paradoxus* Strain DG1768. 84

3.3 Long-Read Genome Assembly of *Saccharomyces uvarum* Strain CBS 7001. 85

3.4 Data availability 89

3.5 Acknowledgements 89

4 Evolution of the Ty4 retrotransposon family in *Saccharomyces* yeasts 90

4.1 Abstract 91

4.2 Introduction 92

4.3 Methods and Materials 95

4.4 Results and Discussion 98

4.5 Conclusion 129

4.6 Acknowledgements 130

5 Conclusions 132

5.1 Summary 132

5.2 Future work 134

Bibliography 137

List of Figures

1.1 Schematic pipeline for coverage-based TE copy number estimation strategy. . 8
1.2 Cladogram of accepted phylogenetic relationships among *Saccharomyces* genus. 9
1.3 Visualization of annotated Ty canonical sequences in *S. cerevisiae* and *S. paradoxus*. 24
2.1 Sample screenshots from the new interactive HTML report in McClintock 2. 40
2.2 Validation of McClintock 2 meta-pipeline and simulation system. 51
2.3 McClintock 2 re-implementation improves CPU efficiency and run time on multi-core architectures. 54
2.4 Positional accuracy of six original McClintock component methods in simulated data generated by the yeast insertion model (Simulation 3). 56
2.5 Positional accuracy of six new McClintock component methods in simulated data generated by the yeast insertion model (Simulation 3). 57
2.6 Positional accuracy of six original McClintock component methods in simulated data generated by the random insertion model (Simulation 4). 58
2.7 Positional accuracy of six new McClintock component methods in simulated data generated by the random insertion model (Simulation 4). 59
2.8 TSD length distributions in simulated data generated by the yeast insertion model (Simulation 3). 61
2.9 TSD length distributions in simulated data generated by the random insertion model (Simulation 4). 62
2.10 Recall of McClintock 2 component methods in simulated data generated by the yeast insertion model (Simulation 3). 63

2.11 Precision of McClintock 2 component methods in simulated data generated by the yeast insertion model (Simulation 3). 63
2.12 Recall of McClintock 2 component methods in simulated data generated by the random insertion model (Simulation 4). 64
2.13 Precision of McClintock 2 component methods in simulated data generated by the random insertion model (Simulation 4). 64
2.14 Performance of McClintock 2 component methods in simulated yeast WGS data 66
2.15 Numbers of Ty elements predicted by McClintock 2 components in a world-wide sample of yeast strains. 71
2.16 Correlation across 12 component methods in all non-reference predictions for 1,011 yeast isolates. 72
2.17 Distribution of estimated copy number of full-length TEs across 1,011 yeast isolates. 73
2.18 Estimated total number of full-length TEs per Ty family across 1,011 *S. cerevisiae* isolates. 75
2.19 *copia*-superfamily retrotransposons show a dyad pattern of insertion in nucleosome-bound regions upstream of yeast tRNA genes. 76
2.20 Predicted TE insertion profiles around tRNA genes in *S. cerevisiae* for the six original McClintock component methods. 78
2.21 Predicted TE insertion profiles around tRNA genes in *S. cerevisiae* for the six new McClintock component methods. 79
4.1 Host phylogeny of *S. paradoxus* annotated with Ty4/Tsu4 copy number estimation. 100
4.2 Host phylogeny of *S. cerevisiae* annotated with the presence or absence of Ty4/Tsu4. 104
4.3 Host phylogeny of *S. uvarum* annotated with Ty4/Tsu4 copy number estimation. 107

4.4 Host phylogeny of *S. eubayanus* annotated with Ty4/Tsu4 copy number estimation. 108
4.5 Host phylogeny of *S. mikatae* annotated with Ty4/Tsu4 copy number estimation. 111
4.6 Re-annotation summary for Ty4/Tsu4 content in *S. cerevisiae* assemblies generated by O'donnell *et al.* [212]. 116
4.7 Phylogenetic analysis for the internal coding regions from full-length Ty4/Tsu4 copies in *Saccharomyces* genomes. 120
4.8 ML phylogeny for the internal coding region from full-length Ty4/Tsu4 copies in *S. paradoxus* and *S. cerevisiae* genomes. 121
4.9 ML phylogeny for the internal coding region from full-length Ty4/Tsu4 copies in *S. uvarum* and *S. eubayanus* genomes. 123
4.10 ML phylogeny for the internal coding region from full-length Ty4/Tsu4 copies in *S. mikatae* and *S. jurei* genomes. 125
4.11 Sequence divergence between full-length elements of Tsu4 in *S. mikatae* and *S. uvarum*, and Ty4 in *S. cerevisiae*. 126
4.12 ML phylogeny for 5' and 3' internal regions from Tsu4 full-length elements in *S. mikatae*, *S. jurei*, and *S. uvarum*, plus representatives of Ty4 elements. . 128

List of Tables

1.1 List of public short-read WGS datasets of *S. cerevisiae*. 11
1.2 List of public long-read WGS datasets of *Saccharomyces* species. 13
1.3 List of public short-read WGS datasets of *S. paradoxus*. 16
1.4 List of public short-read WGS datasets of *S. uvarum*. 19
1.5 List of public short-read WGS datasets of *S. eubayanus*. 20
2.1 Major features that distinguish the original McClintock version reported in Nelson *et al.* [197] and the McClintock 2 system reported in the current study. 37
2.2 Distinguishing features of simulation experiments performed in Nelson *et al.* [197] and this study. 46
3.1 Methods and properties of the PacBio and Illumina whole genome sequencing datasets used to assemble and polish the DG1768 genome. 87
3.2 Statistics for *de novo* assemblies of CBS 7001 87
4.1 Statistics for *de novo* genome assemblies of *S. mikatae* strains IFO 1815 and NBRC 10994. 113
4.2 Ty4 and Tsu4 content in whole genome assemblies of multiple *Saccharomyces* species. 113

Chapter 1

Introduction

1.1 Transposable elements

Transposable elements (TEs) are repetitive DNA sequences that can mobilize and proliferate within host genomes. TEs were firstly described by Dr. Barbara McClintock as "controlling elements" from the maize genome in the 1950s [181]. Since then, TEs have been discovered in almost all eukaryotic organisms, with striking variation in the abundance and diversity across organisms (reviewed in [279]). In some species, TEs can accumulate rapidly to occupy a large proportion of the host genome and thus play an important role in genome size evolution. For instance, the accumulation of TEs can give rise to an expanded genome size in maize, where almost 85% of the host genome are TEs [251]. In humans, TEs occupy about 45% of the genome [155]. TEs have been purged from a few eukaryotic genomes (such as *Plasmodium falciparum* [144]) or suppressed to a relatively low abundance with variable mechanisms to reduce TE load [17]. For instance, in budding yeast *Saccharomyces cerevisiae*, merely 3% of the host genome is TEs [143, 48].

In general, two major classes of eukaryotic TEs are defined based on their transposition intermediate: Class I (Retrotransposons) and Class II (DNA transposons) [88]. Both major classes can be further divided into several subclasses, which could then be distinguished into superfamilies and families in terms of phylogenetic relationships [280]. More specifically, TE families are defined by the "80-80-80" rule, i.e, elements within one family are longer than 80 bp, sharing >80% sequence identity for >80% of the length [280], although the thresholds are relatively fuzzy. In addition, TEs can be classified as autonomous and nonautonomous elements, based on whether or not they are capable of encoding enzymes that meet the

requirement for their own transposition. The abundance of different classes of TEs are also variable across eukaryotic organisms [279].

1.1.1 Class I: Retrotransposons

Class I transposons replicate via RNA intermediates, which will then be reverse-transcribed to cDNA copies before integrating into the host genome. Because the original TE copy is retained after TE replication, retrotransposons are also referred to as "copy-and-paste" elements. Retrotransposons can be classified into two major subclasses based on the presence or absence of flanking long terminal repeats (LTRs): LTR retrotransposons and non-LTR retrotransposons.

LTR retrotransposons share a similar structure and replication mechanism with retroviruses [62], which usually consists of at least two essential sets of open reading frames (ORFs), i.e, *Gag* and *Pol*. While the *Gag* gene is responsible for the formation of virus-like particles (VLPs) (see below), the *Pol* gene encodes multiple proteins (e.g, integrase (IN), reverse transcriptase (RT), Ribonuclease H (RH) and protease (PR)) that are functional during the retrotransposition process. Some LTR elements may also contain a third *env*-like ORF (such as the *Gypsy* element from *Drosophila*) homologous to the retroviral *env* gene, which encodes a transmembrane host receptor-binding protein related to the virus transmission [62].

The mechanism of LTR-retrotransposition has been well-studied and was firstly uncovered with Ty1 in *S. cerevisiae* [32]. Firstly, the TE is transcribed into RNA by RNA polymerase II (Pol II), and then exported from nucleus into the cytoplasm. Secondly, RNA is translated into two major gene products, i.e, *Gag* and *Gag-Pol* proteins. The RNA and two proteins together assemble into VLPs, where *Gag* and *Gag-Pol* proteins are further cleaved, resulting in mature *Gag*, IN, RT, RH and PR proteins. Then, within the VLP, RNA is reverse transcribed by RT using tRNA as primer to form a double-stranded cDNA. Finally, the IN protein integrates the cDNA sequence into the host genome, sometimes using host

factors to select integration sites. During the integration process, IN makes a staggered cut at the targeting site, which is later fixed via DNA repair, introducing target-site duplications (TSDs) at both sides of the new TE copy. The length of TSDs are usually identical across all copies of a family, ranging from 5 to 10 bp. Through this process, the autonomous LTR element gets replicated and inserted to another genomic location, whereas the original copy at the donor site is unchanged, resulting in an increasing copy number in the host genome.

The phylogenetic relationship of RT domain [62] further classifies LTR retrotransposons into several superfamilies. Two of the most abundant and well-studied examples are the Ty1/Copia and Ty3/Gypsy superfamilies, both of which were originally characterized in *Drosophila* [185] and *Saccharomyces* yeasts [32, 95]. LTR retrotransposons are the only TE subclass found in *Saccharomyces* yeasts, which will be described more in detail in Section 1.4.

Non-LTR retrotransposons are another major subclass that use RNA intermediates and reverse transcriptase in their replication process, and include short interspersed nucleotide elements (SINEs) and long interspersed nucleotide elements (LINEs). LINEs are autonomous elements whereas SINEs are nonautonomous elements which rely on proteins encoded by LINEs to mobilize. Autonomous non-LTR elements are usually comprised of two open reading frames (ORFs), i.e, ORF1 and ORF2. While ORF1 is absent in some superfamilies such as R2 [45], ORF2 is more essential since it encodes proteins necessary for reverse transcription [174]. The first step for the retrotransposition of non-LTR elements – production of an RNA intermediate *via* Pol II transcription – is similar to that of LTR retrotranposon. However, non-LTR elements utilize a target-primed reverse transcription (TPRT) process during chromosomal integration. As a consequence of TPRT, the 5' side of the newly integrated copy is often truncated, which might then prevent new transposition activities [62]. Non-LTR retrotransposons are highly active in vertebrates, especially in human, where LINE (L1) and SINE (*Alu*) together make up the majority of TEs in the human genome [123].

1.1.2 Class II: DNA transposons

The majority of DNA transposons replicate with a "cut-and-paste" mechanism via DNA intermediates, in contrast to "copy-and-paste" retrotransposons. Typical "cut-and-paste" elements (e.g, *hAT* and Tc1/mariner superfamilies) usually consist of a single ORF encoding a transposase, flanked by two tandem inverted repeats (TIRs) [238]. The transposition process starts with an "excision" step, which depends nucleophilic reactions to cleave DNA strands at TIRs from host genome, and then ends with an "integration" step, which pastes DNA copies at the targeting site and introduces TSDs on each side through DNA repair (reviewed by [119]). Alternatively, some DNA transposons without TIRs, such as *Helitrons* elements, can mobilize with a "peel-and-paste" mechanism involving a rolling-circle DNA intermediate [137]. Although the "cut-and-paste" process itself is nonreplicative, copy number can still increase since the excised donor site can be repaired via homologous recombination, using a template of either the sister chromatid or the homologous sequence on the same chromosome [187].

1.2 Computational methods to annotate and detect TEs in genome sequences

With the widespread availability of genome data, many approaches have been developed to discover and detect TEs from genome sequences. In general, computational approaches to analyze TEs in genomes can be divided to four categories: (i) *De novo* TE discovery from the genome, aiming to identify consensus sequences of novel TE families; (ii) TE annotation from whole genome assemblies (WGA) using a TE query library, aiming to locate and annotate TE sequences of each element; (iii) Polymorphic TE detection tools utilizing un-assembled sequencing reads, aiming to locate non-reference TE insertions; and (iv) Coverage-based TE abundance estimation with un-assembled sequencing reads, aiming to investigate the presence/absence of TE family and/or TE copy numbers.

High-quality WGA are fundamental to both *de novo* TE discovery and TE annotation. *De novo* genome assemblies generated from short-read sequences are often too fragmented

especially in repetitive elements like TEs, which are either placed in separate short contigs or failed to be assembled at all (reviewed by [266]). Therefore, the value of using short-read in WGA for TE discovery and annotation is relatively limited. The continuity of genome assemblies has been strikingly improved by recent advances in third-generation (long-read) sequencing strategies, which greatly facilitate the identification of structural variants and repetitive regions in the genome. However, technical obstacles still retain in the field of genome assembly even with long reads of multiple kilobases, especially repetitive regions, like regions rich in TEs [215, 266]. Besides, even if extra-long reads are capable of spanning the whole repetitive region and lead to proper assemblies, the base error rates of long-read strategies are usually much higher than that of short-read sequencing (error rate of 1-1.5% for Illumina, 10-15% for PacBio and 5-20% for ONT; reviewed by [221]). Base errors in repeat regions of WGAs are hard to fix due to the repetitive mapping issue for reads that are fully-contained in TE regions, which can't be polished by short-read sequencing data as they can for unique regions [272].

1.2.1 *De novo* TE discovery

De novo TE discovery refers to the process that identifies new TE families from the genome without known information on TE types, structures or sequences, and is fundamental to subsequent TE query-based strategies (sections 1.2.2, 1.2.3, 1.2.4). The major process is to initially detect repeats in the genome assembly using classic sequence searching approaches (e.g, REPuter [154], RepeatFinder [270], BLAST [6], BLAT [142] etc.), followed by sorting repeats into TE families and filtering out non-TE regions (e.g, RECON [13], PILER [**?**], RepeatScout [225], LTRharvest [81] etc; reviewed by [22]). This process can be implemented with an automated and integrated TE family identification and modeling pipeline, RepeatModeler (`http://www.repeatmasker.org/RepeatModeler/` [89]). The *de novo* TE discovery typically needs manual curation [48]. In addition to the WGA-based TE discovery

strategy, RepeatExplorer is also developed as a *de novo* TE identification pipeline that uses un-assembled NGS reads for eukaryotes [205, 207, 206].

1.2.2 Query-based TE annotation from whole genome assemblies

The most straightforward and comprehensive strategy of TE annotation in whole genome assemblies involves RepeatMasker-based pipelines using library of known TE query sequences [48, 292, 64]. In most cases, RepeatMasker (`https://www.repeatmasker.org/` [257]) is firstly performed to characterize and mask Ty elements in the genome assembly based on sequence similarity compared to Repbase or custom TE query library, followed by filtering false-positive hits and fixing mis-annotations. RepeatMasker output can be post-processed automatically by executing REannotate [216] or "One code to find them all" [9]. Custom post-processing of RepeatMasker output is also commonly performed. For example, in Czaja *et al.* [64], preliminary RepeatMasker annotations were improved using custom scripts dealing with the following cases: (i) Removing divergent hits compared to query (>20% sequence divergence); (ii) Fixing mis-labeling of closely related Ty1/2 LTRs; (iii) Splitting tandem repeats sharing a common LTR element; (iv) Splitting mis-joined solo LTRs; and (v) Renaming annotated elements. The hallmark of RepeatMasker-based strategy is that all TE insertions, regardless of their structures and completeness, can be annotated comprehensively, which, on the other hand, requires two key inputs, i.e, high-quality WGA and adequate TE canonical sequences as query.

1.2.3 Polymorphic non-reference TE detection tools with un-assembled reads

Another strategy to annotate the location of TEs with re-sequencing data bypasses the challenges of WGA by using un-assembled high-throughput short-read data. Many such polymorphic TE detection tools have been developed (list of tools can be found in `https://tehub.org/` [82] and [85]). In general, polymorphic TE detection algorithms can be divided

into two major categories, based on the utilized type of evidence: (i) Split reads (SR) refer to reads mapped at the junction of reference genome and non-reference TE insertion region, i.e, a single read is split or clipped into reference genome and TE parts; (ii) Discordant read pairs (RP) contain one read of the pair being uniquely mapped to reference genome while the other being mapped to TE regions (reviewed by [85, 234]). Some methods may combine both types of evidence such as TEMP [295], TEMP2 [289], RelocaTE2 [51], TEBreak [249], while only either type of evidence is used by the other methods such as the SR-only tools ngs_te_mapper [168], ngs_te_mapper2 [110], RelocaTE [235], and the RP-only tools PoPoolationTE [146], PoPoolationTE2 [147], TE-locate [223], RetroSeq [139], TEFLoN [2]. Tools that incorporate SR evidence are usually more accurate in locating non-reference TEs with exact TSD coordinates [197]. Aiming to facilitate the easy installation, execution, output report and so on, the meta-pipeline McClintock [197] has been developed that integrates multiple polymorphic TE detection methods (see details in Chapter 2).

Since polymorphic TE detectors are developed by separate research groups focused on different TE types, specific goals and organisms [234], it would not be surprising that their performance fluctuate greatly under different circumstances. Therefore, comprehensive benchmarking of these tools are practical for researchers to select appropriate tools for their specific purposes, which requires the "gold-standard" simulated or empirical data [57, 289, 269, 277, 234, 178].

1.2.4 Coverage-based TE abundance estimation with high-throughput data

Another assembly-free strategy to investigate TE content in genomes uses a coverage-based copy number estimation approach. The fundamental rationale is competitive read mapping to the reference genome relative to TE canonical sequences, followed by calculating the average depth based on reads aligned to TE sequences and normalized to depth in unique regions of the genome. For instance, this approach has been applied to investigate species-wide Ty abundance for *S. cerevisiae* in Bleykasten-Grosshans *et al.* (2021) [28]. The detailed

coverage profiles of TE regions could be visualized and further inspected to identify potential internal deletions, fixed or polymorphic mutations relative to query using pooled sequencing data as input in DeviaTE [278]. Furthermore, the mapping quality within TE regions could be improved by hard-masking reference TEs in the genome (i.e, by converting A/T/C/G into Ns) using RepeatMasker before read alignment (Fig 1.1) [257], and by splitting query sequences into LTR and internal sequences, both aiming to avoid repetitive or ambiguous read mapping (details in Chapter 2 and 4). Accuracy of the coverage-based approach is affected by the sequence divergence between actual TE copies and query sequences during read mapping.

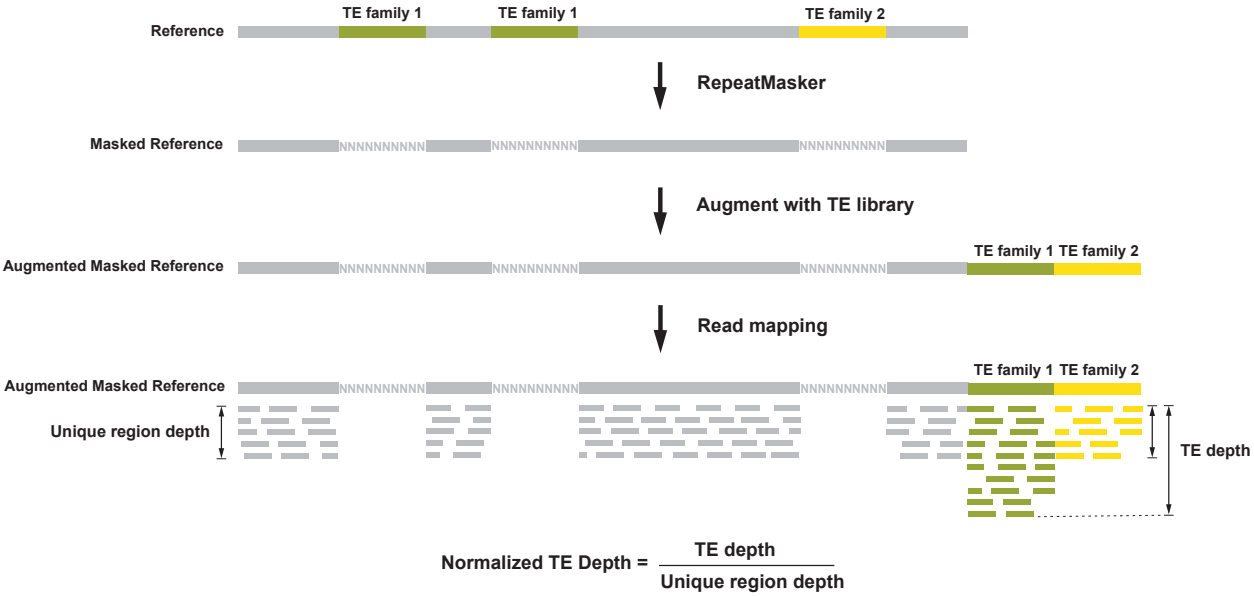

Figure 1.1: **Schematic pipeline for coverage-based TE copy number estimation strategy.**

1.3 *Saccharomyces* YEASTS AND POPULATION STRUCTURE

The genus *Saccharomyces* is comprised of unicellular fungi that can reproduce by budding, originated from ∼10-20 mya [140, 37]. Up to now, eight distinct natural species have been recognized in the *Saccharomyces* genus which are reproductively isolated by post-zygotic

barriers [171, 106]: *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. jurei*, *S. kudriavzevii*, *S. arboricola*, *S. uvarum*, *S. eubayanus* (Fig 1.2). All *Saccharomyces* members share some common features including same number of chromosomes (n=16), compact genome size (∼12 Mbp), easy laboratory propagation, and short generation times.

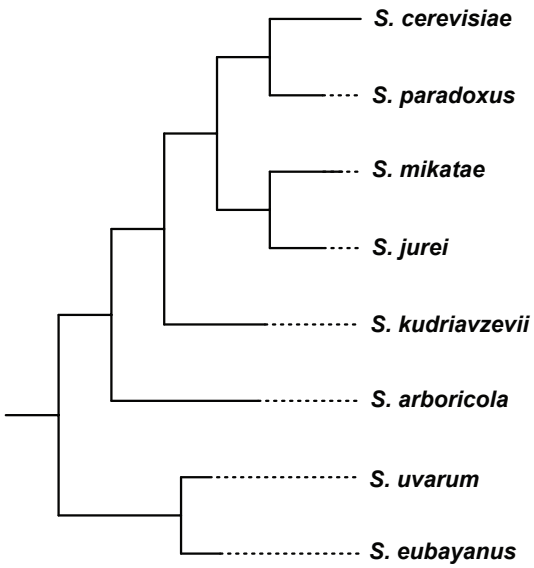

Figure 1.2: **Schematic cladogram of phylogenetic relationships among *Saccharomyces* species.** Topology is adapted from Almeida *et al.* (2014) [4] and Peris *et al.* (2023) [220].

1.3.1 *Saccharomyces cerevisiae*

Saccharomyces cerevisiae, also known as baker's yeast, plays an essential role not only in the fermentation industry involving winemaking, brewing and baking, but also in molecular and evolutionary laboratories. As a leading model eukaryotic organism, the *S. cerevisiae* genome is compact (only $\sim$ 12 Mb), easily-sequenced [102] and well-defined [59], which provides substrate for further genetic experiments. Many key biological mechanisms have been extensively studied in *S. cerevisiae*, so that researchers can control the genetic background and environments, which facilitates the development and validation of new experimental strategies (reviewed by [121]). Specifically, the reference strain S288c and its deriva-

tives have been thoroughly studied in basic research. The *Saccharomyces* Genome Database (SGD, `https://www.yeastgenome.org/`) has comprehensively integrated current experimental results, along with high-throughput sequencing data and well-defined genomic annotations [78]. The latest version of *S. cerevisiae* S288c reference genome (version R64.1.1) is recently updated by SGD project, corresponding to version sacCer3 released by UCSC Genome Browser (`http://genome.ucsc.edu/`) [83].

With the rapid development of sequencing technology in the past 40 years [253], whole genome sequencing (WGS) data have been increasing available, facilitating research on *S. cerevisiae* at the population level. The *Saccharomyces* Genome Resequencing Project (SGRP) was a pioneering project that analyzed Sanger WGS data of more than 30 *S. cerevisiae* strains from divergent domesticated and wild origins, revealing five distinct populations (i.e, Wine/European, West African, Malaysian, North American and Sake) plus several mosaic strains generated by admixture from pure lineages [172]. The SGRP strains were then re-sequenced using next-generation technology in [23]. A similar population structure was later confirmed by "100-genomes" project of *S. cerevisiae*, plus several strains isolated from clinical origin [261]. While the majority of *S. cerevisiae* strains originally included in SGRP are human-associated, an increasing number of populations from wild sources have been identified to resolve questions on species-wide genetic diversity and domestication origins. Wang *et al.* [275] has identified surprisingly diverged *S. cerevisiae* lineages from China designated as CHN I-VIII, with greatest divergence shown in "pure wild" isolates from primeval forest (i.e, CHN I). A novel Mediterranean oak lineage has been discovered as the closest wild origin of the domesticated wine lineage [3]. The oak niche is considered as the candidate natural habitat of *S. cerevisiae* [126, 14]. I note that the terminology "wild" or "natural" is somewhat arbitrary since natural habitats like oaks and the underneath soil may also be affected by human activities.

The most comprehensive survey of *S. cerevisiae* population structure was later performed by Peter *et al.* [222], where 1,011 *S. cerevisiae* isolates from world-wide sources were included

in a phylogenomic analysis. 26 distinct populations, along with several mosaic lineages, were revealed based on genome-wide SNP data. This study highlighted the Taiwanese wild lineage as the most diverged population, supporting the "out-of-China" origin of *S. cerevisiae*. The China/Far East Asia origin was further supported by the analysis of Duan *et al.* [75] which included more WGS data for CHN IX isolates (synonymous with Taiwanese lineage in [222]).

Since the development of next-generation short-read sequencing, ∼2,700 *S. cerevisiae* world-wide isolates have been sequenced with variable instruments (mostly Illumina), lay-outs (mostly paired-end) or read length, providing comprehensive, albeit sampling biased, resources for future population genetics research (summarized in Table 1.1). In addition, the increasing availability of third-generation long-read sequencing technologies including PacBio and Oxford Nanopore Technology (ONT) have given rise to more high-continguity genome assemblies of *S. cerevisiae* facilitating whole genome annotation and structural variation (SV) detection across populations (summarized in Table 1.2). The abundance of public datasets make *S. cerevisiae* a great model species for evolutionary genomic research, including studies on the bioinformatics and evolution of transposable elements.

Table 1.1: **List of public short-read WGS datasets of *S. cerevisiae*.** Only newly sequenced samples in each study are included in this table.

Reference	**Number of strains**	**Major populations**	**Data availability**
Almeida *et al.* (2015) [3]	72	Mediterranean oak, North America/Japan, Wine, Mosaic	PRJEB7601
Barbosa *et al.* (2016) [14]	28	Novel populations from Brazil	PRJEB11698
Barbosa *et al.* (2018) [15]	26	Wine, bread, Brazil	PRJEB24932
Basile *et al.* (2021) [16]	65	European wine	PRJNA632495
Bergström *et al.* (2014) [23]	19	European wine, West African, Malaysian, Sake, North American	PRJEB2299

Table 1.1: **List of public short-read WGS datasets of *S. cerevisiae*.** (Continued)

Reference	**Number of strains**	**Major populations**	**Data availability**
Bigey *et al.* (2021) [25]	17	Mosaic, mixed origin	PRJEB36058
Borneman *et al.* (2016) [36]	198	*Prize de Mousse* (PdM), wine, hybrid Vin7	PRJNA303109
Coi *et al.* (2017) [60]	8	One France wine with its derivatives	PRJEB6586
Drozdova *et al.* (2016) [74]	5	Laboratory	PRJNA296913
Duan *et al.* (2018) [75]	266	Deep lineages from China, Chinese fermentation	PRJNA396809
Fay *et al.* (2019) [86]	120	Wine, beer, bakery, wild, clinical from multi-origins	PRJNA504476
Gallone *et al.* (2016) [93]	157	European wine and beer	PRJNA323691
Gayvskiy *et al.* (2016) [99]	52	New Zealand vineyard lineages	PRJNA247448
Goncalves *et al.* (2016) [103]	39	Wine, beer, bread, Sake	PRJEB13332
Han *et al.* (2021) [110]	126	Chinese alcohol, African beer, African palm wine	PRJNA680387
Istace *et al.* (2017) [129]	22	Distinct lineages from [222]	PRJEB14774
Kang *et al.* (2019) [136]	36	Environmental, Industrial, Clinical, Laboratory	PRJNA412468
Kita *et al.* (2017) [145]	85	European wine, West African, mosaic	PRJNA342356
Legras *et al.* (2018) [162]	31	European wine, West African, Sake	PRJEB8455
Maclean *et al.* (2017) [176]	85	Wine, clinical, bakery, North American oak etc.	PRJNA320792

Table 1.1: **List of public short-read WGS datasets of *S. cerevisiae*.** (Continued)

Reference	Number of strains	Major populations	Data availability
Marsit *et al.* (2015) [180]	58	Wine, bakery, oak in European or North American	PRJEB6529, PRJEB7675
Peter *et al.* (2018) [222]	918	26 world-wide populations	PRJEB13017
Ramazzotti *et al.* (2019) [231]	26	Human feces, wine, insect intestine	PRJEB1871
Skelly *et al.* (2013) [256]	22	Clinical, wild, bakery, fermentation, lab	PRJNA186707
Song *et al.* (2015) [259]	26	Laboratory, West African, Singapore, North American	PRJNA260311
Strope *et al.* (2015) [261]	93	European wine, West African, Malaysian, Sake, North American, mosaic	SRA072302
Yue *et al.* (2017) [291]	7	European wine, West African, Malaysian, Sake, North American	PRJNA340312
Zhu *et al.* (2016) [294]	145	Mostly clinical	PRJNA315044

Table 1.2: **List of public long-read WGS datasets of *Saccharomyces* species.** Only newly sequenced samples in corresponding studies are included in this table.

Species	Reference	Number of Strains	Sequencing Strategy	Data availability
	Basile *et al.* (2021) [16]	11	MinION	PRJNA632495

Species	Reference	Number of Strains	Sequencing Strategy	Data availability
	Bendixsen *et al.* (2021) [18]	8	PacBio Sequel	PRJEB38713
	Czaja *et al.* (2020) [64]	10	PacBio RS II	PRJEB33725
	Istace *et al.* (2017) [129]	22	MinION	PRJEB14774
	Lee *et al.* (2022) [161]	24	GridION	PRJNA755173
	Linder *et al.* (2020) [167]	12	PacBio RS II	PRJNA552112
	O'donnell *et al.* (2023) [212]	96	PromethION, MinION	PRJEB50706
	Yue *et al.* (2017) [291]	7	PacBio RS	PRJEB7245
S. paradoxus	Chen *et al.* (2022) [54]	1	PacBio RS II	PRJNA748953
	Eberlein *et al.* (2019) [80]	6	MinION	PRJNA514804
	Yue *et al.* (2017) [291]	5	PacBio RS	PRJEB7245
S. uvarum	Chen *et al.* (2022) [53]	1	PacBio RS II	PRJNA753102
S. eubayanus	Brickwedde *et al.* (2018) [40]	1	MinION	PRJNA450912
	Brouwers *et al.* (2019) [43]	1	MinION	PRJNA528469
	Mardones *et al.* (2022) [179]	2	MinION	PRJNA666059
S. jurei	Naseeb *et al.* (2018) [190]	2	PacBio RS II	PRJEB24816
S. mikatae	Chen *et al.* (Chapter 4)	2	PacBio RS II	PRJNA934353

Species	Reference	Number of Strains	Sequencing Strategy	Data availability
S. pastorianus	Salazar *et al.* (2019) [242]	1	MinION	PRJNA522669

1.3.2 *Saccharomyces paradoxus*

Saccharomyces paradoxus represents the closest sister species to *S. cerevisiae*. In contrast to complicated domestication processes in *S. cerevisiae*, *S. paradoxus* is rarely found associated with human industry, and, instead, is restricted to natural niches like oak trees, nearby soil or insects [172]. Therefore, the population structure of *S. paradoxus* is delineated more clearly by geographical regions [172]. Based on well-established population structure of both species, the intraspecific genetic distance among *S. paradoxus* populations is strikingly greater than that of *S. cerevisiae* [39, 116, 220].

In general, two major lineages of *S. paradoxus* have been characterized, i.e, Eurasian and American [151, 172, 116]. The Eurasian lineage is comprised of three populations: Europe, Far East [172, 23, 152, 116], and a recently identified Chinese clade [116]. The American lineage includes five major indigenous populations SpB, SpC, SpC*, SpD1 and SpD2 [160, 284, 80] plus a few poorly sampled lineages (see below). In addition, SpA, which is found in North America but clusters within the European lineage, emerged through a recent migration from Europe to North America, but has not yet admixed with other indigenous American lineages [153, 159, 160]. The SpB and SpC populations are considered as incipient species, whereas SpC* was derived from an old hybridization of SpB $\times$ SpC [160]. Subsequently, SpC hybridized with one of its parental lineage (SpB), giving rise to SpD lineages [80]. The formation of SpC* and SpD populations together imply an important role for intraspecific hybridization in the evolutionary history of *S. paradoxus*.

In addition to the major populations mentioned above, a few *S. paradoxus* strains that have been sampled from relatively distinct geographical locations, such as the two South American/Brazilian strains (SGRP UFRJ50816 and UFRJ50791) and one Hawaiian strain (SGRP UWOPS91-917.1). South American strains were originally designated as a separate species - *S. cariocanus* - due to the extremely low post-zygotic spore viability with other SpB strains caused by chromosomal translocations [192, 151, 171]. However, South American strains have been more recently reported as a sub-lineage of SpB in terms of nucleotide sequence similarity [172, 159, 116, 220]. Here, I regard South American *S. paradoxus* strains as part of SpB rather than another species. The geographically isolated Hawaiian strain forms a diverged lineage on its own, which is part of the American lineage, however its phylogenetic relationship to other known American sub-lineages remains uncertain [159, 116, 220].

S. paradoxus is another best-studied species of *Saccharomyces* genus besides *S. cerevisiae*. Until now, ∼400 *S. paradoxus* strains have short-read WGS data from all characterized populations, with SpB as the most heavily sampled sub-lineage (summarized in Table 1.3). More than ten high-quality genome assemblies of diverged *S. paradoxus* strains are available (Table 1.2), including one strain reported as a part of this thesis (Chapter 3).

Table 1.3: **List of public short-read WGS datasets of *S. paradoxus*.** Only newly sequenced samples in corresponding studies are included in this table.

Reference	**Number of strains**	**Major populations**	**Data availability**
Bergström *et al.* (2014) [23]	23	Mostly European, few Far Eastern and American	PRJEB2299
Chen *et al.* (2022) [54]	1	North American SpB	PRJNA748953
Eberlein *et al.* (2019) [80]	74	North American SpA, SpB, SpC, SpC*	PRJNA479851
He *et al.* (2022) [116]	29	China, Far East	PRJNA761003

Table 1.3: **List of public short-read WGS datasets of *S. paradoxus*.** (Continued)

Reference	Number of strains	Major populations	Data availability
Koufopanou *et al.* (2020) [152]	21	Europe	PRJNA635028
Leducq *et al.* (2016) [160]	154	North American SpA, SpB, SpC, SpC*	PRJNA277692
Peris *et al.* (2023) [220]	5	American A, B, C	PRJNA475869
Xia *et al.* (2017) [284]	91	North American a, b, c, d, e	PRJNA324830
Yue *et al.* (2017) [291]	5	Europe, Far East, North/South America	PRJNA340312

1.3.3 *Saccharomyces mikatae*

Saccharomyces mikatae strains were firstly isolated from decaying leaves and soil in Japan [192], with more strains identified later from a same niche in Japan by NITE Biological Resource Centre (NBRC) [194]. Only a very limited number of strains have been sequenced at the genome-wide scale and only the type strain (IFO 1815) is publicly available in complete genome assemblies [59, 140, 246, 220], with one more strain being reported in this thesis (see Table 1.2 and Chapter 4). In a recent study, two *S. mikatae* lineages – Asia A and Asia B – have been characterized with relatively low genetic distance among these lineages [220]. Interestingly, a previous survey of TE contents in *S. mikatae* suggests that *S. mikatae* might be able to shed light on the evolutionary history of transposons among *Saccharomyces* genus [169].

1.3.4 *Saccharomyces jurei*

Saccharomyces jurei is the most recently discovered *Saccharomyces* species which was firstly collected from oak bark in France [191]. Currently, only two *S. jurei* similar strains have

been sampled with high-quality genome assemblies available for the type strain NCYC 3947^T (=CBS 14759^T) [190] (Table 1.2). The sequencing data shows that *S. jurei* is the closest sister species to *S. mikatae* [190, 220]. The geographic distribution of *S. jurei* is unknown.

1.3.5 *Saccharomyces kudriavzevii*

The first *S. kudriavzevii* strain was found from decayed leaves and soil in Japan together with *S. mikatae* [192], followed by the identification of a European population [243]. Two distinct populations were defined according their geographical location, designated as Japanese and Portuguese [122], with available WGS data and assemblies for the type strains IFO 1802 and ZP 591, respectively [59, 140, 246]. A recent population structure survey showed that Japanese *S. kudriavzevii* strains formed two greatly divergent lineages (Asia A and Asia B), whereas the Portuguese (European) lineage is closely related with Asia A [220].

1.3.6 *Saccharomyces arboricola*

Saccharomyces arboricola (also called *S. arboricolus*) was firstly discovered from tree bark (*Fagaceae*) in China [276], with available complete genome assembly for the type strain - H-6^T (=CBS 10644) [170]. *S. arboricola* strains are also found in New Zealand [98]. The global population structure is unclear, but a recent study has revealed two distinct lineages of *S. arboricola*, i.e, Asia A and Oceania [220]. More comprehensive work needs to be done to reveal potential distribution in other continents.

1.3.7 *Saccharomyces uvarum*

The most basal branch in the phylogeny of *Saccharomyces* genus contains two species, *S. uvarum* and *S. eubayanus* (Fig 1.2). The definition of *S. uvarum* used to be controversial, and was originally regarded as the synonym of *S. bayanus* [195]. However, later studies have proved that *S. bayanus* (type strain CBS 380^T) is derived from hybridization between *S. eubayanus* and *S. uvarum* with contributions from *S. cerevisiae* [166]. The fully-sequenced

strain CBS 7001 was originally mis-classified as "*S. bayanus*" but later shown to be a pure type strain of *S. uvarum* [189]. The genome assembly of *S. uvarum* strain CBS 7001 has been reported in several studies [140, 59, 246] containing a large number of gaps, but was recently refined with PacBio sequencing data (Chapter 3 [53]). The global distribution of natural *S. uvarum* was studied in [4], uncovering three major geographic populations: Holarctic (including European and North American), South American and Australasia. *S. uvarum* isolates from South America can be classified into two distinct clades (namely SA-A and SA-B), and the Holarctic lineage is a sub-lineage of the SA-A clade. The Australasian lineage is highly diverged from others [4, 175, 220]. Currently public short-read WGS datasets of *S. uvarum* are listed in Table 1.4.

Table 1.4: **List of public short-read WGS datasets of *S. uvarum*.** Only newly sequenced samples in corresponding studies are included in this table.

Reference	**Number of strains**	**Major populations**	**Data availability**
Almeida *et al.* (2014) [4]	54	South American, Holarctic, Australasia	PRJNA230139, PRJEB5133
Macías *et al.* (2021) [175]	5	South American, Holarctic European	PRJNA471597
Peris *et al.* (2023) [220]	17	South American, Holarctic, Australasia	PRJNA475869
Scannell *et al.* (2011) [246]	1	Holarctic European	SRP006340

1.3.8 *Saccharomyces eubayanus*

Saccharomyces eubayanus is a closely-related sister species to *S. uvarum* (Fig 1.2). *S. eubayanus* was originally proposed to exist as a hypothesized parent of the lager-brewing

hybrid species *S. pastorianus*, contributing to the cryotolerance characteristics required for lager fermentation [166]. Pure representatives of *S. eubayanus* were later isolated from oak of *Nothofagus* in Patagonia region [166]. The complete and annotated genome assembly of the type strain (FM1318 = CRUB 1568^T = PYCC 6148^T = CBS 12357^T) was later published in [10, 40]. More *S. eubayanus* strains were then isolated from the bark and rotten wood of oak in Southwestern China, contributing to three distinct lineages: Tibet/Lager (close to strains from Patagonia and the parent of *S. pastorianus*), West China, and Sichuan [26, 43, 242]. The presence of highly diverged lineages in China provides evidence for China/Far East Asia origin of *S. eubayanus* [26]. Since its initial discovery, a great amount of *S. eubayanus* strains have been isolated from North America, South America, New Zealand and most recently from Europe [219, 217, 157, 19, 98, 179, 182, 198] (Table 1.5). These studies all together give rise to the current knowledge of *S. eubayanus* population structure: (i) Patagonia A; (ii) Patagonia B with "Holarctic" strains (including North America, Tibet/Lager and Europe) as a sub-lineage; and (iii) West China/Sichuan as a highly diverged outgroup [217, 220, 26]. Note that the phylogenetic relationship of West China lineage is inferred solely based on six genes due to the lack of public WGS data [26, 217].

Table 1.5: **List of public short-read WGS datasets of *S. eubayanus*.** Only newly sequenced samples in corresponding studies are included in this table.

Reference	**Number of strains**	**Major populations**	**Data availability**
Almeida *et al.* (2014) [4]	1	Unknown	PRJNA230139
Bergin *et al.* (2022) [19]	2	Holarctic (European)	PRJEB55404
Brickwedde *et al.* (2018) [40]	1	Patagonia	PRJNA450912
Brouwers *et al.* (2019) [43]	2	Holarctic (Tibet/Lager)	PRJNA528469
Langdon *et al.* (2020) [157]	186	Patagonia A, Patagonia B, Holarctic (USA)	PRJNA555221

Table 1.5: **List of public short-read WGS datasets of *S. eubayanus*.** (Continued)

Reference	Number of strains	Major populations	Data availability
Mardones *et al.* (2022) [179]	2	Patagonia B	PRJNA666059
Molinet *et al.* (2022) [182]	3	Patagonia B	PRJNA800259
Nespolo *et al.* (2020) [198]	82	Patagonia B	PRJNA540832
Peris *et al.* (2016) [217]	28	Patagonia A, Patagonia B, Holarctic, Admixed	PRJNA290017
Peris *et al.* (2023) [220]	9	Patagonia A, Patagonia B, Holarctic	PRJNA475869
Salazar *et al.* (2019) [242]	1	Holarctic (Tibet/Lager)	PRJNA522669

1.3.9 Inter-specific hybrids

Numerous examples of inter-specific hybridization have been reported among *Saccharomyces* species, despite post-zygotic reproductive barriers [171]. Hybridization events occasionally occur naturally in closely related species like *S. cerevisiae* × *S. paradoxus* and *S. uvarum* × *S. eubayanus* [14, 79, 166], but are mainly found in industrial environments, where different fermentation temperatures or flavor styles are required. One of the best-characterized hybrid species, *S. pastorianus*, is widely applied in the production of lager beer, which is derived from an inter-specific mating of the fermentation-capable *S. cerevisiae* and the cryotolerant *S. eubayanus* [189, 100, 166]. As more *S. pastorianus* genomes are sequenced, two distinct lineages of hybrids have been identified in terms of the genetic contribution from parent species, i.e, Saaz and Frohberg [92, 156, 77]. The chromosomal-level genome assembly with comprehensive genomic analysis of a Frohberg *S. pastorianus* strain CBS 1483 has been conducted which hypothesized that the two lineages could be emerged from single hybridization followed by separate evolution [242].

In addition to *S. pastorianus*, more inter-species hybridizations have been isolated from different fermentation industries: *S. cerevisiae* × *S. kudriavzevii* hybrids are found in ale beer and wine production [104, 173, 36, 218]; *S. uvarum* × *S. eubayanus* hybridization can occur in the wild, which could be further hybridized with *S. cerevisiae* to produce wine and cider [166, 200, 92, 156]; Other complex hybrids involve three or four parents among *S. cerevisiae*, *S. uvarum*, *S. eubayanus* and *S. kudriavzevii* [255, 92, 156]. These hybridizations provide opportunities for horizontal TE transfer among different *Saccharomyces* species.

1.4 Transposable elements in *Saccharomyces* yeasts

S. cerevisiae has been well-studied as an important model system to address questions about TE biology. For example, the mechanism of mobilization for LTR-retrotransposons was originally discovered in *S. cerevisiae* [31]; Similarly, *S. cerevisiae* was the first completely-sequenced eukaryotic organism, providing the first resources to comprehensively scan TE content from a eukaryotic genome [102].

The first genome-wide survey of TE contents in the *S. cerevisiae* reference genome was conducted in 1998 by Kim *et al.* [143] and later refined by Carr *et al.* [48]. Five distinct TE families of LTR-retrotransposon were originally identified in *S. cerevisiae*, designated Ty1 (transposon yeast) through Ty5. Two superfamilies homologous to other eukaryotic organisms are found in *S. cerevisiae*: Ty1/Copia (including Ty1, Ty2, Ty4 and Ty5) and Ty3/Gypsy (including Ty3). As LTR retrotransposons, Ty families are generally comprised of two ORFs, *Gag* (also known as TYA) and *Pol* (also known as TYB), flanked by two LTRs. The *Pol* gene includes four conserved coding domains, i.e, protease (PR), integrase (IN), reverse transcriptase (RT) and RNase H (RH), which are essential proteins during retrotransposition process (see Section 1.1.1). Ty1/Copia and Ty3/Gypsy members differ in the arrangement of these coding domains [143]. The *Gag* and *Pol* ORFs are slightly overlapping and their expression involves a translational frameshifting [58] in Ty1-4, whereas Ty5 consists of a *Gag-Pol* fusion for the internal coding region instead of separate ORFs.

Under each of the five Ty families, subfamilies are further identified based on phylogenetic analysis of TE sequences, either within *S. cerevisiae* reference genome (e.g, Ty1' [143]), from other populations in the same species (e.g, Ty101 [29]), or from different *Saccharomyces* species (e.g, Ty3_1p [48]). Our lab has recently released an library of TE canonical sequences of *Saccharomyces* genus with structural annotations (`https://github.com/bergmanlab/yeast-transposons`) (Fig 1.3). The discovery of subfamilies will be reviewed more in detail in following subsections. Although our understanding about the diversity of TEs in *Saccharomyces* yeast has grown in recent years, few analysis have been done to reveal the presence and abundance of TE families in a population level within species [28] or at an inter-specific level [169, 199]. As such, the evolutionary mechanisms and history of subfamily formation in the *Saccharomyces* genus is still unclear.

1.4.1 Intra-element LTR-LTR recombination

The internal region of full-length Ty elements can be eliminated from the host genome through a process called LTR-LTR recombination [283, 135]. Two flanking LTRs of a single full-length Ty element is are identical at the time of integration, which provides substrate for them to recombine, giving rise to an external chromosomal circular structure consisting of internal coding region plus one LTR with the other "solo" LTR remaining in the genome. Whereas the complete LTR retrotransposons are autonomous (or "active") elements, solo LTRs clearly lack the ability to replicate and mobilize within host genome on their own (and are therefore "dead"). The LTR-LTR recombination process occurs rapidly after integration, which explains the large number of solo LTRs detected in *S. cerevisiae* reference genome [143, 48]. Solo LTR sequences characterized from the *S. cerevisiae* reference genome are often more divergent than LTRs of full-length elements, indicating that solo LTRs are also undergoing degeneration and may ultimately be eliminated from the host genome [143]. The presence of solo LTRs can be taken as a signal of past activities of a Ty family in *Saccharomyces* species.

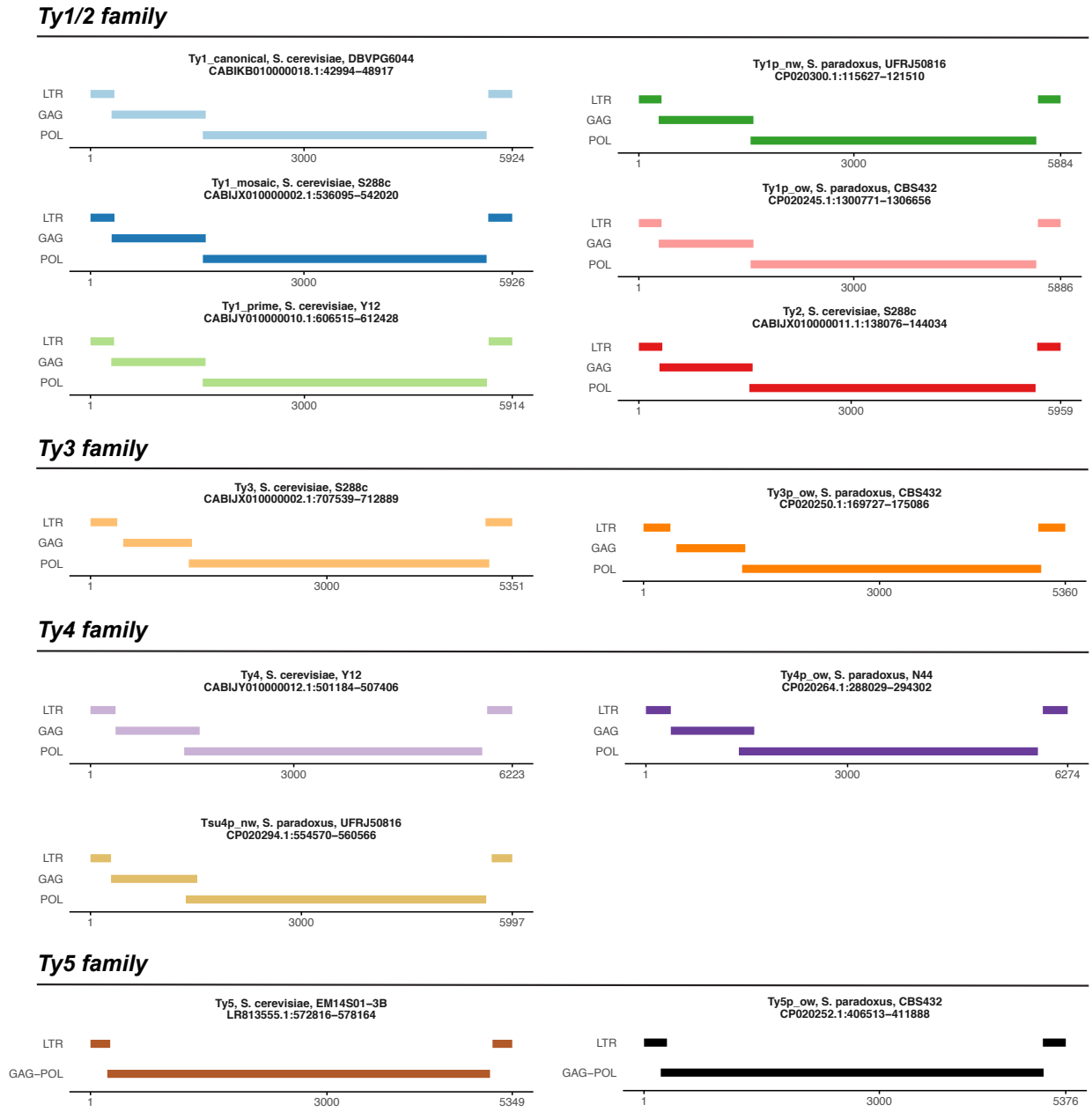

Figure 1.3: **Visualization of all annotated Ty canonical sequences.** The *Gag* and *Pol* genes as well as flanking LTRs are shown. The detailed source of each canonical sequence, including species, strain, Genbank accession and coordinates, is noted as the element title.

1.4.2 Ty1 and Ty2

Ty1 was the first LTR retrotransposon family identified and fully-annotated in *S. cerevisiae* [32, 30], with Ty1-H3 as the most well-studied tester element in the biology of retrotransposition (reviewed in [63]). Ty1 is the most abundant Ty family in *S. cerevisiae* reference strain S288c, with >30 full-length copies being characterized and potentially active in the host genome [143, 48]. Several distinct subfamilies of Ty1 have been identified in *S. cerevisiae*, such as "canonical" Ty1, Ty1' [143], Ty1/2 [133, 134] and Ty101 [29]. Ty1' has been determined as the ancestral Ty1 subfamily based on its association with strains from wild habitats [143, 64]. "Canonical" Ty1 is a recently evolved subfamily derived from a horizontal TE transfer from Eurasian *S. paradoxus* followed by recombination with Ty1' [64]. The Ty1/2 subfamily (such as the well-known Ty1 tester element, Ty1-H3) is also a recombined group arising from complex recombination between "canonical" Ty1 and Ty2 [133, 134, 64]. Finally, Ty101 has been discovered as another recombined subfamily with high *Gag* similarity to American *S. paradoxus* found in multiple strains restricted to South America regions [29, 28]. The ongoing identification of new Ty1 subfamilies reveals complicated evolutionary history of Ty1, involving ongoing intra- and inter-family recombination and horizontal TE transfers, which also indicates that Ty1 family is highly active in several species of *Saccharomyces* genus.

Ty2 is closely related to Ty1, sharing a same type of flanking LTR sequence called the *delta* element and high sequence similarity in coding regions [143]. The number of full-length Ty2 elements is also relatively high (>10 copies) in *S. cerevisiae* strain S288c, yet LTR elements show less sequence divergence comparing with Ty1 family, indicating the more recent origin of Ty2 family in *S. cerevisiae* [143, 48]. One potential model to explain Ty2 origin in *S. cerevisiae* is that it arose through horizontal TE transfer from *S. mikatae*, based on their high nucleotide similarity [169, 48, 105]. More work needs to be done to resolve the origin of Ty2 in *S. cerevisiae* and how to distinguish Ty2 from Ty1 solo LTRs properly given their near identical LTR sequences [237].

1.4.3 Ty3 and Ty3_1p

Ty3 is the sole member of Ty3/Gypsy superfamily found in *S. cerevisiae*, flanked by LTR sequences called *sigma* elements [70, 114, 143]. Like Ty1, Ty3 has been extensively studied at the molecular level (reviewed by [244]). Ty3 is not as transpositionally active as Ty1, with only two full-length elements detected from the reference strain S288c [143, 48]. A Ty3 homologue has been characterized in *S. paradoxus* called Ty3_1p (named after the first Ty3-like element identified in "*paradoxus*") [87]. Distinct Ty3 and Ty3_1p subfamilies that are found in *S. cerevisiae* and *S. kudriavzevii* can be explained by horizontal TE transfers of Ty3_1p from *S. paradoxus* [48, 105].

1.4.4 Ty4

Ty4 is another active yet less-abundant TE family found in *S. cerevisiae*, which is flanked by *tau* LTR elements [264, 263, 130]. Only 3 full-length Ty4 elements are found in *S. cerevisiae* lab strain S288c [143, 48]. A Ty4-like subfamily, called Tsu4 (named after transposon *Saccharomyces uvarum*), has been characterized originally from *S. uvarum* reference strain CBS7001 [199]. Despite of the relatively low copy number of Ty4 in *S. cerevisiae*, an unexpected high copy number of Tsu4 have been identified from a South America *S. paradoxus* strain (UFRJ50816), which reveals another horizontal transfer event that introduced Tsu4 into *S. paradoxus* [20]. In a more recent study, 9 full-length copies of Tsu4 have also been annotated from one *S. cerevisiae* strain [212], indicating an interesting yet unclear evolutionary history of Ty4/Tsu4 family among *Saccharomyces* genus [105]. The evolution history of Ty4 family will be investigated in Chapter 4.

1.4.5 Ty5

Ty5 is the last Ty family identified in *S. cerevisiae* flanked by *omega* LTR elements [271], and is the least abundant family in the *S. cerevisiae* reference genome [143, 48]. The only one "full-length" Ty5 copy identified from *S. cerevisiae* S288c has deletion in its internal coding region,

along with numerous mutations, which is therefore regarded as transpositionally inactive [271, 143, 48]. However, Ty5 elements show active transpositions in *S. paradoxus* [297], which is designated as Ty5p in [143]. In a recent study, multiple full-length Ty5 elements have been characterized from the ancestral Chinese lineage of *S. cerevisiae* [18].

1.4.6 Ty target site selection

New TE insertions are most likely to have deleterious effects on the host genome especially when inserting into essential coding regions. *Saccharomyces* yeast have relatively compact genome arrangement among eukaryotes, with $\sim 70\%$ coding regions and therefore few non-essential DNA regions [102]. The five Ty families in *S. cerevisiae* have evolved different strategies of selecting TE target sites that can help maintain the integrity of host genome (reviewed in [35]).

Both Ty1 and Ty3 have been shown to target RNA polymerase III (Pol III)-transcribed genes, mostly tRNA genes (tDNAs). Ty1 and Ty3 integrase interact with different components of the Pol III transcription system. Ty1 integrase (IN1) interacts with the AC40 sub-unit of Pol III protein, guiding the Ty1 targeting towards tRNA promoter regions [42]. Several studies have reported that majority of Ty1 insertions occur within 1000 bp upstream of tRNA transcription starting sites (TSSs) showing "dyad" distribution patterns near nucleosome-bound regions [186, 227, 11, 41, 42]. Ty2 and Ty4 are predicted to have similar distribution patterns as Ty1, given conserved amino acid sequences of Ty1/2/4 integrases [8]. However, the nucleosomal targeting preference of Ty2 and Ty4 have not yet been confirmed *in vitro*. Ty3 integrase (IN3) is recruited by Pol III transcription factor TFIIIB, more specifically Brf1 and TBP subunits, which localizes Ty3 integrations mostly within 20bp upstream of tRNA TSS [227, 214]. No "dyad" distribution pattern is observed for Ty3.

Ty5, however, has a unique targeting preference that differs from any other Ty families. Instead of interacting with Pol III genes, Ty5 tends to integrate in heterochromatin at silent

mating loci and near telomeric regions, *via* the interaction between Ty5 integrase (IN5) C-terminus and the Sir4 protein [296, 91, 285]. More recently, secondary targeting site bias has been reported that Ty5 targeting is also related to DNase hypersensitivity and nucleosome-free regions, regardless of heterochromatin or euchromatin [12].

1.4.7 Horizontal transfer of TEs in *Saccharomyces* yeasts

TEs are typically vertically inherited from parents to offspring within a species [279]. In addition, TEs are also capable of transmitting between species *via* horizontal transposon transfer (HTT), which has been increasingly recognized as an important phenomenon in shaping eukaryotic genomes [248, 101, 273, 274]. The first HTT event was identified by Daniels *et al.* [68], which reported a *P* element horizontally transferred from *Drosophila willistoni* to *D. melanogaster*. In general, HTT events could be detected using several evolutionary criteria: (i) The phylogenetic incongruence (PI) between the host and TE trees; (ii) Patchy distribution (PD) of a TE family across closely related hosts; or (iii) unexpectedly high sequence similarity (HS) of TE sequences from diverged hosts (reviewed by [273]). Up to 2019, an increasing amount of sequencing data have facilitated the characterization of >5000 HTT events across the tree of life, notably in plants and animals [73, 72]. However, few HTT events were identified in fungi [71, 65, 204, 7, 245].

Over the last decade or so, a growing number of HTT events have been identified among *Saccharomyces* species, which plays a significant role in the successful proliferation of Ty elements and formation of sub-families in yeasts (as mentioned in Section 1.4) [105]. Firstly, "canonical" Ty1 in *S. cerevisiae* arose through a HTT that acquired Ty1 *Gag* sequences from Eurasian *S. paradoxus* followed by further recombination [64]. The formation of Ty101 in *S. cerevisiae* probably resulted from another HTT event from American *S. paradoxus* to *S. cerevisiae* [29, 28]. Secondly, studies reported that Ty2 in *S. cerevisiae* was derived from *S. mikatae* through HTT [169, 48]. Thirdly, the co-existence of Ty3 and Ty3_1p in *S. cerevisiae* and *S. kudriavzevii* could be explained by HTT from *S. paradoxus* [48]. Last but not least,

the proliferation of Tsu4 in *S. paradoxus* was introduced from *S. uvarum* or *S. eubayanus* via HTT [20]. Essentially all active Ty families in the *Saccharomyces* genus are involved in HTT events (except for Ty5), and more events likely remain to be discovered with growing number of comprehensive Ty annotations in yeast genomes.

1.5 Dissertation Outline

The overall topic of my dissertation is to develop new bioinformatics approaches and investigate the evolutionary history of transposable elements in species of the *Saccharomyces* genus.

In Chapter 2, I improved the meta-pipeline for 12 polymorphic TE detectors using short-read whole genome sequencing data called McClintock (`https://github.com/bergmanlab/mcclintock`). In addition, I developed a reproducible and automated simulation framework to comprehensively evaluate the performance of multiple TE detectors, including an example of the *Saccharomyces* yeast model. The simulation analysis identified four "best-in-class" methods based on their precision and recall rates in simulations and consistency of predictions from empirical data. The improved McClintock meta-pipeline was performed to detect non-reference TE insertions from a empirical *S. cerevisiae* dataset containing 1,011 isolates, which provided novel insights into the TE abundance and insertion preference patterns in *S. cerevisiae*. This work provides a user-friendly pipeline for TE detection with short-read WGS data using multiple TE detectors, as well as a reproducible simulation framework, which together should benefit future studies on TE polymorphism in a range of different organisms.

In Chapter 3, I created two high-quality whole genome assemblies of *Saccharomyces* species to support studies on TEs in yeast: the *S. paradoxus* strain DG1768 and the *S. uvarum* strain CBS 7001 using the combination of PacBio and Illumina sequencing data. DG1768 is a lab-derived SpB strain that has been widely studied in molecular biological research especially on Ty1 mobility. My study documented genetic alterations of this key

lab strain and provided resources for further genetic studies. My study also provided an important reference genome for *S. uvarum* type strain CBS 7001 using long-read sequencing data, which should facilitate TE identification of this species.

In Chapter 4, I investigated the evolutionary history of the Ty4 family among *Saccharomyces* species by analyzing large-scale long-read and short-read sequencing data. The species-wide ancestral state of Ty4 family (presence or absence of Ty4 or Tsu4 subfamilies) was studied based on a coverage-based method to estimate TE abundance in several *Saccharomyces* species, including *S. cerevisiae*, *S. paradoxus*, *S. uvarum*, *S. eubayanus*, and *S. mikatae*. In order to facilitate discovery on potential recombinantion or introgression events of TE sequences, I generated new genome assemblies with PacBio sequencing data for two *S. mikatae* strains, which had not been fully sequenced previously. I also integrated currently available genome assemblies of *Saccharomyces* species and comprehensively annotated their TE contents, which generated Ty4/Tsu4 sequences for phylogenetic analysis. Together, my results reported several new horizontal TE transfer events involving the Ty4 family in *Saccharomyces* genus, and added further evidence for the significant role of horizontal TE transfer and recombination in shaping Ty contents for *Saccharomyces* yeasts. This work also provided new genome assemblies resources and novel TE detection strategies using shotgun data that could benefit future studies on evolution of TE content in other species.

Chapter 2

Reproducible evaluation of transposable element detectors with McClintock 2 guides accurate inference of Ty insertion patterns in yeast [1]

[1]Jingxuan Chen, Preston J Basting, Shunhua Han, David J Garfinkel, and Casey M Bergman, Reproducible evaluation of transposable element detectors with McClintock 2 guides accurate inference of Ty insertion patterns in yeast, *Mobile DNA*, 2023, 14 (1): 8;

Reprinted here with permission of the publisher.

2.1 ABSTRACT

Many computational methods have been developed to detect non-reference transposable element (TE) insertions using short-read whole genome sequencing data. The diversity and complexity of such methods often present challenges to new users seeking to reproducibly install, execute or evaluate multiple TE insertion detectors.

We previously developed the McClintock meta-pipeline to facilitate the installation, execution, and evaluation of six first-generation short-read TE detectors. Here, we report a completely re-implemented version of McClintock written in Python using Snakemake and Conda that improves its installation, error handling, speed, stability, and extensibility. McClintock 2 now includes 12 short-read TE detectors, auxiliary pre-processing and analysis modules, interactive HTML reports, and a simulation framework to reproducibly evaluate the accuracy of component TE detectors. When applied to the model microbial eukaryote *Saccharomyces cerevisiae*, we find substantial variation in the ability of McClintock 2 components to identify the precise locations of non-reference TE insertions, with RelocaTE2 showing the highest recall and precision in simulated data. We find that RelocaTE2, TEMP, TEMP2 and TEBreak provide consistent estimates of ∼50 non-reference TE insertions per strain and that Ty2 has the highest number of non-reference TE insertions in a species-wide panel of ∼1000 yeast genomes. Finally, we show that best-in-class predictors for yeast applied to resequencing data have sufficient resolution to reveal a dyad pattern of integration in nucleosome-bound regions upstream of yeast tRNA genes for Ty1, Ty2, and Ty4, allowing us to extend knowledge about fine-scale target preferences revealed previously for experimentally-induced Ty1 insertions to spontaneous insertions for other *copia*-superfamily retrotransposons in yeast.

McClintock (`https://github.com/bergmanlab/mcclintock/`) provides a user-friendly pipeline for the identification of TEs in short-read WGS data using multiple TE detectors, which should benefit researchers studying TE insertion variation in a wide range of different organisms. Application of the improved McClintock system to simulated and empirical yeast genome data reveals best-in-class methods and novel biological insights for one of the most

widely-studied model eukaryotes and provides a paradigm for evaluating and selecting non-reference TE detectors in other species.

2.2 Introduction

Transposable elements (TEs) are mobile, repetitive DNA sequences that occupy large fractions of most eukaryotic genomes and can cause mutations of large effect [38]. Since their discovery, a variety of cytological, molecular and genomic techniques have been developed to identify TE insertions in genomes at varying levels of resolution [24, 290, 22, 241, 232, 286, 85]. In principle, the most comprehensive method for detecting new TE insertions in a genome is whole genome sequencing (WGS) followed by *de novo* genome assembly and systematic genome-wide annotation of TEs. However, the difficulty of accurately assembling repetitive DNA sequences using short reads [266, 215] and the challenges of generating high-quality long-read genome sequencing data hinder using assembly-based methods for genome-wide detection of TE insertions, especially for large population samples [69].

Currently, the most widely-used approach for detection of new TE insertions involves mapping unassembled short-read WGS data to a reference genome and inferring the location and family using split-read or read-pair information [85]. Over the last decade, many bioinformatic methods have been developed for the reference-based detection of TE insertions using short-read WGS data (`https://tehub.org/en/resources/repeat_tools`) [82]. Most of these methods were initially developed for a specific organism with its unique TE composition, and thus have different design goals and features. The general performance of short-read TE detectors is typically unknown because of the "self-assessment trap" and limited number of methods compared in primary studies [203]. A few independent performance evaluation studies have been conducted comparing subsets of short-read TE detectors, but these studies differ in their evaluation frameworks, selection of tools, and focal organisms [234, 197, 269]. Crucially, no study to date has provided a reproducible evaluation of multiple TE detectors, making it difficult to validate published results or extend findings beyond the

context of the initial evaluation study. Despite these issues, one emerging theme from current evaluation studies is that considerable variation in performance exists among different short-read TE detectors, across organisms and evaluation frameworks [85, 234, 197, 269]. As such, there is not yet any clear basis to select the best short-read TE detector for a new organism, and thus researchers must base tool choice by extrapolation from performance on other taxa or employ multiple methods to ensure robust biological conclusions [21, 177, 66].

Previously, we developed a meta-pipeline called McClintock [197] to facilitate the installation and execution of six first-generation short-read TE detectors (ngs_te_mapper [168], PoPoolationTE [146], RelocaTE [235], RetroSeq [139], TE-locate [223], and TEMP [295]). The original McClintock system automated installation of these six "component" TE detection methods, provided a common interface to run all components, reduced the number of shared input files, and generated a standard set of output files [197]. Since its initial development, the McClintock system has been used to support detection of TE insertions and enable biological discoveries in a variety of organisms and biological contexts [197, 269, 21, 177, 66, 127, 239, 76, 17, 149, 132, 128, 209, 265, 111, 287, 112, 113, 196, 115, 120, 117, 247, 184] and to facilitate comparative evaluation of multiple TE detectors [197, 269, 267].

A long-term aim of the McClintock project was to develop a flexible framework to incorporate additional TE detectors as they were published, allowing researchers to run the most relevant set of methods and compare performance on their genomes of interest. While the original pipeline made it easier to install and run multiple TE detectors and compare their output, its initial design had several limitations that made it difficult to achieve this long-term aim. Most importantly, the original McClintock pipeline was implemented in Bash with minimal use of functions or classes to encapsulate, abstract, or modularize the code, and most variables were defined in the global scope. These shortcomings made it difficult to modify and expand the original codebase without compromising existing functionality, and thus hindered the addition of new TE detectors to the original framework.

The original McClintock system also had several limitations related to its installation. Component TE detectors were installed automatically by McClintock, however the software dependencies for each component had to be installed manually, which required substantial effort on the part of the user. Moreover, all components in the original McClintock system ran in a single computing environment, which necessitated a mutually-compatible set of software dependencies to be installed. This fragile configuration also caused compromises in the versions of software dependencies that were used and, in some cases, locked the original system to increasingly out-of-date versions of software dependencies.

Even after successful installation, the original McClintock system had a number of limitations regarding its usability. Notably, error handling was largely absent from the original codebase. Thus, when failures occurred during the execution of component TE detectors or their software dependencies, McClintock would continue to run, often producing a cascade of error messages from downstream processes that made it difficult for users to know how and why the pipeline failed. Similarly, the original pipeline had strict input file formatting requirements (e.g., requiring unzipped fastq files), but which would provide no warnings or messages if the input files did not comply to required specifications. Moreover, the original McClintock system hard-coded some input and post-processing parameters that were not easily modified by users but impacted the performance of component methods. The original McClintock pipeline also output TE insertion predictions only in Browser Extensible Data (BED) format (`https://samtools.github.io/hts-specs/BEDv1.pdf`), rather than the Variant Call Format (VCF) [67] that is standard for reporting genetic variation relative to a reference genome. Finally, the original pipeline also did not produce plots or tables to facilitate comparison of predictions across component methods.

To address these limitations, we completely re-implemented McClintock in Python leveraging Conda (`https://github.com/conda/conda`) and Snakemake [150] to improve the installation, error handling, speed, stability, and extensibility of the pipeline. We also incorporated six new TE detectors (ngs_te_mapper2 [111], PoPoolationTE2 [147], RelocaTE2 [51],

TEBreak [249], TEFLoN [2], and TEMP2 [289]), doubling the total number of components available in the new McClintock system. In addition, the updated McClintock system now includes new modules to automate read trimming, provide estimates of TE family abundance from depth of coverage, and produce an integrated HTML report summarizing predictions made by all components. Importantly, McClintock now also provides a new reproducible simulation framework that allows users to evaluate the performance of component TE detectors in a flexible manner across different organismal contexts. In this report, we describe the structure and function of the updated McClintock system, highlight its improvements and new features, evaluate component TE detector performance using simulated and empirical yeast genome data, and use the updated McClintock system to provide new biological insights into the pattern of TE insertion in *S. cerevisiae*.

2.3 Implementation

2.3.1 Re-implementation of McClintock in Python using Conda and Snakemake

Here we describe the major features distinguishing the re-implemented McClintock 2 system from the original version reported in Nelson *et al.* [197] (Table 2.1). We initially sought to further develop the original Bash-based McClintock system reported in Nelson *et al.* [197] by improving its installation using the Conda package and environment management system (`https://github.com/conda/conda`). The use of Conda allows the creation of distinct environments (both for McClintock itself and for the component methods wrapped in McClintock) that can be reproducibly generated on different computing systems without the need for root privileges. This modification to the original McClintock system permitted external software dependencies to be automatically installed separately for each component, and each component to be executed in its own isolated environment. While solving some of the problems with the installation and versioning of external dependencies in the original system, this development work led to an endpoint because it did not solve most of the

software engineering and usability limitations described above. Therefore, the final version of the Bash-based McClintock system extended to use Conda (referred to in this report as McClintock 1) was retired.

Table 2.1: Major features that distinguish the original McClintock version reported in Nelson *et al.* [197] and the McClintock 2 system reported in the current study.

	Nelson *et al.* [197]	McClintock 2
McClintock implementation	Bash	Python
Workflow management	N.A.	Snakemake
Dependency installation	Manual	Conda (automated)
# Component methods	6	12 (with extensibility)
Read trimming and QC	N.A.	trimgalore module
TE copy number estimation	N.A.	coverage module
Output format	.BED	.VCF and .BED
Summary report	.CSV	.HTML (interactive)
Reproducible simulation system	N.A.	Built-in (with documentation)

N.A. indicates that the feature is not available.

We next completely rewrote the McClintock system in Python 3 using Conda and Snakemake [150], with the initial aim of recapitulating the general functionality in McClintock 1. The new McClintock 2 codebase is heavily modularized and leverages the native error handling functions in Python. Snakemake was used to simplify and fully automate installation of all components and their software dependencies (either from Conda channels like Bioconda [108] or directly from project repositories), and to automatically create independent Conda environments for each component. Additionally, because the "meta-pipeline" architecture of McClintock is naturally suited to a rule-based workflow, we encapsulated of each component as an individual rule in a Snakefile that is executed by Snakemake inside a specific Conda environment whose dependencies do not conflict with any other component/rule.

Using the Snakemake workflow manager to underpin the McClintock 2 meta-pipeline provided several advantages. Since the Snakemake engine determines the rules that are run

to create the requested outputs, no additional coding logic is required in McClintock itself to manage the complexities of which components were run and in which order, or how files shared by multiple components are created and used (e.g., reference genomes, trimmed read files, BAM files, etc.). Importantly, this virtue of Snakemake applies regardless of how many components are executed at run time or incorporated into McClintock 2 in the future, permitting flexible execution of various components by users and easy addition of new components by developers. Finally, the Snakemake engine can run multiple independent rules simultaneously if enough central processing unit (CPU) cores are available, potentially allowing for multiple rules to be run in parallel on multi-core systems. In the re-implemented system, if $n>1$ cores are requested, McClintock 2 by default now allocates a maximum of $n/2$ (for an even number of cores) or $(n-1)/2$ (for an odd number of cores) to each rule which allows multiple rules to be executed in parallel. Optionally, users can specify the "–serial" flag which prevents multiple rules from being run in parallel and allows rules that use multiple cores to have access to maximal number of cores requested.

After integrating the original six components (ngs_te_mapper [168], PoPoolationTE [146], RelocaTE [235], RetroSeq [139], TE-locate [223], and TEMP [295]) into the McClintock 2 system, we cross-validated its accuracy versus the original McClintock 1 system using simulated genomes generated by our new reproducible simulation platform (see below and Supplemental Text). We then incorporated six more recently-developed TE detectors that fulfilled our original inclusion criteria [197] into the re-engineered McClintock system: ngs_te_mapper2 [111], PoPoolationTE2 [147], RelocaTE2 [51], TEBreak [249], TEFLoN [2], and TEMP2 [289].

In addition to providing a full re-implementation and additional TE detectors, McClintock 2 has new functionality not present in the original system (Table 2.1). First, McClintock 2 now has a preprocessing module to perform read trimming and quality control (QC) using TrimGalore (`https://github.com/FelixKrueger/TrimGalore`) and produce a read QC report using MultiQC [84]. If the read trimming option is set (either optionally in

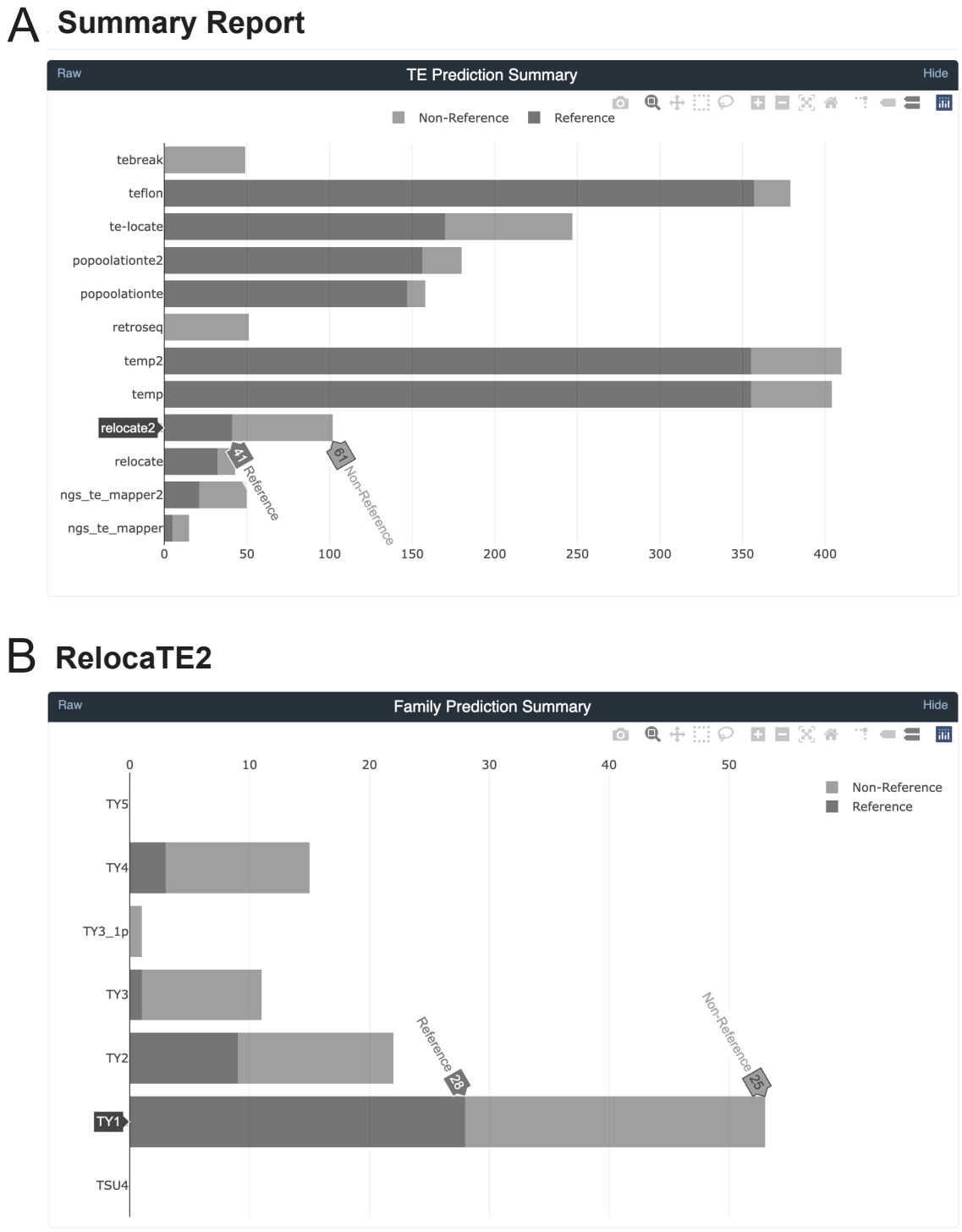

Figure 2.1: (Description see the next page.) **Sample screenshots from the new interactive HTML report in McClintock 2.**

Figure 2.1: (Figure see the previous page.) **Sample screenshots from the new interactive HTML report in McClintock 2.** The HTML report generates summary information for the McClintock run including interactive bar plots for: (A) the number of reference, non-reference, and total number of predictions made across all TE families by all 12 component methods; and (B) the number of reference, non-reference, and total number of predictions made for a specific component method (e.g., RelocaTE2). Barplots from the report shown were generated by a complete McClintock run (revision d2b819a18b2a549be483fdcc948e1346e589a4cb) applied to Illumina 101-bp paired-end sequences for *S. cerevisiae* strain YJM1460 (SRA: SRR800842), down-sampled to 50× fold-coverage.

conjunction with user-specified components, or in a default run that executes all 12 components), then trimmed reads are used as input for other components in McClintock 2. Second, McClintock 2 has a new "coverage module" that can generate estimates of TE copy number by computing relative depth-of-coverage for each query TE sequence normalized by depth-of-coverage in non-repetitive regions of the genome. The coverage module also produces plots for each query that allow users to inspect variation in read depth across TE sequences, which may be caused by the existence of multiple TE sub-families that differ by structural variants (e.g., [21, 278]). Third, McClintock 2 reports non-reference TE insertion variants predicted by each component in VCF format [67] in addition to BED files containing standardized information about reference and non-reference TE insertions. Finally, McClintock 2 is able to generate an interactive HTML report for each run that allows users to view, sort, and filter results for each component and TE family (Fig. 2.1).

2.3.2 A reproducible simulation system for evaluating McClintock component performance

In addition to improvements in the meta-pipeline itself, McClintock 2 now also provides code to automatically simulate TE insertions in any user-supplied genome and generate synthetic WGS datasets that can be used as input to evaluate component TE detector performance.

Previously, we developed a "single synthetic insertion" framework for evaluating the performance of McClintock components to detect non-reference TE insertions and used this framework to evaluate the performance of the six TE detectors in McClintock 1 [197]. This evaluation framework created a synthetic genome comprised of a single non-reference TE and its corresponding target site duplication (TSD) inserted into in an otherwise-unmodified reference genome, then simulated a corresponding paired-end WGS dataset that could be used as input for McClintock. The Bash code for this simulation framework was not released as part of the original McClintock 1 system, did not run inside a controlled environment, was hard-coded to simulate TE insertion preferences for only one species (*S. cerevisiae*), only evaluated only one WGS fold-coverage (100×), and did not generate standard performance metrics like precision and recall.

To overcome these limitations, we re-implemented a flexible and fully-reproducible version of this single synthetic insertion simulation framework in Python 3 using Conda and Snakemake. The new simulation framework in McClintock 2 allows customization for different organisms, and consists of three major parts that run under controlled Conda environments:

(i) A Python 3 script that creates a configuration file with information about the simulation experiment to be run in JavaScript Object Notation (JSON) format. Required inputs to the configuration script include user-supplied files containing the reference genome, TE library, and permissible locations for TE insertions. The configuration script also allows users to modify the number of replicates in the experiment, properties of the synthetic genomes (e.g., TSD length and strand of the TE insertion), properties of the simulated WGS datasets (e.g., read length, single or paired end reads, insertion sizes for paired-end reads, read error rates, fold-coverage), the read simulator used to generate WGS datasets (ART [124] or wgsim [164]), and computing resources for each cluster job (e.g., number of threads, amount of memory).

(ii) A primary Snakemake workflow that reads input parameters from the JSON file created by the configuration script, then submits and manages jobs on a high-performance

computing (HPC) cluster. Currently our simulation framework is designed for use with cluster systems running the SLURM workload manager, but it can be adapted to other HPC systems using alternative Snakemake cluster profiles (`https://github.com/Snakemake-Profiles/`). Each submitted job runs a Python script that creates a new synthetic genome with one additional insertion, simulates a corresponding WGS dataset from this synthetic genome, then runs the McClintock pipeline on the simulated WGS dataset. Replicates for each combination of parameters are submitted as individual cluster jobs to maximize available cluster resources, and any potential failed jobs are automatically re-submitted by the workflow to minimize effort needed to monitor the simulation experiment.

(iii) A secondary Snakemake workflow that is automatically executed after all cluster jobs submitted by the primary workflow are completed, which summarizes results and generates preliminary plots from all replicates of the overall simulation experiment. This workflow runs several scripts to calculate performance metrics (e.g., mean numbers of non-reference TE predictions, precision, recall) across replicates, generate precision/recall curves as a function of fold-coverage, and generate positional accuracy and predicted TSD length plots for each fold-coverage level. Performance metrics are automatically generated by the simulation system and calculated as follows. Non-reference TE predictions with the identical family, chromosome, and strand, with start and end locations falling within an N bp window from the synthetic insertion (N=0, 5, 100, 300, or 500) are labeled as "within-N" predictions. Within-0 bp predictions have the identical start and end coordinates as the synthetic insertion in that simulated sample and are considered "exact" predictions. The mean number of reference and non-reference TE predictions (overall, and at various window sizes) per synthetic genome are then calculated across all replicates. To calculate precision and recall, non-reference TE predictions within-N bp from the insertion site that match the TE family and strand of the synthetic insertion are regarded as true positives (TP), while all other predictions

are regarded as false positives (FP). If more than one non-reference prediction for the correct TE family and strand was made within-N bp from the synthetic insertion, only one was considered a TP and all others are considered FPs. If a replicate had zero non-reference TE predictions within-N bp from the synthetic insertion, it was regarded as a false-negative (FN). For each N bp window size, recall was calculated as TP/(TP+FN) and precision was calculated as TP/(TP+FP).

2.4 Materials and Methods

2.4.1 Evaluation of McClintock component performance on simulated data

We conducted a series of four simulation experiments to validate aspects of the McClintock re-implementation, demonstrate the utility of our new reproducible simulation framework, and evaluate the ability of McClintock components to predict non-reference TE insertions in *S. cerevisiae*. The different simulation experiments varied in the version of McClintock used, which component methods were run, their placement of synthetic insertions, the read simulator used to generate WGS datasets, and performance metrics used. Simulations 1 and 2 were designed to cross-validate the McClintock 2 meta-pipeline and simulation system at a fixed coverage (100×), while Simulations 3 and 4 were designed to evaluate component method performance at range of fold-coverages (3×, 6×, 12×, 25×, 50×, 100×). In all cases, synthetic genomes were created with the new reproducible simulation framework available in McClintock 2 (see Implementation above), the UCSC sacCer2 version of the *S. cerevisiae* S288c reference genome (to allow cross-validation with results in Nelson *et al.* [197]), and 5-bp TSDs for all Ty families [197, 90, 233, 49, 143]. All simulations and McClintock jobs run on simulated data used the following input files (provided in `https://github.com/bergmanlab/mcclintock/blob/master/test/`): the UCSC sacCer2 version of the *S. cerevisiae* S288c reference genome (sacCer2.fasta, commit SHA 53a5e8f); reference TE annotations (reference_TE_locations.gff, commit SHA 571ef0e), taxonomy files (sac_cer_te_families.tsv, commit SHA 571ef0e), and canonical sequences for *S. cerevisiae* Ty

elements (sac_cer_TE_seqs.fasta, commit SHA d649713) from [48]. Details of the four simulation experiments are as follows:

Simulation 1 aimed replicate previous results reported in the original McClintock paper [197] using predictions from McClintock 1 on WGS datasets created by the new simulation framework in McClintock 2. Simulation parameters were the same as in Nelson *et al.* [197] and are based on the well-established preferences of Ty elements to insert upstream of genes transcribed by RNA Pol III such as tRNA genes (reviewed in [35]). To create synthetic genomes, a single Ty element was sequentially chosen from the list of Ty families that are active in S288c (Ty1, Ty2, Ty3, and Ty4) and inserted upstream one of the 299 tRNA genes in reference genome. Potential insertion sites were set to be 195-200 bp upstream of tRNA genes for Ty1, Ty2 and Ty4, and 12-17 bp upstream of tRNA genes for Ty3. In total, 598 synthetic genomes were generated: 299 genomes each having one insertion upstream of one tRNA gene in the forward strand, and 299 genomes each having one insertion upstream one tRNA gene in the reverse strand. As in Nelson *et al.* [197], the wgsim read simulator (v1, build h7132678_5) [164] was then used to create WGS paired-end datasets with a fold-coverage of 100×, read lengths of 101 bp, median insert sizes of 300 bp and a base error rate of 1%. Performance for the six original components on simulated WGS datasets was evaluated using McClintock 1 (commit SHA ad0c6cb).

Simulation 2 aimed to cross-validate the accuracy of the McClintock re-implementation using predictions from McClintock 2 for the six original components available in McClintock 1. Construction of synthetic genomes and simulation of WGS read datasets was the same as in Simulation 1. Performance for the six original components on simulated WGS datasets was evaluated using predictions from McClintock 2 (commit SHA 7aa5298).

Simulation 3 aimed to evaluate the ability of all 12 components in McClintock 2 to detect TE insertions placed in biologically-realistic locations in the *S. cerevisiae* genome over a range of WGS fold-coverages. Simulation 3 used modified Ty targeting preferences and a different read simulator than in Simulations 1 and 2. First, rather than fixing the location of insertions

for Ty1, Ty2, and Ty4 to be at one position (195-200 bp upstream of tRNA genes) as in Nelson *et al.* [197], we allowed synthetic insertions for these families to be placed in a broader region (50-800 bp upstream of tRNA genes) that better reflects patterns observed for both spontaneous and induced insertions [197, 131, 11, 186]. As above, potential insertion sites for Ty3 were set to be 12-17 bp upstream of tRNA genes as is observed for both spontaneous and induced insertions [197, 49, 50, 227]. Second, the tRNA gene and Ty family was chosen randomly from the list of tRNA genes and Ty families, rather than sequentially as in Nelson *et al.* [197]. Also, because of incompatibilities between the format of fastq sequences generated by wgsim [164] and one of the new components in McClintock 2 (TEbreak), we used the ART read simulator (v2016.06.05 build h874f42a_8) [124] for Simulation 3. Simulated WGS paired end datasets were generated by ART with a built-in error profile for the Illumina HiSeq 2500 instrument, read lengths of 101 bp, median insert sizes of 300 bp, and a standard deviation of simulated DNA fragment sizes of 10 bp (options: -ss HS25 -sam -p -s 10 -l 101 -m 300). For each synthetic genome, six sets of simulated reads were produced at different fold-coverages (3×, 6×, 12×, 25×, 50× and 100×). We also rounded the number of replicates per strand to 300 for each fold-coverage. Performance for all 12 components on simulated WGS datasets was evaluated using predictions from McClintock 2 (commit SHA 7aa5298).

Simulation 4 aimed to evaluate the performance of all 12 components in McClintock 2 to detect TE insertions in random, unique genomic regions over a range of WGS fold-coverages. Because Ty targeting of RNA Pol III transcribed regions has caused a historical accumulation of Ty fragments upstream of tRNA genes [143, 48], many synthetic non-reference Ty insertions in Simulations 1-3 are placed into fragments of pre-existing Ty sequences found in the reference genome. Simulation 4 therefore allows us to gain insight into component method performance for TE families that do not target repetitive DNA in the reference genome (unlike Ty elements in *S. cerevisiae*), and by contrasting with results from Simulation 3 to understand the effects that insertion into repetitive DNA has on TE detector performance in yeast. Unique regions were defined as the complement of regions annotated as Ty elements

in the UCSC sacCer2 version of the *S. cerevisiae* S288c reference genome [48]. Aside from differences in the location of potential insertion sites, construction of synthetic genomes and simulation of WGS read datasets was the same as in Simulation 3. Performance for all 12 components on simulated WGS datasets was evaluated using predictions from McClintock 2 (commit SHA 7aa5298).

Table 2.2: Distinguishing features of simulation experiments performed in Nelson *et al.* [197] and this study.

	Nelson *et al.* [197]	Simulation 1	Simulation 2	Simulation 3	Simulation 4
McClintock implementation	Bash	Bash	Python	Python	Python
McClintock git revision	e945d20	ad0c6cb	7aa5298	7aa5298	7aa5298
Simulation implementation	Bash	Python	Python	Python	Python
Simulation git revision	N.A.	7aa5298	7aa5298	7aa5298	7aa5298
TE insertion targets	tRNA promoter	tRNA promoter	tRNA promoter	tRNA promoter	Non-repetitive DNA
Insertion position	Fixed distance	Fixed distance	Fixed distance	Specified range	Specified range
Candidate selection	Serial	Serial	Serial	Random	Random
Read simulator	wgsim	wgsim	wgsim	ART	ART
Simulated fold-coverage	100×	100×	100×	3× to 100×	3× to 100×
Replicates per setting	299	299	299	300	300

2.4.2 Evaluation of McClintock CPU efficiency and run time on empirical data

To test whether Snakemake's ability to run multiple rules simultaneously could improve run-time performance in the new McClintock system, we applied the original McClintock 1 (commit SHA 714fe6d) and the new McClintock 2 (commit SHA f766e73) systems to a 101-bp paired-end Illumina HiSeq 2000 dataset for *S. cerevisiae* strain YJM1460 [261] that was down-sampled to 50× and 100× coverage using seqtk (v1.3) [163]. To compare run times for McClintock 1 and McClintock 2, we invoked both versions of the system only with the original

six components present in McClintock 1 ("-m ngs_te_mapper,relocate,temp,retroseq,popoolationte,te-locate"). To evaluate the impact on run time caused by addition of the six new components in McClintock 2, we also invoked McClintock 2 using all 12 components ("-m ngs_te_mapper,ngs_te_mapper2,relocate, relocate2,temp,temp2,retroseq,popoolationte, popoolationte2,te-locate,teflon,tebreak"). In both cases, we varied the number of cores requested (1, 5, 10, 15, 20, 25, 30) but kept the amount of RAM constant (20 Gb). For all runs, we used unzipped, untrimmed fastq reads (as required by McClintock 1) and the following input files (provided in `https://github.com/bergmanlab/mcclintock/blob/master/test/`): the UCSC sacCer2 version of the *S. cerevisiae* S288c reference genome (sacCer2.fasta, commit SHA 53a5e8f); reference TE annotations (reference_TE_locations.gff, commit SHA 571ef0e), taxonomy files (sac_cer_te_families.tsv, commit SHA 571ef0e), and canonical sequences for *S. cerevisiae* Ty elements (sac_cer_TE_seqs.fasta, commit SHA d649713) from [48]. Trials were run serially on the same cluster node (an AMD EPYC processor with a total of 32 cores and 256GB RAM, operating system Linux 3.10.0-1160.36.2.el7.x86_64). The average and standard deviation of CPU efficiency and run times was calculated from five replicates of each setting (defined as the combination of number of cores and McClintock run options). All possible settings were run as a batch for each replicate to control for differences in cluster node performance over the duration of the experiment.

2.4.3 Analysis of TE insertions in a diverse panel of wild and domesticated yeast strains

To demonstrate the utility of the McClintock 2 system for large-scale mining of empirical short-read WGS data, we applied our pipeline to Illumina paired-end datasets for 1,011 *S. cerevisiae* isolates [261, 222] sequenced to variable depths (from 50× to 900× fold-coverage). Illumina WGS data were downloaded from NCBI Sequence Read Archive and converted to fastq format using SRA toolkit (v2.10.8) [1]. To reduce the influence

of variable coverage and computing resources, if the coverage of the original dataset was greater than 50×, we down-sampled it to 50× fold-coverage using seqtk (v1.3) [163]. The full McClintock 2 pipeline using all 12 components plus the coverage and trimgalore module ("-m ngs_te_mapper,ngs_te_mapper2,relocate, relocate2,temp,temp2,retroseq,popoolationte,popoolationte2,telocate,teflon,tebreak, coverage,trimgalore") (commit SHA 7aa5298) was applied to this yeast resequencing dataset using the following input files (provided in `https://github.com/bergmanlab/mcclintock/blob/master/test/`): the UCSC sacCer2 version of the *S. cerevisiae* S288c reference genome (sacCer2.fasta, commit SHA 53a5e8f); reference TE annotations (reference_TE_locations.gff, commit SHA 571ef0e), taxonomy files (sac_cer_te_families.tsv, commit SHA 571ef0e), and canonical sequences for *S. cerevisiae* Ty elements (sac_cer_TE_seqs.fasta, commit SHA d649713) from [48].

We identified non-reference TE insertions in the vicinity of tRNA genes using BEDtools (v2.30.0,"window -u -sw -l 1000 -r 500") [228]. BEDtools ("closest -D b") was used to calculate the genomic distance between the start position of each non-reference TE insertion and its closest tRNA transcription start site (TSS). To analyze the relationship between the distribution of non-reference TE insertions and nucleosome occupancy upstream of tRNA genes, we mapped micrococcal nuclease digestion with deep sequencing (MNase-seq) data from [208] to the sacCer2 version of the *S. cerevisiae* S288c reference genome using Bowtie (v1.2.3 [158]) and generated genome-wide nucleosome occupancy profiles using NUCwave [229]. Whole genome nucleosome occupancy profiles in .wig format were converted to .bw format using wigToBigWig (v377) [141], and then bwtool (v20170428, "aggregate 2000:500") [224] was used to calculate nucleosome occupancy in the region 2 kb upstream to 500 bp downstream of all tRNA TSSs.

2.4.4 Data analysis and visualization

Data were analyzed and visualized in R (v3.6.3) using the ggplot2 (v3.3.3) [281] and GGally (v2.1.1, https://ggobi.github.io/ggally/) packages. UpSet plots were generated by comparing

overlaps among true-positive predictions across component methods using UpSetR (v1.4.0) [61]. Kernel density plots showing distributions of Ty insertion sites around tRNA genes were created with the ggplot2 function "geom_density" with Gaussian kernel and 0.4× default bandwidth. In contrast to Nelson *et al.* [197], predictions from TEMP [295] in the current study were not partitioned into those with split-read versus read-pair evidence.

2.5 Results and Discussion

2.5.1 Validation of McClintock 2 implementation and simulation system

Here we present a Python-based re-implemention and extension of the McClintock TE detection meta-pipeline, as well as a reproducible simulation system to test the performance of component TE detectors integrated into McClintock. The major features that distinguish the new McClintock 2 system from the original version reported in Nelson *et al.* [197] are described in the Implementation section above and are summarized in Table 2.1.

To validate the new McClintock implementation and simulation system, we used results from Simulations 1-3 (see Section 2.4.1). To be able to compare with previous results, we used the same approach as in Nelson *et al.* [197] to summarize the six original component methods' performance across simulated samples. Namely, we calculated the average number of non-reference TE predictions made by each component overall and at different levels of positional accuracy determined by varying window sizes (within 0, 5, 100, 300 or 500 bp of the synthetic insertion). Because differences in component performance on the positive and negative strands were minimal, here we simplified reporting of summary metrics relative to Nelson *et al.* [197] by averaging results from forward and reverse strands across simulations (Fig 2.2). The expected value of non-reference TE insertions per simulation is one, with values less than one representing a tendency for a TE detector to make false negative predictions and values greater than one representing a tendency for false positive predictions.

We initially asked whether we could replicate performance results for the original six McClintock components reported in Nelson *et al.* [197] using the new reproducible simu-

lation system in McClintock 2 (Simulation 1). This simulation used identical TE insertion settings but a more recent version of McClintock 1 and different simulation implementation relative to Nelson *et al.* [197] (Table 2.2). As shown in Fig. 2.2, results from Simulation 1 generated very similar numbers (±0.09) of average non-reference TE predictions as in Nelson *et al.* [197], with the exception of the total number of non-reference TE predictions made by RelocaTE being substantially lower (∼25%) in the current study. This difference could arise from many sources (including different git commit SHAs for McClintock 1, component method dependencies, computing environments, simulation code, or run-to-run variation in component methods) and underscores the difficulty in fully reproducing previous TE detector benchmarking studies. Nevertheless, the general consistency between results reported previously by Nelson *et al.* [197] and those generated independently here (Simulation 1) indicates that our new Python-based reproducible simulation system yields similar results as the old Bash-based simulation system, and that performance estimates of the original six component methods generated by McClintock 1 are broadly replicable.

Next, we sought to test whether the new Python-based McClintock 2 meta-pipeline generates similar performance results as the original Bash-based McClintock system. Because of the differences observed between Nelson *et al.* [197] and Simulation 1 (Fig. 2.2), we chose to use the reproducible results for McClintock 1 generated here (Simulation 1) as a baseline to compare with the McClintock 2 meta-pipeline. Also, using Simulation 1 as a baseline fully controls for differences between the new Python-based simulation system and the Bash-based simulation system used in Nelson *et al.* [197], and therefore allows us to isolate any potential performance differences to the McClintock meta-pipeline itself. As shown in Fig. 2.2, no major differences in performance of the six original component methods (±0.065) are observed for the Python-based (Simulation 2) and Bash-based (Simulation 1) implementations of the McClintock meta-pipeline. The largest difference observed is for RetroSeq, where this component exhibits improved performance in McClintock 2. Overall, this analysis demonstrates that performance estimates generated by the Python-based McClintock 2 are

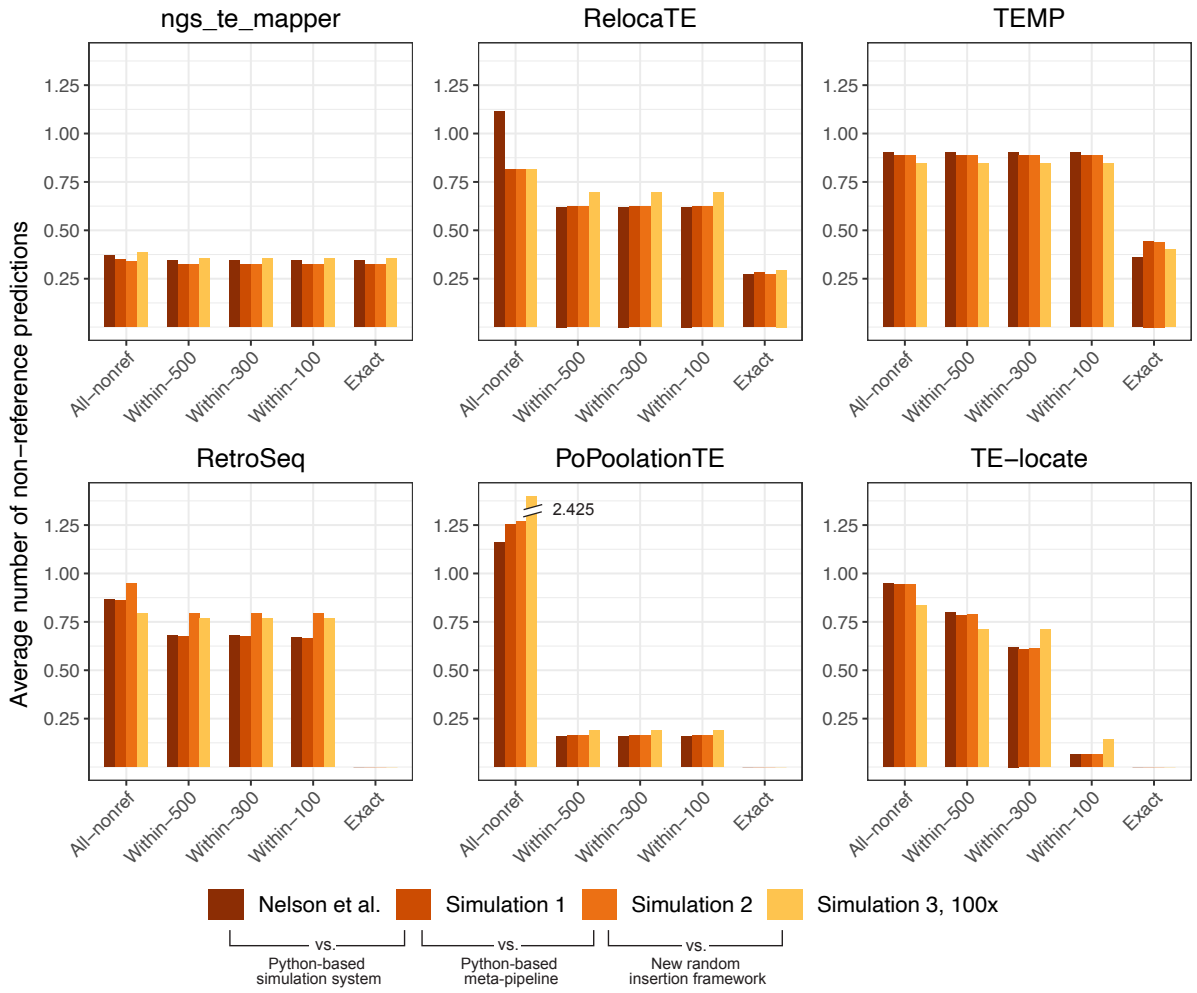

Figure 2.2: **Validation of McClintock 2 meta-pipeline and simulation system.** The bar plot shows the average number of all non-reference predictions and true positive predictions falling in different window sizes (exact, within 100 bp, within 300 bp and within 500 bp) across all simulated replicates, including old McClintock data [197], simulation 1, simulation 2 and simulation 3 (100×). Values shown in this figure are the average number of the forward and reverse strand since two strands were run separately in our simulation framework. Only six original methods included in McClintock are shown here for validation. The PoPoolationTE value exceeds y-axis limitation and thus is annotated with a break and the actual number.

the same or better as those from the Bash-based McClintock 1 and indicates that no major differences should be experienced by users migrating from McClintock 1 to McClintock 2.

The approach used to simulate non-reference TE insertions in Nelson *et al.* [197] had a number of hard-coded constraints tailored for the yeast genome that made it unsuitable for application to other organismal contexts. Namely, the TE family and location of simulated insertions were generated sequentially from a candidate set of TE families and tRNA gene promoters, and the insertion positions were at fixed distances upstream of tRNA genes. To allow more general modeling of insertion preferences, we designed the new McClintock 2 simulation system to generate an instance of a randomly selected TE family to be randomly inserted into a set of arbitrary genomic locations specified in a configurable JSON file. By modifying parameters in the JSON file, the new simulation system is flexible enough to permit implementation of the Nelson *et al.* [197] approach as well as more biologically-realistic models of TE insertion, such as Ty insertion over a range of positions upstream of tRNA genes (see Materials and Methods for details). We tested if this new biologically-realistic insertion framework (Simulation 3) gave similar results at the same fold-coverage (100×) as the original sequential insertion framework with fixed Ty insertion locations (Simulation 2). As shown in Fig. 2.2, both the random (Simulation 3) and sequential (Simulation 2) insertion frameworks give similar results (±0.17) with the exception of PoPoolationTE. We speculate that the unusually high number of total non-reference predictions made by PoPoolationTE in Simulation 3 may be caused by misclassification of reference TE insertions as non-reference TE insertions (see also [197]).

In summary, we can conclude that key improvements in McClintock 2 including (i) the Python-based implementation of the simulation system (Simulation 1 vs. Nelson *et al.* [197]), (ii) the Python-based implementation of the meta-pipeline (Simulation 2 vs. Simulation 1), and (ii) the random insertion model in the single insertion simulation framework (Simulation 3 vs. Simulation 2) together generate results that broadly replicate those published previously in Nelson *et al.* [197].

2.5.2 Parallelization by Snakemake in McClintock 2 improves multi-core job run times

McClintock 2 uses Snakemake as a workflow manager, which allows parallelization of multiple rules when enough CPU cores are available. To investigate potential improvements in computational efficiency due to Snakemake job management in McClintock 2, we executed McClintock 1 and McClintock 2 (using either the six original or all 12 components) with a variable number of cores under controlled computing environments. For this analysis, we used a 101-bp paired-end Illumina dataset from the *S. cerevisiae* domesticated palm-wine strain Y12 (YJM1460; SRA: SRR800842) downsampled to either 50× or 100× (see Materials and Methods for details). In general, we observe a decreasing trend in CPU efficiency for all McClintock run configurations as the number of CPU cores increases (Fig. 2.3A). However, when more than one core is used, we found that McClintock 2 has better CPU efficiency than McClintock 1 for the six original components, and that CPU efficiency is highest for all numbers of cores used when executing all 12 McClintock 2 components. Improved CPU efficiency for McClintock 2 is observed at both 50× and 100× coverage.

When comparing run times for the six original components at both 50× and 100× coverage (Fig. 2.3B), we found that McClintock 2 was slightly slower (∼10%) than McClintock 1 when using only a single core, however McClintock 2 finished faster than McClintock 1 when five or more cores were available. Run times for McClintock 2 decreased up to 15 CPU cores when executing either the original six or all 12 components, while no major benefits in run time were observed beyond 5 CPU cores for McClintock 1. Interestingly, when 15 or more cores are allocated, executing McClintock 2 with all 12 components is faster than running McClintock 1 using only six components. Together, these results indicate that parallelization afforded by using Snakemake in McClintock 2 results in better utilization of computing resources in multi-core computing systems relative to McClintock 1.

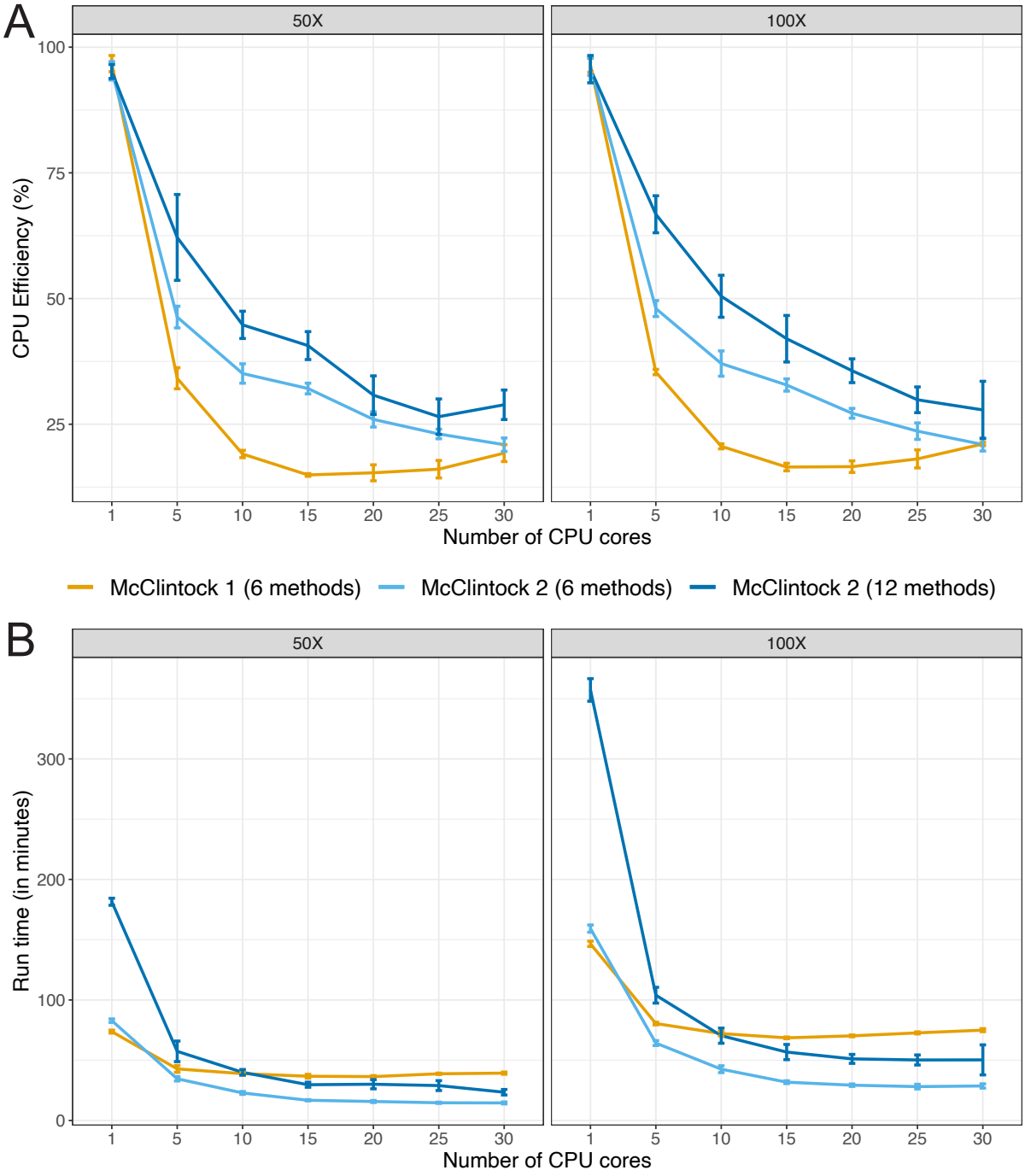

Figure 2.3: **McClintock 2 re-implementation improves CPU efficiency and run time on multi-core architectures.** Shown are average (A) CPU efficiency and (B) run times across 5 replicates of McClintock 1 (six component methods, orange line) or McClintock 2 (same six component methods as for McClintock 1, light blue line; all 12 component methods in McClintock 2, dark blue line) applied to 50× and 100× Illumina 101-bp paired-end sample for *S. cerevisiae* strain YJM1460 (SRA: SRR800842). Error bars indicate standard deviations across replicates. To allow compatibility with McClintock 1, all runs were performed on unzipped, untrimmed fastq files and thus run times do not include these processes.

2.5.3 Evaluation of McClintock 2 component TE detectors on simulated yeast data

McClintock 2 now includes six additional component methods for detecting non-reference TE insertions using short-read WGS data that are not available in McClintock 1: ngs_te_mapper2 [111], PoPoolationTE2 [147], RelocaTE2 [51], TEBreak [249], TEFLoN [2], and TEMP2 [289]). Four of the new components (ngs_te_mapper2, PoPoolationTE2, RelocaTE2, TEMP2) are "second-generation" versions of tools previously incorporated in McClintock 1, and the other two (TEBreak and TEFLoN) represent new TE detection strategies not represented in McClintock 1.

To evaluate the relative performance of these 12 TE detection strategies, we used the new reproducible single insertion simulation approach available in McClintock 2. The evaluation of McClintock 2 component methods performed here differs from the approach originally used in Nelson *et al.* [197] in several ways. First, we used a more biologically realistic model to randomly select the family and location of simulated non-reference TE insertions over a range of genomic positions. Second, we investigated the effects of component performance over a range of fold-coverages, which provides practical guidance for users to optimize data generation. Third, the summary statistics used in Nelson *et al.* [197] are not standard evaluation metrics, and therefore difficult to compare with other evaluation studies (e.g., [234, 269]). Thus, here we used the standard evaluation metrics of recall and precision derived from the number of true-positive, false-positive, and false-negative predictions at various window sizes (see Implementation for details). Finally, here we evaluated performance using simulations that model yeast TE target preferences (Simulation 3) as in Nelson *et al.* [197], as well as random insertion into non-repetitive DNA (Simulation 4). Simulation 4 helps us interpret how insertion into repetitive DNA in the reference genome influences component performance in yeast and gives insight into potential component performance in other organisms that don't have strong TE targeting preferences.

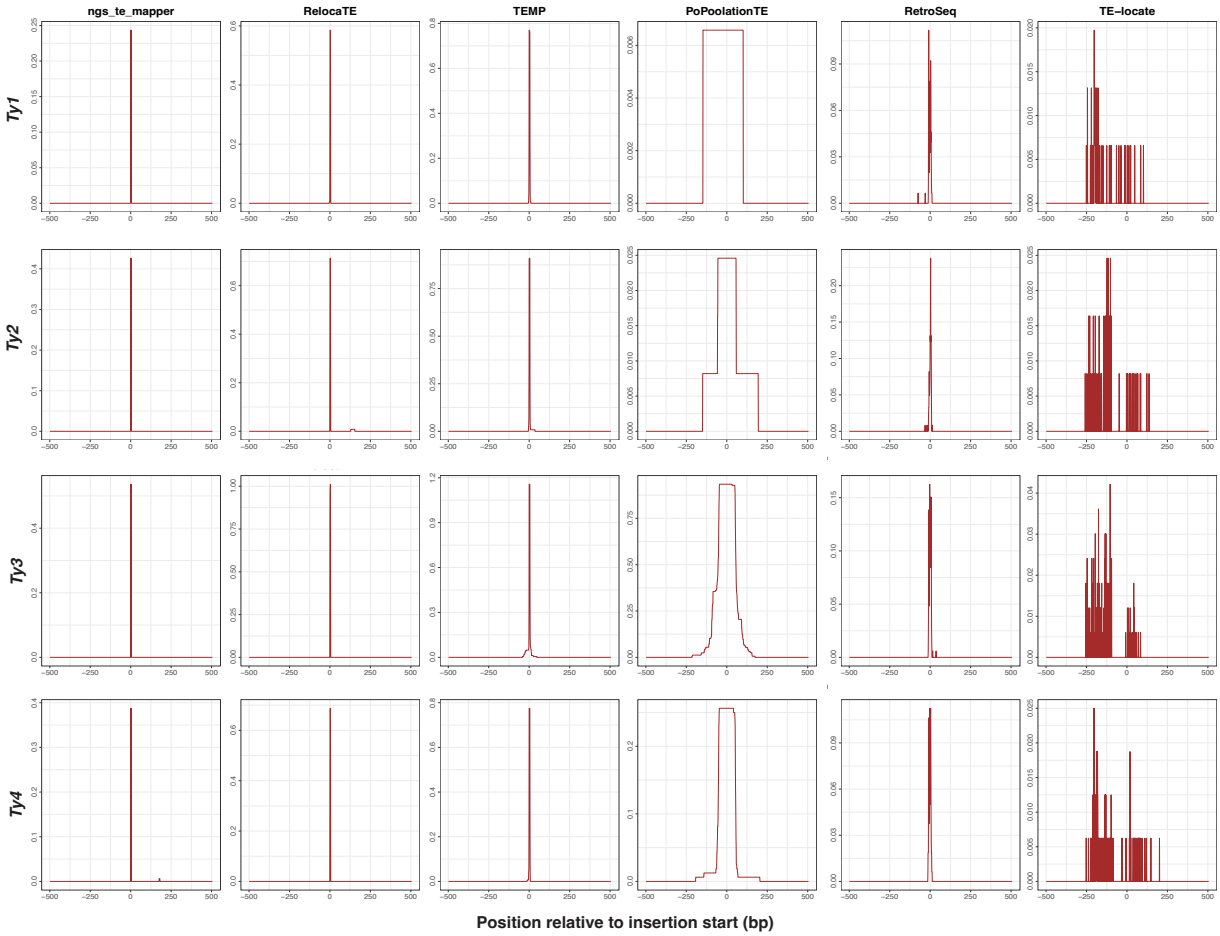

Figure 2.4: **Positional accuracy of six original McClintock component methods in simulated data generated by the yeast insertion model (Simulation 3).** Plots show coverage of all non-reference TE predictions relative to the coordinate of the start position of synthetic insertions in simulated datasets with 50× fold-coverage averaged across all 600 replicate simulations. Predictions more than 500 bp from the start position of synthetic insertions are not shown.

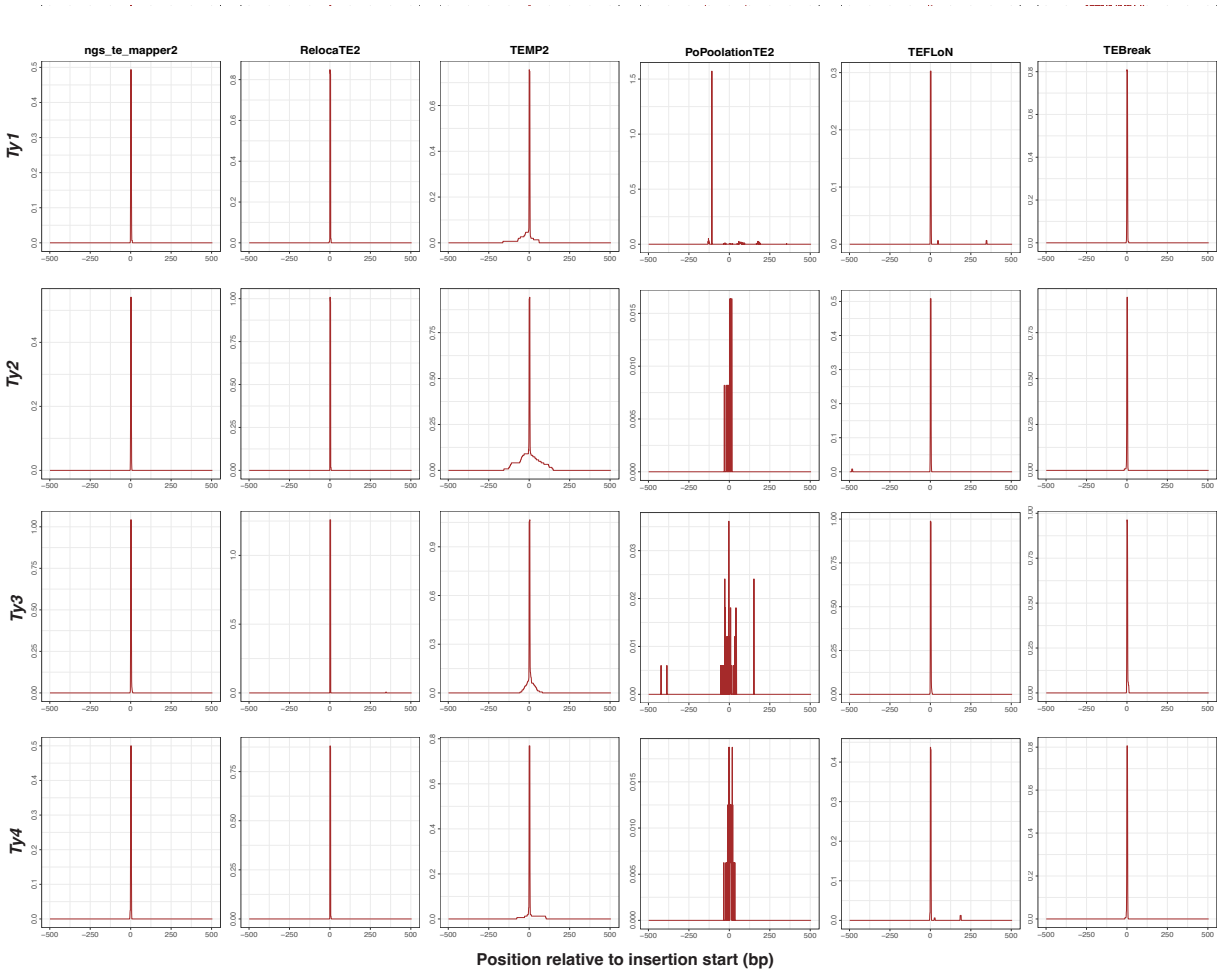

Figure 2.5: **Positional accuracy of six new McClintock component methods in simulated data generated by the yeast insertion model (Simulation 3).** Plotting criteria are the same as in Figure 2.4.

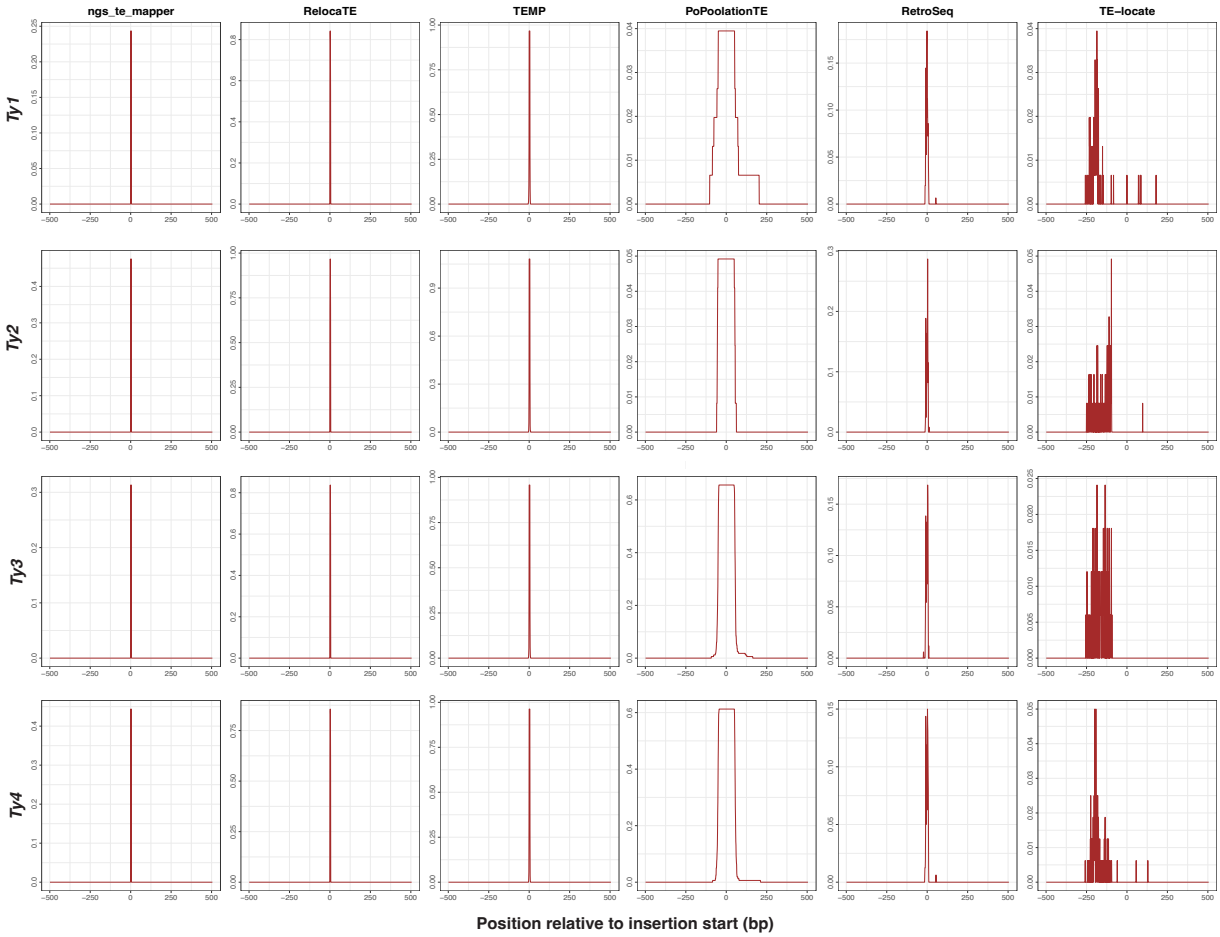

Figure 2.6: **Positional accuracy of six original McClintock component methods in simulated data generated by the random insertion model (Simulation 4).** Plotting criteria are the same as in Figure 2.4.

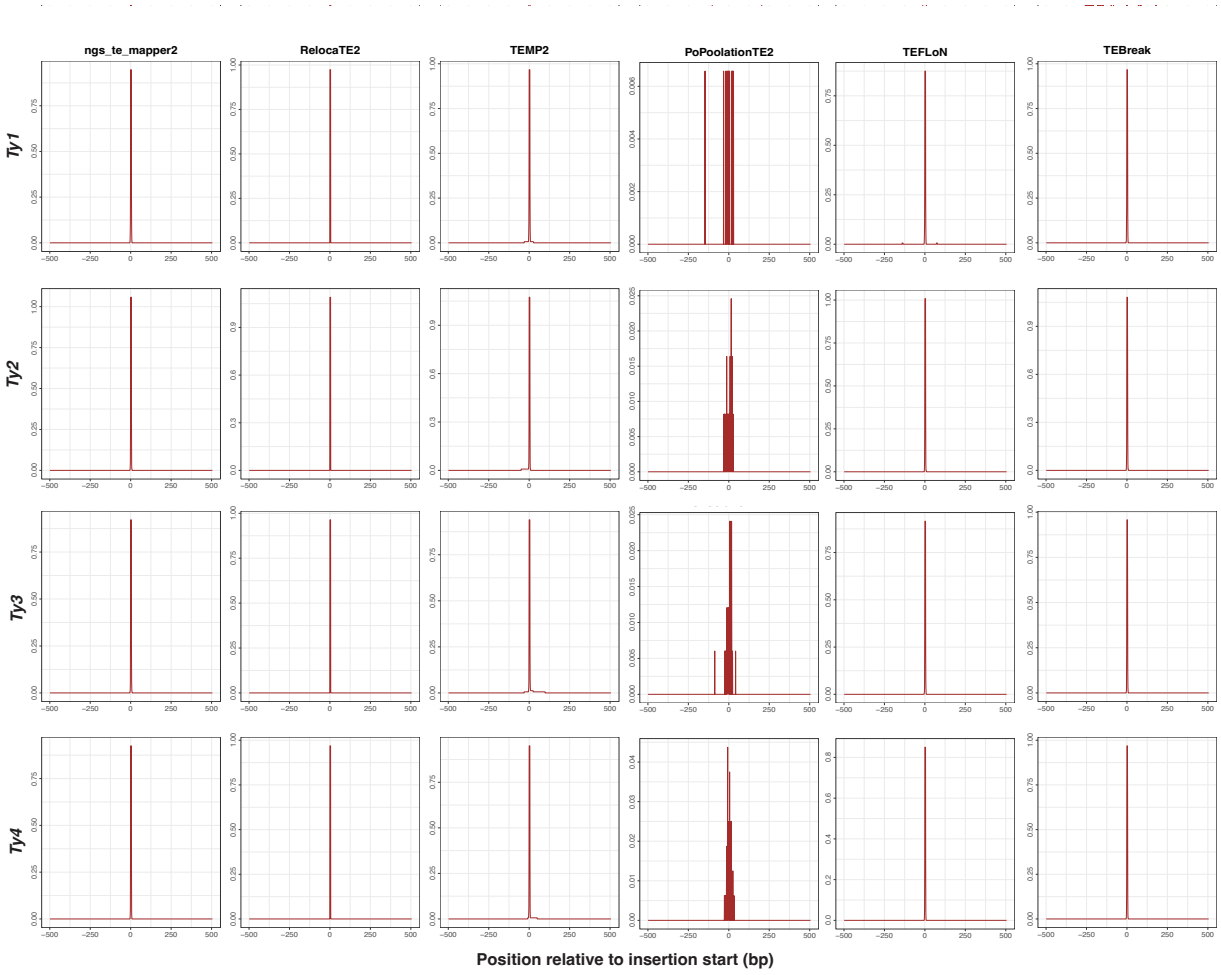

Figure 2.7: **Positional accuracy of six new McClintock component methods in simulated data generated by the random insertion model (Simulation 4).** Plotting criteria are the same as in Figure 2.4.

As shown previously for the original six components in McClintock 1 [197], different methods in McClintock 2 vary in their ability to accurately predict the position of synthetic TE insertions in simulated data (Figs. 2.4, 2.5, 2.6, 2.7). Several components (i.e., ngs_te_mapper, ngs_te_mapper2, RelocaTE, RelocaTE2, TEFLoN and TEBreak) make most predictions very close the expected location of the synthetic insertion, whereas other components (i.e., PoPoolationTE, PoPoolationTE2, and TE-locate) predict insertions over a wider distance around the synthetic insertion (∼500bp). Even for components that attempt to identify insertion breakpoints at nucleotide resolution using split-read information, only ngs_te_mapper and RelocaTE2 have predictions with the expected 5-bp TSD length for all Ty families (Figs. 2.8, 2.9). This method-dependent variation in positional accuracy makes it difficult to fairly compare performance of different components at a single window size. Therefore, we calculated performance metrics over a range of window sizes for overlaps between predicted and simulated insertions (exact, within-5, within-100, within-300, and within-500 bp). Inspection of the complete set of results for all window sizes (Figs. 2.10, 2.11, 2.12, 2.13) revealed that two window sizes (exact and within-100 bp) could illustrate the majority of key features in McClintock 2 component performance across the insertion models and range of fold-coverages investigated here.

To guide selection of TE detectors in the model species *S. cerevisiae*, we first analyzed the performance of McClintock 2 components using data from Simulation 3 where TEs were inserted in their biologically realistic locations in yeast upstream of tRNA genes [227, 11, 41, 186, 214] (Fig. 2.14; purple lines). Three general trends can be observed in the recall and precision curves for McClintock 2 components for non-reference TEs in tRNA promoter regions. First, component performance is higher when allowing a more relaxed window size (within-100 bp) to classify TP predictions (dashed lines) than when requiring TP predictions to be exact (solid lines), consistent with most methods making many predictions that are not precise to exact nucleotide coordinates. Only ngs_te_mapper makes all its predictions exactly, and thus performance curves are overlayed for both window sizes shown

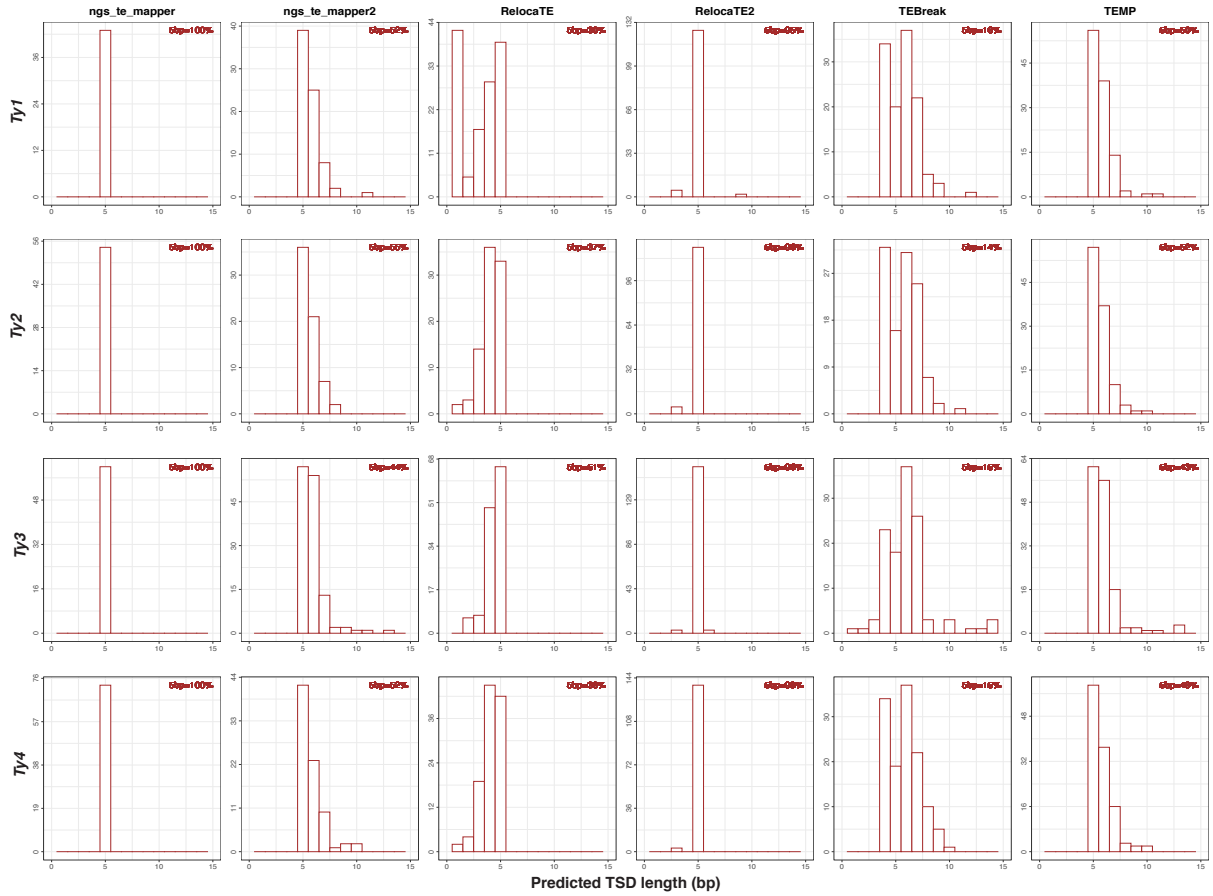

Figure 2.8: **TSD length distributions in simulated data generated by the yeast insertion model (Simulation 3).** Plots show histograms of TSD lengths for predictions in simulated datasets with 50× fold-coverage that have split-read evidence. Shown are all predictions made by ngs_te_mapper, ngs_te_mapper2, RelocaTE, RelocaTE2, and TEBreak and the subset of TEMP predictions with split-read evidence. Percentages indicate the proportion of predictions with the expected 5-bp TSD length.

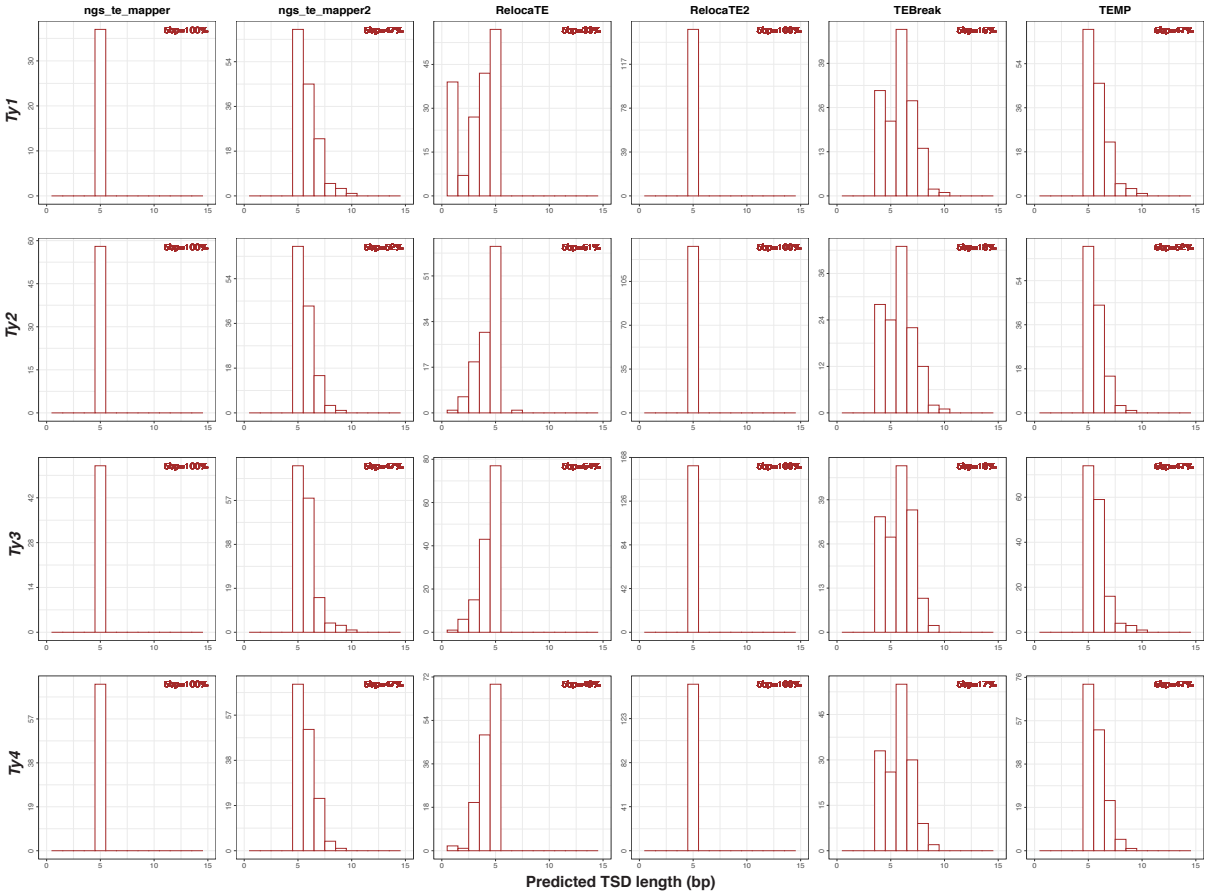

Figure 2.9: **TSD length distributions in simulated data generated by the random insertion model (Simulation 4).** Plotting criteria are the same as in Figure 2.8.

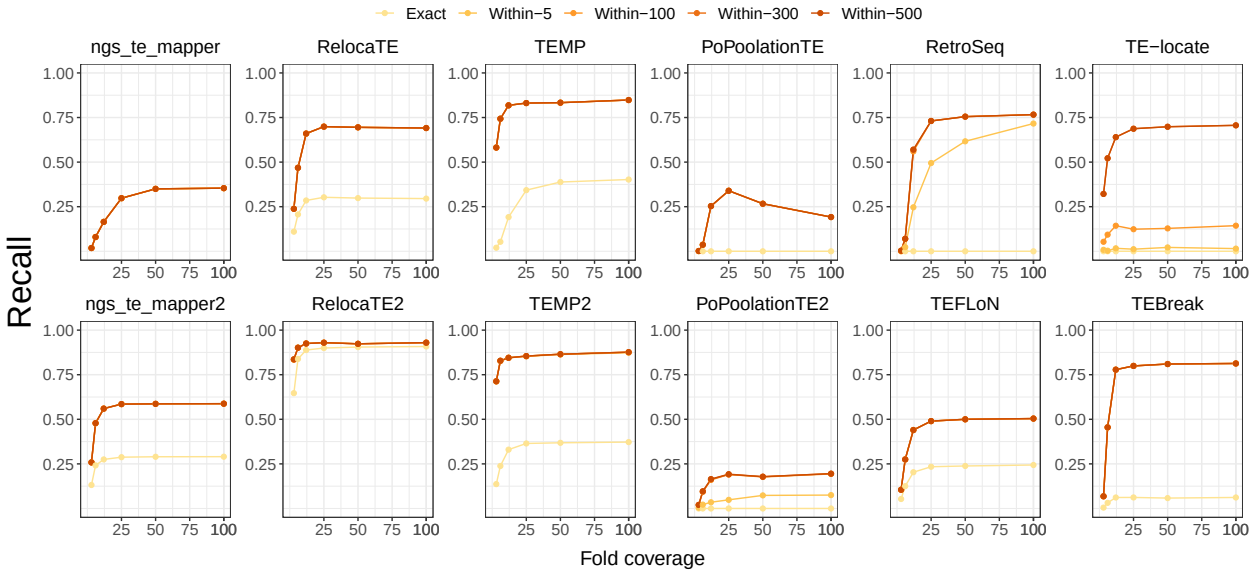

Figure 2.10: **Recall of McClintock 2 component methods in simulated data generated by the yeast insertion model (Simulation 3).** Shown are the recall of 12 component methods across variable fold-coverage (i.e., 3×, 6×, 12×, 25×, 50× and 100×). Curves in different colors indicate variable overlapping window sizes to identify true-positive predictions, including exact coordinates, within 5 bp, within 100 bp, within 300 bp, and within 500 bp. The six original McClintock component methods are shown on the first row and the six newly-integrated methods are shown on the second row.

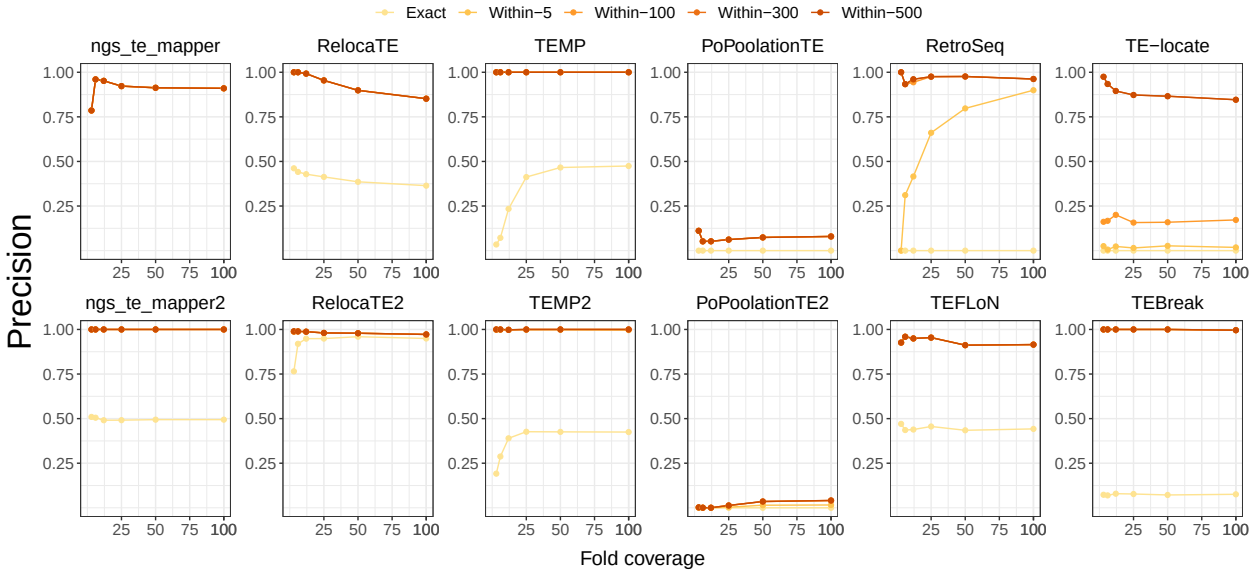

Figure 2.11: **Precision of McClintock 2 component methods in simulated data generated by the yeast insertion model (Simulation 3).** Plotting criteria are the same as in Figure 2.10.

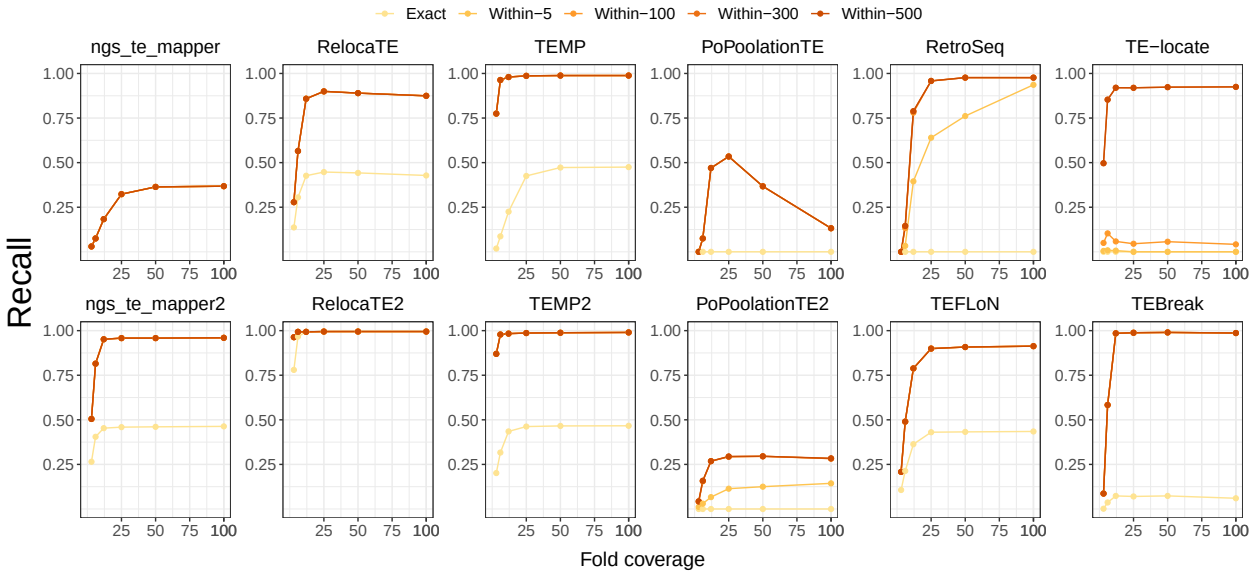

Figure 2.12: **Recall of McClintock 2 component methods in simulated data generated by the random insertion model (Simulation 4).** Plotting criteria are the same as in Figure 2.10.

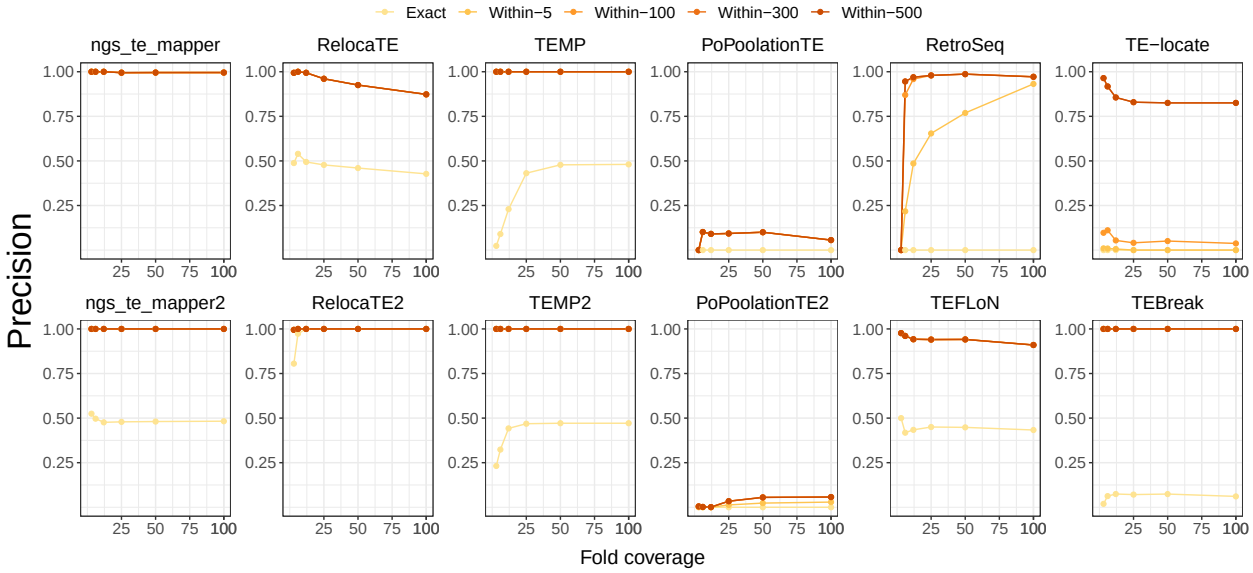

Figure 2.13: **Precision of McClintock 2 component methods in simulated data generated by the random insertion model (Simulation 4).** Plotting criteria are the same as in Figure 2.10.

for this method. Second, component performance typically increases with sequencing depth and then plateaus at a method-specific coverage. Some exceptions to this trend are observed for PoPoolationTE recall and RelocaTE and TEFLoN precision, which show decreasing performance with increasing coverage. Performance curves for most components suggest that 25× is the most cost-effective depth of sequencing coverage, and 50× is sufficient to optimize performance for most component methods. Third, second generation versions of TE detectors – apart from PoPoolationTE2 – typically show improved performance relative to their first generation counterparts, suggesting that gains in TE detector performance can be made by continued method development. Overall, results from Simulation 3 suggest that RelocaTE2 has remarkably high recall (∼92%) and precision (∼98%) to predict non-reference TE insertions in tRNA promoter regions of the yeast genome (at 50× coverage and ≥5 bp window size). The recall and precision of TEMP, TEMP2, RetroSeq and TEBreak to detect TE insertions in yeast promoter regions are all both greater than 75% when allowing non-exact predictions in WGS datasets with greater than 50× coverage.

To guide selection of TE detectors in the model species *S. cerevisiae*, we first analyzed the performance of McClintock 2 components using data from Simulation 3 where TEs were inserted in their biologically realistic locations for yeast in regions upstream of tRNA genes [227, 11, 41, 186, 214] (Fig. 2.14; purple lines). Three general trends can be observed in the recall and precision curves for McClintock 2 components for non-reference TEs in tRNA promoter regions. First, component performance is higher when allowing a more relaxed window size (within-100 bp) to classify TP predictions (dashed lines) than when requiring TP predictions to be exact (solid lines), consistent with most methods making many predictions that are not precise to exact nucleotide coordinates. Only ngs_te_mapper makes all its predictions exactly, and thus performance curves are overlayed for both window sizes shown for this method. Second, component performance typically increases with sequencing depth and then plateaus at a method-specific coverage. Some exceptions to this trend are observed for PoPoolationTE recall and RelocaTE and TEFLoN precision, which show decreasing per-

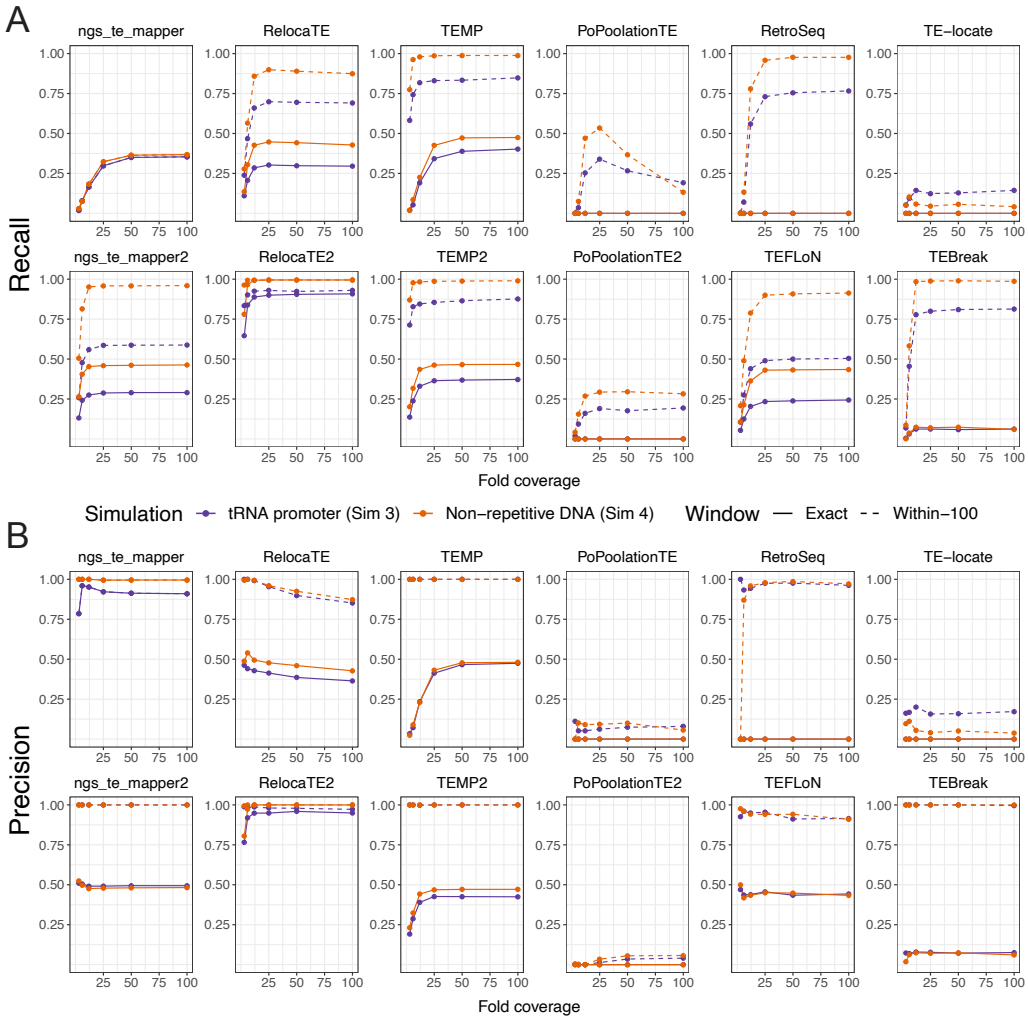

Figure 2.14: **Performance of McClintock 2 component methods in simulated yeast WGS data.** Shown are the (A) recall and (B) precision across different fold-coverage for individual compnent methods to detect single synthetic insertions in an otherwise unmodified *S. cerevisiae* reference genome. Purple lines (Simulation 3) model the biologically realistic insertion preferences of yeast TEs, with synthetic Ty insertions created upstream of tRNA genes in regions that often have fragments of prior TE insertions in the reference genome. Orange lines (Simulation 4) model random insertions in non-repetitive regions, which allows insight into the effects of insertion within repetitive DNA and component's performance for organisms without strong TE targeting preferences. Points indicate tested fold-coverage configurations, i.e, 3×, 6×, 12×, 25×, 50× and 100×. Solid lines represent performance estimates for non-reference TE predictions made at the exact site of the synthetic insertion. Dashed lines represent performance estimates for non-reference TE predictions made within 100 bp surrounding the synthetic insertion site. The six original component methods in McClintock 1 are on the top row of each panel, and the six new methods in McClintock 2 are on the second row of each panel.

formance with increasing coverage. Performance curves for most components suggest that 25× is the most cost-effective depth of sequencing coverage, and 50× is sufficient to optimize performance for most component methods. Third, second generation versions of TE detectors – apart from PoPoolationTE2 – typically show improved performance relative to their first generation counterparts, suggesting that gains in TE detector performance can be made by continued method development. Overall, results from Simulation 3 suggest that RelocaTE2 has remarkably high recall (∼92%) and precision (∼98%) to predict non-reference TE insertions in tRNA promoter regions of the yeast genome (at 50× coverage and ≥5 bp window size). The recall and precision of TEMP, TEMP2, RetroSeq and TEBreak to detect TE insertions in yeast promoter regions are all both greater than 75% when allowing non-exact predictions in WGS datasets with greater than 50× coverage.

While biologically relevant for *S. cerevisiae*, Simulation 3 potentially underestimates component performance since synthetic insertions in the yeast insertion model are often placed into fragments of TE sequences that occur upstream of tRNA genes in the reference genome [48, 143]. To gain insight into other organismal contexts and understand how insertion into reference TE sequences may impact McClintock 2 component performance in yeast, we also investigated a model of random insertion into unique regions of the *S. cerevisiae* genome (Simulation 4). Recall and precision curves for the random insertion model (Fig. 2.14; orange lines) show the same general trends but have consistently higher performance relative to results for the same component in the tRNA promoter insertion model (Fig. 2.14; purple lines). General performance improvements in Simulation 4 relative to Simulation 3 indicate that prediction of non-reference TE insertions into locally TE-rich regions (like tRNA promoters in yeast) is more challenging for most component methods than for those in non-repetitive DNA. Consistent with results from the tRNA promoter model, RelocaTE2 has the highest recall and precision of any component in McClintock 2, with essentially perfect performance in non-repetitive regions of the *S. cerevisiae* genome (∼100%) even at exact base-pair resolution. At the within-100 bp window size, TEMP, TEMP2, RetroSeq

and TEBreak also perform very well at identifying non-reference insertions in unique regions. We note that two additional methods – ngs_te_mapper2 and TEFLoN– show dramatically better recall for non-exact predictions in non-repetitive regions relative to tRNA promoter regions, putting these methods in a similar performance class as TEMP, TEMP2, RetroSeq and TEBreak for insertions in non-repetitive DNA.

In summary, results from our reproducible simulation experiments show that RelocaTE2 is the best-performing component method currently in the McClintock 2 meta-pipeline to detect non-reference TE insertions in *S. cerevisiae* under both the biologically relevant tRNA promoter and non-repetitive insertion models. If exact base-pair positional accuracy is not required, TEMP, TEMP2, RetroSeq and TEBreak also exhibit high performance to detect TE insertions in *S. cerevisiae* tRNA promoter regions at 50× coverage. Furthermore, our results suggest that at 50× coverage TEMP, TEMP2, RetroSeq, TEBreak, ngs_te_mapper2, and TEFLoN may have high performance to detect TE insertions in non-repetitive regions of other genomes. Finally, our results indicate that users should provide McClintock 2 with WGS datasets of at least 25× fold-coverage to generate the best performance from most component methods, but that coverage higher than 50× may not lead to further performance benefits.

2.5.4 Best-in-class component methods reveal consistent insights into patterns of TE insertion in yeast.

Our reproducible simulation system allowed us to validate the McClintock 2 re-implementation and evaluate component methods under ideal conditions, where only one TE insertion per strain is to be detected and all sequencing data is generated in an identical way. However, analysis of large-scale empirical re-sequencing data is expected to present additional challenges to TE detectors, such as more complex insertion patterns in a given strain (e.g., multiple insertions, tandem elements) and variability in WGS dataset composition across strains (e.g., fold-coverage or median insert size). To assess component method performance

under a more realistic population genomic scenario, we applied the complete McClintock 2 system (all 12 components plus the coverage module) to empirical WGS datasets from a world-wide sample of 1,011 *S. cerevisiae* strains isolated from diverse environments and geographical locations [261, 222]. Since these WGS datasets varied significantly in their fold-coverage (50× to 900×), we down-sampled all WGS datasets to 50× prior to running McClintock 2 based on simulation results above that indicated 50× coverage is sufficient to maximize component method performance (Fig. 2.14). Our goals for this analysis were to evaluate whether McClintock 2 component methods: (i) predict consistent numbers of non-reference TE insertions per strain (overall and by Ty family), and (ii) have sufficient positional accuracy to identify expected patterns of targeting in yeast tRNA promoter regions.

To evaluate consistency among component methods on empirical WGS data, we first quantified the distribution of non-reference TE predictions made by each method across 1,011 yeast strains (Fig. 2.15A). We observe substantial variation in median values of non-reference Ty insertions across all 12 methods, as well as in the correlation among methods in the number of predicted non-reference Ty insertions (Fig. 2.16). Interestingly, median values of ∼50 non-reference Ty insertions per strain are consistently predicted by the five component methods (RelocaTE2, TEMP, TEMP2, RetroSeq and TEBreak) (Fig. 2.15A; bold outlines) that showed the best performance (when considering non-exact predictions) in simulated data under the biologically relevant tRNA promoter model (Fig. 2.14A; purple dotted lines). These five components also have the highest pairwise correlation of non-reference Ty counts per strain across methods (Fig. 2.16). We note that the original six components do not show the general pattern of ∼50 non-reference Ty insertions per strain (Fig 2.15A; top row) and have some of the lowest correlation of non-reference Ty counts across methods (Fig. 2.16), underscoring the value of incorporating additional recently-developed methods into McClintock 2 to reveal consistent biological patterns that could not be revealed by the McClintock 1 system. Since McClintock 2 component methods do not distinguish full-length

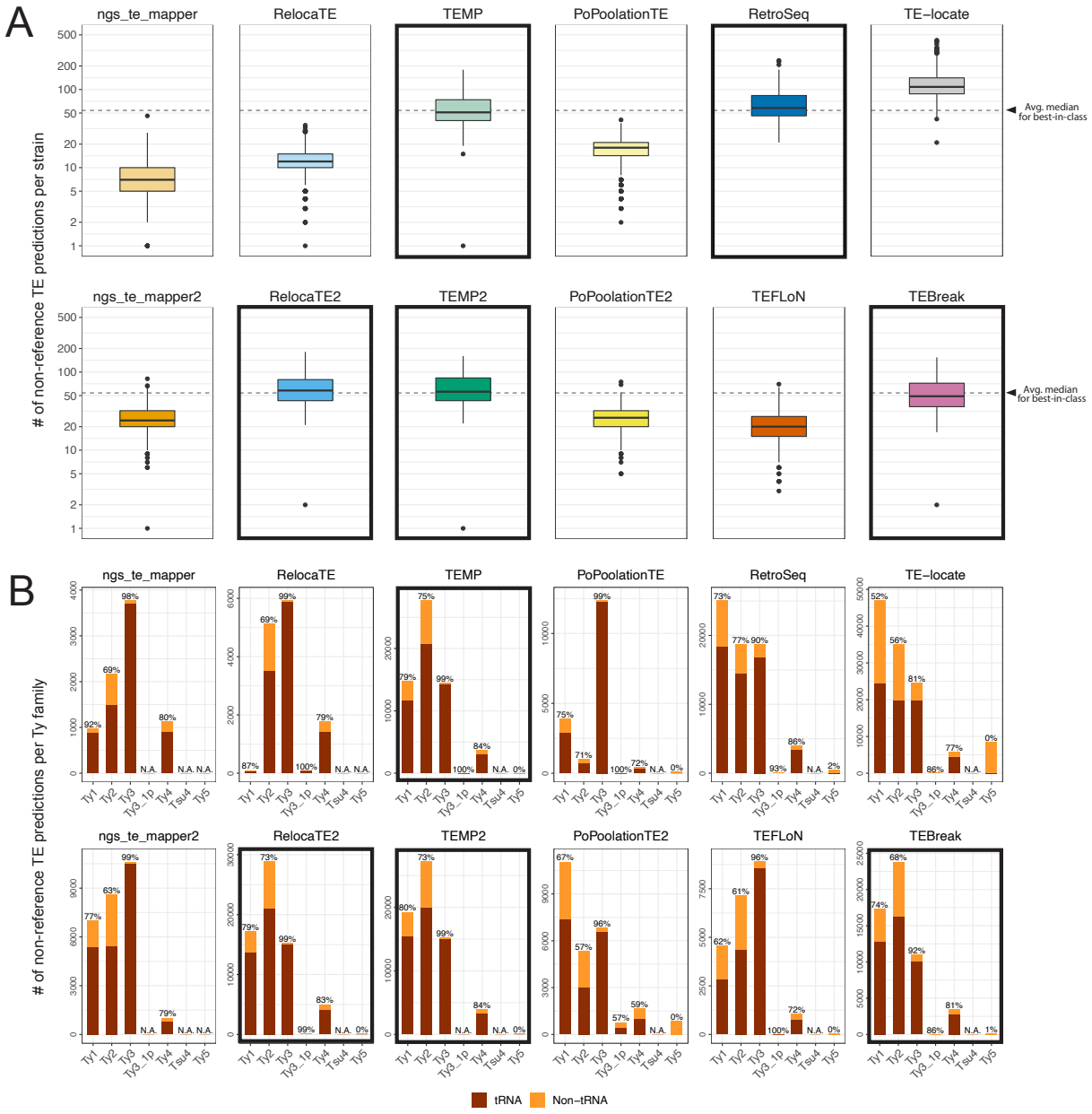

Figure 2.15: (Description see the next page.) **Numbers of Ty elements predicted by McClintock 2 components in a world-wide sample of yeast strains.**

Figure 2.15: (Figure see the previous page.) **Numbers of Ty elements predicted by McClintock 2 components in a world-wide sample of yeast strains.** (A) Numbers of non-reference TE predictions per strain (summed over all Ty families) and (B) numbers of non-reference TE predictions across Ty families (summed over all strains) in 1,011 *S. cerevisiae* WGS samples [261, 222], down-sampled to 50× fold-coverage. In panel (A), lines inside boxes indicate median values, colored boxes show interquartile ranges (IQR), whiskers show values 1.5×IQR of the upper or lower quartiles, and the dots indicate outliers that beyond 1.5×IQR. Components with bold outlines in panel (A) have have median values of ∼50 non-reference Ty insertions per strain, as well as recall and precision both >75% in tRNA promoter insertion simulations when allowing non-exact predictions in WGS datasets with >50× coverage (see Fig. 2.14). We note that the y-axis is on a $\log_{10}$ scale, and that 16 zero-count data points and one extreme TE-locate data point (count=749) is removed to aid with visualization. In panel (B) total numbers of non-reference TE predictions are partitioned as "tRNA" (dark red) if they are located between 1000 bp upstream and 500 bp downstream of tRNA genes, or "non-tRNA" (orange) if outside these windows. Note that the y-scale varies for each component method. The percentage of near tRNA gene predictions is annotated at the top of each bar. "N.A." means no such Ty family was found using that component. Components with bold outlines in panel (B) predict consistent relative TE family abundance and also have properties of components with bold outlines in panel (A), and thus we designate them as "best-in-class" methods for predicting non-reference TE insertions in *S. cerevisiae*. Dashed lines in panel (A) represent the average of the median number of non-reference TE insertions across the four best-in-class methods (n=54).

elements from solo LTRs, the estimate of ∼50 non-reference Ty insertions per yeast strain is likely to represent a combination of both structural types. Indeed, applying the McClintock 2 coverage module to internal coding regions of all Ty families, we estimate that the number of full-length elements per strain is ∼20 (Fig. 2.17), suggesting that many non-reference Ty insertions detected by McClintock 2 components in this population sample are polymorphic solo LTRs [48, 29].

Next, we assessed the consistency among component methods to classify the TE family of non-reference predictions across all 1,011 isolates (Fig. 2.15B). As for estimates of overall non-reference Ty abundance per strain, we observe variable patterns in relative Ty family abundance across component methods, with most methods being able to differentiate Ty

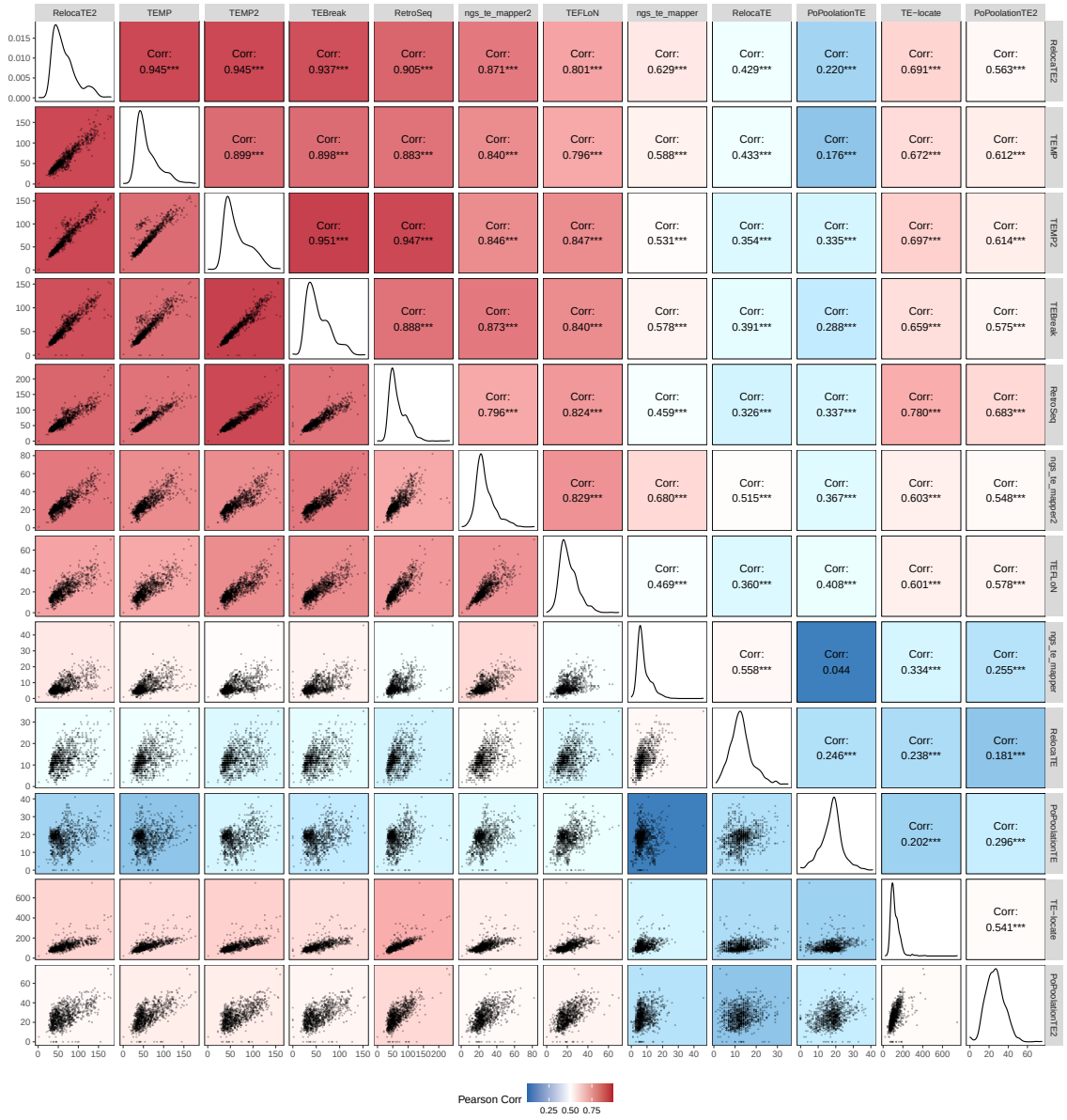

Figure 2.16: **Correlation across 12 component methods in all non-reference predictions for 1,011 yeast isolates.** The number of non-reference predictions for each yeast sample was plotted and compared across 12 component methods. WGS data are from [261, 222] and down-sampled to 50× fold-coverage. The panels in lower-left triangle are scatter plots on total TE count of each isolate between the corresponding pairs of methods. Panels in the upper-right triangle reveal the Pearson correlation between pairs of methods. Background colors for each panel are selected related to the Pearson correlation statistic, where higher and lower correlations are plotted in red and blue, respectively. Diagonal panels show the distribution of TE counts made by each component method. The component methods are ordered according to the general performance revealed in our simulation results.

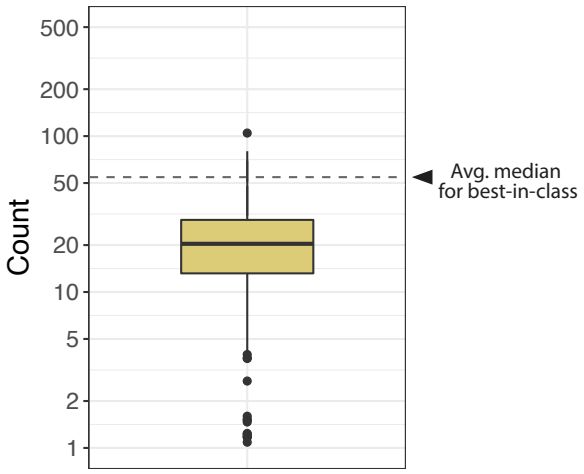

Figure 2.17: **Distribution of estimated copy number of full-length TEs across 1,011 yeast isolates.** Shown is the distribution of copy number of full-length TEs per strain (summed over all Ty families) estimated by the McClintock 2 coverage module using internal coding regions of all Ty families. Estimated full-length TE copy number includes both reference and non-reference TEs, but does not include solo LTRs. All WGS samples from [261, 222] were down-sampled to 50× fold-coverage. The y-axis is transferred to a $\log_{10}$ scale. The line inside the box indicates the median value, the colored box shows the interquartile range (IQR), whiskers show values 1.5×IQR of the upper or lower quartiles, and the dots indicate outliers that beyond 1.5×IQR. The dashed line indicates the average of median numbers non-reference Ty insertions per strain predicted by the four "best-in-class" component methods (RelocaTE2, TEMP, TEMP2 and TEBreak) (see Main Text for details).

families that are known to be active (Ty1, Ty2, Ty3 and Ty4) from those that are inactive or low-abundance (Ty3_1p, TSU4, Ty5). As observed for numbers of non-reference TEs per strain above, we note that no clear pattern of Ty family abundance can be discerned across the original six components (Fig. 2.15B; top row). However, a consistent pattern of relative Ty family abundance is observed for four of the five component methods (RelocaTE2, TEMP, TEMP2 and TEBreak) that show both consistent non-reference TE abundance per strain and high performance in simulated data. These four methods all indicate that Ty2 has the highest overall number of non-reference Ty insertions in this sample of *S. cerevisiae* strains. While Ty1 is often cited as the most abundant Ty family in *S. cerevisiae* because of its high copy number in the reference strain S288c [143, 48], the finding that Ty2 has the highest number of non-reference insertions in this diverse worldwide sample of *S. cerevisiae* strains is supported by orthogonal copy-number estimates from the McClintock 2 coverage module (Fig. 2.18) and a similar depth-based approach used in recent independent study [28]. In contrast, the fifth component method that shows consistent Ty abundance per strain and high performance in simulated data – RetroSeq – predicts more Ty1 and Ty2 insertions in empirical yeast data, which we infer to be misidentification because of the similarity in LTR sequences for these two Ty families [237, 143].

Previously [197], we used the well-established pattern that Ty1, Ty2, Ty3 and Ty4 non-randomly target promoters of genes transcribe by RNA polymerase III such as tRNAs (reviewed in [35]) to validate non-reference TE predictions made by McClintock 1 component methods in small sample of 93 *S. cerevisiae* genomes [197, 261]. Here, we confirm that the majority of non-reference predictions made by all component methods in McClintock 2 for these four Ty families are in the expected vicinity of tRNA genes in the species-wide *S. cerevisiae* resequencing dataset analyzed here (Fig. 2.15B) [261, 222]. Moreover, the four methods (RelocaTE2, TEMP, TEMP2 and TEBreak) that consistently predict similar numbers and families of non-reference Ty insertions in empirical yeast genome data, also predict similar proportions of insertions in tRNA genes. Combined with their high performance in simulated

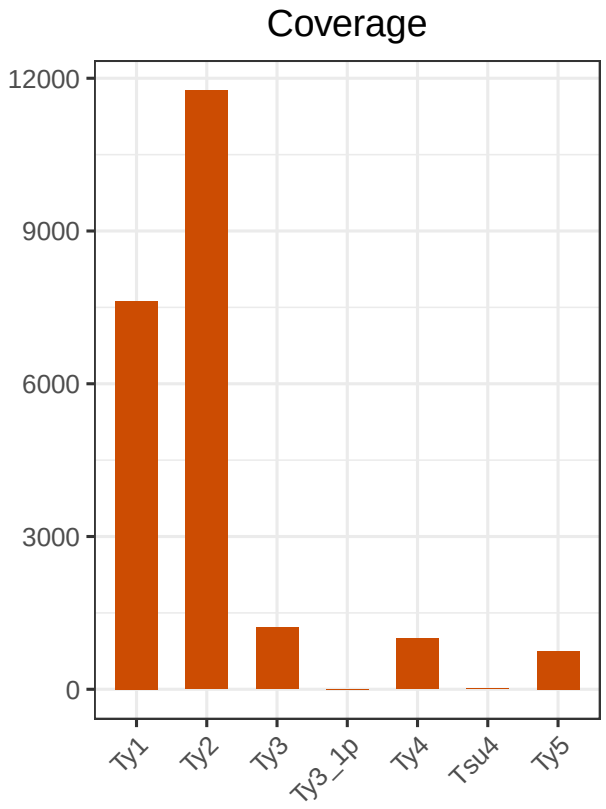

Figure 2.18: **Estimated total number of full-length TEs per Ty family across 1,011 *S. cerevisiae* isolates.** Shown is the total copy number of full-length TEs (summed across all strains) estimated by the McClintock 2 coverage module using internal coding regions of all Ty families. Estimated full-length TE copy number includes both reference and non-reference TEs, but does not include solo LTRs. All WGS samples from [261, 222] were down-sampled to 50× fold-coverage.

data, we conclude that RelocaTE2, TEMP, TEMP2 and TEBreak are the "best-in-class" methods for predicting Ty insertions in *S. cerevisiae* WGS resequencing data.

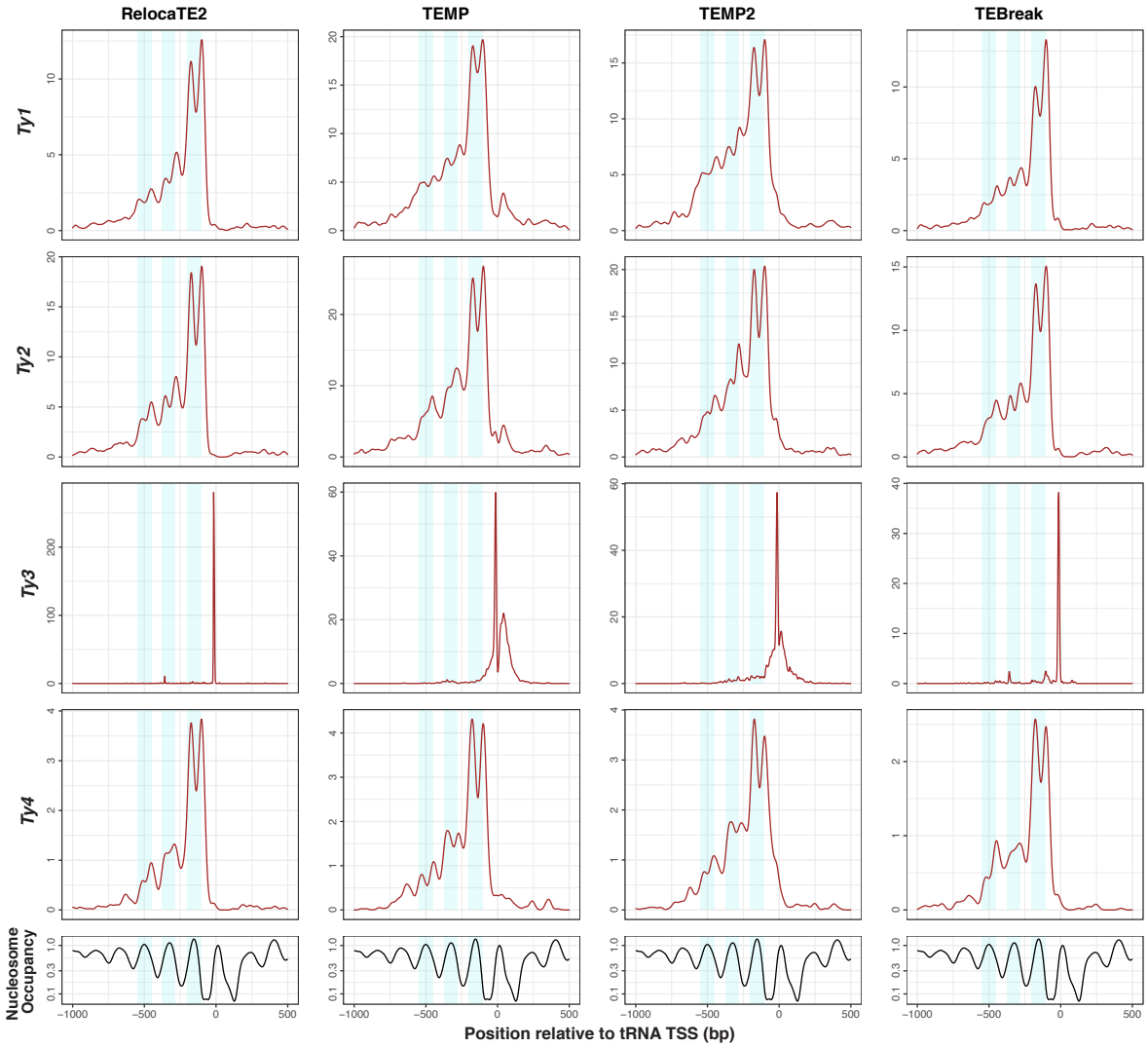

Figure 2.19: ***copia*-superfamily retrotransposons show a dyad pattern of insertion in nucleosome-bound regions upstream of yeast tRNA genes.** The top four rows show density profiles of non-redundant insertion sites for non-reference Ty predictions made by best-in-class McClintock 2 components (RelocaTE2, TEMP, TEMP2 and TEBreak) in tRNA promoter regions in a panel of 1,011 *S. cerevisiae* WGS samples [261, 222], down-sampled to 50× fold-coverage. Only the four Ty familes (Ty1, Ty2, Ty3 and Ty4) that are know to non-randomly target tRNA genes are included in this analysis. The bottom row shows nucleosome occupancy inferred using MNase-seq data from [208]. Light blue shaded areas indicate 100-bp regions surrounding peaks of nucleosome occupancy.

Finally, we sought to test whether predictions in species-wide *S. cerevisiae* WGS data made by the best-in-class components in McClintock 2 have sufficient positional accuracy to recapitulate the association between Ty1 insertion profiles and nucleosome occupancy in tRNA promoters observed for experimentally-induced insertions [11, 186, 42]. Recent work by Hays *et al.* [115] using WGS data from experimentally-evolved genomes has shown that Ty insertions predicted by RelocaTE2, aggregated across all Ty families, exhibit periodicity upstream of tRNA genes in *S. cerevisiae* [115]. However, their analysis does not partition insertions to individual Ty families, show a direct correlation with nucleosome profiles, or demonstrate that this signal is not an artifact of a single TE detection system. As shown in Fig. 2.19, spontaneous Ty1 insertions predicted in species-wide *S. cerevisiae* resequencing data by the four "best-in-class" component methods show a clear profile of dyad peaks on the first nucleosome-bound regions upstream of tRNA TSS with weaker dyad signals on the second and third upstream nucleosome-bound regions, as observed for experimentally-induced Ty1 insertions [11, 186, 42]. Intriguingly, spontaneous Ty2 and Ty4 insertions predicted by all four best-in-class methods (and by RetroSeq, ngs_te_mapper2 and TEFLoN; Figs. 2.20 and 2.21) also have dyad peaks on the first nucleosome-bound region upstream of *S. cerevisiae* tRNA genes. These results are consistent with the observation that the domain of Ty1 integrase shown to be responsible for targeting nucleosomes upstream of tRNA genes is conserved in Ty2 and Ty4 [8, 202]. In contrast, spontaneous Ty3 insertions show no correlation with nucleosomal profiles and for RelocaTE2 and TEBreak are restricted to ∼15 bp upstream of tRNA TSSs as seen for experimentally-induced insertions [214]. Profiles of spontaneous Ty3 insertions predicted by TEMP and TEMP2 are more variable than the other best-in-class methods, which can be explained by lower positional accuracy for a proportion of Ty3 predictions made by these methods (see Fig 2.4 and 2.6). A proportion of lower resolution predictions for TEMP may also explain the additional peak of insertions for Ty1 and Ty2 downstream of the tRNA TSS, which is not observed for spontaneous insertions predicted by RelocaTE2, TEMP2, or TEBreak or in experimentally-induced Ty1 insertions

[11, 186, 42]. Overall, these results demonstrate that the insertion profiles for Ty1 and Ty3 upstream of tRNA genes are not artifacts of experimental induction, and suggest that the molecular mechanisms responsible for the targeting of Ty1 to nucleosomes are conserved in the related *copia*-superfamily retrotransposons Ty2 and Ty4.

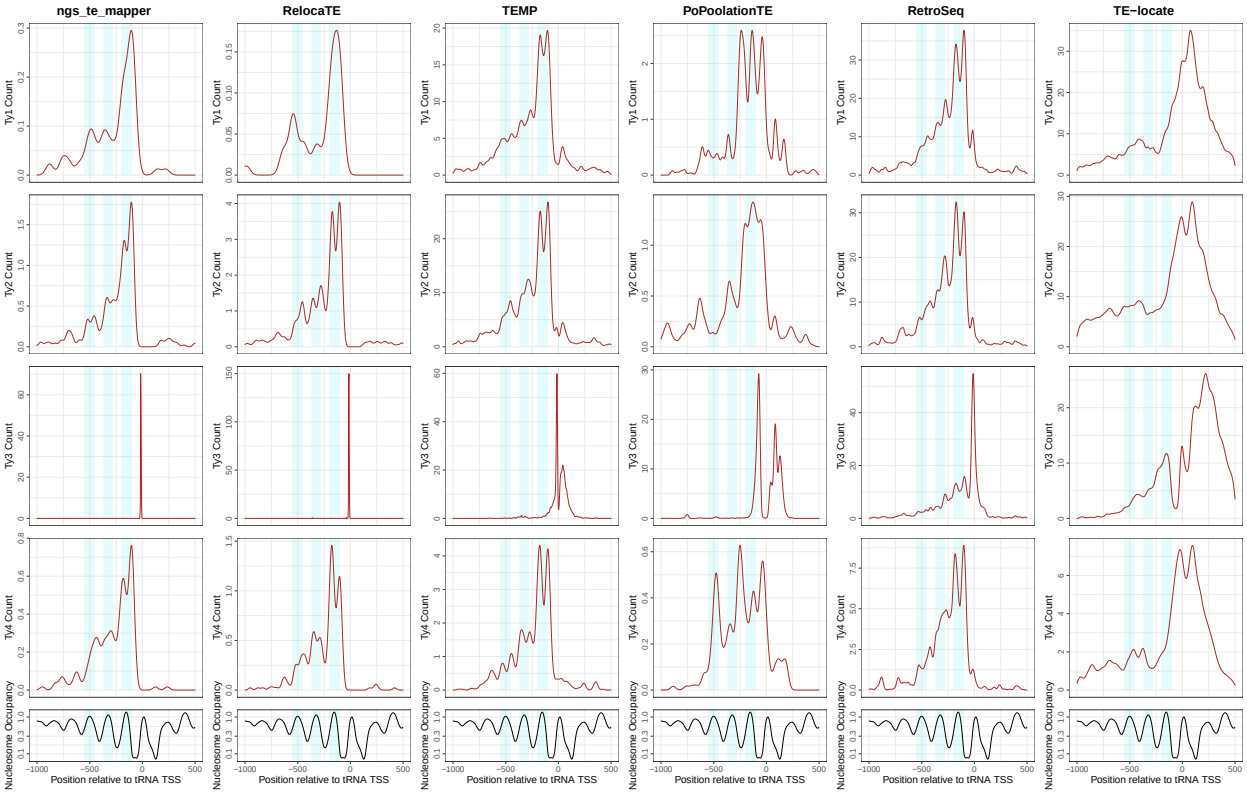

Figure 2.20: **Predicted TE insertion profiles around tRNA genes in *S. cerevisiae* for the six original McClintock component methods.** The first four rows show non-reference TE predictions made by six original methods in McClintock [197] from 1000 bp upstream and 500 bp downstream of tRNA genes partitioned by Ty family (summed over all strains) in 1,011 *S. cerevisiae* WGS samples [261, 222], down-sampled to 50× fold-coverage. Only the four active Ty families are included in this analysis. The bottom row shows nucleosome occupancy inferred using MNase-seq data from [208]. Light blue shading indicates 100-bp regions surrounding peaks of nucleosome occupancy.

2.6 Conclusions

In this report we present a re-implemented version of the McClintock TE detection system with improved installation, error handling, speed, stability, and extensibility. Relative to the

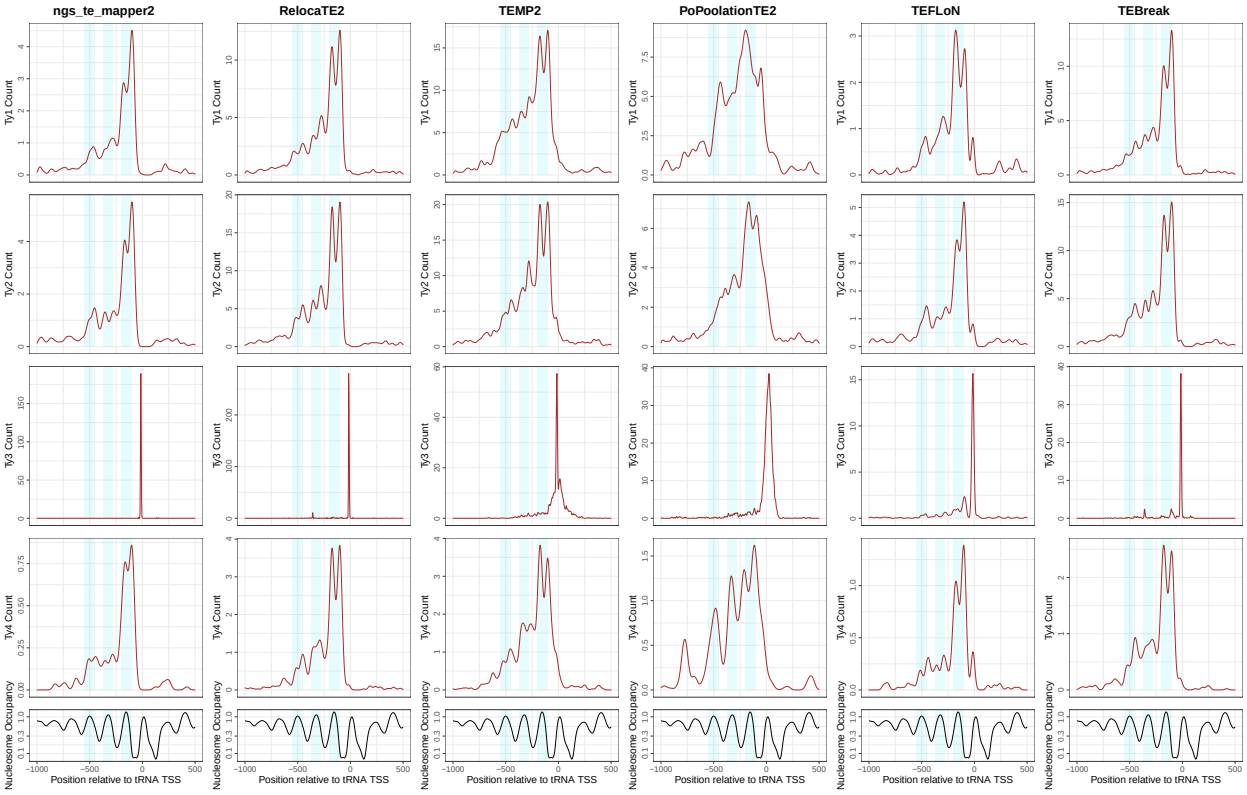

Figure 2.21: **Predicted TE insertion profiles around tRNA genes in *S. cerevisiae* for the six new McClintock component methods.** Plotting criteria are the same as in Figure 2.20.

original system [197], McClintock 2 includes six new short-read TE detectors, pre-processing and analysis modules, interactive HTML reports, and a reproducible simulation system that we used here to cross-validate the re-implemented system and evaluate component TE detectors. We acknowledge that not all available short-read TE detectors are currently incorporated into the McClintock 2 meta-pipeline, however use of Conda and Snakemake in the re-implemented meta-pipeline puts us in a better position to integrate additional or improved components in the future. Other potential improvements include creating a Bioconda package for McClintock 2 itself, which could aid installation and enhance portability by the automatic construction of Docker/Singularity containers containing McClintock 2, its component methods, and all relevant dependencies [107]. We are also considering modifying McClintock 2 to allow installation of TE detectors from user-downloaded executables or source code to be able to integrate high-performing methods like MELT [94] which have restrictive licenses that prevent automatic installation.

A major improvement in McClintock 2 is the availability of a reproducible simulation framework that allows evaluation of component method performance in a wide range of organismal contexts besides *S. cerevisiae*, including humans and other model organisms. In the future, we plan to modify the simulation framework to allow multiple insertions in a single synthetic genome to reduce the number of replicates needed to analyze larger genomes that have longer run times. We also plan to implement a complementary evaluation system that uses TE annotations in long-read assemblies as gold standards to empirically benchmark TE detectors using short-read WGS data from the same strain [269, 289].

Based on performance evaluations in simulated WGS data, we conclude that RelocaTE2 is the best McClintock 2 component to detect non-reference TE insertions in *S. cerevisiae* at base-pair accuracy, which supports the use of this method in recent yeast experimental evolution studies [55, 115]. However, we caution against the general conclusion that RelocaTE2 is the best McClintock 2 component for other species, since this method can have long run times that precludes its use in larger genomes [269, 289]. By combining results from simulated

data with consistency analyses in empirical WGS data, we also identify TEMP, TEMP2 and TEBreak as additional "best-in-class" methods for predicting Ty insertions in *S. cerevisiae* WGS resequencing data with 50× fold-coverage or higher, albeit with lower positional accuracy than RelocaTE2. Results from the four best-in-class McClintock 2 components for *S. cerevisiae* provide additional support to the emerging view that Ty2 is the most abundant TE in this species [28], although we acknowledge that non-independence of strains and non-random sampling in the strain panel analyzed [261, 222] may bias this conclusion. Finally, we show that best-in-class McClintock 2 components for *S. cerevisiae* generate reasonable non-reference TE predictions with sufficient resolution to reveal a dyad pattern of integration in nucleosome-bound regions upstream of yeast tRNA genes for Ty1, Ty2, and Ty4. This finding allows knowledge about fine-scale target preferences revealed in experimentally-induced Ty1 insertions [11, 186, 42] to be extended to natural Ty1 insertions and related *copia*-superfamily retrotransposons in yeast. Together with the new bioinformatics resources provided in McClintock 2, our work provides novel biological insights about TEs in one of the best-studied model systems and a paradigm for selecting optimal non-reference TE detectors in a diversity of organisms.

2.7 Data availability

The McClintock TE detection meta-pipeline and reproducible simulation system are available at: `https://github.com/bergmanlab/mcclintock/`. Output from McClintock runs on simulated and empirical yeast data can be found in Additional Files 2-5 at the *Mobile DNA* website. Scripts to generate results presented in this report are available in Additional File 6 at the *Mobile DNA* website.

2.8 Acknowledgements

We thank members of the Bergman and Garfinkel Labs for helpful discussions throughout the project; the BioConda community for contributing software recipes for software dependencies

used in this project; and the Georgia Advanced Computing Resource Center for computing resources and HPC support.

2.8.1 Abbreviations

BED: browser extensible data; CPU: central processing unit; FN: false negative; FP: false positive; HPC: high-performance computing; IQR: inter-quartile range; JSON: JavaScript object notation; QC: quality control; TE: transposable element; TP: true positive; TSD: target site duplication; TSS: transcription start site; VCF: variant call format; WGS: whole genome sequencing.

Chapter 3

High-quality genome assemblies of *S. paradoxus*[1] and *S. uvarum*[2]

[1]Jingxuan Chen, Holly McQueary, David W Hall, Peter Philippsen, David J Garfinkel, and Casey M Bergman, Genome Assembly of the Ty1-Less *Saccharomyces paradoxus* Strain DG1768, *Microbiology Resource Announcements*, 2022, 11 (1): e00868-21;

Reprinted here with permission of the publisher.

[2]Jingxuan Chen, David J Garfinkel, and Casey M Bergman, Long-Read Genome Assembly of *Saccharomyces uvarum* Strain CBS 7001, *Microbiology Resource Announcements*, 2022, 11 (1): e00972-21;

Reprinted here with permission of the publisher.

3.1 Abstract

Here, we report finished genome assemblies for the Ty1-less *Saccharomyces paradoxus* strain DG1768 (derivative of strain 337) based on PacBio and Illumina shotgun sequence data, and the *S. uvarum* strain CBS 7001 based on PacBio whole genome shotgun sequence data. We document the genetic alterations that make the *S. paradoxus* strain DG1768 a key resource for Ty1 mobility studies. Our *S. uvarum* CBS 7001 assembly provides an improved reference genome for an important yeast species in the *Saccharomyces sensu stricto* clade.

3.2 Genome Assembly of the Ty1-Less *Saccharomyces paradoxus* Strain DG1768.

Saccharomyces yeasts have played essential roles in understanding the biology of eukaryotic retrotransposons [47, 33, 63, 244]. Here we present a *de novo* genome assembly for *S. paradoxus* strain DG1768, a representative of a "Ty1-less" lineage frequently used to study Ty1 retrotransposition [282, 96, 27, 183, 97, 240, 64].

To create DG1768 [*MATα ura3-100 Ty1-155-Δura3(Scer)-101 his3-Δ200hisG*], a wild *Saccharomyces* collected in California in 1951 (UCDFST 51-186) was obtained by Peter Philippsen from the Ron Davis lab in ∼1980 (renamed 155). Strain 155 was selected to create a Ty1-less yeast because it had only one full-length Ty1 element (Ty1-155). Michael Ciriacy mated strain 155 (assumed at the time to be *S. cerevisiae*) with *S. cerevisiae* GRF18, then repeatedly backcrossed to strain 155 to create a nearly-isogenic strain (called 276) carrying *ho(Scer)*, Ty1-155, and a *ura3* mutation at the endogenous *URA3* locus. Homologous recombination was used to replace Ty1-155 in strain 276 with *S. cerevisiae URA3* plus ∼600 bp of vector sequence, followed by selection of a *ura3* mutation in the new heterologous *URA3* locus to create the $\rho-$ strain 337. Finally, *HIS3* was deleted (replaced with *Salmonella HisG*) and *S. cerevisiae* functional mitochondria were introduced *via* cytoduction to create strain DG1768 [96]. Subsequently, DG1768 was shown to be *S. paradoxus* [183].

To prepare DNA for sequencing, single colonies of strain DG1768 were inoculated in 7 ml of YPD liquid broth and cultured for ∼24 hours at 30°C. Table 3.1 provides details on DNA isolation, library preparation, sequencing, and read statistics for PacBio and Illumina datasets. Assembly was performed with HGAP 4 (smrtlink-release_5.1.0.26412) [56] using unfiltered PacBio reads. The PacBio-only assembly was polished five times with unfiltered PE150 Illumina data using Pilon (v1.24) [272]. Polished contigs were scaffolded with RagTag (v2.0.1) [5] using *S. paradoxus* YPS138 as reference [291]. Assembly statistics were calculated using QUAST (v5.0.2) [109] and BUSCO (v5.0.0, saccharomycetes_odb10) [254]. Ty elements were annotated as in [64].

Our DG1768 assembly is 11,932,051 bp in length (scaffold N50 of 917,222 bp) with an overall GC content of 38.43%. The assembly has 18 contigs in 17 scaffolds: 15 chromosomes and the mtDNA are completely assembled in scaffolds with one contig each; ChrXII is assembled in one scaffold with two contigs separated by a gap at the rDNA locus. 97.5% of *Saccharomycetes* BUSCOs are present and single copy. We confirmed that no full-length copies of Ty1 are present in the DG1768 genome. The scar of the Ty1-155 knock-out was localized to chrVI:292,634-294,347. The scar of the *HIS3* knockout was localized to chrXV:686,961-688,113. We also identified a segment of *S. cerevisiae* DNA spanning chrIV:1-135,828 caused by introduction of the *ho* allele, and confirmed that DG1768 mtDNA is from *S. cerevisiae*. The complete provenance and genome assembly for DG1768 presented here should provide a useful resource for future studies on Ty1 retrotransposon biology.

3.3 Long-Read Genome Assembly of *Saccharomyces uvarum* Strain CBS 7001.

S. uvarum is a globally-distributed cryotolerant yeast species that can be commonly extracted from low temperature fermentations and other natural substrates [230, 226, 201, 166, 4, 236]. Several efforts have been made to assemble the genome for a *S. uvarum* reference strain called CBS 7001 (also called MCYC 623) [140, 59, 246]. Draft assemblies based on low-coverage Sanger sequences were first reported in 2003 for MCYC 623 [140] and

a spore clone called 623-6c [59]. Later, Scannell *et al.* [246] generated a chromosome-level scaffolded assembly for CBS 7001 by combining public Sanger data with high-coverage Illumina short-read sequences. All previous CBS 7001 assemblies contain a large number of gaps (Table 3.2), which limit the comprehensive analysis of transposable element (TE) sequences [246, 215]. Here we generated a complete *de novo* genome assembly for strain CBS 7001 using PacBio long-read sequences to improve the *S. uvarum* reference genome and support analysis of TE sequences in this important yeast species.

	PacBio	Illumina
DNA isolation method	Wizard Genomic DNA Purification Kit (Promega)	YeaSTAR Genomic DNA Kit (Zymo Research)
Library preparation method	SMRTbell Express Template Prep Kit (>15 kb protocol)	Urich *et al.* [268] with no bisulfite conversion
Sequencing chemistry	Sequel Sequencing Kit v2.1	S4 Reagent Kit (300 cycles)
Sequencing instrument	PacBio Sequel II	NovaSeq 6000
Average read length (bp)	5,912	150
Number of reads	615,827	2,357,654

Table 3.1: **Methods and properties of the PacBio and Illumina whole genome sequencing datasets used to assemble and polish the DG1768 genome.**

Assembly	Accession number	Contigs	Scaffolds	Total Length (bp)	Contig N50 (bp)	Complete BUSCOs (%)	Full-length *Tsu4*
Kellis *et al.* [140]	GCA_000166995.1	1,098	1,098	11,475,890	4,795	88.3	0
Cliften *et al.* [59]	GCA_000167035.1	588	586	11,862,889	19,032	94.2	3
Scannell *et al.* [246]	N.A.[†]	363	16	11,502,913	18,172	97.1	0
This study	JAIPTR000000000	17	16	11,968,998	836,618	97.3	9

Table 3.2: **Statistics for *de novo* assemblies of CBS 7001**. Contigs within scaffold were defined by 10 consecutive N's. [†]The CBS 7001 genome assembly from Scannell *et al.* [246] is available at: `http://www.saccharomycessensustricto.org/`.

To prepare DNA for PacBio sequencing, a single colony of strain CBS 7001 (NCYC 2909) was inoculated in 7 mL of YPD liquid broth and cultured for ∼24 hours at 30°C. DNA was isolated using the Wizard Genomic DNA Purification Kit (Promega) and a PacBio library was prepared using the SMRTbell Express Template Prep Kit (>15 kb protocol). PacBio sequencing was performed with the Sequel II instrument (Sequencing Kit v2.1) yielding 655,409 reads (average length: 8,490 bp). HGAP 4 (smrtlink-release_5.1.0.26412) [56] was used for *de novo* assembly and the resulting PacBio contigs were polished five times with Pilon (v1.24) [272] using 33,377,822 51-bp Illumina GAII reads for CBS 7001 (SRR173084) from [246]. Polished contigs were scaffolded with RagTag (v2.0.1) [5] using as reference a version of CBS 7001 ultra-scaffolds from [246] that corrected a name swap for ChrX and ChrXII noted in [4]. Assembly statistics were calculated using "stats.sh" from BBMap (v38.90) [46] and BUSCO (v5.0.0, saccharomycetes_odb10) [254]. Yeast TEs were annotated as in [64].

Our CBS 7001 assembly is 11,968,998 bp in length (contig N50 of 836,618 bp) with an overall GC content of 40.18% (Table 3.2). 15/16 chromosomes are completely assembled in scaffolds with one contig each; ChrX is assembled in one scaffold with two contigs separated by a gap at the rDNA locus. 97.3% of *Saccharomycetes* BUSCOs are present and single copy. Our assembly was co-linear with all chromosome scaffolds in the Scannell *et al.* [246] assembly, except for one discrepancy at a sequencing gap in the Scannell *et al.* [246] assembly (ChrIII:219,401-219,500). Aligned PacBio reads aligned to the Scannell *et al.* [246] assembly did not span this gap, but did align continuously through the corresponding region in our assembly. Our assembly corrects this mis-assembly, the ChrX↔ChrXII name swap, and closes over 340 sequencing gaps in the Scannell *et al.* [246] assembly. Our complete assembly allows confirmation that *Tsu4* is the only TE with full-length copies in the CBS 7001 genome [34, 199, 169, 258]: we find nine full-length *Tsu4* copies in our assembly, tripling previous estimates [59, 20] (Table 3.2). The complete genome assembly for CBS 7001 presented here should provide a useful resource for future genomic studies in *S. uvarum* and other *Saccharomyces* species.

3.4 Data availability

PacBio data used to generate and Illumina data used to polish the DG1768 assembly are available under SRA project PRJNA748953. The DG1768 assembly was deposited at NCBI under GenBank accessions CP081969-CP081986. PacBio data used to generate the CBS 7001 assembly is available under SRA project PRJNA753102. The CBS 7001 assembly produced here was deposited at NCBI under accession JAIPTR000000000.

3.5 Acknowledgements

For *S. paradoxus* DG1768 genome announcement, we thank Michael Ciriacy, who performed all genetic steps essential for the Ty1-155 knock out experiment at the University of Düsseldorf and tragically died before seeing the fruition of his work. We thank Kyria Boundy-Mills (University of California-Davis) and Jasper Rine (University of California-Berkeley) for assistance determining the provenance of strain 155. We thank Marissa Howard, Noah Workman, and Magdy Alabady at the University of Georgia Genomics and Bioinformatics Core for assistance with PacBio library preparation and sequencing, Tyler Earp for assistance with Illumina library preparation, GeneWiz for Illumina sequencing, and the Georgia Advanced Computing Resource Center for computational resources. For *S. uvarum* CBS 7001 genome announcement, we thank Marissa Howard, Noah Workman, and Magdy Alabady at the University of Georgia Genomics and Bioinformatics Core for assistance with PacBio library preparation and sequencing, the Georgia Advanced Computing Resource Center for computational resources, and Caiti Smukowski Heil (North Carolina State University) for helpful comments. This work was funded by the University of Georgia Research Foundation (CMB) and NIH grant R01GM124216 (DJG and CMB).

Chapter 4

Evolution of the Ty4 retrotransposon family in *Saccharomyces* yeasts driven by horizontal TE transfer and recombination [1]

[1] Jingxuan Chen, David J Garfinkel and Casey M Bergman, Evolution of the Ty4 retrotransposon family in *Saccharomyces* yeasts;
To be submitted to *Genome Biology Evolution*.

4.1 Abstract

Horizontal transposon transfer (HTT) has been increasingly reported to play a significant role in the evolution of eukaryotic genomes. HTT events in *Saccharomyces* yeasts involving different Ty families have been previously reported, including a HTT that introduced a subfamily of the Ty4 family called Tsu4 into *S. paradoxus*. However, the distribution of the Ty4 subfamilies among different *Saccharomyces* species as well as the evolutionary history of this Ty family remain unclear. In this study, we conducted a comprehensive survey on the evolution of the Ty4 family in *Saccharomyces* yeasts. First, Ty4/Tsu4 subfamily content was investigated among multiple *Saccharomyces* species by performing a coverage-based copy number estimation approach on unassembled whole genome sequencing (WGS) datasets. Additionally, we annotated TE elements from a collection of whole genome assemblies (WGA) for *Saccharomyces* species, which allowed us to cross-validate the observations from the Ty4/Tsu4 short-read content scan, and generate Ty4/Tsu4 sequences for further phylogenetic analysis. We conclude that the Ty4 subfamily is ancestral to the *S. paradoxus*/*S. cerevisiae* species pair, whereas the Tsu4 subfamily is present in the ancestor of *S. uvarum*/*S. eubayanus*. We identified multiple HTT events involving the Ty4 family among *Saccharomyces* species: (i) a HTT that initially introduced the Tsu4 subfamily into American *S. paradoxus*, plus a possible second HTT of Tsu4 into South American *S. paradoxus*; (ii) two separate HTTs of Tsu4 from *S. paradoxus* to *S. cerevisiae*; (iii) a possible HTT that introduced a distinct Tsu4 sub-type into Holarctic *S. eubayanus*; and (iv) a HTT that introduced Tsu4 into the *S. mikatae*/*S. jurei* species pair. Notably, we first investigated Ty4/Tsu4 sequences from two new long-read genome assemblies of *S. mikatae* strains, which allowed us to identify a "recombinant" Tsu4 clade in *S. mikatae* generated through recombination between Ty4 and Tsu4 sequences. Our results together reveal a complex evolutionary history of the Ty4 family among *Saccharomyces* species, and highlight the active role of HTT and recombination events in shaping TE contents in yeasts.

4.2 Introduction

Transposable elements (TEs) are repetitive DNA sequences that can mobilize within host genomes, and have been found ubiquitous among eukaryotes. Typically, TEs can be vertically inherited from parents to offspring within a species [279], whereas the transmission of TEs between different species have also been identified through horizontal transposon transfer (HTT) [68]. With the increasing availability of whole genome sequencing data, HTT has been recognized as an important phenomenon in the evolution of eukaryotic genomes [248, 101, 273, 274]. HTT events are identified using a variety of evolutionary evidence including: phylogenetic incongruence between host genomes and TE sequences, unexpectedly high sequence similarity between TEs from divergent species, or patchy distribution of a TE family across a set of closely-related lineages [273]. Based on these criteria, an increasing number of HTT events have been identified across the tree of life, mainly in plants and animals [73]. Although the number of HTT events detected in fungi is still relatively low [71, 65, 204, 7, 245], a growing number of HTT events have been identified among *Saccharomyces* yeast species [64, 20, 48, 169, 105]. As in other taxa, previous studies in yeast have typically focused on detecting the existence of HTT events, however the time scale, geographic location, and specific donor/recipient lineages for these events often remain unknown. In recent years, the global population structure of several *Saccharomyces* species – including *S. paradoxus* [160, 116], *S. cerevisiae* [222], *S. uvarum* [4], *S. eubayanus* [217], and *S. mikatae* [220] – has been revealed through short-read sequencing of large population samples. Moreover, a growing number of high-quality genome assemblies have been generated using third-generation long-read sequencing in the *Saccharomyces* genus [291] that allow comprehensive analysis of TE sequences [64]. Together, these resources make *Saccharomyces* yeasts an excellent model system to provide a high-resolution understanding of how HTT events occur in nature.

A promising model for studying the process of HTT events among *Saccharomyces* yeasts involves the long-terminal repeat (LTR) retrotransposon Ty4. The Ty4 family is a member

of the Ty1/Copia superfamily first identified from *S. cerevisiae* [264], and contains elements with two overlapping ORFs (*Gag* and *Pol*) flanked by LTRs [130, 263]. Neuvéglise *et al.* [199] proposed that a Ty4-like element was acquired in an ancestor of *S. cerevisiae* and *S. uvarum* (and was therefore present in the common ancestor of all *Saccharomyces sensus stricto* yeasts), which diverged into the Ty4 subfamily in part of the *Saccharomyces* genus that includes *S. cerevisiae*, and into the Tsu4 subfamily in another branch of the *Saccharomyces* genus that includes *S. uvarum*. More recently, Bergman [20] reported the unexpected presence of the Tsu4 subfamily in *S. paradoxus* and *S. cerevisiae* and presented evidence for multiple HTT events involving the Tsu4 subfamily. The strongest evidence for Tsu4 HTT was found in *S. paradoxus* based on (i) its patchy distribution among global *S. paradoxus* lineages, (ii) an unexpectedly high similarity between Tsu4 elements from *S. paradoxus* and those of its distant congeners (*S. uvarum* and *S. eubayanus*), and (iii) discordance between the Tsu4 phylogeny and host tree of *Saccharomyces* sensus stricto species [20]. Bergman [20] also reported a possible second Tsu4 HTT event involving *S. cerevisiae* based on the presence of a single full-length Tsu4 element in one *S. cerevisiae* strain (called 245) and high sequence similarity between this *S. cerevisiae* Tsu4 element and the Tsu4 sequences recently introduced into *S. paradoxus*. Subsequently, O'donnell *et al.* [212] confirmed the presence of Tsu4 in a different strain of *S. cerevisiae* (called CQS) and suggested that Tsu4 was active in *S. cerevisiae* based on the presence of multiple full-length elements in this strain. Bergman [20] also provided evidence for a possible third Tsu4 HTT event in *S. mikatae* based on phylogenetic discordance of a single *S. mikatae* Tsu4 element relative to the accepted host species tree. Bergman [20] proposed a scenario whereby either *S. uvarum*, *S. eubayanus*, or another related species first donated Tsu4 to an ancestral lineage of *S. paradoxus* most likely in the Americas, with this primary HTT event followed by a secondary HTT event from *S. paradoxus* to *S. cerevisiae*. No evidence was provided to connect the potential HTT event in *S. mikatae* to those in *S. paradoxus* or *S. cerevisiae*.

As with most cases of HTT, many aspects of these Tsu4 HTT events in *Saccharomyces* species are currently unresolved because of limitations in the quantity and quality of sequence data available at the time of the initial study by Bergman [20]. First, the *S. paradoxus* and *S. cerevisiae* genomes previously studied did not sample the global diversity in these species. Second, evidence that the primary HTT into *S. paradoxus* could have come from *S. eubayanus* was made indirectly using genome data from the hybrid yeast species *S. pastorianus*, which contains subgenomes from *S. eubayanus* and *S. cerevisiae* [166, 10, 210]. Third, data for many strains analyzed were based on fragmented whole genome assemblies (WGAs) made using first generation sequencing data that often have incompletely reconstructed TE sequences [188]. As a consequence, many questions related to the Tsu4 HTT events in *Saccharomyces* remain unanswered, but can now be addressed using improved genome data.

In this work, we investigated the history of the Ty4 family in *Saccharomyces sensu stricto* species as a model to understand the process of HTT in microbial eukaryotes. Firstly, we comprehensively surveyed the Ty4/Tsu4 subfamilies content among multiple well-studied *Saccharomyces* species using short-read WGS datasets, which allowed us to establish the ancestral state (i.e., presence or absence) of Ty4/Tsu4 subfamilies in *S. paradoxus*, *S. cerevisiae*, *S. uvarum*, *S. eubayanus*, and *S. mikatae*. Next, we compiled available long-read genome assemblies for multiple *Saccharomyces* species (including *S. paradoxus*, *S. cerevisiae*, *S. uvarum*, *S. eubayanus*, and *S. jurei*), generated long-read assemblies of two *S. mikatae* strains (IFO 1815 and NBRC 10994) for the first time, and then annotated TE content from genomes. The result of TE annotation in WGAs cross-validated our observations on the presence/absence of Ty4/Tsu4 subfamilies and provided a rich source of full-length elements for phylogenetic analysis. Third, we reconstructed maximum-likelihood phylogenetic trees for the internal coding sequences from full-length Ty4/Tsu4 copies annotated from whole genome assemblies, which revealed a complex history of HTT events involving the Ty4 family among the *Saccharomyces* genus. Together these results add to an increasingly complex model for Ty family evolution in yeast that is shaped by host population structure, recurrent HTT

events and extinction events, and creation of new subfamilies by recombination among Ty elements which co-exist because of HTT.

4.3 Methods and Materials

4.3.1 Genome sequence datasets

We compiled a collection of public short-read paired-end WGS datasets of multiple *Saccharomyces* species to survey Ty4/Tsu4 subfamily content across the genus. The compiled collection of WGS datasets for each *Saccharomyces* species were cleaned and normalized using the following steps: (i) Raw reads with identical BioSample accession, strain name and sequencing strategy (i.e, sequencing instrument and layout) were merged according to the metatable provided by NCBI SRA; (ii) For the NCBI BioSample accessions that have multiple records after merging, only the record with the highest sequencing depth was retained; and (iii) All samples with sequencing depth <10× were removed. After these quality control processes, the integrated WGS dataset used in this study includes 371 *S. paradoxus* isolates [23, 160, 284, 220, 80, 152, 116], 2,404 *S. cerevisiae* isolates [222, 3, 14, 15, 16, 23, 25, 60, 74, 75, 86, 93, 99, 103, 110, 129, 136, 145, 162, 176, 180, 231, 256, 259, 261, 291, 294], 62 *S. uvarum* isolates [220, 4, 246, 175], 292 *S. eubayanus* isolates [220, 19, 43, 157, 179, 182, 198, 217, 242], and 18 *S. mikatae* isolates [220].

In addition, we complied public high-quality genome assemblies of *Saccharomyces* species generated mostly with long-read sequencing data (PacBio or ONT) plus a few WGAs for *S. cerevisiae* generated with short-read data that had prior evidence of having full-length Tsu4 elements. In total, we analyzed 12 *S. paradoxus* assemblies [80, 291, 54], 210 *S. cerevisiae* assemblies [291, 212, 180, 222], 1 *S. uvarum* assembly from [53], 4 *S. eubayanus* assemblies [43, 179, 40], 2 *S. jurei* assemblies from [190], and 2 new *S. mikatae* assemblies generated here (see Section 4.3.2). Note that we included three *S. cerevisiae* assemblies generated with short-read sequencing data that showed evidence of Tsu4 internal regions (strain 245 from [180]; strains AFQ and CDM from [222]).

The Reference genomes used in this study are as follows: *S. cerevisiae* laboratory strain S288c (UCSC version sacCer2); *S. paradoxus* European strain CBS432 (GCA_002079055.1) [291]; *S. uvarum* European strain CBS 7001 (GCA_019953615.1) [53]; *S. eubayanus* Patagonia strain FM 1318 (CBS 12357^T, GCA_001298625.1) [10]; *S. mikatae* Japanese strain IFO 1815 (JARBHO000000000) generated in this study (see Section 4.3.2).

4.3.2 DNA preparation, PacBio sequencing and genome assembly of *S. mikatae* strains IFO 1815 and NBRC 10994

To prepare DNA for PacBio sequencing, single colonies of strain IFO 1815 (NCYC 2888) and NBRC 10994 was inoculated in 7 ml yeast extract-peptone-dextrose (YPD) liquid broth and cultured for ∼24 hours at 30°C. DNA was isolated using the Wizard genomic DNA purification kit (Promega), and a PacBio library was prepared using the SMRTbell Express template prep kit (Pacific Biosciences), following the >15-kb size-selection protocol that includes Covaris g-TUBE shearing. PacBio sequencing was performed using the Sequel II instrument (sequencing kit v2.1). Draft genome assemblies were generated by performing Flye (v2.9) [148] with parameters "–pacbio-raw -g 12m." Raw PacBio reads and genome assemblies of two *S. mikatae* strains have been submitted to NCBI under BioProject PRJNA934353.

4.3.3 Ty4/Tsu4 copy number estimation

Copy number of Ty elements was estimated using the coverage module of McClintock 2 (see details in Chapter 2 [52]) using public short-read WGS datasets compiled above. For this analysis, we used McClintock revision 7aa5298 with parameters "--keep_intermediate minimal,coverage -m coverage" [52]. Input files used to perform McClintock coverage module includes: (i) the reference genome for each *Saccharomyces* species as stated Section 4.3.1; (ii) a library of yeast TE query sequences (`https://github.com/bergmanlab/mcclintock/blob/master/test/sac_cer_TE_seqs.fasta`); and (iii) TE query library split into LTR and internal regions derived from [48], which includes additional sequences for the Tsu4

subfamily. The edge-trimming option in the McClintock coverage module was disabled by specifying "omit_edges" as "False" in the configuration file "config/coverage/coverage.py". To reduce the influence of variable coverage and computing resources, samples with original fold-coverage greater than 100× were down-sampled to 100× using seqtk (v1.3) [163].

4.3.4 Phylogenetic analysis of host species

For each species, multi-sample variant calling was performed with BCFtools (v1.16, "bcftools mpileup -a 'FORMAT/DP' -Q 20 -q 20"; "bcftools call -f GQ,GP -mv –skip-variants indels") using BAM files generated by McClintock [165]. Alignments with mapping quality less than 20, or bases with quality score less than 20, were removed. All indels were excluded from variant calling. Subsequently, the SNP matrix was filtered with BCFtools filter (v1.16) to discard sites with polymorphic probabilities under 99%; or genotypes with average supporting read depth less than 10×. Vcf2phylip (revision 0eb1b80) [211] was executed to create multi-sequence alignment from the filtered VCF file. For all *Saccharomyces* species other than *S. cerevisiae* (which contains over 2,000 strains), maximum likelihood (ML) phylogenetic analysis was performed with RAxML (v8.2.12, "-f a -x 23333 -p 2333 –no-bfgs") [260] applying GTRGAMMA model and 100 times of bootstrap resampling. R packages SNPRelate (v1.32.0) [293] and ape (v5.7_1) [213] were executed to perform neighbor joining (NJ) phylogenetic analysis of host species. For *S. cerevisiae*, we followed the workflow in Peter *et al.* [222]. R functions "snpgdsVCF2GDS" and "snpgdsDiss" were used to create the distance matrix from SNP data, and then "bionj" was used to reconstruct the NJ tree. The host species trees were midpoint rooted and then visualized in R (v4.2.2) using the ape (v5.7_1) [213] and ggtree (v3.6.0) [288] packages.

4.3.5 Annotation and sequence analysis of full-length TEs from complete genome assemblies

Ty elements were annotated from whole genome assemblies using a RepeatMasker-based pipeline previously described in Czaja *et al.* [64]. Multi-sequence alignments of annotated full-length elements were generated using MAFFT (v7.508, default parameters) [138]. Specific sub-regions of full-length elements in alignments were identified by aligning the "Tsu4p_nw"' sequence from a public database of canonical yeast transposons (https://github.com/bergmanlab/yeast-transposons) with the set of full-length elements, then selecting regions with seqkit subseq (v0.16.1) [252]. ML phylogenetic trees of internal coding regions were reconstructed using RAxML (v8.2.12, "-f a -x 23333 -p 2333 –no-bfgs") [260] with GTRGAMMA model and 100 bootstrap replicates. Phylogenetic network analysis was performed with SplitsTree4 (v4.15.1) [125] applying the "Uncorrect_P" model and "NeighborNet" method. Pairwise sequence divergence was calculated based on Kimura's 2-parameter substitution model with 50-bp sliding window size and 10-bp step size with R package spider (GitHub revision e93c5b4) [44] and phangorn (v2.11.1) [250] in R (v4.2.2).

4.4 Results and Discussion

4.4.1 Survey of Ty4/Tsu4 subfamilies in global populations of *Saccharomyces* species

The number of high-quality WGAs for *Saccharomyces* species (with the exception of *S. cerevisiae*) is limited due to the high cost of third-generation sequencing strategies (including PacBio and Oxford Nanopore). On the other hand, a large amount of short-read WGS datasets have been published for many *Saccharomyces* species in recent years which often represent global samples. To further broaden our understanding of the global distribution of Tsu4 and Ty4 subfamilies across *Saccharomyces* species, we analyzed unassembled WGS reads to estimate Ty subfamily abundance in the context of host species phylogenies. First, we

compiled large-scale Illumina paired-end datasets of yeast isolates that span the known global diversity in each species involoving *S. paradoxus*, *S. cerevisiae*, *S. uvarum*, *S. eubayanus*, and *S. mikatae* (datasets described in Section 4.3). Next, we used the McClintock coverage module to align WGS reads against a modified species-specific reference genome in which all TE sequences in the reference genome were hard masked (converted to N's) and representative "canonical" TE sequences for each known Ty family were attached as separate "chromosomes". Mapped reads were used by the McClintock coverage module to estimate Ty copy number based on read depth in canonical Ty sequences normalized to non-repetitive regions in the reference genome. We also used mapped reads in BAM files generated by McClintock to call variants for each strain, then reconstructed the phylogeny for each *Saccharomyces* species based on SNPs in non-repetitive regions of the host genome. Finally, we plotted the estimated copy number of Ty4/Tsu4 subfamilies on the reconstructed host phylogenies. Because recombination among LTRs within full-length elements frequently excises internal Ty sequences [135], we separately estimated copy number for LTR and internal regions, which allows us to interpret the presence of internal sequences as evidence of recent Ty insertion, and the presence of solo LTRs as evidence of past Ty activity in a strain. Using the above pipeline, we were able to conduct a comprehensive survey for the ancestral state and current activity of Ty4/Tsu4 subfamilies across well-studied *Saccharomyces* species including *S. paradoxus*, *S. cerevisiae*, *S. uvarum*, *S. eubayanus*, and *S. mikatae*. In the following subsections, we describe Ty4/Tsu4 subfamily content in the context of host biogeography for each of these *Saccharomyces* species.

S. paradoxus

Bergman [20] previously reported a HTT event that introduced the Tsu4 subfamily into *S. paradoxus* somewhere outside of Eurasia, based on the complete absence of Tsu4 sequences in two Eurasian *S. paradoxus* strains (CBS432 and N44) and the presence of Tsu4 LTR and internal sequences in three American strains (UWOPS91-917.1, YPS138, and UFRJ50816).

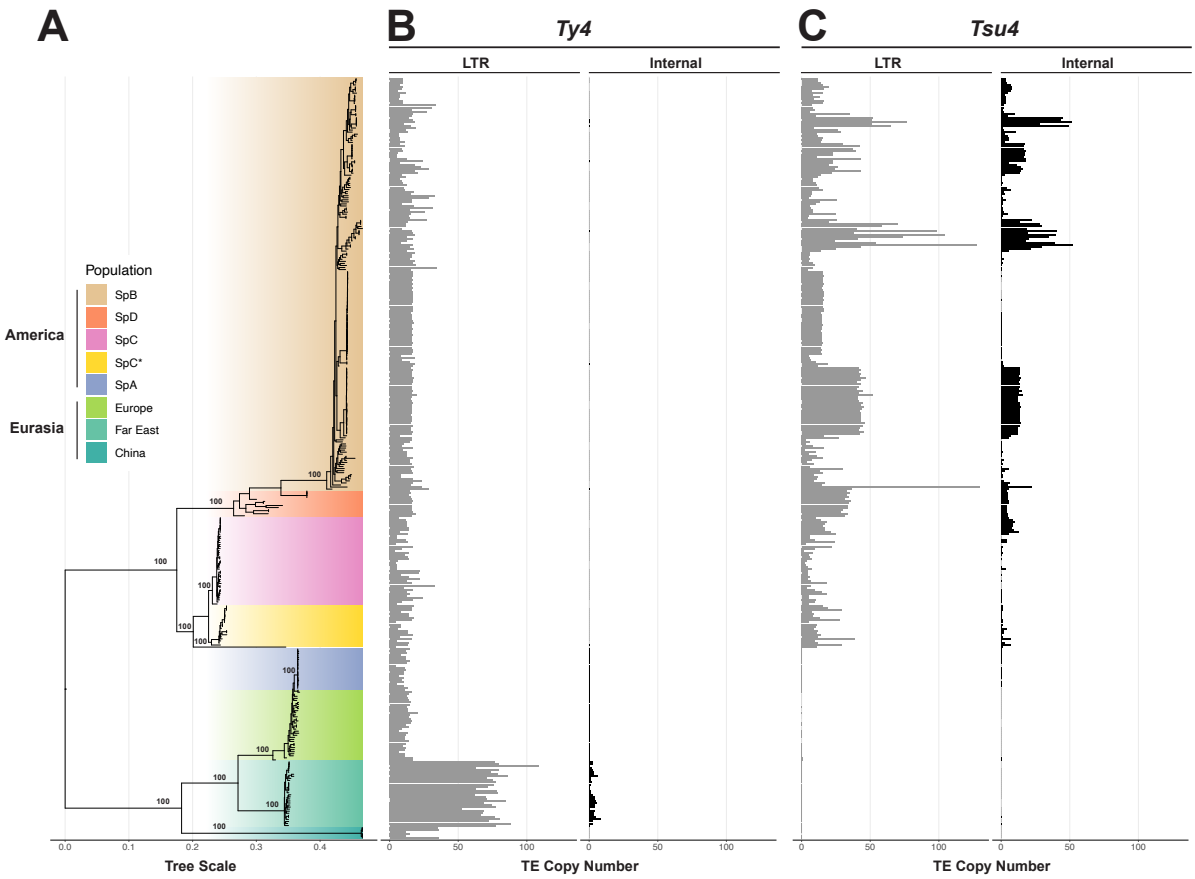

Figure 4.1: **Host phylogeny of *S. paradoxus* annotated with Ty4/Tsu4 copy number estimation.** Shown are copy number estimates for Tsu4 and Ty4 subfamilies from worldwide *S. paradoxus* populations. (A) shows the maximum likelihood phylogenetic tree of *S. paradoxus* host strains integrating several public WGS datasets (see Section 4.3 for details). The ML tree is reconstructed using 713,524 genome-wide SNPs from 371 isolates and midpoint rooted. Bootstrap support is annotated for key nodes. The population structure of *S. paradoxus* is colored according to [80, 116]. Copy number estimates for Ty4 and Tsu4 subfamilies are shown with bar plots in panel (B) and (C), respectively. Gray bars are copy number estimates for LTR elements, whereas black bars show copy number for internal coding regions.

Likewise, the Ty4 subfamily was inferred to be the ancestral representative of the Ty4 family in *S. paradoxus* based on its presence in the same set of five strains from Eurasia and America. These conclusions were based on a small sample size of *S. paradoxus* WGAs (n=5 [291]), and did not fully reveal the species-wide distribution of the Ty4/Tsu4 subfamilies in *S. paradoxus*.

The population structure of *S. paradoxus* has been previously studied in multiple publications [160, 80, 116]. Here, we integrated 371 *S. paradoxus* strains sequenced in previous studies and re-constructed a maximum-likelihood phylogenetic tree based on 713,524 genome-wide SNPs (Figure 4.1A), which confirmed the major known lineages and their relationships in this species [193, 151, 153, 172, 160, **?**, 80, 116]. Worldwide diversity in *S. paradoxus* splits into two major lineages: (i) the Eurasian lineage, which includes sub-lineages with multiple strains indigenous to Europe, Far East Asia, and China; and (ii) the American lineage, which includes several indigenous North American sub-lineages each with multiple strains (SpB, SpC, SpC* and SpD), as well a lineage with a single strain from Hawaii. SpC* and SpD are hybrid lineages derived from crosses between SpB and SpC, and between SpB and SpC*, respectively [160, **?**, 80]. The European lineage also includes a sub-lineage (SpA) composed of strains from North America that descend from a recent trans-oceanic migration event [153, 160]. SpA shows no evidence of admixture with indigenous American lineages after secondary contact [153, 126, 159, 160, 116]. The single known strain of *S. paradoxus* from Hawaii is a divergent lineage that has been reported to share similarity with either the SpB [159] or SpC/SpC* lineages [116, 220]. Our analysis places the Hawaiian lineage as an outgroup to SpC/SpC* lineages. Finally we note that *S. paradoxus* strains isolated from S. America, which were formerly considered a distinct species called *S. cariocanus* [192], are not a distinct lineage but rather are closely related to the North American SpB sub-lineage at the sequence level [151, 171, 172, 126, 159, 116].

Mapping estimated Ty content onto the global host strain tree for *S. paradoxus*, we observe that Ty4 LTRs are detected across all *S. paradoxus* strains from both Eurasia and

America (Figure 4.1B). Notably, strains from the Far East Asia show a significantly higher Ty4 LTR copy number in comparison to other populations. Ty4 internal regions are essentially absent across the species except for the LTR-rich Far East population (Figure 4.1B).

In contrast to the prevalent distribution of Ty4 subfamily in *S. paradoxus*, we find clear evidence that Tsu4 sequences are only found in indigenous American strains and never found in strains with a Eurasian origin (Figure 4.1C). Tsu4 LTRs are present in all indigenous American *S. paradoxus* strains. Internal Tsu4 sequences are found in all recognized American lineages (SpB, SpC, SpC*, SpD, and Hawaii) with variable copy numbers across strains including many strains where Tsu4 internal regions are absent (Figure 4.1C). Notably, we found no Tsu4 sequences in the Eurasian-derived North American SpA lineage, indicating that any potential admixture after secondary contact in North America between SpA and indigenous American strains has been too low to permit invasion of Tsu4 into the SpA lineage [153]. We found no evidence for the current co-existence of Ty4/Tsu4 internal regions in any *S. paradoxus* strain.

These data support the following parsimonious scenario to explain the distribution of Ty4/Tsu4 subfamilies in *S. paradoxus*. The Ty4 subfamily was present in the common ancestor of all *S. paradoxus* lineages prior to global dispersal and represents the original Ty4 subfamily in *S. paradoxus*. Ty4 went extinct in all recognized lineages except for Far East Asia, including the ancestor of all American lineages. In contrast, the complete lack of Tsu4 sequences in Eurasian *S. paradoxus* indicates this subfamily never existed in Eurasian lineages and therefore was not present in the common ancestor of all *S. paradoxus* strains. A HTT event led to Tsu4 being acquired in an ancestor of all American lineages (where the Ty4 subfamily had already likely gone extinct) after the divergence of American from Eurasian lineages but prior to dispersal throughout the Americas. Tsu4 is currently active in all recognized American lineages, but there are many isolates in which it has secondarily gone extinct. The widespread distribution of Tsu4 in the American lineages is compatible with Tsu4 HTT occurring anywhere in the American region, or prior to its expansion in

Americas. More complex scenarios are possible, such as acquisition of Tsu4 in a specific location after *S. paradoxus* dispersed in the Americas followed by transmission throughout the American region enabled by mating. Also, the possibility of multiple HTT events in the Americas cannot be ruled out using Ty content interpreted in a host phylogenetic context alone, however parsimony would still invoke at least one HTT event in the ancestor of all American lineages.

S. cerevisiae

Originally, only the Ty4 subfamily was found in the *S. cerevisiae* reference genome, and was regarded as the only active Ty4 family in *S. cerevisiae* and thus thought to be ancestral [143, 48]. More recently, Bergman [20] found evidence for the presence of the Tsu4 subfamily in *S. cerevisiae*, but in only one strain (called 245) from the West Indies by scanning a dataset of 336 shotgun assemblies [180]. Later on, O'donnell *et al.* [212] confirmed that Tsu4 was generally absent from a sample of 142 complete genomes selected from the collection of Peter *et al.* [222]. However, 9 complete Tsu4 copies were characterized in one strain (called CQS) of *S. cerevisiae* from French Guiana.

To confirm the ancestral state of Ty4 subfamily and potential invasion of Tsu4 subfamily into *S. cerevisiae* using a worldwide sample, we integrated WGS datasets containing 2,404 *S. cerevisiae* isolates from multiple publications (datasets see Section 4.3), and scanned Ty4/Tsu4 subfamily content using the same coverage-based strategy described above. The host phylogeny of 2,404 *S. cerevisiae* strains was reconstructed with neighbor-joining approach based on 2,787,384 genome-wide SNPs (Figure 4.2). Despite of a much larger collection of strains, our integrated host phylogenetic tree shows a similar topology as Peter *et al.* [222], who identified a complex population structure including more than 26 lineages plus many mosaic lineages arisen from admixture. Additional strains in our integrated dataset that are not present in Peter *et al.* [222] – such as Duan *et al.* and Barbosa *et al.* [75, 14] – cluster with known populations that have been characterized in Peter *et al.* [222]. For

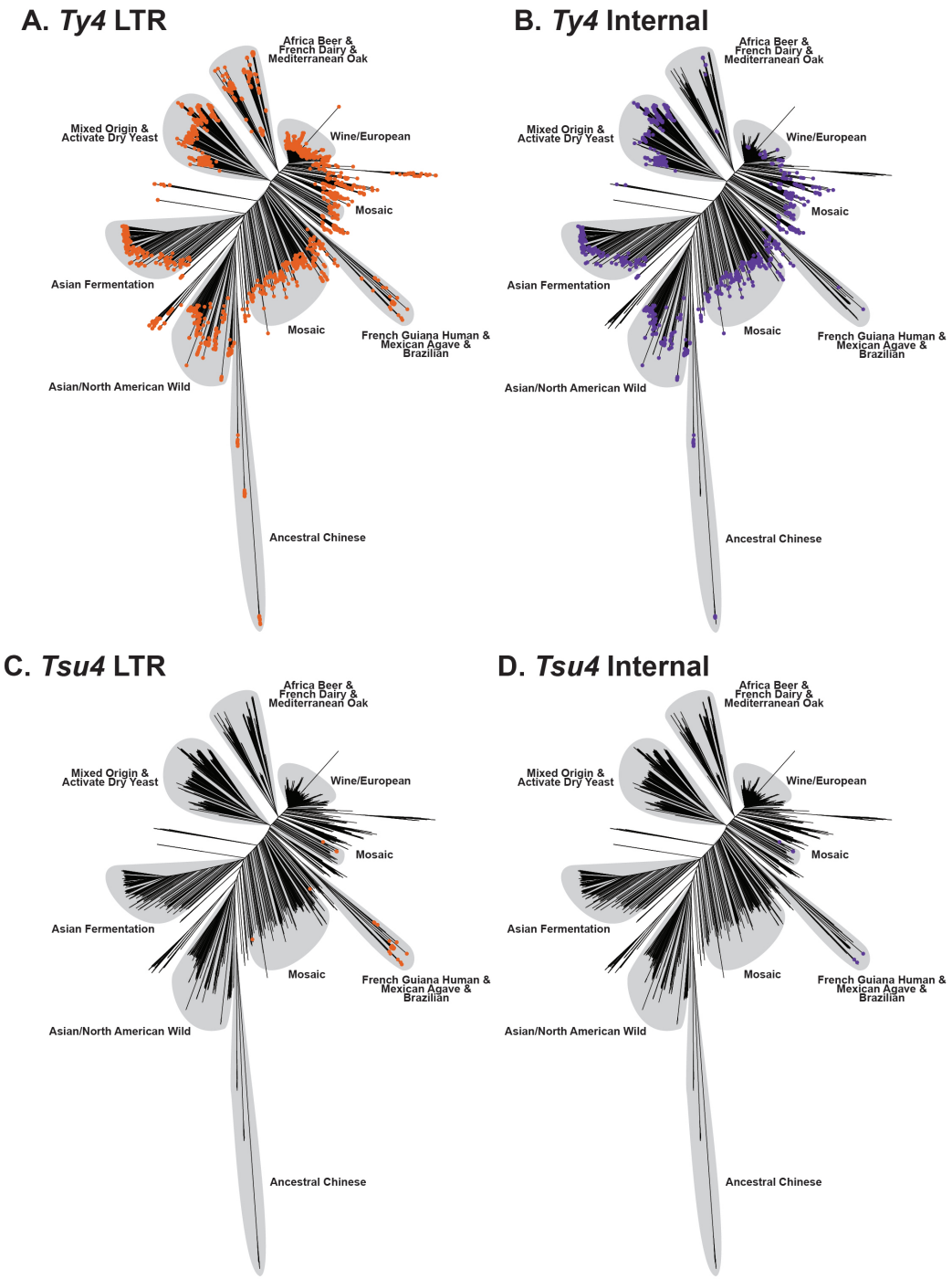

Figure 4.2: **Host phylogeny of *S. cerevisiae* annotated with the presence or absence of Ty4/Tsu4.** Shown are the presence/absence estimates for Tsu4 and Ty4 subfamilies from worldwide *S. cerevisiae* populations. In all four panels, the neighbor joining phylogenetic tree was reconstructed based on 2,787,384 genome-wide SNPs from 2,404 *S. cerevisiae* isolates integrating several public WGS datasets (see Section 4.3.1). Major lineages are annotated in light gray shadows based on the population structure reported in [222, 75]. Colored dots indicate the presence of Ty4 (panel A and B) and Tsu4 (panel C and D) for each *S. cerevisiae* strain. The presence of a Ty subfamily is determined by copy number estimates that are >1 for LTR (annotated with orange dots in panel A and C) and >0.5 for internal region (annotated with purple dots in panel B and D).

instance, the "activate dry yeast" clade from Duan *et al.* [75] is closely related to "mixed origin" from Peter *et al.* [222] (Figure 4.2).

After mapping the presence/absence of Ty4/Tsu4 subfamilies onto the host phylogeny of *S. cerevisiae*, we find that Ty4 LTRs are prevalent in all *S. cerevisiae* populations (Figure 4.2A), although Ty4 internal sequences are absent in some lineages or strains (Figure 4.2B). In contrast, Tsu4 signals are restricted to lineages including "French Guiana human," the related "Mexican agave" clade, and few mosaic strains from Brazil (Figure 4.2C and D). Notably, all *S. cerevisiae* isolates showing Tsu4 signal are from a restricted geographical region, i.e, Central/South America (specifically French Guiana, Mexico, Brazil, and the West Indies). Tsu4 elements are absent from most other *S. cerevisiae* lineages, including the most ancestral Chinese lineages. Interestingly, we identified three *S. cerevisiae* isolates sequenced in Peter *et al.* [222] (CQS of French Guiana human, AFQ and CDM of mosaic strains in Brazil) that contain internal coding sequences for both Ty4 and Tsu4 subfamilies, suggesting that Ty4 subfamilies can co-exist in the same strain of *Saccharomyces* species.

The prevalence of the Ty4 subfamily across most populations including ancient Chinese lineages indicates that Ty4 subfamily is ancestral in *S. cerevisiae*, as is also the case in its sister species, *S. paradoxus* (see above). Therefore, the Ty4 subfamily was likely present in the common ancestor of *S. cerevisiae* and *S. paradoxus* prior to their speciation. The highly restricted phylogenetic and geographical distribution of Tsu4 subfamily in *S. cerevisiae* strongly indicates at least one HTT introduced Tsu4 into *S. cerevisiae* in Central/South America and that Tsu4 has not dispersed widely despite human activities influence the biogeography of *S. cerevisiae* [222]. The distribution of Tsu4 in *S. cerevisiae* can be explained by two hypothetical models: (i) A single HTT event introduced Tsu4 into French Guiana *S. cerevisiae* strains, followed by admixture to explain its presence in mosaic strains; or (ii) Separate HTT events introduced Tsu4 into French Guiana and mosaic lineages, respectively.

S. uvarum and *S. eubayanus*

The Tsu4 subfamily was first characterized in the *S. uvarum* reference genome CBS 7001 [199]. Subsequently, Tsu4 elements were found in the *S. eubayanus* subgenome of *S. pastorianus* [20], although no direct evidence for the presence of Tsu4 in a pure *S. eubayanus* genome has been reported to date. Tsu4 elements from the hybrid *S. pastorianus* genome were most closely related to those from *S. paradoxus*, and thus *S. eubayanus* was hypothesized as a potential donor of the HTT event into *S. paradoxus*. To investigate the global presence/absence of Ty4/Tsu4 subfamilies in *S. uvarum* and *S. eubayanus*, we complied WGS shotgun datasets for identified populations (datasets see Section 4.3), reconstructed comprehensive ML phylogenetic trees, and scanned Ty content from shotgun reads using the same coverage-based strategy used for *S. paradoxus*/*S. cerevisiae* (Figures 4.3 and 4.4). We decided to exclude *S. pastorianus* from this analysis because: (i) *S. pastorianus* is a recent human-associated hybrid between *S. eubayanus* and *S. cerevisiae* used in lager beer brewing industry [166], and thus unlikely to be the wild donor of Tsu4 into *S. paradoxus*; (ii) The *S. eubayanus* portion of *S. pastorianus* is originated from Holarctic/Tibet population (strain CDFM21L.1) [242], and pure strains from this *S. eubayanus* population are included in our analysis; and (iii) Analysis of hybrid genomes like *S. pastorianus* are more complicated than pure genomes, especially since we have detected Tsu4 elements from some *S. cerevisiae* strains (see Section 4.4.1).

As shown in Figure 4.3A, the integrated ML host phylogeny of *S. uvarum* including 62 isolates based on 254,300 SNPs reveals four major lineages consistent with previous studies [4, 220]: South American A and B, Holarctic, and Australasia. The Holarctic clade is a sub-group of the South American A lineage, and the Australasia group is highly diverged from all other populations (Figure 4.3A). Unlike the general presence of Ty4 subfamily in *S. paradoxus* and *S. cerevisiae*, Ty4 LTR or internal regions are absent in all identified *S. uvarum* lineages (Figure 4.3B). In contrast, Tsu4 sequences are distributed across all lineages within *S. uvarum* (Figure 4.3C). Tsu4 internal abundance varies for each *S. uvarum* strain,

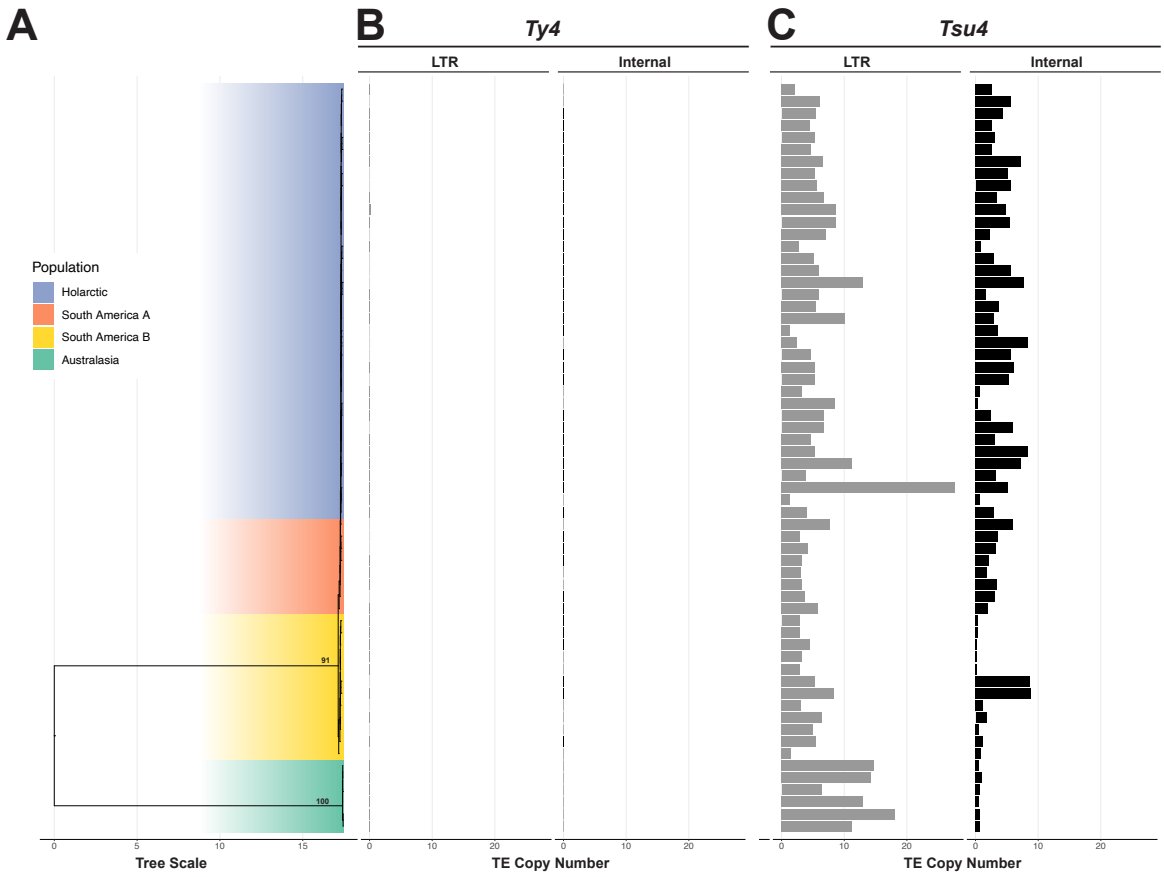

Figure 4.3: **Host phylogeny of *S. uvarum* annotated with Ty4/Tsu4 copy number estimation.** Shown are the copy number estimates for Ty4 (B) and Tsu4 (C) subfamilies from worldwide *S. uvarum* populations. The ML tree is reconstructed using 254,300 genome-wide SNPs from 62 *S. uvarum* isolates and midpoint rooted. Plotting details are identical as described in Figure 4.1. The population structure of *S. uvarum* is colored according to [4].

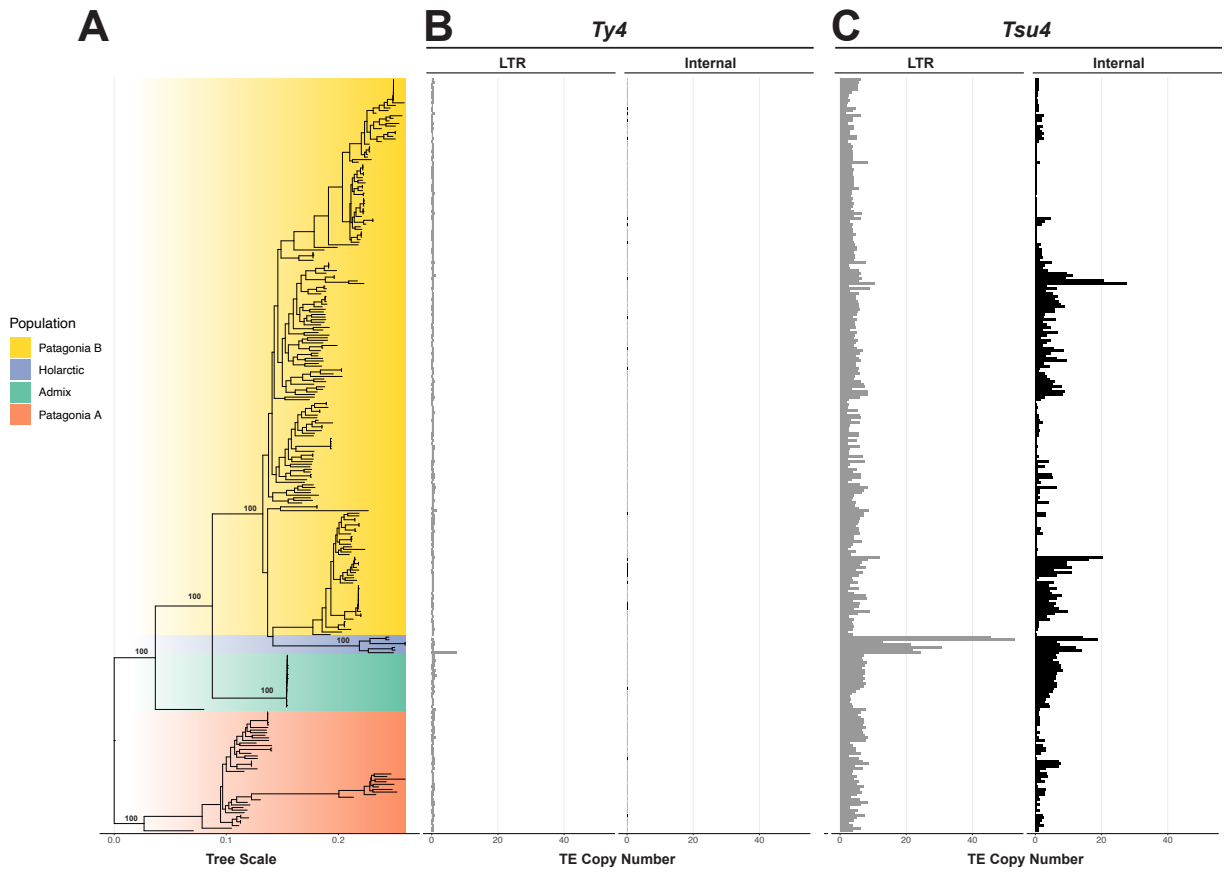

Figure 4.4: **Host phylogeny of *S. eubayanus* annotated with Ty4/Tsu4 copy number estimation.** Shown are the copy number estimates for Ty4 (B) and Tsu4 (C) subfamilies from worldwide *S. eubayanus* populations. The ML tree is reconstructed using 321,442 genome-wide SNPs from 292 *S. eubayanus* isolates and midpoint rooted. Plotting details are identical as described in Figure 4.1. The population structure of *S. eubayanus* is colored according to [217].

and is absent in some strains, yet shows no correlation with population structure (Figure 4.3C). These results indicate a relatively simple history of the Ty4 family in *S. uvarum* based on current samples: the Tsu4 subfamily represents the ancestral state and the only active subfamily in *S. uvarum*, with no evidence of Ty4 ever having been present in this species.

Figure 4.4A presents the integrated host tree for *S. eubayanus* reconstructed using maximum-likelihood strategy, involving 292 isolates and 321,442 genome-wide SNPs (datasets see Section 4.3), which reveals the four major lineages previously identified [217]: Patagonia/South America A and B, Holarctic, and a clade of strains derived from admixture between two Patagonia clades [217, 220]. The Holarctic lineage is a sub-clade of Patagonia B. The genetic diversity of the available *S. eubayanus* lineages in our sample is much lower than that of *S. uvarum* (Figures 4.3A and 4.4A), which could be partially explained by the current lack of WGS data for the deepest West China/Sichuan lineage [26, 217]. The overall Ty4/Tsu4 content of *S. eubayanus* is similar to *S. uvarum*, showing prevalence of the Tsu4 subfamily yet near complete absence of Ty4 subfamily (Figure 4.4B and C). Notably, the abundance of Tsu4 LTRs is significantly higher for the Holarctic lineage than the others (Figure 4.4C). Additionally, the complete "absence" of Ty4 in *S. eubayanus* is not as robust as *S. uvarum*, with a few *S. eubayanus* strains showing signal of Ty4 LTR sequences (Figure 4.4B). Nevertheless, our results confirm that the Tsu4 subfamily is the ancestral subfamily in both *S. uvarum* and *S. eubayanus*, with relatively high Tsu4 abundance in the Holarctic lineage of *S. eubayanus*. Whereas the Ty4 subfamily is inferred to be ancestral in *S. cerevisiae* and *S. paradoxus* clade, the Tsu4 subfamily is inferred to be ancestral in the *S. uvarum* and *S. eubayanus* clade.

S. mikatae

Bergman [20] identified an unusual Tsu4 element from one *S. mikatae* strain that is closely related with Tsu4 elements from *S. uvarum* and *S. eubayanus*. This observation is inconsistent with the widely accepted phylogenetic relationship of *Saccharomyces* species which

indicates a closer relationship between *S. mikatae* and *S. cerevisiae*/*S. paradoxus* [4, 220]. Assuming the host species tree is correct, the expectation is that *S. mikatae* should have a Ty4-like subfamily, rather than a Tsu4-like subfamily. Such inconsistency between phylogenies of TE sequences and host genomes suggested another possible HTT event involving the Ty4 family in *S. mikatae*.

The global population structure of *S. mikatae* has been reported recently, which reveals two major lineages both sampled from Japan: Asia A and Asia B [220]. However, *S. mikatae* is poorly sampled compared to the four *Saccharomyces* species investigated above. Our reconstructed ML tree consisting of 18 isolates based on 13,335 genome-wide SNPs confirms the two lineages reported in Peris *et al.* [220] (Figure 4.5A). Ty4/Tsu4 copy number estimates indicate that both subfamilies are present across the species (Figure 4.5B and C). The copy number of Ty4 internal regions is low in all strains, while the copy number of Tsu4 internal region is relatively high in two Asia A and all Asia B *S. mikatae* isolates, suggesting that Tsu4 is currently the most active subfamily in *S. mikatae*. From these data alone, we are not able to conclude whether the Ty4 or Tsu4 subfamilies represent the ancestral state in *S. mikatae*.

We unexpectedly observe that the abundance of the Tsu4 internal region is greater than estimated LTR abundance in some strains, which could be explained by two hypotheses: (i) The presence of truncated Tsu4 elements that have internal regions but lack flanking LTRs; or (ii) The query sequence used to estimate abundance of Ty4/Tsu4 LTRs is too divergent in *S. mikatae* for accurate short-read mapping. The first hypothesis is less likely given the Ty4/Tsu4 annotation from whole genome assemblies (see below; Table 4.2). If the second hypothesis is true, it is possible that a divergent Ty4/Tsu4 subfamily is present in *S. mikatae*.

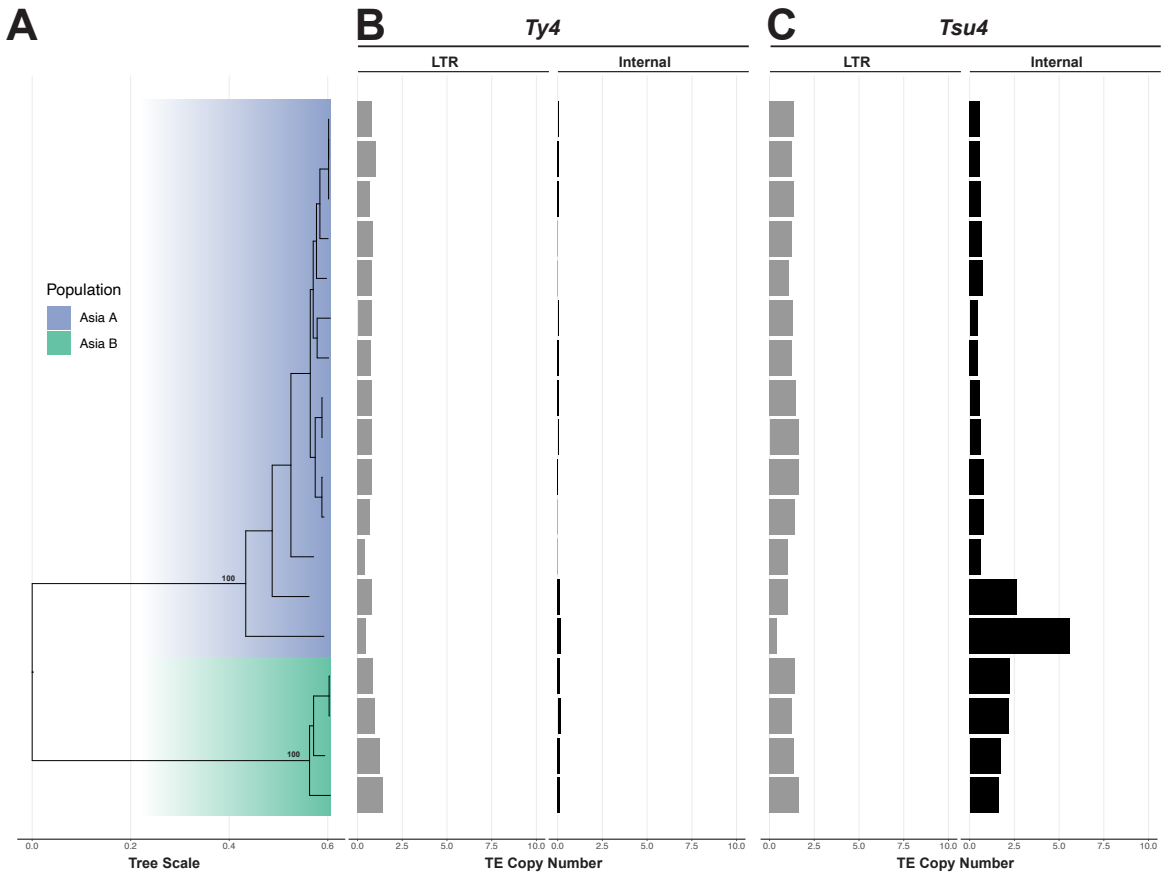

Figure 4.5: **Host phylogeny of *S. mikatae* annotated with Ty4/Tsu4 copy number estimation.** Shown are the copy number estimates for Ty4 (B) and Tsu4 (C) subfamilies from identified *S. mikatae* populations. The ML tree is reconstructed using 13,335 genome-wide SNPs from 18 *S. mikatae* isolates and midpoint rooted. Plotting details are identical as described in Figure 4.1. The population structure of *S. mikatae* is colored according to [220].

4.4.2 Survey of Ty4/Tsu4 subfamilies in high-quality whole genome assemblies of *Saccharomyces* species

The short-read Ty abundance estimation strategy allowed us to investigate the presence/absence and relative abundance of TE content in large-scale samples from multiple species, but did not distinguish full-length element from solo LTRs, identify the exact genomic location of TEs, or generate sequences for each element. Annotation of TEs in whole genome assemblies could provide substrate for more detailed TE content survey and cross-validation of the conclusions made by unassembled short-read scan results. To do this, we integrated recently published whole genome assemblies generated using long-read sequencing data for diverse *Saccharomyces* species including *S. paradoxus* [291, 80, 54], *S. cerevisiae* [212, 222, 291], *S. uvarum* [53], *S. eubayanus* [43, 179], and *S. jurei* [190]. In addition, we included short-read assemblies for three *S. cerevisiae* strains (245 [180], AFQ and CDM [222]) that were identified to contain Tsu4 internal sequences according to our unassembled short-read scan (Figure 4.2). *S. jurei* is the sister species to *S. mikatae* and included here for TE annotation but not in the short-read scan, due to the lack of short-read WGS datasets of this species. In addition, we sequenced and assembled two *S. mikatae* strains (IFO 1815 and NBRC 10994) using PacBio long reads. The basic statistics of two new *S. mikatae* genome assemblies can be found in Table 4.1. We then annotated TE content in all genome assemblies with a RepeatMasker-based pipeline (details described in [64]). The count of annotated Ty4/Tsu4 elements from whole genome assemblies are summarized in Table 4.2.

Table 4.1: **Statistics for *de novo* genome assemblies of *S. mikatae* strains IFO 1815 and NBRC 10994 that are newly generated by this study.**

	IFO 1815	**NBRC 10994**
# Contigs	22	25
Total Length (bp)	12,243,664	12,361,694
Contig N50 (bp)	778,614	771,552
L50	7	7
GC content (%)	37.8	37.99
# tRNAs	285	274
Complete BUSCOs (%)	97.19	97.23
Data availability	JARBHO000000000	JARBHP000000000

Table 4.2: **Ty4 and Tsu4 content in whole genome assemblies of multiple *Saccharomyces* species.**

Species	**Strain**	**Ref.**	**Population**	**Strategy**	**# Ty4 full**	**# Ty4 trun.**	**# Ty4 solo**	**# Tsu4 full**	**# Tsu4 trun.**	**# Tsu4 solo**
S. paradoxus	CBS432	[291]	EU	PacBio	0	2	46	0	0	0
S. paradoxus	N44	[291]	FE	PacBio	3	3	106	0	0	0
S. paradoxus	LL2012_001	[80]	SpA	ONT	0	2	45	0	0	0
S. paradoxus	DG1768	[54]	SpB	PacBio	0	1	48	0	0	17
S. paradoxus	LL2011_012	[80]	SpC	ONT	0	1	47	1	0	26
S. paradoxus	LL2012_016	[80]	SpC*	ONT	0	1	46	1	1	33
S. paradoxus	MSH-604	[80]	SpB	ONT	0	1	46	4	0	19
S. paradoxus	R23	[80]	SpD	ONT	0	1	51	4	0	29
S. paradoxus	WX20	[80]	SpD	ONT	0	1	49	4	0	26
S. paradoxus	UFRJ50816	[291]	SpB (SA)	PacBio	0	1	45	22	2	105
S. paradoxus	UWOPS91-917.1	[291]	Hawaii	PacBio	0	0	40	1	1	88
S. paradoxus	YPS138	[291]	SpB	PacBio	0	0	49	1	0	18
S. cerevisiae	245	[180]	Mosaic region 2	Illumina	0	0	12	1	0	3
S. cerevisiae	AFQ	[222]	Mosaic region 2	Illumina	0	10	24	1	0	1
S. cerevisiae	CDM	[222]	Mosaic region 2	Illumina	0	7	32	0	1	3

Table 4.2: **Ty4 and Tsu4 content in whole genome assemblies of multiple *Saccharomyces* species (Continued).**

Species	Strain	Ref.	Population	Strategy	# Ty4 full	# Ty4 trun.	# Ty4 solo	# Tsu4 full	# Tsu4 trun.	# Tsu4 solo
S. cerevisiae	CQS	[212]	French Guiana Human	ONT	1	1	71	9	0	36
S. cerevisiae	DBVPG6044	[291]	West African	PacBio	0	0	18	0	0	0
S. cerevisiae	DBVPG6765	[291]	Wine/European	PacBio	0	0	5	0	0	0
S. cerevisiae	S288C	[291]	Laboratory	PacBio	3	0	14	0	0	0
S. cerevisiae	SK1	[291]	Mosaic	PacBio	0	0	14	0	0	0
S. cerevisiae	UWOPS03-461.4	[291]	Malaysian	PacBio	0	0	60	0	0	0
S. cerevisiae	Y12	[291]	Asian fermentation	PacBio	4	1	14	0	0	0
S. cerevisiae	YPS128	[291]	North American	PacBio	4	0	13	0	0	0
S. uvarum	CBS7001	[53]	European	PacBio	0	0	0	9	7	71
S. eubayanus	CBS12357	[40]	PB-1	ONT	0	0	1	0	6	98
S. eubayanus	CDFM21L.1	[43]	Holarctic	ONT	0	0	1	14	4	160
S. eubayanus	CL216.1	[179]	PB-3	ONT	0	0	1	0	7	100
S. eubayanus	CL450.1	[179]	PB-1	ONT	0	0	1	4	10	107
S. mikatae	IFO 1815	This study	Japan	PacBio	0	2	5	15	11	207
S. mikatae	NBRC 10994	This study	Japan	PacBio	0	1	10	26	8	234
S. jurei	NCYC 3947	[190]	N.A.	PacBio	0	0	7	4	2	137
S. jurei	NCYC 3962	[190]	N.A.	PacBio	0	0	6	3	2	138

Ty4 sequences were identified in all 12 assembled *S. paradoxus* genomes in this study, whereas full-length copies are only found in the Far East strain N44 (n=3). In contrast, Tsu4 sequences are identified in 9 out of 12 *S. paradoxus* strains from indigenous American strains, but are completely absent from the other three strains from either Eurasian lineage or the recently migrated SpA (CBS432, N44, and LL2012_001) (Table 4.2). At least one full-length Tsu4 copy is identified in all indigenous American *S. paradoxus* strains with WGAs

except for the SpB strain DG1768 [54]. As reported previously [20], the Tsu4 copy number in UFRJ50916 – the South American SpB strain – is much higher than other *S. paradoxus* strains with including 22 full-length copies, 2 truncated elements (with one near-complete copy), and >100 solo LTRs.

The Ty4 subfamily is identified in every *S. cerevisiae* strain with WGAs. In contrast, the Tsu4 subfamily is essentially absent from most *S. cerevisiae* genome assemblies [291, 212] generated using PacBio or Oxford Nanopore long reads (Table 4.2 and Figure 4.6). For each of the three genomes we included that were based on short-read data (245, AFQ, and CDM), only one full-length or truncated Tsu4 copy was found, allowing proper assembly of these Tsu4 elements using short-read data (Table 4.2). Nine Tsu4 full-length copies are annotated from the CQS genome using our pipeline, which confirms the observation in O'donnell *et al.* [212]. All four of these *S. cerevisiae* strains are genetically related with French Guiana Human or Mosaic region 2 and geographically restricted to South America, validating our observation in short-read scan (Figure 4.2A and B). Notably, full-length copies of both Tsu4 and Ty4 subfamilies are found in the CQS genome, which proves that these two subfamilies can co-exist in the same *Saccharomyces* strain.

The Ty4 subfamily is absent in the only *S. uvarum* whole genome assembly (CBS 7001) in our dataset and nearly absent in all four *S. eubayanus* genomes. Interestingly, we annotate one Ty4 solo LTR from each of the four *S. eubayanus* genome assemblies included in this study at essentially the same genomic location on Chromosome 3 (Chr 3: 80,000, Table 4.2). The common location of this Ty4 LTR in *S. eubayanus* was possibly caused by an ancient HTT event in the common ancestor of this species that did not lead to successful invasion. In contrast, the reference genome of *S. uvarum* contains 9 full-length Tsu4 copies across the genome [53], plus several truncated elements and ∼70 solo LTRs, suggesting recent activity of the Tsu4 subfamily in this strain (Table 4.2). Similar to *S. uvarum*, all four *S. eubayanus* genomes contain relatively high copy number of Tsu4 elements, especially in the Holarctic isolate CDFM21L.1, where 14 full-length Tsu4 copies are annotated. The high

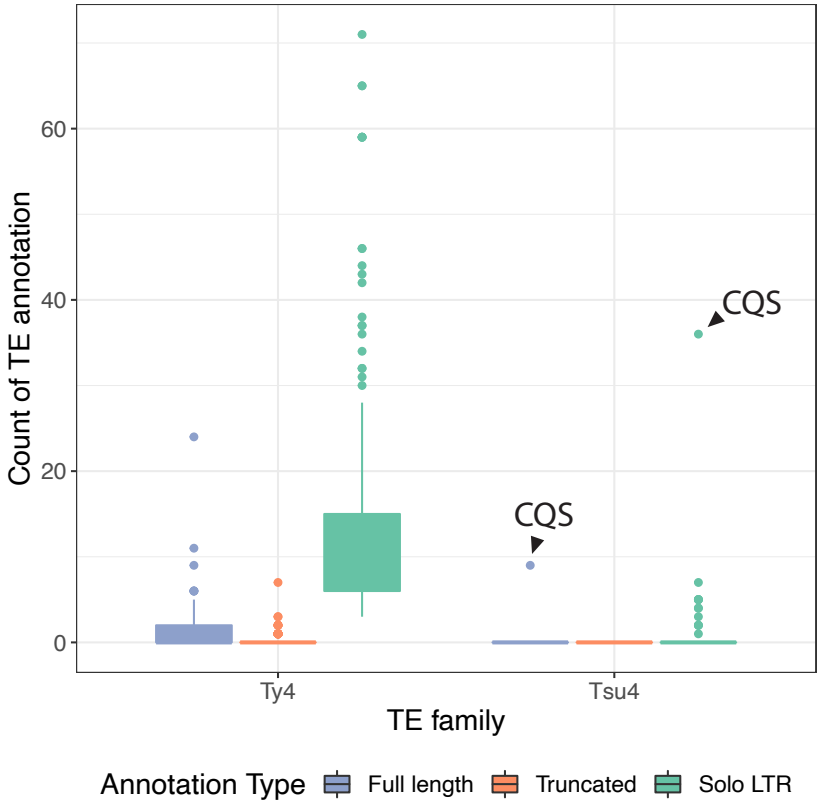

Figure 4.6: **Re-annotation summary for Ty4/Tsu4 content in *S. cerevisiae* assemblies generated by O'donnell *et al.* [212].** Shown are the box plots for numbers of Ty4 or Tsu4 subfamilies annotated in each *S. cerevisiae* genome from O'donnell *et al.* [212]. The color of boxes indicates the type of elements: blue, red, and green boxes indicate full-length elements, truncated elements, and solo LTRs, respectively. Colored boxes show interquartile ranges (IQR), whiskers show values 1.5×IQR of the upper or lower quartiles, and the dots indicate outliers that beyond 1.5×IQR. Outliers for Tsu4 elements in strain CQS are annotated. Other outliers for Tsu4 LTRs indicate strains from either French Guiana or mosaic lineages.

Tsu4 abundance in Holarctic *S. eubayanus* is also consistent with results from the short-read scan (see Section 4.4.1).

Truncated elements and solo LTRs of the Ty4 subfamily are found in both *S. mikatae* genomes, whereas the Ty4 subfamily is only present as LTRs in *S. jurei*. Unlike the Ty4 subfamily, both *S. mikatae* strains contain an extremely high abundance of full-length Tsu4 elements (n=15 and 26 for IFO 1815 and NBRC 10994, respectively), plus ∼10 truncated copies and >200 solo LTRs, suggesting that the Tsu4 subfamily is highly active in these strains (Table 4.2). Complete, truncated and solo LTRs for the Tsu4 subfamily are also found in both *S. jurei* genomes, suggesting that both Ty4 and Tsu4 subfamilies were present in the ancestor of these two species.

In conclusion, annotation of TEs in genome assemblies provides similar results to those observed from TE content scan of unassembled short reads, not only in the overall presence/absence of Ty4 subfamilies for *Saccharomyces* species, but also in observations on relative abundance across populations. These results cross validate both approaches, and allow us to use sequences of TEs extracted from this limited number of complete genome assemblies to reflect species-wide patterns.

4.4.3 Phylogenetic analysis of Ty4/Tsu4 sequences from *Saccharomyces* genome assemblies

In previous sections, we used unassembled short-read data and long-read WGAs, and resolved the distribution and ancestral states of Ty4/Tsu4 subfamilies in multiple *Saccharomyces* species. To better understand the evolutionary history of these Ty4 subfamilies, we performed phylogenetic analysis on Ty4/Tsu4 full-length elements annotated from *Saccharomyces* genomes in Section 4.4.2. To do this, we extracted full-length Ty4/Tsu4 copies from our integrated dataset of *S. paradoxus*, *S. cerevisiae*, *S. uvarum*, *S. eubayanus*, *S. mikatae*, and *S. jurei* genome assemblies shown in Table 4.2 and Figure 4.6. We then created multiple sequence alignments from the collection of full-length elements (n=400), and reconstructed

a ML phylogenetic tree based on the internal coding regions of full-length copies including 3,739 distinct alignment sites (see Section 4.3.5). Flanking LTR sequences of each full-length element were excluded because the LTR sequences could be poorly annotated due to their high sequence divergence between species.

As shown in Figure 4.7A, two distinct subfamilies of the Ty4 family are reconstructed, i.e., the Ty4 subfamily and the Tsu4 subfamily. The Ty4 subfamily shows a relatively simple topology and is only identified in two *Saccharomyces* species – *S. paradoxus* and *S. cerevisiae*. Each forms a well-supported species-specific monophyletic group (Figure 4.7A), despite of the high copy number of annotated Ty4 full-length elements (n=276). This observation is expected and consistent with our conclusion in Section 4.4.1 that Ty4 is the ancestral subfamily in *S. paradoxus* and *S. cerevisiae*, and was present in the common ancestor of both species.

The Tsu4 subfamily, on the other hand, reveals a much more complex topology with at least 10 well-supported monophyletic groups involving all six inspected *Saccharomyces* species (Figure 4.7A). Clades of Tsu4 sequences from *S. paradoxus* and *S. cerevisiae* (1, 2, 3, and 4 in Figure 4.7A) are clustered together with no clear boundary between species. *S. eubayanus* genomes contain clades of Tsu4 sequences (5 and 6) that are closest to the Tsu4 clades in *S. paradoxus* and *S. cerevisiae*, followed by a clade of Tsu4 sequences from *S. uvarum* (7 and 8). These observations support results reported by Bergman [20] that Tsu4 sequences from *S. eubayanus* are most closely related to Tsu4 in *S. paradoxus* using direct genomic evidence from pure *S. eubayanus* genomes rather than indirect data from the hybrid species *S. pastorianus*. Consistent with Bergman [20], we also identified a clade of Tsu4 sequences from *S. mikatae* (10) that are clustered as an outgroup to those from *S. uvarum* and *S. eubayanus*. A clade of *S. jurei* Tsu4 sequences (9) form a monophyletic group which is even more similar to the *S. uvarum*/*S. eubayanus* Tsu4. A second distinct group of Tsu4 sequences is identified in *S. mikatae* (11), which shows evidence of recombination between the Ty4 and Tsu4 subfamilies (Figure 4.7B). In the following subsections, we divide

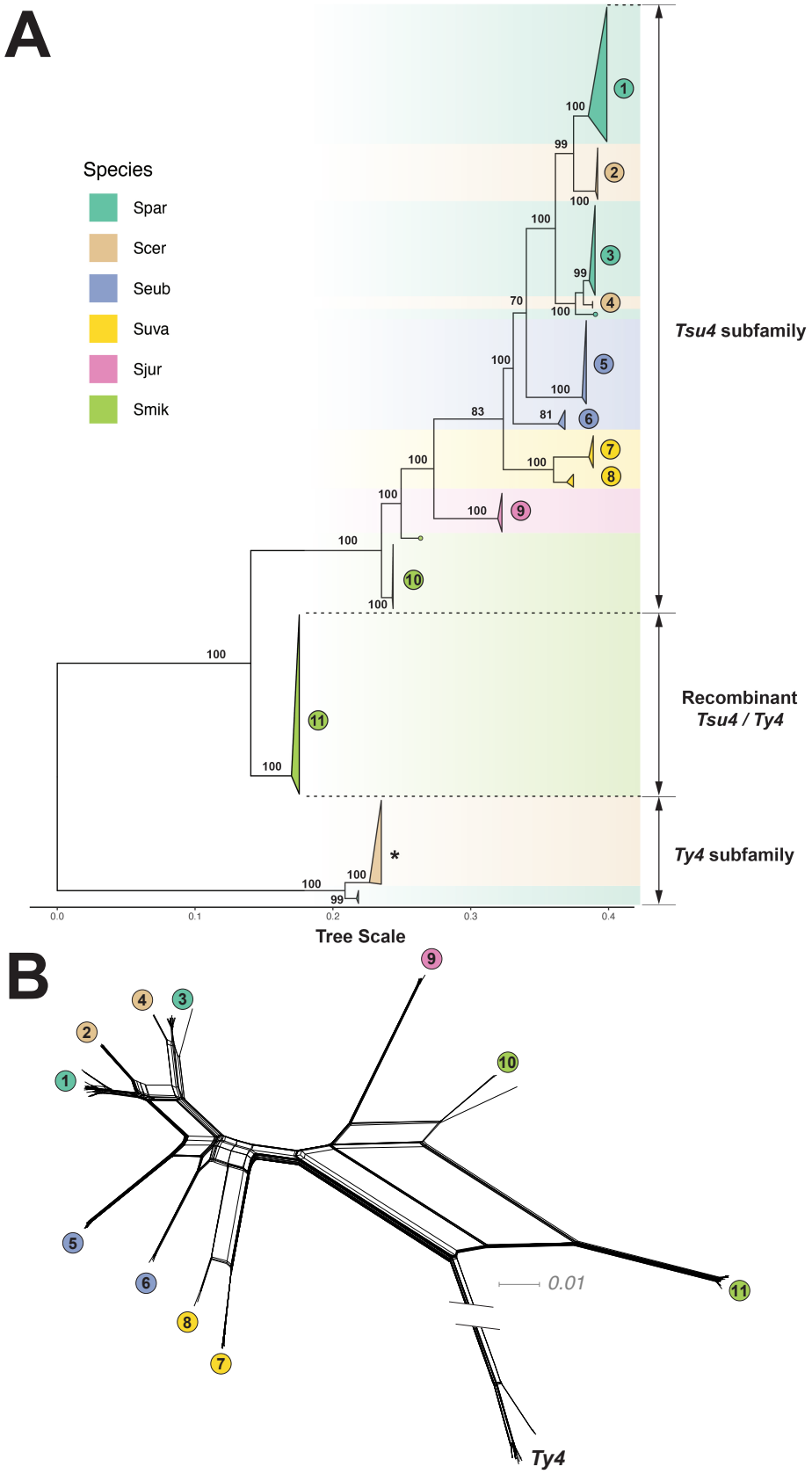

Figure 4.7: (Description see the next page.) **Phylogenetic analysis for the internal coding regions from full-length Ty4/Tsu4 copies in *Saccharomyces* genomes.**

Figure 4.7: (Figure see the previous page.) **Phylogenetic analysis of the internal coding regions from full-length Ty4/Tsu4 copies in *Saccharomyces* genomes.** (A) shows the ML phylogeny of all internal coding regions from full-length Ty4/Tsu4 elements annotated from *Saccharomyces* genome assemblies based on 3,739 distinct alignment sites. 100 bootstrap replicates are performed and bootstrap support is annotated for each key clade. The ML phylogeny is midpoint rooted. The tree scale bar for branch lengths is in units of substitutions per site. All monophyletic groups are collapsed as triangles. Two single/ungrouped Tsu4 elements (f267 from Hawaiian *S. paradoxus* strain UWOPS91-917.1 and f256 from *S. mikatae* strain NBRC 10994) are noted as dots at tips. Triangles, tip dots, and ranges are colored for each species. Vertical heights of triangles are proportional to the number of taxa. Horizontal widths of triangles are equal to the maximum branch length within the clade. Note that the monophyletic clade for the Ty4 subfamily from *S. cerevisiae* (annotated with asterisk) is re-scaled to 5% of the real size both horizontally and vertically, due to the large number of Ty4 sequences (n=273) in *S. cerevisiae* genomes. (B) shows the phylogenetic network for internal regions of full-length Tsu4 copies with NeighborNet algorithm. Here, the Ty4 subfamily is included as an outgroup and branch lengths are truncated for visualization. Only Ty4 subfamily elements annotated in genome assemblies from Yue *et al.* [291] are included in this network for simplicity. Lineages in the network are labeled according to monophyletic groups identified in Panel (A).

the Tsu4 subfamily among *Saccharomyces* species according to pairs of sister host species (i.e, *S. paradoxus*/*S. cerevisiae*, *S. uvarum*/*S. eubayanus*, and *S. mikatae*/*S. jurei*) to investigate the history of the Ty4 family in detail for each species pair separately.

S. paradoxus and *S. cerevisiae*

Our comprehensive survey of Ty4/Tsu4 content using unassembled and assembled WGS data consistently indicated that the Tsu4 subfamily was derived in *S. paradoxus* and *S. cerevisiae*, consistent with its introduction by HTT. As discussed in Section 4.4.1, different models may require one or more HTT events of Tsu4 subfamily into *S. paradoxus* and *S. cerevisiae*. Figure 4.8 shows a magnified version for *S. paradoxus*/*S. cerevisiae* clades of Ty4/Tsu4 sequences phylogeny (Figure 4.7A).

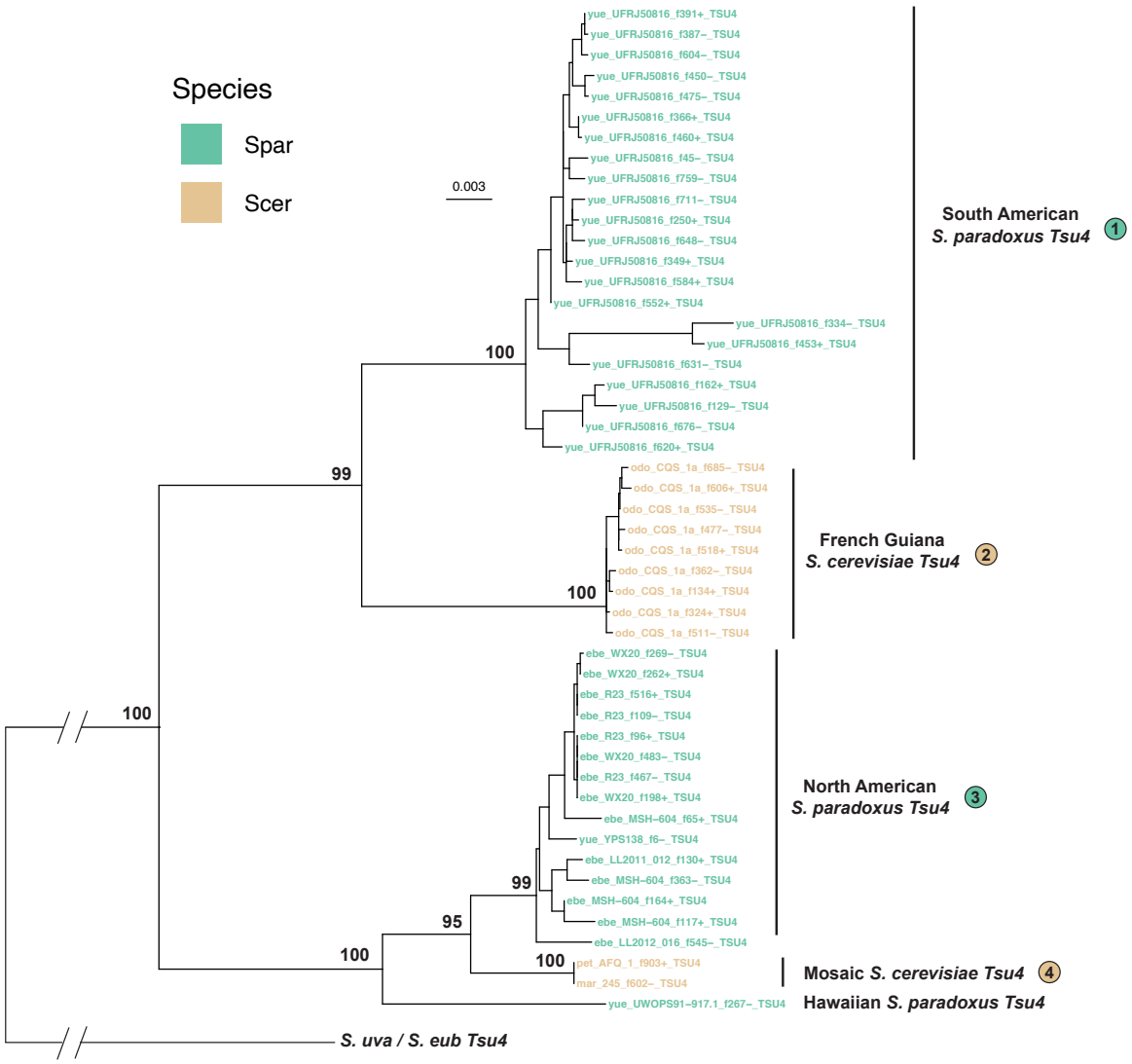

Figure 4.8: **ML phylogeny for the internal coding region from full-length Ty4/Tsu4 copies in *S. paradoxus* and *S. cerevisiae* genomes.** The phylogeny shown here is modified and rescaled from Figure 4.7A, only showing the Tsu4 groups from *S. paradoxus* and *S. cerevisiae* labeled with same clade numbers. Bootstrap supports are labeled for each identified monophyletic group. The tree scale bar for branch lengths is in units of substitutions per site. The isolation source of each Tsu4 clade is annotated.

As shown in Figure 4.8, two major well-supported Tsu4 sub-clades are characterized in *S. paradoxus*: (i) the "South American" clade contains Tsu4 sequences from the SpB strain UFRJ50816; and (ii) the "North American" clade consisting of Tsu4 sequences from SpB, SpC, SpC*, SpD, and Hawaiian strains. While two clades of Tsu4 sequences in *S. paradoxus* are divided geographically, this topology is inconsistent with the host phylogeny, because the South American strain UFRJ50816 is predicted to be a sublineage of SpB rather than an outgroup (see Section 4.4.1). This incongruence between Tsu4 and host phylogeny in *S. paradoxus* may imply more than one Tsu4 transfers into *S. paradoxus*, including one initial HTT into the ancestor of North American strains and a separate HTT into South American *S. paradoxus*. Alternatively, this TE/host tree incongruence can also be explained by "incomplete lineage sorting", where divergent clades of Tsu4 evolved in allopatry after an initial HTT event, with one lineage gone extinct in one region while staying active in the other geographical region, and vice versa.

For *S. cerevisiae*, our phylogenetic analysis reveals two clear lineages for Tsu4 sequences, one from two mosaic strains (245 and AFQ, Tsu4 sequence n=1 for each), and another from the French Guiana strain CQS (n=9). The Tsu4 sequences found in mosaic strains were introduced into *S. cerevisiae via* a recent HTT event from the North American *S. paradoxus* lineages. The French Guiana Tsu4 clade of *S. cerevisiae* is not directly derived from an extant *S. paradoxus* lineage but is closely related with the Tsu4 clade of South American *S. paradoxus*, indicating that it was either acquired from the South American *S. paradoxus* a long time ago, or from another related unknown donor species that donated Tsu4 to both South American *S. paradoxus* and French Guiana *S. cerevisiae*. Regardless, these results clearly show two HTT events for Tsu4 into *S. cerevisiae*.

S. uvarum and *S. eubayanus*

Bergman [20] previously suggested that *S. eubayanus* has the most closely related Tsu4 sequences to those in *S. paradoxus*, based on indirect evidence of Tsu4 sequences from the

hybrid species *S. pastorianus*. To confirm this hypothesis using pure *S. eubayanus* genome data, we include Tsu4 sequences annotated from several *S. eubayanus* genome assemblies which represent identified populations of *S. eubayanus*, as well as Tsu4 sequences from a Holarctic *S. uvarum* reference genome (Figure 4.9).

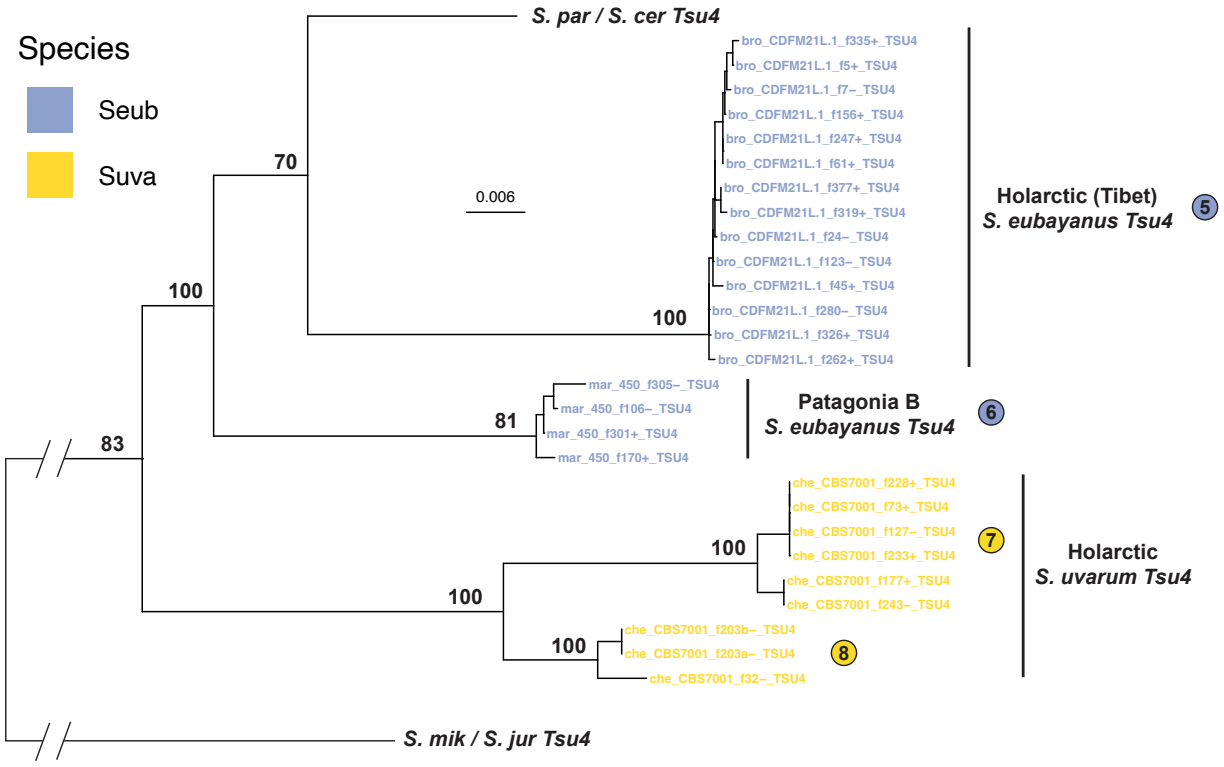

Figure 4.9: **ML phylogeny for the internal coding region from full-length Ty4/Tsu4 copies in *S. uvarum* and *S. eubayanus* genomes.** The phylogeny shown here is modified and rescaled from Figure 4.7A, only showing the Tsu4 groups from *S. uvarum* and *S. eubayanus* labeled with same clade numbers. Bootstrap supports are labeled for each identified monophyletic group. The tree scale bar for branch lengths is in units of substitutions per site. The isolation source of each Tsu4 clade is annotated.

We observe two monophyletic groups of Tsu4 in *S. eubayanus*, including one group from the Holarctic strain (CDFM21L.1) and a distinct group from the Patagonia B strain (450). The host phylogeny indicates that Holarctic populations of *S. eubayanus* should be a sublineage derived from Patagonia B (Section 4.4.1), which is not observed in the Tsu4 phylogeny. Such TE/host incongruence could be explained by a HTT event from an unknown donor to

Holarctic *S. eubayanus*, or incomplete lineage sorting, similar to the case we observed for Tsu4 sequences in South American *S. paradoxus*. Tsu4 elements annotated from *S. uvarum* cluster into one monophyletic group, but can be divided into two sub-groups, indicating potential HTT or recombination events. Since both lineages of Tsu4 in *S. uvarum* are found in the same strains, their existence cannot be explained by reciprocal lineage sorting. Our results confirm that Tsu4 sequences of Holarctic *S. eubayanus* strains are the closest outgroup to the *S. paradoxus*/*S. cerevisiae* Tsu4 clade (Figure 4.9). However, this observation does not necessarily imply that extant Holarctic *S. eubayanus* is the donor that directly introduced the Tsu4 subfamily into *S. paradoxus*. The observed pattern can be explained by two possibilities: (i) The HTT of Tsu4 subfamily from Holarctic *S. eubayanus* to *S. paradoxus* happened a long time ago and that ancestor cannot be currently sampled in Holarctic *S. eubayanus* populations; or (ii) An unknown donor species introduced Tsu4 into both Holarctic *S. eubayanus* and *S. paradoxus*.

Because of all of this complexity, specifically TE/host phylogeny incongruence observed in multiple *Saccharomyces* species and the lack of direct donor strains, it is challenging to identify the donor species and/or population that introduced the Tsu4 subfamily into *S. paradoxus* and secondarily into French Guiana *S. cerevisiae* based on current data.

S. mikatae* and *S. jurei

In addition to Tsu4 HTT events identified in *S. paradoxus* and *S. cerevisiae*, another possible HTT event was reported in *S. mikatae* by Bergman [20] based on phylogenetic discordance observed in one Tsu4 sequence. Specifically, the identified *S. mikatae* Tsu4 element was closely related with *S. uvarum*/*S. eubayanus* Tsu4 element, whereas *S. mikatae* is the closest relative to *S. cerevisiae*/*S. paradoxus* in the host phylogeny. To resolve the evolutionary history of Tsu4-like elements in *S. mikatae*, we analyzed *S. mikatae* and *S. jurei* in the phylogenetic tree of Ty4/Tsu4 elements (Figure 4.10). Assuming the Tsu4 subfamily originated in the *S. uvarum*/*S. eubayanus* clade, the fact that Tsu4 sequences from both *S.*

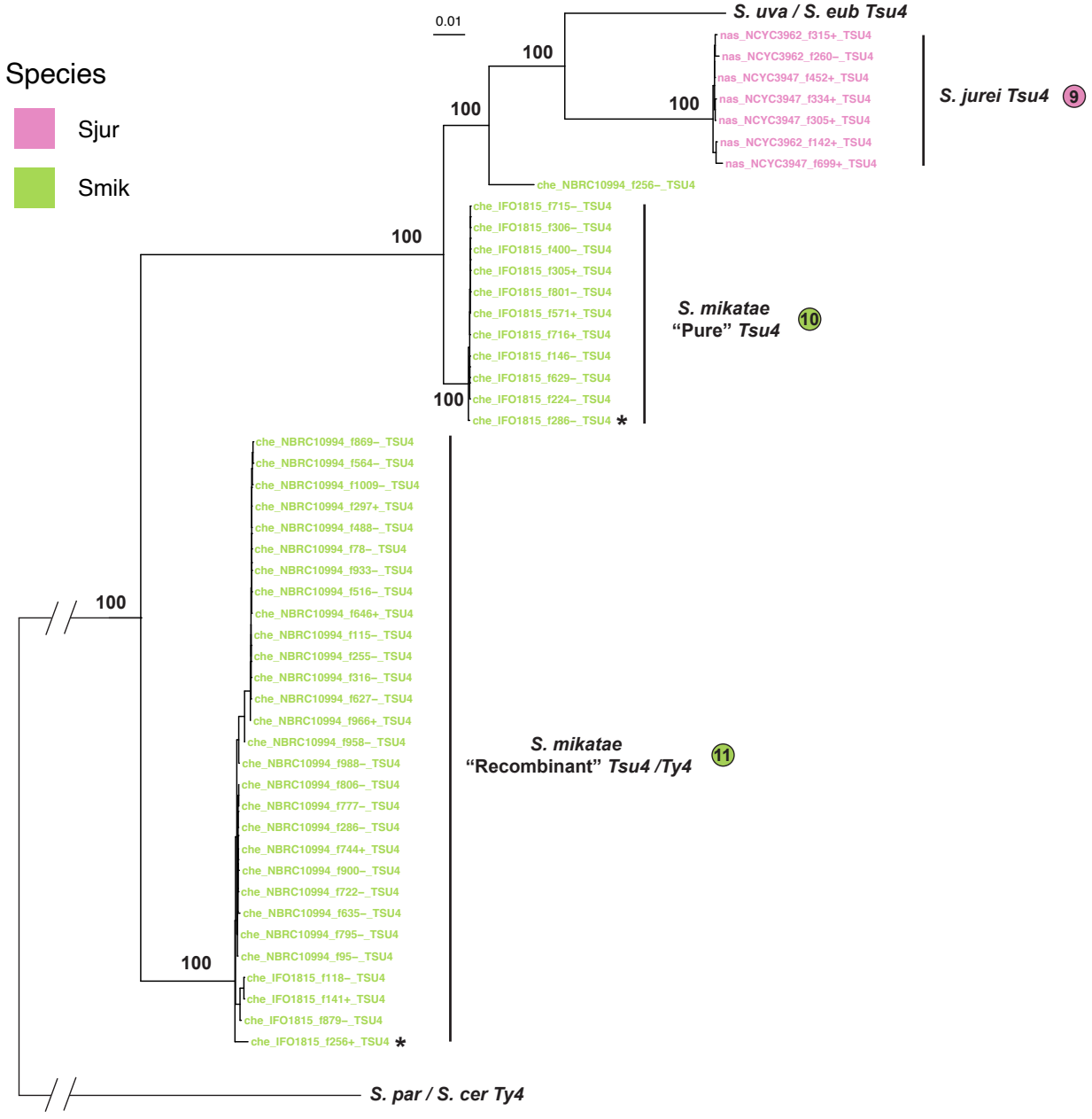

Figure 4.10: **ML phylogeny for the internal coding region from full-length Ty4/Tsu4 copies in *S. mikatae* and *S. jurei* genomes.** The phylogeny shown here is modified and rescaled from Figure 4.7A, only showing the Tsu4 groups from *S. mikatae* and *S. jurei* labeled with same clade numbers. Bootstrap supports are labeled for each identified monophyletic group. The tree scale bar for branch lengths is in units of substitutions per site. The isolation source of each Tsu4 clade is annotated. Tsu4 sequences of *S. mikatae* selected for sequence divergence analysis are labeled with asterisks.

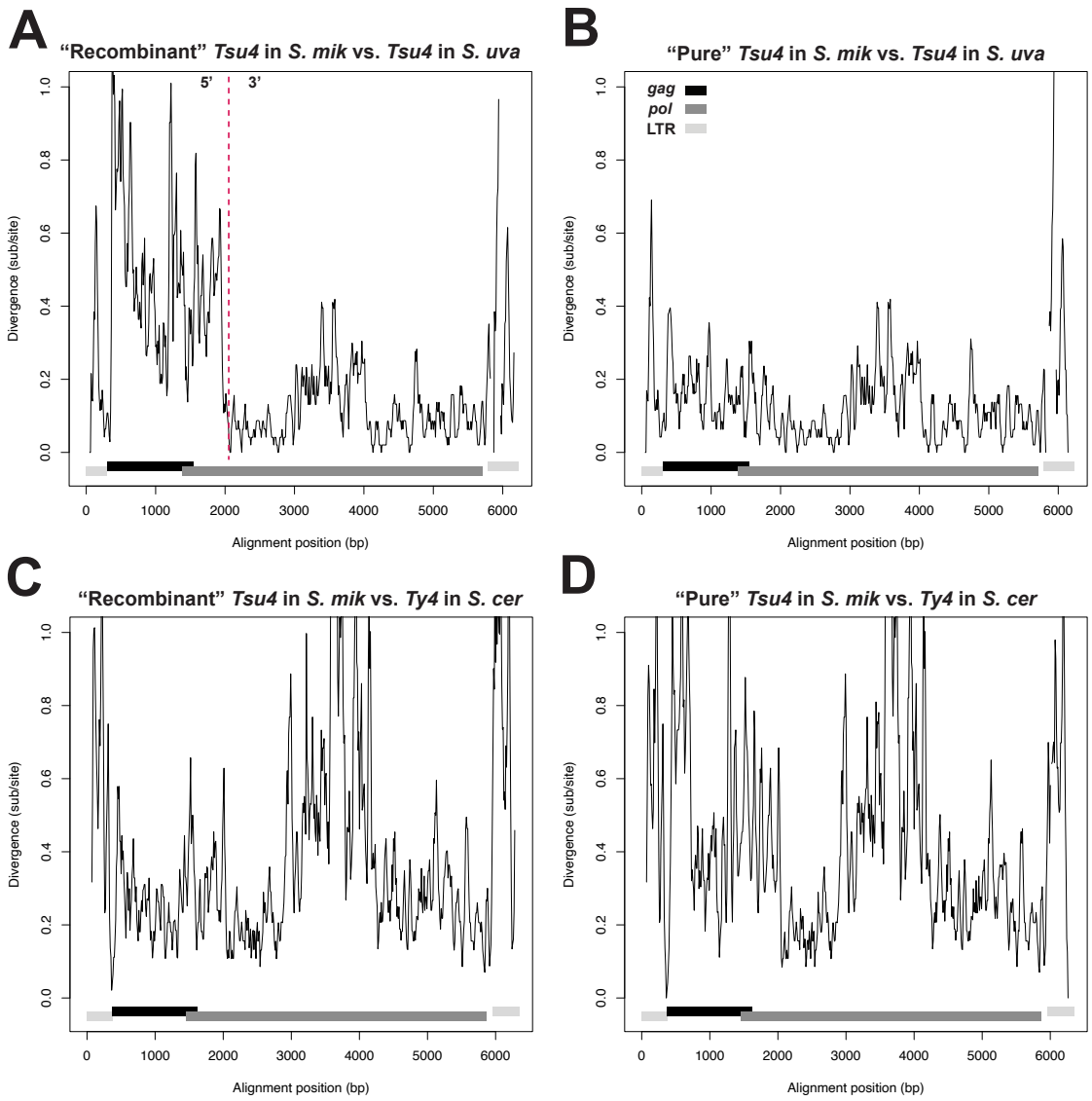

Figure 4.11: **Sequence divergence between full-length elements of Tsu4 in *S. mikatae* and *S. uvarum*, and Ty4 in *S. cerevisiae*.** Shown are sliding window analysis of pairwise sequence divergence between (A) "Recombinant" Tsu4 in *S. mikatae* vs. Tsu4 in *S. uvarum*; (B) "Pure" Tsu4 in *S. mikatae* vs. Tsu4 in *S. uvarum*; (C) "Recombinant" Tsu4 in *S. mikatae* vs. Ty4 in *S. cerevisiae*; (D) "Pure" Tsu4 in *S. mikatae* vs. Ty4 in *S. cerevisiae*. Structure of a full-length Ty4 element is annotated at the bottom of each panel in colored rectangles (*Gag* in black, *Pol* in darker gray, and LTRs in lighter gray). The dashed red line in panel (A) indicates the boundary of 5' and 3' internal region which is later used for partitioning the Ty4/Tsu4 phylogeny in Figure 4.12. Elements used in this analysis include IFO1815_f256 for "Recombinant" Tsu4 in *S. mikatae*, IFO1815_f286 for "Pure" Tsu4 in *S. mikatae*, CBS7001_f32 for Tsu4 in *S. uvarum*, YPS128_f49 for Ty4 in *S. cerevisiae*. IFO 1815 assembly was generated in this study. CBS 7001 and YPS128 assemblies were previous published in Chen *et al.* [53] and Yue *et al.* [291], respectively. Divergence measured in substitutions per site was calculated using a Kimura 2-parameter model in overlapping 50 bp windows with a 10 bp step size.

mikatae and *S. jurei* cluster closer to the Tsu4 subfamily suggests another HTT event into the ancestor of *S. mikatae* and *S. jurei*. Interestingly, two distinct clades of Tsu4 are observed in *S. mikatae* (10 and 11 in Figure 4.10), one of which (11) shows evidence of recombination with the Ty4 subfamily in the phylogenetic network (Figure 4.7B). To explore differences between these two clades, we selected one Tsu4 element from each clade of *S. mikatae* (clade 10 f286 and clade 11 f256 from the IFO 1815 genome, labeled with asterisks in Figure 4.10) plus one representative element of Tsu4 from *S. uvarum* (f32 from the CBS7001 genome) and Ty4 from *S. cerevisiae* (f49 from the YPS128 genome), and inspected the pairwise sequence divergence of *S. mikatae* Tsu4 clades vs. pure Tsu4 or Ty4 copies in sliding windows (Figure 4.11). Our result clearly suggests that the first $\sim$1/3 of the internal coding region (referred to as the 5' internal region) – including the complete *Gag* gene and the first 500bp of *Pol* gene – is significantly divergent between the "recombinant" Tsu4 lineage of *S. mikatae* and Tsu4 of *S. uvarum*, whereas the "pure" Tsu4 of *S. mikatae* shows more uniform sequence divergence across the complete TE (Figure 4.11 A and B). Conversely, the 5' internal region shows low sequence divergence between the recombinant Tsu4 of *S. mikatae* and Ty4, relative to the pure Tsu4 lineage (Figure 4.11C and D).

Recombination between the Ty4 and Tsu4 subfamilies can explain the basal placement of clade 11 in the Ty4 family phylogeny (Figure 4.7). To confirm this possibility, we partitioned the multiple sequence alignment of Tsu4 sequences into two segments according to the border of high sequence divergence as shown in Figure 4.11A, i.e, 5' and 3' internal regions, and then reconstructed ML phylogenies for the two partitions (Figure 4.11). A striking discordance can be observed in phylogenies reconstructed from 5' and 3' internal regions of Tsu4 sequences. In the 5' partition, two divergent groups of "Tsu4" elements are present in *S. mikatae*, including one group closely related to Tsu4 from *S. jurei*, *S. uvarum*, and *S. eubayanus*, and the other becoming a well-supported sister group to Ty4 from *S. paradoxus*/*S. cerevisiae*. In contrast, both clades of Tsu4 sequences from *S. mikatae* form a single monophyletic group in the tree reconstructed from the 3' region, which are both closely related to *S. uvarum*

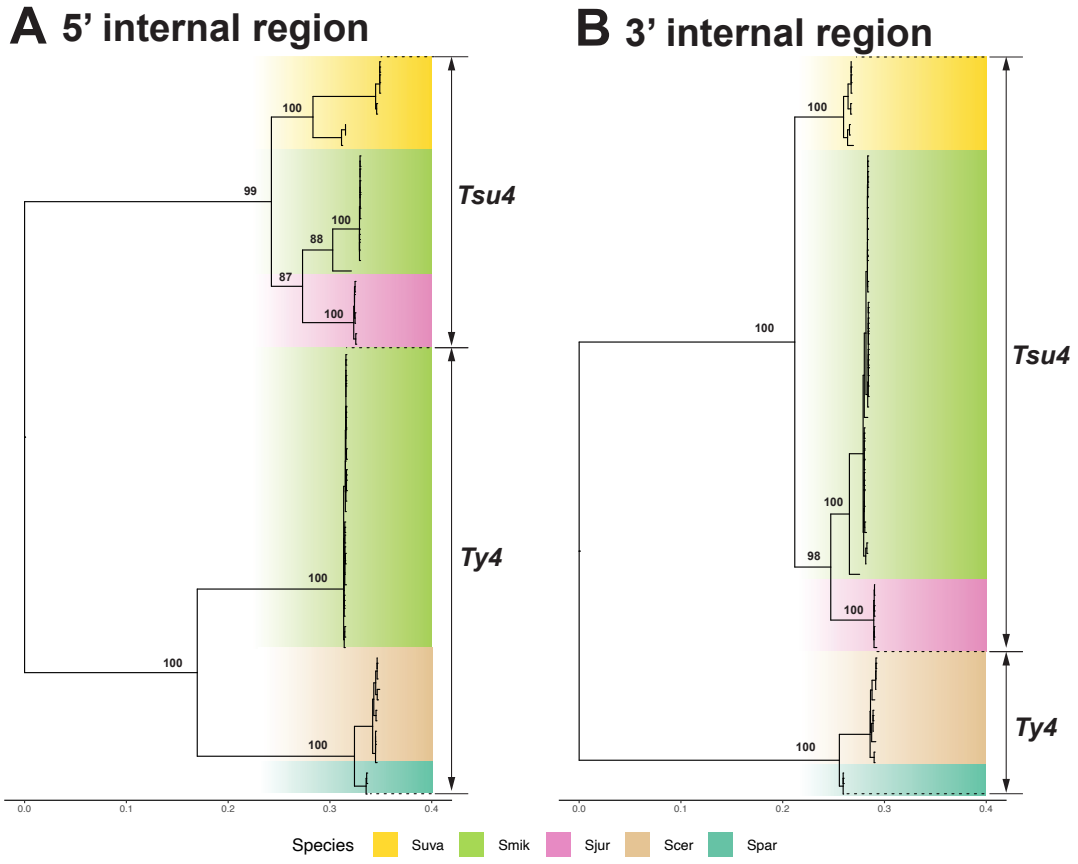

Figure 4.12: **ML phylogeny for 5' and 3' internal regions from Tsu4 full-length elements in *S. mikatae*, *S. jurei*, and *S. uvarum*, plus representatives of Ty4 elements.** Panel (A) shows the ML phylogeny for 5' internal region containing 459 distinct alignment sites; Panel (B) shows 3' internal region containing 455 distinct alignment sites. Both trees are midpoint rooted, and visualized in the same tree scale which is shown in units of substitutions per site. Bootstrap supporting values are annotated for key nodes.

Tsu4 sequences. Together with the results from Ty content scan (Section 4.4.1), we propose the following possible scenario in *S. mikatae*: The Ty4 subfamily was previously active in *S. mikatae*, but subsequently went extinct prior to the speciation of *S. jurei*. A HTT event introduced the Tsu4 subfamily at a time when the Ty4 internal sequences were still present in the *S. mikatae* genome. Recombination between the Ty4 and Tsu4 subfamilies created the "recombinant" Tsu4 clade, which is still active and coexists with the "pure" Tsu4 clade in *S. mikatae* genomes, as shown by their presence in the reference genome of *S. mikatae* (IFO 1815).

4.5 Conclusion

HTT events have been increasing reported by recent studies that involve multiple TE families in yeasts [20, 48, 64]. In this study, we comprehensively investigated evolutionary history of the Ty4 family among *Saccharomyces* species using multiple data mining approaches using both short-read and long-read WGS datasets.

First, we surveyed the Ty4/Tsu4 content from unassembled WGS reads of multiple *Saccharomyces* species using a short-read-based strategy facilitated by McClintock coverage module. This analysis allowed us to investigate the distribution and ancestral states of the Ty4/Tsu4 subfamilies in *Saccharomyces* species. We conclude that the Ty4 subfamily is ancestral in *S. paradoxus* and *S. cerevisiae*, based on its presence across all identified populations in these two species. In contrast, distributions of the Tsu4 subfamily are restricted to specific lineages of *S. paradoxus* and *S. cerevisiae*, suggesting that the Tsu4 subfamily was later introduced into *S. paradoxus*/*S. cerevisiae* from other species *via* HTT. We also identified general presence of the Tsu4 subfamily and near complete absence of the Ty4 subfamily in *S. uvarum* and *S. eubayanus*, which indicates that the Tsu4 subfamily is ancestral in these two species. Our result for *S. mikatae* suggested the presence of both Ty4 and Tsu4 sequences across all strains.

Second, we annotated Ty4/Tsu4 content from a compiled collection of high-quality public WGAs for *Saccharomyces* species, plus two new PacBio assemblies of *S. mikatae* strains generated in this study. Ty4/Tsu4 content inferred from Ty annotation results cross-validated our observations on the distribution of Ty4/Tsu4 content inferred with unassembled short-read scan. Furthermore, Ty annotation of WGAs were capable of generating sequences for each Ty element, which allowed us to infer detailed evolutionary history of the Ty4 family using phylogenetic analysis of full-length Ty4/Tsu4 sequences. We identified Ty4 subfamily sequences only from *S. paradoxus* and *S. cerevisiae* genomes, with monophyletic groups for each species. In contrast, the Tsu4 subfamily was identified in all *Saccharomyces* species investigated in our analysis, with a more complex evolutionary history. In *S. paradoxus*, we identified a possible second HTT of Tsu4 into South American strains in addition to the initial HTT which introduced the Tsu4 subfamily into American *S. paradoxus* reported by Bergman [20]; In *S. cerevisiae*, we found two separate HTT events of Tsu4 from *S. paradoxus* into *S. cerevisiae*; In *S. eubayanus*, we provided evidence for a potential HTT of a distinct Tsu4 clade into Holarctic strains; In *S. mikatae*, we identified a recombinant Tsu4 clade shaped by HTT of Tsu4 into *S. mikatae* followed by recombination with the Ty4 subfamily.

Overall, this work resolves the distribution and ancestral states of Ty4/Tsu4 subfamilies in global samples of *Saccharomyces* species. Additionally, this work reports unexpectedly frequent HTT events between different yeast species, which further highlights the importance of HTTs in the evolution of Ty elements in *Saccharomyces* [64, 28]. Finally, this work generates new genomic resources for *S. mikatae*, which should benefit future TE evolution studies among the *Saccharomyces* genus.

4.6 Acknowledgements

We thank members of the Bergman and Garfinkel Labs for inspiring discussion and comments along the project. We thank Marissa Howard, Noah Workman, and Magdy Alabady at the University of Georgia Genomics and Bioinformatics Core for assistance with PacBio library

preparation and sequencing, and Georgia Advanced Computing Resource Center for technical supporting and computational resources. This work was funded by the University of Georgia Research Foundation (CMB) and NIH grant R01GM124216 (DJG and CMB).

Chapter 5

Conclusions

5.1 Summary

Transposable elements (TEs) are repetitive DNA sequences and capable of mobilizing and propagating within the host genome. The repetitive nature of TE sequences has caused technical challenges on genome assembly and TE annotation, which motivated the development of many computational methods to detect non-reference TE insertions using unassembled short-read sequencing data. A meta-pipeline called McClintock containing six such TE detection tools was previously developed for easy installation and execution. In Chapter 2, I presented McClintock 2, an updated meta-pipeline integrating 12 TE detectors with improved installation, error handling, speed, stability, and extensibility. Additionally, I developed a reproducible and automated simulation framework for evaluating the performance of component methods in McClintock 2. Based on simulation results modeled with *S. cerevisiae* biological features, I characterized five "best-in-class" methods (RelocaTE2, TEMP, TEMP2, TEBreak and RetroSeq) that had both high precision and recall. Furthermore, I assessed component methods of McClintock 2 in a realistic and population genomic scale using an empirical *S. cerevisiae* WGS dataset of 1,011 isolates. "Best-in-class" methods were able to generate similar results, which cross-validated simulation conclusions. Moreover, "best-in-class" components revealed reasonable Ty integration preferences that distributed in a dayd pattern at nucleosome-bound regions upstream of yeast tRNA genes for Ty1, Ty2 and Ty4. All together, this work serves as a practical tool that promotes the selection of optimal non-reference TE detectors in a diversity of organisms, and provides novel biological insights about TEs in one of the best-studied model species.

The *Saccharomyces* genus, especially *S. cerevisiae* and its sister species *S. paradoxus* are important model organisms for TE biology studies on mobility and evolution. However, the complete genome sequences for most species other than the well-studied model species *S. cerevisiae* were not available at the beginning of my thesis research. In Chapter 3, I generated PacBio long-read sequencing data and first high-quality *de novo* genome assemblies for two essential *Saccharomyces* strains: (i) A lab strain of *S. paradoxus* DG1768 that has been widely used in Ty mobility experiments; and (ii) The type strain of *S. uvarum* CBS 7001 that has been fully-sequenced whereas current assemblies are fragmented. This work also documents the complete provenance of *S. paradoxus* strain DG1768 and shows the power of long-read assemblies in facilitating TE annotation, which together provide genomic resources for future Ty biology studies in *Saccharomyces* genus.

Horizontal TE transfer (HTT) events have been increasing reported by recent studies to play an important role in the evolution of Ty sub-families in *Saccharomyces* species. A previous study confirmed a HTT event involving Ty4 family which introduced Tsu4 from *S. uvarum* to *S. paradoxus*. However, the detailed evolutionary history of this HTT event or the Ty4/Tsu4 content in general among *Saccharomyces* species are poorly understood. In Chapter 4, I resolved the ancestral states of the Ty4 subfamilies for multiple *Saccharomyces* species using the coverage-based strategy developed in Chapter 2, which revealed that the Ty4 subfamily was ancestral to *S. cerevisiae* and *S. paradoxus* clade and that Tsu4 was evolved prior to he divergence of *S. uvarum* and *S. eubayanus*. To discover the underlying mechanism and history of Ty4 family evolution, I integrated currently available genome assemblies for *Saccharomyces* species (including two strains generated in Chapter 3 and two new *S. mikatae* assemblies in this Chapter) and annotated TE contents from the genome. Results indicated a complicated evolutionary history of Ty4 family among *Saccharomyces* species involving multiple HTT events and recombinations between sub-families: (i) Multiple HTT events of Tsu4 from *S. paradoxus* to *S. cerevisiae*; (ii) HTT of Tsu4 from unknown donor to *S. paradoxus* and *S. eubayanus*; and (iii) the coexistence and recombination of Ty4

and Tsu4 sub-families in *S. mikatae*. Overall, this work generates new genomic resources for *S. mikatae*, resolves the history of the Ty4 family, and recovers unexpected frequent HTT events and gene flow among *Saccharomyces* species.

5.2 Future work

In Chapter 2, I introduced an updated meta-pipeline of polymorphic TE detectors called McClintock 2, which made progress towards easy installation and execution, stability, extensibility and better result reports. Despite these improvements, this work can be further developed in the following directions. Firstly, it's not surprising that better algorithms for TE detection will give rise to more accurate tools, which can be incorporated into the meta-pipeline in the future. Our McClintock 2 meta-pipeline re-implemented in Python leveraging Snakemake promotes future extensibility, and provides substrate for integrating new component methods. For example, our team has considered to add Jitterbug [118] and TEPID [262] into the meta-pipeline. Next, although the installation process of McClintock 2 has been automated using Conda, dependency conflict issues are still possible due to occasional changes on versions of dependencies made by Bioconda community. Thus, future improvements can be made on creating a more robust computing environment using static containerized virtual environments like Docker or Singularity.

In addition, I presented a reproducible and automated simulation framework for McClintock 2 meta-pipeline which helps users to identify the optimized component for different purposes in Chapter 2. Currently, the simulation framework only supports a "single insertion" mode, where only one synthetic insertion can be created and inserted into the reference genome. Although candidate insertion sites could be restricted by users, "single insertion" simulation is still a over-simplified model compared to realistic biology. In future works, the simulation framework can be extended that allows users to specify the number of synthetic insertions to create in one genome and the expected distribution of new insertions. Such improvement should benefit different custom simulation configuration and thus increase the

confidence of the simulation results. Besides, in my analysis (Chapter 2), the evaluation of optimized TE detector was only performed with one paradigm, the *S. cerevisiae* model. More evaluations in other model organisms with larger genome sizes, such as plants, animals or human, can be implemented in future works, which would provide useful guidance for TE studies on other model species.

In Chapter 3, 4 and several recent studies [291, 212], the utilization of long-read sequencing technology has strikingly improved the continuity and completeness of whole genome assemblies for *Saccharomyces* species. While short-read strategies only allow TE location identification and/or TE abundance estimation, long-read genome assemblies provide substrate for complete TE annotation so that sequences for all TE elements in the genome are retrievable. However, higher cost and sequencing error rates are the two major limitations of long-read strategies that needs to be resolved in future works. The most recent development of PacBio HiFi sequencing has greatly improved the accuracy of long-read sequencing, served as the cutting-edge yet high-cost approach to resolve genome assembly problems. Another well-known long-read sequencing technology, ONT, is more affordable and supports ultra-long reads, however the sequencing error rates for ONT are higher than PacBio and short-read strategies, which may be improved with better base-calling algorithms in the near future. With the development of lower cost for sequencing and better computational algorithms for base-calling, we should expect an increasingly amount of telomere-to-telomere genome assemblies in yeast and other species for TE studies in the near future.

Further sampling and sequencing of *Saccharomyces* yeasts in a world-wide scale is essential to facilitate population genetic studies in the future. Current sampling for *Saccharomyces* species are highly biased. For instance, the majority of *S. cerevisiae* isolates are associated with fermentation industries especially the wine production in Europe, whereas wild *S. cerevisiae* strains are poorly isolated compared to human associated ones. In general, *Saccharomyces* isolates are biased to geographic regions including North America, Europe

or Far East Asia, where most related research are conducted [220]. *S. paradoxus* have been firstly heavily sampled from several locations in North America, Europe and then in China. However, the two interesting populations revealed in Chapter 4 – South American and Hawaiian *S. paradoxus* – are less available with only one or two representatives. Isolation and sequencing of more strains from divergent populations – especially for those with very few representatives – should benefit future studies on the evolution of host species and TEs. In Chapter 4, I put forward several putative horizontal Tsu4 transfers from hidden donor populations to Holarctic *S. eubayanus* and South American *S. paradoxus*. The unknown donors in such HTT events could be potentially resolved by whole genome sequencing of the West China *S. eubayanus* lineage [26] or future characterization of other related populations. Currently characterized distribution of *S. jurei*, *S. mikatae*, *S. kudriavzevii* and *S. arboricola* are restricted to certain geographical locations, which awaits future works to reveal biological questions including their population structure, genetic diversity, and TE evolution.

The quantity of high-quality genome assemblies for *Saccharomyces* species is insufficient to reveal a comprehensive evolutionary history of TE families. Only the best-studied model species *S. cerevisiae* is extensively sequenced with long-read strategy in a recent study [212]. Complete genome assemblies for *S. paradoxus* strains are available for one representative of each population [80, 291]. For other less-studied or newly-characterized *Saccharomyces* species, only the type strain and/or few strains from different population have been assembled completely (summarized in Table 1.2). My dissertation moves a few steps forward by generating long-read assemblies for *S. paradoxus*, *S. uvarum* and *S. mikatae* strains. The more long-read genome assemblies and TE annotations are performed, the more Ty subfamilies could be identified and their evolutionary history can be more accurately traced.

Bibliography

[1] SRAtools.

[2] Jeffrey R. Adrion, Michael J. Song, Daniel R. Schrider, Matthew W. Hahn, and Sarah Schaack. Genome-wide estimates of transposable element insertion and deletion rates in Drosophila melanogaster. *Genome Biol Evol*, 9(5):1329–1340, May 2017.

[3] Pedro Almeida, Raquel Barbosa, Polona Zalar, Yumi Imanishi, Kiminori Shimizu, Benedetta Turchetti, Jean-Luc Legras, Marta Serra, Sylvie Dequin, Arnaud Couloux, Julie Guy, Douda Bensasson, Paula Goncalves, and Jose Paulo Sampaio. A population genomics insight into the Mediterranean origins of wine yeast domestication. *Mol Ecol*, 24(21):5412–5427, November 2015.

[4] Pedro Almeida, Carla Goncalves, Sara Teixeira, Diego Libkind, Martin Bontrager, Isabelle Masneuf-Pomarede, Warren Albertin, Pascal Durrens, David James Sherman, Philippe Marullo, Chris Todd Hittinger, Paula Goncalves, and Jose Paulo Sampaio. A Gondwanan imprint on global diversity and domestication of wine and cider yeast Saccharomyces uvarum. *Nat Commun*, 5:4044, June 2014.

[5] Michael Alonge, Sebastian Soyk, Srividya Ramakrishnan, Xingang Wang, Sara Goodwin, Fritz J. Sedlazeck, Zachary B. Lippman, and Michael C. Schatz. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol*, 20(1):224, October 2019.

[6] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–10, October 1990.

[7] Stefan G. Amyotte, Xiaoping Tan, Kayla Pennerman, Maria del Mar Jimenez-Gasco, Steven J. Klosterman, Li-Jun Ma, Katherine F. Dobinson, and Paola Veronese. Transposable elements in phytopathogenic Verticillium spp.: insights into genome evolution and inter- and intra-specific diversification. *BMC Genomics*, 13:314, July 2012.

[8] Amna Asif-Laidin, Christine Conesa, Amandine Bonnet, Camille Grison, Indranil Adhya, Rachid Menouni, Hélène Fayol, Noé Palmic, Joël Acker, and Pascale Lesage. A small targeting domain in Ty1 integrase is sufficient to direct retrotransposon integration upstream of tRNA genes. *The EMBO Journal*, 39(17):e104337, September 2020.

[9] Marc Bailly-Bechet, Annabelle Haudry, and Emmanuelle Lerat. "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. *Mob DNA*, 5(1):13, May 2014.

[10] EmilyClare Baker, Bing Wang, Nicolas Bellora, David Peris, Amanda Beth Hulfachor, Justin A. Koshalek, Marie Adams, Diego Libkind, and Chris Todd Hittinger. The genome sequence of Saccharomyces eubayanus and the domestication of lager-brewing yeasts. *Mol Biol Evol*, 32(11):2818–2831, November 2015.

[11] Joshua A. Baller, Jiquan Gao, Radostina Stamenova, M. Joan Curcio, and Daniel F. Voytas. A nucleosomal surface defines an integration hotspot for the Saccharomyces cerevisiae Ty1 retrotransposon. *Genome Res*, 22(4):704–713, April 2012.

[12] Joshua A. Baller, Jiquan Gao, and Daniel F. Voytas. Access to DNA establishes a secondary target site bias for the yeast retrotransposon Ty5. *Proc Natl Acad Sci USA*, 108(51):20351–20356, December 2011.

[13] Zhirong Bao and S R Eddy. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res*, 12(8):1269–76, August 2002.

[14] Raquel Barbosa, Pedro Almeida, Silvana V. B. Safar, Renata Oliveira Santos, Paula B. Morais, Lou Nielly-Thibault, Jean-Baptiste Leducq, Christian R. Landry, Paula Goncalves, Carlos A. Rosa, and Jose Paulo Sampaio. Evidence of natural hybridization in Brazilian wild lineages of Saccharomyces cerevisiae. *Genome Biol Evol*, 8(2):317–329, February 2016.

[15] Raquel Barbosa, Ana Pontes, Renata O. Santos, Gabriela G. Montandon, Camila M. de Ponzzes-Gomes, Paula B. Morais, Paula Gonçalves, Carlos A. Rosa, and José Paulo Sampaio. Multiple rounds of artificial selection promote microbe secondary domestication—the case of cachaça yeasts. *Genome Biol Evol*, 10(8):1939–1955, August 2018.

[16] Arianna Basile, Fabio De Pascale, Federico Bianca, Alessandro Rossi, Martina Frizzarin, Nicola De Bernardini, Matteo Bosaro, Anna Baldisseri, Paolo Antoniali, Raffaele Lopreiato, Laura Treu, and Stefano Campanaro. Large-scale sequencing and comparative analysis of oenological Saccharomyces cerevisiae strains supported by nanopore refinement of key genomes. *Food Microbiol*, 97:103753, August 2021.

[17] Jens Bast, Kamil S Jaron, Donovan Schuseil, Denis Roze, and Tanja Schwander. Asexual reproduction reduces transposable element load in experimental yeast populations. *eLife*, 8:e48548, September 2019.

[18] Devin P Bendixsen, Noah Gettle, Ciaran Gilchrist, Zebin Zhang, and Rike Stelkens. Genomic Evidence of an Ancient East Asian Divergence Event in Wild Saccharomyces cerevisiae. *Genome Biol Evol*, 13(2):evab001, February 2021.

[19] Sean A Bergin, Stephen Allen, Conor Hession, Eoin O Cinneide, Adam Ryan, Kevin P Byrne, Tadhg O Croinín, Kenneth H Wolfe, and Geraldine Butler. Identification of European isolates of the lager yeast parent Saccharomyces eubayanus. *FEMS Yeast Res*, 22(1):foac053, December 2022.

[20] Casey M. Bergman. Horizontal transfer and proliferation of Tsu4 in Saccharomyces paradoxus. *Mob DNA*, 9:18, June 2018.

[21] Casey M. Bergman, Shunhua Han, Michael G. Nelson, Vladyslav Bondarenko, and Iryna Kozeretska. Genomic analysis of P elements in natural populations of Drosophila melanogaster. *PeerJ*, 5:e3824, September 2017.

[22] Casey M. Bergman and Hadi Quesneville. Discovering and detecting transposable elements in genome sequences. *Brief Bioinformatics*, 8(6):382–92, November 2007.

[23] Anders Bergstrom, Jared T. Simpson, Francisco Salinas, Benjamin Barre, Leopold Parts, Amin Zia, Nguyen Ba, Alex N, Alan M. Moses, Edward J. Louis, Ville Mustonen, Jonas Warringer, Richard Durbin, and Gianni Liti. A high-definition view of functional genetic variation from natural yeast genomes. *Mol Biol Evol*, 31(4):872–888, April 2014.

[24] Christian Biemont, Laurence Monti-Dedieu, and Francoise Lemeunier. Detection of transposable elements in Drosophila salivary gland polytene chromosomes by in situ hybridization. *Methods Mol Biol*, 260:21–28, 2004.

[25] Frederic Bigey, Diego Segond, Anne Friedrich, Stephane Guezenec, Aurélie Bourgais, Lucie Huyghe, Nicolas Agier, Thibault Nidelet, and Delphine Sicard. Evidence for two main domestication trajectories in saccharomyces cerevisiae linked to distinct bread-making processes. *Curr Biol*, 31(4):722–732.e5, February 2021.

[26] Jian Bing, Pei-Jie Han, Wan-Qiu Liu, Qi-Ming Wang, and Feng-Yan Bai. Evidence for a Far East Asian origin of lager beer yeast. *Curr Biol*, 24(10):R380–381, May 2014.

[27] Victoria M. Blanc and Julian Adams. Ty1 insertions in intergenic regions of the genome of Saccharomyces cerevisiae transcribed by RNA polymerase III have no detectable selective effect. *FEMS Yeast Res*, 4(4-5):487–491, January 2004.

[28] Claudine Bleykasten-Grosshans, Romeo Fabrizio, Anne Friedrich, and Joseph Schacherer. Species-wide transposable element repertoires retrace the evolutionary history of the Saccharomyces cerevisiae host. *Mol Biol Evol*, 38(10):4334–4345, October 2021.

[29] Claudine Bleykasten-Grosshans, Anne Friedrich, and Joseph Schacherer. Genome-wide analysis of intraspecific transposon diversity in yeast. *BMC Genomics*, 14:399, 2013.

[30] J D Boeke, D Eichinger, D Castrillon, and G R Fink. The Saccharomyces cerevisiae genome contains functional and nonfunctional copies of transposon Ty1. *Mol Cell Biol*, 8(4):1432–1442, April 1988.

[31] J. D. Boeke, D. J. Eichinger, and G. Natsoulis. Doubling Ty1 element copy number in Saccharomyces cerevisiae: host genome stability and phenotypic effects. *Genetics*, 129(4):1043–1052, December 1991.

[32] J. D. Boeke, D. J. Garfinkel, C. A. Styles, and G. R. Fink. Ty elements transpose through an RNA intermediate. *Cell*, 40(3):491–500, March 1985.

[33] Jef D. Boeke. Retrotransposons of fungi. *Encyclopedia of Virology*, pages 1503–1507, 1999.

[34] Elisabeth Bon, Cécile Neuvéglise, Serge Casaregola, François Artiguenave, Patrick Wincker, Michel Aigle, and Pascal Durrens. Genomic Exploration of the Hemiascomycetous Yeasts: 5. Saccharomyces bayanus var. uvarum. *FEBS Letters*, 487(1):37–41, 2000. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1016/S0014-5793%2800%2902276-6.

[35] Amandine Bonnet and Pascale Lesage. Light and shadow on the mechanisms of integration site selection in yeast Ty retrotransposon families. *Curr Genet*, 67(3):347–357, June 2021.

[36] Anthony R. Borneman, Angus H. Forgan, Radka Kolouchova, James A. Fraser, and Simon A. Schmidt. Whole genome comparison reveals high levels of inbreeding and strain redundancy across the spectrum of commercial wine strains of Saccharomyces cerevisiae. *G3*, 6(4):957–971, April 2016.

[37] Anthony R. Borneman and Isak S. Pretorius. Genomic insights into the Saccharomyces sensu stricto complex. *Genetics*, 199(2):281–291, February 2015.

[38] Guillaume Bourque, Kathleen H. Burns, Mary Gehring, Vera Gorbunova, Andrei Seluanov, Molly Hammell, Michaël Imbeault, Zsuzsanna Izsvák, Henry L. Levin, Todd S. Macfarlan, Dixie L. Mager, and Cédric Feschotte. Ten things you should know about transposable elements. *Genome Biol*, 19(1):199, November 2018.

[39] Primrose J. Boynton and Duncan Greig. The ecology and evolution of non-domesticated Saccharomyces species. *Yeast*, 31(12):449–462, 2014.

[40] Anja Brickwedde, Nick Brouwers, Marcel van den Broek, Joan S. Gallego Murillo, Julie L. Fraiture, Jack T. Pronk, and Jean-Marc G. Daran. Structural, Physiological and Regulatory Analysis of Maltose Transporter Genes in Saccharomyces eubayanus CBS 12357T. *Front Microbiol*, 9, 2018.

[41] Antoine Bridier-Nahmias and Pascale Lesage. Two large-scale analyses of Ty1 LTR-retrotransposon de novo insertion events indicate that Ty1 targets nucleosomal DNA near the H2A/H2B interface. *Mob DNA*, 3(1):22, December 2012.

[42] Antoine Bridier-Nahmias, Aurélie Tchalikian-Cosson, Joshua A. Baller, Rachid Menouni, Hélène Fayol, Amando Flores, Ali Saïb, Michel Werner, Daniel F. Voytas, and Pascale Lesage. An RNA polymerase III subunit determines sites of retrotransposon integration. *Science*, 348(6234):585–588, May 2015.

[43] Nick Brouwers, Anja Brickwedde, Arthur R. Gorter de Vries, Marcel van den Broek, Susan M. Weening, Lieke van den Eijnden, Jasper A. Diderich, Feng-Yan Bai, Jack T.

Pronk, and Jean-Marc G. Daran. Himalayan Saccharomyces eubayanus genome sequences reveal genetic markers explaining heterotic maltotriose consumption by Saccharomyces pastorianus hybrids. *Appl Environ Microbiol*, 85(22):e01516–19, November 2019.

[44] Samuel D. J. Brown, Rupert A. Collins, Stephane Boyer, Marie-Caroline Lefort, Jagoba Malumbres-Olarte, Cor J. Vink, and Robert H. Cruickshank. Spider: An R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Mol Ecol Resour*, 12(3):562–565, 2012.

[45] W. D. Burke, C. C. Calalang, and T. H. Eickbush. The site-specific ribosomal insertion element type II of Bombyx mori (R2Bm) contains the coding sequence for a reverse transcriptase-like enzyme. *Mol Cell Biol*, 7(6):2221–2230, June 1987.

[46] Brian Bushnell. BBMap: a fast, accurate, splice-aware aligner. Technical Report LBNL-7065E, Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States), March 2014.

[47] J. R. Cameron, E. Y. Loh, and R. W. Davis. Evidence for transposition of dispersed repetitive DNA families in yeast. *Cell*, 16(4):739–751, April 1979.

[48] Martin Carr, Douda Bensasson, and Casey M. Bergman. Evolutionary genomics of transposable elements in Saccharomyces cerevisiae. *PLOS ONE*, 7(11):e50978, November 2012.

[49] D. L. Chalker and S. B. Sandmeyer. Transfer RNA genes are genomic targets for de novo transposition of the yeast retrotransposon Ty3. *Genetics*, 126(4):837–850, December 1990.

[50] D. L. Chalker and S. B. Sandmeyer. Ty3 integrates within the region of RNA polymerase III transcription initiation. *Genes Dev*, 6(1):117–128, January 1992.

[51] Jinfeng Chen, Travis R. Wrightsman, Susan R. Wessler, and Jason E. Stajich. RelocaTE2: a high resolution transposable element insertion site mapping tool for population resequencing. *PeerJ*, 5:e2942, 2017.

[52] Jingxuan Chen, Preston J. Basting, Shunhua Han, David J. Garfinkel, and Casey M. Bergman. Reproducible evaluation of transposable element detectors with McClintock 2 guides accurate inference of Ty insertion patterns in yeast. *Mob DNA*, 14(1):8, July 2023.

[53] Jingxuan Chen, David J. Garfinkel, and Casey M. Bergman. Long-read genome assembly of Saccharomyces uvarum strain CBS 7001. *Microbiol Resour Announc*, 11(1):e00972–21, January 2022.

[54] Jingxuan Chen, Holly McQueary, David W. Hall, Peter Philippsen, David J. Garfinkel, and Casey M. Bergman. Genome assembly of the Ty1-less Saccharomyces paradoxus strain DG1768. *Microbiol Resour Announc*, 11(1):e0086821, January 2022.

[55] Piaopiao Chen and Jianzhi Zhang. Asexual experimental evolution of yeast does not curtail transposable elements. *Mol Biol Evol*, 38(7):2831–2842, July 2021.

[56] Chen-Shan Chin, David H. Alexander, Patrick Marks, Aaron A. Klammer, James Drake, Cheryl Heiner, Alicia Clum, Alex Copeland, John Huddleston, Evan E. Eichler, Stephen W. Turner, and Jonas Korlach. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods*, 10(6):563–569, June 2013.

[57] Chong Chu, Rebeca Borges-Monroy, Vinayak V. Viswanadham, Soohyun Lee, Heng Li, Eunjung Alice Lee, and Peter J. Park. Comprehensive identification of transposable element insertions using multiple sequencing technologies. *Nat Commun*, 12(1):3836, June 2021.

[58] J. J. Clare, M. Belcourt, and P. J. Farabaugh. Efficient translational frameshifting occurs within a conserved sequence of the overlap between the two genes of a yeast Ty1 transposon. *Proc Natl Acad Sci USA*, 85(18):6816–6820, September 1988.

[59] Paul Cliften, Priya Sudarsanam, Ashwin Desikan, Lucinda Fulton, Bob Fulton, John Majors, Robert Waterston, Barak A. Cohen, and Mark Johnston. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science*, 301(5629):71–76, July 2003.

[60] A. L. Coi, F. Bigey, S. Mallet, S. Marsit, G. Zara, P. Gladieux, V. Galeote, M. Budroni, S. Dequin, and J. L. Legras. Genomic signatures of adaptation to wine biological ageing conditions in biofilm-forming flor yeasts. *Mol Ecol*, 26(7):2150–2166, April 2017.

[61] Jake R Conway, Alexander Lex, and Nils Gehlenborg. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33(18):2938–2940, September 2017.

[62] Nancy L. Craig. *Mobile DNA II*. ASM Press, Washington, D.C., 2002.

[63] M. Joan Curcio, Sheila Lutz, and Pascale Lesage. The Ty1 LTR-retrotransposon of budding yeast, Saccharomyces cerevisiae. *Microbiol Spectr*, 3(2):1–35, April 2015.

[64] Wioletta Czaja, Douda Bensasson, Hyo Won Ahn, David J. Garfinkel, and Casey M. Bergman. Evolution of Ty1 copy number control in yeast by horizontal transfer and recombination. *PLOS Genetics*, 16(2):e1008632, February 2020.

[65] Marie-Josee Daboussi, Jean-Michel Daviere, Stephane Graziani, and Thierry Langin. Evolution of the Fot1 transposons in the genus Fusarium: discontinuous distribution and epigenetic inactivation. *Mol Biol Evol*, 19(4):510–520, April 2002.

[66] Austin T Daigle, Thaddeus C Deiss, Robert H Melde, Ulfar Bergthorsson, and Vaishali Katju. Bergerac strains of Caenorhabditis elegans revisited: expansion of Tc1 elements imposes a significant genomic and fitness cost. *G3*, 12(11):jkac214, November 2022.

[67] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, and Richard Durbin. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, August 2011.

[68] S. B. Daniels, K. R. Peterson, L. D. Strausbaugh, M. G. Kidwell, and A. Chovnick. Evidence for horizontal transmission of the P transposable element between Drosophila species. *Genetics*, 124(2):339–55, 1990.

[69] Wouter De Coster, Matthias H. Weissensteiner, and Fritz J. Sedlazeck. Towards population-scale long-read sequencing. *Nat Rev Genet*, 22(9):572–587, September 2021.

[70] F J del Rey, T F Donahue, and G R Fink. sigma, a repetitive element found adjacent to tRNA genes of yeast. *Proc Natl Acad Sci USA*, 79(13):4138–4142, July 1982.

[71] K. F. Dobinson, R. E. Harris, and J. E. Hamer. Grasshopper, a long terminal repeat (LTR) retroelement in the phytopathogenic fungus Magnaporthe grisea. *Mol Plant Microbe Interact*, 6(1):114–126, February 1993.

[72] Bruno Reis Dotto, Evelise Leis Carvalho, Alexandre Freitas da Silva, Filipe Zimmer Dezordi, Paulo Marcos Pinto, Tulio de Lima Campos, Antonio Mauro Rezende, and Gabriel da Luz Wallau. HTT-DB: new features and updates. *Database (Oxford)*, 2018, January 2018.

[73] Bruno Reis Dotto, Evelise Leis Carvalho, Alexandre Freitas Silva, Duarte Silva, Luiz Fernando, Paulo Marcos Pinto, Mauro Freitas Ortiz, and Gabriel Luz Wallau. HTT-DB: Horizontally transferred transposable elements database. *Bioinformatics*, 31(17):2915–2917, September 2015.

[74] Polina B. Drozdova, Oleg V. Tarasov, Andrew G. Matveenko, Elina A. Radchenko, Julia V. Sopova, Dmitrii E. Polev, Sergey G. Inge-Vechtomov, and Pavel V. Dobrynin.

Genome Sequencing and Comparative Analysis of Saccharomyces cerevisiae Strains of the Peterhof Genetic Collection. *PLOS ONE*, 11(5):e0154722, 2016.

[75] Shou-Fu Duan, Pei-Jie Han, Qi-Ming Wang, Wan-Qiu Liu, Jun-Yan Shi, Kuan Li, Xiao-Ling Zhang, and Feng-Yan Bai. The origin and adaptive evolution of domesticated populations of yeast from Far East Asia. *Nat Commun*, 9(1):2690, July 2018.

[76] Celine Duc, Marianne Yoth, Silke Jensen, Nolwenn Mouniee, Casey M. Bergman, Chantal Vaury, and Emilie Brasset. Trapping a somatic endogenous retrovirus into a germline piRNA cluster immunizes the germline against further invasion. *Genome Biol*, 20(1):127, June 2019.

[77] Barbara Dunn and Gavin Sherlock. Reconstruction of the genome origins and evolution of the hybrid lager yeast Saccharomyces pastorianus. *Genome Res*, 18(10):1610–1623, October 2008.

[78] S. S. Dwight, R. Balakrishnan, K. R. Christie, M. C. Costanzo, K. Dolinski, S. R. Engel, B. Feierbach, D. G. Fisk, J. Hirschman, E. L. Hong, L. Issel-Tarver, R. S. Nash, A. Sethuraman, B. Starr, C. L. Theesfeld, R. Andrada, G. Binkley, Q. Dong, C. Lane, M. Schroeder, S. Weng, D. Botstein, and J. M. Cherry. Saccharomyces genome database: underlying principles and organisation. *Brief Bioinform*, 5(1):9–22, 2004.

[79] Melania D'Angiolo, Matteo De Chiara, Jia-Xing Yue, Agurtzane Irizar, Simon Stenberg, Karl Persson, Agnès Llored, Benjamin Barré, Joseph Schacherer, Roberto Marangoni, Eric Gilson, Jonas Warringer, and Gianni Liti. A yeast living ancestor reveals the origin of genomic introgressions. *Nature*, pages 1–6, November 2020.

[80] Chris Eberlein, Mathieu Henault, Anna Fijarczyk, Guillaume Charron, Matteo Bouvier, Linda M. Kohn, James B. Anderson, and Christian R. Landry. Hybridization is a recurrent evolutionary stimulus in wild yeast speciation. *Nat Commun*, 10(1):923, February 2019.

[81] David Ellinghaus, Stefan Kurtz, and Ute Willhoeft. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, 9(1):18, January 2008.

[82] Tyler A. Elliott, Tony Heitkam, Robert Hubley, Hadi Quesneville, Alexander Suh, Travis J. Wheeler, and The TE Hub Consortium. TE Hub: A community-oriented space for sharing and connecting tools, data, resources, and methods for transposable element annotation. *Mob DNA*, 12(1):16, June 2021.

[83] Stacia R. Engel, Fred S. Dietrich, Dianna G. Fisk, Gail Binkley, Rama Balakrishnan, Maria C. Costanzo, Selina S. Dwight, Benjamin C. Hitz, Kalpana Karra, Robert S. Nash, Shuai Weng, Edith D. Wong, Paul Lloyd, Marek S. Skrzypek, Stuart R. Miyasato, Matt Simison, and J. Michael Cherry. The Reference Genome Sequence of Saccharomyces cerevisiae: Then and Now. *G3*, 4(3):389–398, December 2013.

[84] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, October 2016.

[85] Adam D. Ewing. Transposable element detection from whole genome sequence data. *Mob DNA*, 6:24, 2015.

[86] Justin C. Fay, Ping Liu, Giang T. Ong, Maitreya J. Dunham, Gareth A. Cromie, Eric W. Jeffery, Catherine L. Ludlow, and Aimee M. Dudley. A polyploid admixed origin of beer yeasts derived from European and Asian wine populations. *PLOS Biology*, 17(3):e3000147, March 2019.

[87] Ethan G. Fingerman, Peter G. Dombrowski, Chantal A. Francis, and Paul D. Sniegowski. Distribution and sequence analysis of a novel Ty3-like element in natural Saccharomyces paradoxus isolates. *Yeast*, 20(9):761–70, July 2003.

[88] David J. Finnegan. Eukaryotic transposable elements and genome evolution. *Trends in Genetics*, 5:103–107, January 1989.

[89] Jullien M. Flynn, Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G. Clark, Cédric Feschotte, and Arian F. Smit. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA*, 117(17):9451–9457, April 2020.

[90] J. Gafner and P. Philippsen. The yeast transposon Ty1 generates duplications of target DNA on insertion. *Nature*, 286(5771):414–418, July 1980.

[91] X. Gai and D. F. Voytas. A single amino acid change in the yeast retrotransposon Ty5 abolishes targeting to silent chromatin. *Mol Cell*, 1(7):1051–1055, June 1998.

[92] Brigida Gallone, Jan Steensels, Stijn Mertens, Maria C. Dzialo, Jonathan L. Gordon, Ruben Wauters, Florian A. Theßeling, Francesca Bellinazzo, Veerle Saels, Beatriz Herrera-Malaver, Troels Prahl, Christopher White, Mathias Hutzler, Franz Meußdoerffer, Philippe Malcorps, Ben Souffriau, Luk Daenen, Guy Baele, Steven Maere, and Kevin J. Verstrepen. Interspecific hybridization facilitates niche adaptation in beer yeast. *Nat Ecol Evol*, pages 1–14, October 2019.

[93] Brigida Gallone, Jan Steensels, Troels Prahl, Leah Soriaga, Veerle Saels, Beatriz Herrera-Malaver, Adriaan Merlevede, Miguel Roncoroni, Karin Voordeckers, Loren Miraglia, Clotilde Teiling, Brian Steffy, Maryann Taylor, Ariel Schwartz, Toby Richardson, Christopher White, Guy Baele, Steven Maere, and Kevin J. Verstrepen. Domestication and divergence of Saccharomyces cerevisiae beer yeasts. *Cell*, 166(6):1397–1410.e16, September 2016.

[94] Eugene J. Gardner, Vincent K. Lam, Daniel N. Harris, Nelson T. Chuang, Emma C. Scott, W. Stephen Pittard, Ryan E. Mills, The 1000 Genomes Project Consortium, and

Scott E. Devine. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res*, 27(11):1916–1929, November 2017.

[95] D. J. Garfinkel, J. D. Boeke, and G. R. Fink. Ty element transposition: reverse transcriptase and virus-like particles. *Cell*, 42(2):507–517, September 1985.

[96] David J. Garfinkel, Katherine Nyswaner, Jun Wang, and Jae-Yong Cho. Post-transcriptional cosuppression of Ty1 retrotransposition. *Genetics*, 165(1):83–99, September 2003.

[97] David J. Garfinkel, Katherine M. Nyswaner, Karen M. Stefanisko, Caroline Chang, and Sharon P. Moore. Ty1 copy number dynamics in Saccharomyces. *Genetics*, 169(4):1845–1857, April 2005.

[98] Velimir Gayevskiy and Matthew R. Goddard. Saccharomyces eubayanus and Saccharomyces arboricola reside in North Island native New Zealand forests. *Environ Microbiol*, 18(4):1137–1147, April 2016.

[99] Velimir Gayevskiy, Soon Lee, and Matthew R. Goddard. European derived Saccharomyces cerevisiae colonisation of New Zealand vineyards aided by humans. *FEMS Yeast Res*, 16(7), November 2016.

[100] Brian Gibson and Gianni Liti. Saccharomyces pastorianus: genomic insights inspiring innovation for industry. *Yeast*, 32(1):17–27, January 2015.

[101] Clément Gilbert and Cédric Feschotte. Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. *Curr Opin Genetics Dev*, 49:15–24, April 2018.

[102] A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. Life with 6000 genes. *Science*, 274(5287):546, 563–7, October 1996.

[103] Margarida Goncalves, Ana Pontes, Pedro Almeida, Raquel Barbosa, Marta Serra, Diego Libkind, Mathias Hutzler, Paula Goncalves, and Jose Paulo Sampaio. Distinct domestication trajectories in top-fermenting beer yeasts and wine yeasts. *Curr Biol*, 26(20):2750–2761, October 2016.

[104] Sara S. Gonzalez, Eladio Barrio, and Amparo Querol. Molecular characterization of new natural hybrids of Saccharomyces cerevisiae and S. kudriavzevii in brewing. *Appl Environ Microbiol*, 74(8):2314–2320, April 2008.

[105] Cooper A. Grace. *Evolutionary Genomics of Transposable Elements in the Saccharomyces sensu lato Complex*. doctoral, University of Huddersfield, September 2018.

[106] D. Greig. Reproductive isolation in Saccharomyces. *Heredity (Edinb)*, 102(1):39–44, January 2009.

[107] Bjorn Gruening, Olivier Sallou, Pablo Moreno, Felipe da Veiga Leprevost, Hervé Ménager, Dan Søndergaard, Hannes Röst, Timo Sachsenberg, Brian O'Connor, Fábio Madeira, Victoria Dominguez Del Angel, Michael R. Crusoe, Susheel Varma, Daniel Blankenberg, Rafael C. Jimenez, and Yasset Perez-Riverol. Recommendations for the packaging and containerizing of bioinformatics software. *F1000Res*, 7:ISCB Comm J–742, March 2019.

[108] Bjorn Gruning, Ryan Dale, Andreas Sjodin, Brad A. Chapman, Jillian Rowe, Christopher H. Tomkins-Tinch, Renan Valieris, and Johannes Koster. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods*, 15(7):475–476, July 2018.

[109] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, April 2013.

[110] Da-Yong Han, Pei-Jie Han, Karl Rumbold, Anbessa Dabassa Koricha, Shou-Fu Duan, Liang Song, Jun-Yan Shi, Kuan Li, Qi-Ming Wang, and Feng-Yan Bai. Adaptive gene content and allele distribution variations in the wild and domesticated populations of Saccharomyces cerevisiae. *Front Microbiol*, 12, 2021.

[111] Shunhua Han, Preston J. Basting, Guilherme B. Dias, Arthur Luhur, Andrew C. Zelhof, and Casey M. Bergman. Transposable element profiles reveal cell line identity and loss of heterozygosity in Drosophila cell culture. *Genetics*, 219(2):iyab113, October 2021.

[112] Shunhua Han, Guilherme B Dias, Preston J Basting, Michael G Nelson, Sanjai Patel, Mar Marzo, and Casey M Bergman. Ongoing transposition in cell culture reveals the phylogeny of diverse Drosophila S2 sublines. *Genetics*, 221(3):iyac077, July 2022.

[113] Shunhua Han, Guilherme B Dias, Preston J Basting, Raghuvir Viswanatha, Norbert Perrimon, and Casey M Bergman. Local assembly of long reads enables phylogenomics of transposable elements in a polyploid cell line. *Nucleic Acids Res*, 50(21):e124, November 2022.

[114] L. J. Hansen, D. L. Chalker, and S. B. Sandmeyer. Ty3, a yeast retrotransposon associated with tRNA genes, has homology to animal retroviruses. *Mol Cell Biol*, 8(12):5245–5256, December 1988.

[115] Michelle Hays, Katja Schwartz, Danica T. Schmidtke, Dimitra Aggeli, and Gavin Sherlock. Paths to adaptation under fluctuating nitrogen starvation: The spectrum of adaptive mutations in Saccharomyces cerevisiae is shaped by transposons and microhomology-mediated recombination. *bioRxiv*, page 2022.05.11.491585, January 2022.

[116] Peng-Yu He, Xu-Qian Shao, Shou-Fu Duan, Da-Yong Han, Kuan Li, Jun-Yan Shi, Ri-Peng Zhang, Pei-Jie Han, Qi-Ming Wang, and Feng-Yan Bai. Highly diverged lineages

of Saccharomyces paradoxus in temperate to subtropical climate zones in China. *Yeast*, 39(1-2):69–82, 2022.

[117] Lucas W Hemmer, Sherif Negm, Xuewen Geng, Cecile Courret, Beatriz Navarro-Dominguez, Iain Speece, Xiaolu Wei, Eddyson Altidor, James Chaffer, John S. Sproul, and Amanda M Larracuente. Centromere-associated retroelement evolution in Drosophila melanogaster reveals an underlying conflict. *bioRxiv*, page 2022.11.25.518008, January 2022.

[118] Elizabeth Henaff, Luís Zapata, Josep M. Casacuberta, and Stephan Ossowski. Jitterbug: somatic and germline transposon insertion detection at single-nucleotide resolution. *BMC Genomics*, 16:768, 2015.

[119] Alison B. Hickman and Fred Dyda. DNA Transposition at Work. *Chem Rev*, 116(20):12758–12784, October 2016.

[120] Markus Hiltunen, Sandra Lorena Ament-Velasquez, Martin Ryberg, and Hanna Johannesson. Stage-specific transposon activity in the life cycle of the fairy-ring mushroom Marasmius oreades. *Proc Natl Acad Sci USA*, 119(46):e2208575119, November 2022.

[121] Chris Todd Hittinger. Saccharomyces diversity and evolution: a budding model genus. *Trends Genet*, 29(5):309–317, May 2013.

[122] Chris Todd Hittinger, Paula Gonçalves, José Paulo Sampaio, Jim Dover, Mark Johnston, and Antonis Rokas. Remarkably ancient balanced polymorphisms in a multi-locus gene network. *Nature*, 464(7285):54–58, March 2010.

[123] Cheng Ran Lisa Huang, Kathleen H. Burns, and Jef D. Boeke. Active Transposition in Genomes. *Annual Review of Genetics*, 46(1):651–675, 2012.

[124] Weichun Huang, Leping Li, Jason R. Myers, and Gabor T. Marth. ART: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, February 2012.

[125] Daniel H. Huson and David Bryant. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*, 23(2):254–267, February 2006.

[126] Katie E. Hyma and Justin C. Fay. Mixing of vineyard and oak-tree ecotypes of Saccharomyces cerevisiae in North American vineyards. *Mol Ecol*, 22(11):2917–2930, June 2013.

[127] Artem A. Ilyin, Sergei S. Ryazansky, Semen A. Doronin, Oxana M. Olenkina, Elena A. Mikhaleva, Evgeny Y. Yakushev, Yuri A. Abramov, Stepan N. Belyakin, Anton V. Ivankin, Alexey V. Pindyurin, Vladimir A. Gvozdev, Mikhail S. Klenov, and Yuri Y. Shevelyov. Piwi interacts with chromatin at nuclear pores and promiscuously binds nuclear transcripts in Drosophila ovarian somatic cells. *Nucleic Acids Res*, 45(13):7666–7680, July 2017.

[128] Artem A. Ilyin, Anastasia D. Stolyarenko, Mikhail S. Klenov, and Yuri Y. Shevelyov. Various modes of HP1a interactions with the euchromatic chromosome arms in Drosophila ovarian somatic cells. *Chromosoma*, 129(3):201–214, December 2020.

[129] Benjamin Istace, Anne Friedrich, Léo d'Agata, Sébastien Faye, Emilie Payen, Odette Beluche, Claudia Caradec, Sabrina Davidas, Corinne Cruaud, Gianni Liti, Arnaud Lemainque, Stefan Engelen, Patrick Wincker, Joseph Schacherer, and Jean-Marc Aury. De novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *Gigascience*, 6(2):1–13, February 2017.

[130] B. Janetzky and L. Lehle. Ty4, a new retrotransposon from Saccharomyces cerevisiae, flanked by tau-elements. *J Biol Chem*, 267(28):19798–19805, October 1992.

[131] H. Ji, D. P. Moore, M. A. Blomberg, L. T. Braiterman, D. F. Voytas, G. Natsoulis, and J. D. Boeke. Hotspots for unselected Ty1 transposition events on yeast chromosome III are near tRNA genes and LTR sequences. *Cell*, 73(5):1007–1018, June 1993.

[132] Jaime Jimenez-Ruiz, Jorge A. Ramirez-Tejero, Noe Fernandez-Pozo, María de la O. Leyva-Perez, Haidong Yan, Raul de la Rosa, Angjelina Belaj, Eva Montes, Mª Oliva Rodriguez-Ariza, Francisco Navarro, Juan Bautista Barroso, Carmen R. Beuzon, Victoriano Valpuesta, Aureliano Bombarely, and Francisco Luque. Transposon activation is a major driver in the genome evolution of cultivated olive trees (Olea europaea L.). *The Plant Genome*, 13(1):e20010, 2020.

[133] I. K. Jordan and J. F. McDonald. Evidence for the role of recombination in the regulatory evolution of Saccharomyces cerevisiae Ty elements. *Journal of Molecular Evolution*, 47(1):14–20, July 1998.

[134] I. K. Jordan and J. F. McDonald. Phylogenetic perspective reveals abundant Ty1/Ty2 hybrid elements in the Saccharomyces cerevisiae genome. *Mol Biol Evol*, 16(3):419–422, March 1999.

[135] I. K. Jordan and J. F. McDonald. Tempo and mode of Ty element evolution in Saccharomyces cerevisiae. *Genetics*, 151(4):1341–51, April 1999.

[136] Kang Kang, Basti Bergdahl, Daniel Machado, Laura Dato, Ting-Li Han, Jun Li, Silas Villas-Boas, Markus J Herrgård, Jochen Förster, and Gianni Panagiotou. Linking genetic, metabolic, and phenotypic diversity among Saccharomyces cerevisiae strains using multi-omics associations. *GigaScience*, 8(4):giz015, April 2019.

[137] V. V. Kapitonov and J. Jurka. Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci USA*, 98(15):8714–9, 2001.

[138] Kazutaka Katoh and Daron M. Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, 30(4):772–780, April 2013.

[139] Thomas M. Keane, Kim Wong, and David J. Adams. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics*, 29(3):389–390, February 2013.

[140] Manolis Kellis, Nick Patterson, Matthew Endrizzi, Bruce Birren, and Eric S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–54, May 2003.

[141] W. J. Kent, A. S. Zweig, G. Barber, A. S. Hinrichs, and D. Karolchik. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, 26(17):2204–2207, September 2010.

[142] W. James Kent. BLAT—The BLAST-Like Alignment Tool. *Genome Res*, 12(4):656–664, April 2002.

[143] J. M. Kim, S. Vanguri, J. D. Boeke, A. Gabriel, and D. F. Voytas. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. *Genome Res*, 8(5):464–478, May 1998.

[144] Jessica C. Kissinger and Jeremy DeBarry. Genome cartography: charting the apicomplexan genome. *Trends in Parasitology*, 27(8):345–354, August 2011.

[145] Ryosuke Kita, Sandeep Venkataram, Yiqi Zhou, and Hunter B. Fraser. High-resolution mapping of cis-regulatory variation in budding yeast. *Proc Natl Acad Sci USA*, 114(50):E10736–E10744, December 2017.

[146] Robert Kofler, Andrea J. Betancourt, and Christian Schlotterer. Sequencing of pooled DNA samples (pool-seq) uncovers complex dynamics of transposable element insertions in Drosophila melanogaster. *PLOS Genet*, 8(1):e1002487, January 2012.

[147] Robert Kofler, Daniel Gomez-Sanchez, and Christian Schlotterer. PoPoolationTE2: comparative population genomics of transposable elements using pool-seq. *Mol Biol Evol*, 33(10):2759–2764, October 2016.

[148] Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin, and Pavel A. Pevzner. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*, 37(5):540–546, May 2019.

[149] Maria Kordyukova, Olesya Sokolova, Valeriya Morgunova, Sergei Ryazansky, Natalia Akulenko, Sergey Glukhov, and Alla Kalmykova. Nuclear Ccr4-Not mediates the degradation of telomeric and transposon transcripts at chromatin in the Drosophila germline. *Nucleic Acids Res*, 48(1):141–156, January 2020.

[150] Johannes Koster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, October 2012.

[151] Vassiliki Koufopanou, Joseph Hughes, Graham Bell, and Austin Burt. The spatial scale of genetic differentiation in a model organism: The wild yeast Saccharomyces paradoxus. *Philos Trans R Soc Lond B Biol Sci*, 361(1475):1941–1946, November 2006.

[152] Vassiliki Koufopanou, Susan Lomas, Olga Pronina, Pedro Almeida, Jose Paulo Sampaio, Timothy Mousseau, Gianni Liti, and Austin Burt. Population size, sex and purifying selection: comparative genomics of two sister taxa of the wild yeast saccharomyces paradoxus. *Genome Biol Evol*, 12(9):1636–1645, September 2020.

[153] Heidi A. Kuehne, Helen A. Murphy, Chantal A. Francis, and Paul D. Sniegowski. Allopatric divergence, secondary contact, and genetic isolation in wild yeast populations. *Curr Biol*, 17(5):407–411, March 2007.

[154] S. Kurtz, J. V. Choudhuri, E. Ohlebusch, C. Schleiermacher, J. Stoye, and R. Giegerich. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res*, 29(22):4633–42, 2001.

[155] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R.

McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[156] Quinn K. Langdon, David Peris, EmilyClare P. Baker, Dana A. Opulente, Huu-Vang Nguyen, Ursula Bond, Paula Gonçalves, José Paulo Sampaio, Diego Libkind, and Chris Todd Hittinger. Fermentation innovation through complex hybridization of wild and domesticated yeasts. *Nat Ecol Evol*, pages 1–11, October 2019.

[157] Quinn K. Langdon, David Peris, Juan I. Eizaguirre, Dana A. Opulente, Kelly V. Buh, Kayla Sylvester, Martin Jarzyna, María E. Rodríguez, Christian A. Lopes, Diego Libkind, and Chris Todd Hittinger. Postglacial migration shaped the genomic diversity and global distribution of the wild ancestor of lager-brewing hybrids. *PLOS Genetics*, 16(4):e1008680, April 2020.

[158] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, March 2009.

[159] Jean-Baptiste Leducq, Guillaume Charron, Pedram Samani, Alexandre K. Dube, Kayla Sylvester, Brielle James, Pedro Almeida, Jose Paulo Sampaio, Chris Todd Hittinger, Graham Bell, and Christian R. Landry. Local climatic adaptation in a widespread microorganism. *Proc Biol Sci*, 281(1777), February 2014.

[160] Jean-Baptiste Leducq, Lou Nielly-Thibault, Guillaume Charron, Chris Eberlein, Jukka-Pekka Verta, Pedram Samani, Kayla Sylvester, Chris Todd Hittinger, Graham Bell, and Christian R. Landry. Speciation driven by hybridization and chromosomal plasticity in a wild yeast. *Nature Microbiology*, 1(1):15003, January 2016.

[161] Tracy J. Lee, Yuching Liu, Wei-An Liu, Yu-Fei Lin, Hsin-Han Lee, Huei-Mien Ke, Jen-Pen Huang, Mei-Yeh Jade Lu, Chia-Lun Hsieh, Kuo-Fang Chung, Gianni Liti, and Isheng J. Tsai. Extensive sampling of Saccharomyces cerevisiae in Taiwan reveals ecology and evolution of predomesticated lineages. *Genome Res*, page gr.276286.121, March 2022.

[162] Jean-Luc Legras, Virginie Galeote, Frédéric Bigey, Carole Camarasa, Souhir Marsit, Thibault Nidelet, Isabelle Sanchez, Arnaud Couloux, Julie Guy, Ricardo Franco-Duarte, Marina Marcet-Houben, Toni Gabaldon, Dorit Schuller, José Paulo Sampaio, Sylvie Dequin, and Patricia Wittkopp. Adaptation of S. cerevisiae to fermented food environments reveals remarkable genome plasticity and the footprints of domestication. *Mol Biol Evol*, 35(7):1712–1727, July 2018.

[163] Heng Li. seqtk.

[164] Heng Li. wgsim.

[165] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, November 2011.

[166] Diego Libkind, Chris Todd Hittinger, Elisabete Valerio, Carla Goncalves, Jim Dover, Mark Johnston, Paula Goncalves, and Jose Paulo Sampaio. Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast. *Proc Natl Acad Sci USA*, 108(35):14539–14544, August 2011.

[167] Robert A. Linder, Arundhati Majumder, Mahul Chakraborty, and Anthony Long. Two Synthetic 18-Way Outcrossed Populations of Diploid Budding Yeast with Utility for Complex Trait Dissection. *Genetics*, 215(2):323–342, June 2020.

[168] Raquel S. Linheiro and Casey M. Bergman. Whole genome resequencing reveals natural target site preferences of transposable elements in Drosophila melanogaster. *PLOS One*, 7(2):e30008, February 2012.

[169] G. Liti, A. Peruffo, S. A. James, I. N. Roberts, and E. J. Louis. 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, 2005.

[170] Gianni Liti, Alex N Nguyen Ba, Martin Blythe, Carolin A Müller, Anders Bergström, Francisco A Cubillos, Felix Dafhnis-Calas, Shima Khoshraftar, Sunir Malla, Neel Mehta, Cheuk C Siow, Jonas Warringer, Alan M Moses, Edward J Louis, and Conrad A Nieduszynski. High quality de novo sequencing and assembly of the Saccharomyces arboricolus genome. *BMC Genomics*, 14:69, January 2013.

[171] Gianni Liti, David B H Barton, and Edward J Louis. Sequence diversity, reproductive isolation and species concepts in Saccharomyces. *Genetics*, 174(2):839–850, October 2006.

[172] Gianni Liti, David M. Carter, Alan M. Moses, Jonas Warringer, Leopold Parts, Stephen A. James, Robert P. Davey, Ian N. Roberts, Austin Burt, Vassiliki Koufopanou, Isheng J. Tsai, Casey M. Bergman, Douda Bensasson, Michael J. T.

O'Kelly, Alexander van Oudenaarden, David B. H. Barton, Elizabeth Bailes, Alex N. Nguyen, Matthew Jones, Michael A. Quail, Ian Goodhead, Sarah Sims, Frances Smith, Anders Blomberg, Richard Durbin, and Edward J. Louis. Population genomics of domestic and wild yeasts. *Nature*, 458(7236):337–41, March 2009.

[173] Christian A. Lopes, Eladio Barrio, and Amparo Querol. Natural hybrids of S. cerevisiae × S. kudriavzevii share alleles with European wild populations of Saccharomyces kudriavzevii. *FEMS Yeast Res*, 10(4):412–421, June 2010.

[174] D D Luan, M H Korman, J L Jakubczak, and T H Eickbush. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell*, 72(4):595–605, February 1993.

[175] Laura G. Macias, Melisa González Flores, Ana Cristina Adam, María E. Rodriguez, Amparo Querol, Eladio Barrio, Christian Ariel Lopes, and Roberto Perez-Torrado. Convergent adaptation of Saccharomyces uvarum to sulfite, an antimicrobial preservative widely used in human-driven fermentations. *PLOS Genet*, 17(11):e1009872, November 2021.

[176] Calum J. Maclean, Brian P. H. Metzger, Jian-Rong Yang, Wei-Chin Ho, Bryan Moyers, and Jianzhi Zhang. Deciphering the genic basis of yeast fitness variation by simultaneous forward and reverse genetics. *Mol Biol Evol*, 34(10):2486–2502, October 2017.

[177] Manee M Manee, John Jackson, and Casey M Bergman. Conserved noncoding elements influence the transposable element landscape in Drosophila. *Genome Biol Evol*, 10(6):1533–1545, May 2018.

[178] Serghei Mangul, Lana S. Martin, Brian L. Hill, Angela Ka-Mei Lam, Margaret G. Distler, Alex Zelikovsky, Eleazar Eskin, and Jonathan Flint. Systematic benchmarking of omics computational tools. *Nat Commun*, 10(1):1393, March 2019.

[179] Wladimir Mardones, Carlos A. Villarroel, Valentina Abarca, Kamila Urbina, Tomás A. Peña, Jennifer Molinet, Roberto F. Nespolo, and Francisco A. Cubillos. Rapid selection response to ethanol in Saccharomyces eubayanus emulates the domestication process under brewing conditions. *Microb Biotechnol*, 15(3):967–984, 2022.

[180] Souhir Marsit, Adriana Mena, Frederic Bigey, François-Xavier Sauvage, Arnaud Couloux, Julie Guy, Jean-Luc Legras, Eladio Barrio, Sylvie Dequin, and Virginie Galeote. Evolutionary advantage conferred by an eukaryote-to-eukaryote gene transfer event in wine yeasts. *Mol Biol Evol*, 32(7):1695–1707, July 2015.

[181] B McClintock. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci USA*, 36(6):344–55, June 1950.

[182] Jennifer Molinet, Kamila Urbina, Claudia Villegas, Valentina Abarca, Christian I. Oporto, Pablo Villarreal, Carlos A. Villarroel, Francisco Salinas, Roberto F. Nespolo, and Francisco A. Cubillos. A Saccharomyces eubayanus haploid resource for research studies. *Sci Rep*, 12(1):5976, April 2022.

[183] Sharon P. Moore, Gianni Liti, Karen M. Stefanisko, Katherine M. Nyswaner, Caroline Chang, Edward J. Louis, and David J. Garfinkel. Analysis of a Ty1-less variant of Saccharomyces paradoxus: the gain and loss of Ty1 elements. *Yeast*, 21(8):649–660, June 2004.

[184] Naomi G. Moresi, Renee C. Geck, Ryan Skophammer, Dennis Godin, yEvo Students, M. Bryce Taylor, and Maitreya J. Dunham. Caffeine-tolerant mutations selected through an at-home yeast experimental evolution teaching lab. *microPublication Biology*, February 2023.

[185] S. M. Mount and G. M. Rubin. Complete nucleotide sequence of the Drosophila transposable element copia: homology between copia and retroviral proteins. *Mol Cell Biol*, 5(7):1630–8, 1985.

[186] Loris Mularoni, Yulian Zhou, Tyson Bowen, Sunil Gangadharan, Sarah J. Wheelan, and Jef D. Boeke. Retrotransposon Ty1 integration targets specifically positioned asymmetric nucleosomal DNA segments in tRNA hotspots. *Genome Res*, 22(4):693–703, April 2012.

[187] Martín Muñoz-López and José L. García-Pérez. DNA transposons: Nature and applications in genomics. *Curr Genomics*, 11(2):115–128, April 2010.

[188] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H. H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, and J. C. Venter. A whole-genome assembly of Drosophila. *Science*, 287(5461):2196–204, March 2000.

[189] Y. Nakao, T. Kanamori, T. Itoh, Y. Kodama, S. Rainieri, N. Nakamura, T. Shimonaga, M. Hattori, and T. Ashikari. Genome sequence of the lager brewing yeast, an interspecies hybrid. *DNA Res*, 16(2):115–29, April 2009.

[190] Samina Naseeb, Haya Alsammar, Tim Burgis, Ian Donaldson, Norman Knyazev, Christopher Knight, and Daniela Delneri. Whole genome sequencing, de novo assembly and phenotypic profiling for the new budding yeast species Saccharomyces jurei. *G3*, 8:2967–2977, August 2018.

[191] Samina Naseeb, Stephen A. James, Haya Alsammar, Christopher J. Michaels, Beatrice Gini, Carmen Nueno-Palop, Christopher J. Bond, Henry McGhie, Ian N. Roberts, and Daniela Delneri. 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, June 2017.

[192] G. I. Naumov, S. A. James, E. S. Naumova, E. J. Louis, and I. N. Roberts. Three new species in the Saccharomyces sensu stricto complex: Saccharomyces cariocanus, Saccharomyces kudriavzevii and Saccharomyces mikatae. *Int J Syst Evol Microbiol*, 50:1931–1942, September 2000.

[193] G. I. Naumov, E. S. Naumova, and P. D. Sniegowski. Differentiation of European and Far East Asian populations of Saccharomyces paradoxus by allozyme analysis. *Int J Syst Bacteriol*, 47(2):341–344, April 1997.

[194] Gennadi I. Naumov, Ching-Fu Lee, and Elena S. Naumova. Molecular genetic diversity of the Saccharomyces yeasts in Taiwan: Saccharomyces arboricola, Saccharomyces cerevisiae and Saccharomyces kudriavzevii. *Antonie Van Leeuwenhoek*, 103(1):217–228, January 2013.

[195] Gl Naumov. Genetic identification of biological species in the Saccharomyces sensu stricto complex. *Journal of Industrial Microbiology & Biotechnology*, 17(3):295–302, September 1996.

[196] Beatriz Navarro-Dominguez, Ching-Ho Chang, Cara L Brand, Christina A Muirhead, Daven C Presgraves, and Amanda M Larracuente. Epistatic selection on a selfish Segregation Distorter supergene – drive, recombination, and genetic load. *eLife*, 11:e78981, April 2022.

[197] Michael G. Nelson, Raquel S. Linheiro, and Casey M. Bergman. McClintock: an integrated pipeline for detecting transposable element insertions in whole-genome shotgun sequencing data. *G3*, 7:2749–2762, August 2017.

[198] Roberto F. Nespolo, Carlos A. Villarroel, Christian I. Oporto, Sebastián M. Tapia, Franco Vega-Macaya, Kamila Urbina, Matteo De Chiara, Simone Mozzachiodi, Ekaterina Mikhalev, Dawn Thompson, Luis F. Larrondo, Pablo Saenz-Agudelo, Gianni

Liti, and Francisco A. Cubillos. An Out-of-Patagonia migration explains the worldwide diversity and distribution of Saccharomyces eubayanus lineages. *PLOS Genetics*, 16(5):e1008777, May 2020.

[199] Cécile Neuveglise, Horst Feldmann, Elisabeth Bon, Claude Gaillardin, and Serge Casaregola. Genomic evolution of the long terminal repeat retrotransposons in hemiascomycetous yeasts. *Genome Res*, 12(6):930–43, June 2002.

[200] Huu-Vang Nguyen and Teun Boekhout. Characterization of Saccharomyces uvarum (Beijerinck, 1898) and related hybrids: assessment of molecular markers that predict the parent and hybrid genomes and a proposal to name yeast hybrids. *FEMS Yeast Res*, 17(2), 2017.

[201] Huu-Vang Nguyen and Claude Gaillardin. 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):471–483, February 2005.

[202] Phong Quoc Nguyen, Sonia Huecas, Amna Asif-Laidin, Adrián Plaza-Pegueroles, Beatrice Capuzzi, Noé Palmic, Christine Conesa, Joël Acker, Juan Reguera, Pascale Lesage, and Carlos Fernández-Tornero. Structural basis of Ty1 integrase tethering to RNA polymerase III for targeted retrotransposon integration. *Nat Commun*, 14(1):1729, March 2023.

[203] Raquel Norel, John Jeremy Rice, and Gustavo Stolovitzky. The self-assessment trap: can we all be better than average? *Molecular Systems Biology*, 7(1):537, January 2011. Publisher: John Wiley & Sons, Ltd.

[204] Olga Novikova, Victor Fet, and Alexander Blinov. Non-LTR retrotransposons in fungi. *Funct Integr Genomics*, 9(1):27–42, February 2009.

[205] Petr Novák, Pavel Neumann, and Jiří Macas. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, 11(1):378, July 2010.

[206] Petr Novák, Pavel Neumann, and Jiří Macas. Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. *Nat Protoc*, 15(11):3745–3776, November 2020. Number: 11 Publisher: Nature Publishing Group.

[207] Petr Novák, Pavel Neumann, Jiří Pech, Jaroslav Steinhaisl, and Jiří Macas. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, 29(6):792–793, March 2013.

[208] Elisa Oberbeckmann, Michael Wolff, Nils Krietenstein, Mark Heron, Jessica L. Ellins, Andrea Schmid, Stefan Krebs, Helmut Blum, Ulrich Gerland, and Philipp Korber. Absolute nucleosome occupancy map for the Saccharomyces cerevisiae genome. *Genome Res*, 29(12):1996–2009, December 2019.

[209] Ursula Oggenfuss, Thomas Badet, Thomas Wicker, Fanny E Hartmann, Nikhil Kumar Singh, Leen Abraham, Petteri Karisto, Tiziana Vonlanthen, Christopher Mundt, Bruce A McDonald, and Daniel Croll. A population-level invasion by transposable elements triggers genome expansion in a fungal pathogen. *eLife*, 10:e69249, September 2021.

[210] Miki Okuno, Rei Kajitani, Rie Ryusui, Hiroya Morimoto, Yukiko Kodama, and Takehiko Itoh. Next-generation sequencing analysis of lager brewing yeast strains reveals the evolutionary history of interspecies hybridization. *DNA Res*, 23(1):67–80, February 2016.

[211] Edgardo M. Ortiz. vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis., January 2019.

[212] Samuel O'Donnell, Jia-Xing Yue, Omar Abou Saada, Nicolas Agier, Claudia Caradec, Thomas Cokelaer, Matteo De Chiara, Stéphane Delmas, Fabien Dutreux, Téo Fournier, Anne Friedrich, Etienne Kornobis, Jing Li, Zepu Miao, Lorenzo Tattini, Joseph Schacherer, Gianni Liti, and Gilles Fischer. 142 telomere-to-telomere assemblies reveal the genome structural landscape in Saccharomyces cerevisiae. *bioRxiv*, page 2022.10.04.510633, October 2022.

[213] Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290, January 2004.

[214] Kurt Patterson, Farbod Shavarebi, Christophe Magnan, Ivan Chang, Xiaojie Qi, Pierre Baldi, Virginia Bilanchone, and Suzanne B. Sandmeyer. Local features determine Ty3 targeting frequency at RNA polymerase III transcription start sites. *Genome Res*, 29(8):1298–1309, August 2019.

[215] Valentina Peona, Mozes P. K. Blom, Luohao Xu, Reto Burri, Shawn Sullivan, Ignas Bunikis, Ivan Liachko, Tri Haryoko, Knud A. Jonsson, Qi Zhou, Martin Irestedt, and Alexander Suh. Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Mol Ecol Resour*, 21(1):263–286, January 2021.

[216] V. Pereira. Automated paleontology of repetitive DNA with REANNOTATE. *BMC Genomics*, 9:614, 2008.

[217] David Peris, Quinn K. Langdon, Ryan V. Moriarty, Kayla Sylvester, Martin Bontrager, Guillaume Charron, Jean-Baptiste Leducq, Christian R. Landry, Diego Libkind, and Chris Todd Hittinger. Complex ancestries of lager-brewing hybrids were shaped by standing variation in the wild yeast Saccharomyces eubayanus. *PLOS Genet*, 12(7):e1006155, July 2016.

[218] David Peris, Roberto Pérez-Torrado, Chris Todd Hittinger, Eladio Barrio, and Amparo Querol. On the origins and industrial applications of Saccharomyces cerevisiae × Saccharomyces kudriavzevii hybrids. *Yeast*, 35(1):51–69, January 2018.

[219] David Peris, Kayla Sylvester, Diego Libkind, Paula Gonçalves, José Paulo Sampaio, William G. Alexander, and Chris Todd Hittinger. Population structure and reticulate evolution of Saccharomyces eubayanus and its lager-brewing hybrids. *Mol Ecol*, 23(8):2031–2045, April 2014.

[220] David Peris, Emily J. Ubbelohde, Meihua Christina Kuang, Jacek Kominek, Quinn K. Langdon, Marie Adams, Justin A. Koshalek, Amanda Beth Hulfachor, Dana A. Opulente, David J. Hall, Katie Hyma, Justin C. Fay, Jean-Baptiste Leducq, Guillaume Charron, Christian R. Landry, Diego Libkind, Carla Gonçalves, Paula Gonçalves, José Paulo Sampaio, Qi-Ming Wang, Feng-Yan Bai, Russel L. Wrobel, and Chris Todd Hittinger. Macroevolutionary diversity of traits and genomes in the model yeast genus Saccharomyces. *Nat Commun*, 14(1):690, February 2023.

[221] Muhammad Tariq Pervez, Mirza Jawad ul Hasnain, Syed Hassan Abbas, Mahmoud F. Moustafa, Naeem Aslam, and Syed Shah Muhammad Shah. A Comprehensive Review of Performance of Next-Generation Sequencing Platforms. *BioMed Research International*, 2022:e3457806, September 2022.

[222] Jackson Peter, Matteo De Chiara, Anne Friedrich, Jia-Xing Yue, David Pflieger, Anders Bergstrom, Anastasie Sigwalt, Benjamin Barre, Kelle Freel, Agnès Llored, Corinne Cruaud, Karine Labadie, Jean-Marc Aury, Benjamin Istace, Kevin Lebrigand, Pascal Barbry, Stefan Engelen, Arnaud Lemainque, Patrick Wincker, Gianni Liti, and Joseph Schacherer. Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature*, 556(7701):339–344, April 2018.

[223] Alexander Platzer, Viktoria Nizhynska, and Quan Long. TE-Locate: a tool to locate and group transposable element occurrences using paired-end next-generation sequencing data. *Biology*, 1(2):395–410, September 2012.

[224] Andy Pohl and Miguel Beato. bwtool: a tool for bigWig files. *Bioinformatics*, 30(11):1618–1619, June 2014.

[225] A. L. Price, N. C. Jones, and P. A. Pevzner. De novo identification of repeat families in large genomes. *Bioinformatics*, 21 Suppl 1:i351–i358, 2005.

[226] A. Pulvirenti, H. Nguyen, C. Caggia, P. Giudici, S. Rainieri, and C. Zambonelli. Saccharomyces uvarum, a proper species within Saccharomyces sensu stricto. *FEMS Microbiol Lett*, 192(2):191–196, November 2000.

[227] Xiaojie Qi, Kenneth Daily, Kim Nguyen, Haoyi Wang, David Mayhew, Paul Rigor, Sholeh Forouzan, Mark Johnston, Robi David Mitra, Pierre Baldi, and Suzanne Sandmeyer. Retrotransposon profiling of RNA polymerase III initiation sites. *Genome Res*, 22(4):681–92, April 2012.

[228] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, March 2010.

[229] Luis Quintales, Enrique Vázquez, and Francisco Antequera. Comparative analysis of methods for genome-wide nucleosome cartography. *Briefings in Bioinformatics*, 16(4):576–587, July 2015.

[230] S. Rainieri, C. Zambonelli, J. E. Hallsworth, A. Pulvirenti, and P. Giudici. Saccharomyces uvarum, a distinct group within Saccharomyces sensu stricto. *FEMS Microbiol Lett*, 177(1):177–185, August 1999.

[231] Matteo Ramazzotti, Irene Stefanini, Monica Di Paola, Carlotta De Filippo, Lisa Rizzetto, Luisa Berná, Leonardo Dapporto, Damariz Rivero, Noemi Tocci, Tobias Weil,

Marcello S. Lenucci, Paolo Lionetti, and Duccio Cavalieri. Population genomics reveals evolution and variation of Saccharomyces cerevisiae in the human and insects gut. *Environ Microbiol*, 21(1):50–71, 2019.

[232] David A. Ray and Mark A. Batzer. Reading TE leaves: New approaches to the identification of transposable element insertions. *Genome Res*, 21(6):813–820, June 2011.

[233] L. A. Rinckel and D. J. Garfinkel. Influences of histone stoichiometry on the target site preference of retrotransposons Ty1 and Ty2 in Saccharomyces cerevisiae. *Genetics*, 142(3):761–776, March 1996.

[234] Lavanya Rishishwar, Leonardo Marino-Ramirez, and I. King Jordan. Benchmarking computational tools for polymorphic transposable element detection. *Brief Bioinformatics*, 18(6):908–918, November 2017.

[235] Sofia M. C. Robb, Lu Lu, Elizabeth Valencia, James M. Burnette, Yutaka Okumoto, Susan R. Wessler, and Jason E. Stajich. The use of RelocaTE and unassembled short reads to produce high-resolution snapshots of transposable element generated diversity in rice. *G3*, 3(6):949–957, June 2013.

[236] M. Eugenia Rodríguez, Laura Pérez-Través, Marcela P. Sangorrín, Eladio Barrio, and Christian A. Lopes. Saccharomyces eubayanus and Saccharomyces uvarum associated with the fermentation of Araucaria araucana seeds in Patagonia. *FEMS Yeast Res*, 14(6):948–965, September 2014.

[237] G. S. Roeder, P. J. Farabaugh, D. T. Chaleff, and G. R. Fink. The origins of gene instability in yeast. *Science*, 209(4463):1375–1380, September 1980.

[238] G. M. Rubin, M. G. Kidwell, and P. M. Bingham. The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. *Cell*, 29(3):987–94, 1982.

[239] Sergei Ryazansky, Elizaveta Radion, Anastasia Mironova, Natalia Akulenko, Yuri Abramov, Valeriya Morgunova, Maria Y. Kordyukova, Ivan Olovnikov, and Alla

Kalmykova. Natural variation of piRNA expression affects immunity to transposable elements. *PLOS Genetics*, 13(4):e1006731, April 2017.

[240] Agniva Saha, Jessica A. Mitchell, Yuri Nishida, Jonathan E. Hildreth, Joshua A. Ariberre, Wendy V. Gilbert, and David J. Garfinkel. A trans-dominant form of Gag restricts Ty1 retrotransposition and mediates copy number control. *J Virol*, 89(7):3922–3938, April 2015.

[241] Surya Saha, Susan Bridges, Zenaida V. Magbanua, and Daniel G. Peterson. Computational approaches and tools used in identification of dispersed repetitive DNA sequences. *Tropical Plant Biol*, 1(1):85–96, March 2008.

[242] Alex N. Salazar, Arthur R. Gorter de Vries, Marcel van den Broek, Nick Brouwers, Pilar de la Torre Cortes, Niels G. A. Kuijpers, Jean-Marc G. Daran, and Thomas Abeel. Chromosome level assembly and comparative genome analysis confirm lager-brewing yeasts originated from a single hybridization. *BMC Genomics*, 20(1):916, December 2019.

[243] Jose Paulo Sampaio and Paula Goncalves. 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):2144–2152, April 2008.

[244] Suzanne Sandmeyer, Kurt Patterson, and Virginia Bilanchone. Ty3, a position-specific retrotransposon in budding yeast. *Microbiol Spectr*, 3(2):MDNA3–0057–2014, April 2015.

[245] Véronique Sarilar, Claudine Bleykasten-Grosshans, and Cecile Neuveglise. Evolutionary dynamics of hAT DNA transposon families in Saccharomycetaceae. *Genome Biol Evol*, 7(1):172–190, January 2015.

[246] Devin R. Scannell, Oliver A. Zill, Antonis Rokas, Celia Payen, Maitreya J. Dunham, Michael B. Eisen, Jasper Rine, Mark Johnston, and Chris Todd Hittinger. The awesome

power of yeast evolutionary genetics: New genome sequences and strain resources for the Saccharomyces sensu stricto genus. *G3*, 1(1):11–25, June 2011.

[247] Virginia T Scarlett, John T Lovell, Mingqin Shao, Jeremy Phillips, Shengqiang Shu, Joanna Lusinska, David M Goodstein, Jerry Jenkins, Jane Grimwood, Kerrie Barry, Boulos Chalhoub, Jeremy Schmutz, Robert Hasterok, Pilar Catalán, and John P Vogel. Multiple origins, one evolutionary trajectory: gradual evolution characterizes distinct lineages of allotetraploid Brachypodium. *Genetics*, 223(2):iyac146, February 2023.

[248] Sarah Schaack, Clément Gilbert, and Cédric Feschotte. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol Evol*, 25(9):537–546, September 2010.

[249] Stephanie N. Schauer, Patricia E. Carreira, Ruchi Shukla, Daniel J. Gerhardt, Patricia Gerdes, Francisco J. Sanchez-Luque, Paola Nicoli, Michaela Kindlova, Serena Ghisletti, Alexandre Dos Santos, Delphine Rapoud, Didier Samuel, Jamila Faivre, Adam D. Ewing, Sandra R. Richardson, and Geoffrey J. Faulkner. L1 retrotransposition is a common feature of mammalian hepatocarcinogenesis. *Genome Res*, 28(5):639–653, May 2018.

[250] Klaus Peter Schliep. phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593, February 2011.

[251] Patrick S. Schnable, Doreen Ware, Robert S. Fulton, Joshua C. Stein, Fusheng Wei, Shiran Pasternak, Chengzhi Liang, Jianwei Zhang, Lucinda Fulton, Tina A. Graves, Patrick Minx, Amy Denise Reily, Laura Courtney, Scott S. Kruchowski, Chad Tomlinson, Cindy Strong, Kim Delehaunty, Catrina Fronick, Bill Courtney, Susan M. Rock, Eddie Belter, Feiyu Du, Kyung Kim, Rachel M. Abbott, Marc Cotton, Andy Levy, Pamela Marchetto, Kerri Ochoa, Stephanie M. Jackson, Barbara Gillam, Weizu Chen, Le Yan, Jamey Higginbotham, Marco Cardenas, Jason Waligorski, Elizabeth Applebaum, Lindsey Phelps, Jason Falcone, Krishna Kanchi, Thynn Thane, Adam Scimone,

Nay Thane, Jessica Henke, Tom Wang, Jessica Ruppert, Neha Shah, Kelsi Rotter, Jennifer Hodges, Elizabeth Ingenthron, Matt Cordes, Sara Kohlberg, Jennifer Sgro, Brandon Delgado, Kelly Mead, Asif Chinwalla, Shawn Leonard, Kevin Crouse, Kristi Collura, Dave Kudrna, Jennifer Currie, Ruifeng He, Angelina Angelova, Shanmugam Rajasekar, Teri Mueller, Rene Lomeli, Gabriel Scara, Ara Ko, Krista Delaney, Marina Wissotski, Georgina Lopez, David Campos, Michele Braidotti, Elizabeth Ashley, Wolfgang Golser, HyeRan Kim, Seunghee Lee, Jinke Lin, Zeljko Dujmic, Woojin Kim, Jayson Talag, Andrea Zuccolo, Chuanzhu Fan, Aswathy Sebastian, Melissa Kramer, Lori Spiegel, Lidia Nascimento, Theresa Zutavern, Beth Miller, Claude Ambroise, Stephanie Muller, Will Spooner, Apurva Narechania, Liya Ren, Sharon Wei, Sunita Kumari, Ben Faga, Michael J. Levy, Linda McMahan, Peter Van Buren, Matthew W. Vaughn, Kai Ying, Cheng-Ting Yeh, Scott J. Emrich, Yi Jia, Ananth Kalyanaraman, An-Ping Hsia, W. Brad Barbazuk, Regina S. Baucom, Thomas P. Brutnell, Nicholas C. Carpita, Cristian Chaparro, Jer-Ming Chia, Jean-Marc Deragon, James C. Estill, Yan Fu, Jeffrey A. Jeddeloh, Yujun Han, Hyeran Lee, Pinghua Li, Damon R. Lisch, Sanzhen Liu, Zhijie Liu, Dawn Holligan Nagel, Maureen C. McCann, Phillip SanMiguel, Alan M. Myers, Dan Nettleton, John Nguyen, Bryan W. Penning, Lalit Ponnala, Kevin L. Schneider, David C. Schwartz, Anupma Sharma, Carol Soderlund, Nathan M. Springer, Qi Sun, Hao Wang, Michael Waterman, Richard Westerman, Thomas K. Wolfgruber, Lixing Yang, Yeisoo Yu, Lifang Zhang, Shiguo Zhou, Qihui Zhu, Jeffrey L. Bennetzen, R. Kelly Dawe, Jiming Jiang, Ning Jiang, Gernot G. Presting, Susan R. Wessler, Srinivas Aluru, Robert A. Martienssen, Sandra W. Clifton, W. Richard McCombie, Rod A. Wing, and Richard K. Wilson. The B73 maize genome: complexity, diversity, and dynamics. *Science*, 326(5956):1112–1115, November 2009.

[252] Wei Shen, Shuai Le, Yan Li, and Fuquan Hu. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE*, 11(10):e0163962, October 2016.

[253] Jay Shendure, Shankar Balasubramanian, George M. Church, Walter Gilbert, Jane Rogers, Jeffery A. Schloss, and Robert H. Waterston. DNA sequencing at 40: past, present and future. *Nature*, 550(7676):345–353, October 2017.

[254] Felipe A. Simao, Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212, October 2015.

[255] Matthias Sipiczki. Interspecies hybridization and recombination in Saccharomyces wine yeasts. *FEMS yeast research*, 8(7):996–1007, November 2008.

[256] Daniel A. Skelly, Gennifer E. Merrihew, Michael Riffle, Caitlin F. Connelly, Emily O. Kerr, Marnie Johansson, Daniel Jaschob, Beth Graczyk, Nicholas J. Shulman, Jon Wakefield, Sara J. Cooper, Stanley Fields, William S. Noble, Eric G. D. Muller, Trisha N. Davis, Maitreya J. Dunham, Michael J. MacCoss, and Joshua M. Akey. Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. *Genome Res*, 23(9):1496–1504, September 2013.

[257] A.F.A. Smit, R. Hubley, and P. Green. RepeatMasker, 2013.

[258] Caiti Smukowski Heil, Kira Patterson, Angela Shang-Mei Hickey, Erica Alcantara, and Maitreya J Dunham. Transposable element mobilization in interspecific yeast hybrids. *Genome Biol Evol*, 13(3):evab033, February 2021.

[259] Giltae Song, Benjamin J. A. Dickins, Janos Demeter, Stacia Engel, Jennifer Gallagher, Kisurb Choe, Barbara Dunn, Michael Snyder, and J. Michael Cherry. AGAPE (Automated Genome Analysis PipelinE) for pan-genome analysis of Saccharomyces cerevisiae. *PLOS ONE*, 10(3):e0120671, 2015.

[260] Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, May 2014.

[261] Pooja K. Strope, Daniel A. Skelly, Stanislav G. Kozmin, Gayathri Mahadevan, Eric A. Stone, Paul M. Magwene, Fred S. Dietrich, and John H. McCusker. 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):762–774, May 2015.

[262] Tim Stuart, Steven Eichten, Jonathan Cahn, Yuliya Karpievitch, Justin Borevitz, and Ryan Lister. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *eLife*, 5:e20777, December 2016.

[263] R. Stucka, C. Schwarzlose, H. Lochmüller, U. Häcker, and H. Feldmann. Molecular analysis of the yeast Ty4 element: homology with Ty1, copia, and plant retrotransposons. *Gene*, 122(1):119–128, December 1992.

[264] Rolf Stucka, Harms Lochmäller, and Horst Feldmann. 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):4993–5002, July 1989.

[265] Matthew Bryce Taylor, Ryan Skophammer, Alexa R Warwick, Renee C Geck, Josephine M Boyer, yEvo Students, Margaux Walson, Christopher R L Large, Angela Shang-Mei Hickey, Paul A Rowley, and Maitreya J Dunham. yEvo: experimental evolution in high school classrooms selects for novel mutations that impact clotrimazole resistance in Saccharomyces cerevisiae. *G3*, 12(11):jkac246, November 2022.

[266] Todd J. Treangen and Steven L. Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*, 13(1):36–46, January 2012. Number: 1 Publisher: Nature Publishing Group.

[267] Panisa Treepong, Christophe Guyeux, Alexandre Meunier, Charlotte Couchoud, Didier Hocquet, and Benoit Valot. panISa: ab initio detection of insertion sequences in

bacterial genomes from short read sequence data. *Bioinformatics*, 34(22):3795–3800, November 2018.

[268] Mark A Urich, Joseph R Nery, Ryan Lister, Robert J Schmitz, and Joseph R Ecker. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nat Protoc*, 10(3):475–483, March 2015.

[269] Pol Vendrell-Mir, Fabio Barteri, Miriam Merenciano, Josefa González, Josep M. Casacuberta, and Raúl Castanera. A benchmark of transposon insertion detection tools using real data. *Mob DNA*, 10(1):53, December 2019.

[270] N. Volfovsky, B. J. Haas, and S. L. Salzberg. A clustering method for repeat analysis in DNA sequences. *Genome Biol*, 2(8):RESEARCH0027, 2001.

[271] D. Voytas and J. Boeke. Yeast retrotransposon revealed. *Nature*, 358(6389):717, August 1992.

[272] Bruce J. Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K. Young, and Ashlee M. Earl. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE*, 9(11):e112963, 2014.

[273] Gabriel Luz Wallau, Mauro Freitas Ortiz, and Elgion Lucio Silva Loreto. Horizontal transposon transfer in eukarya: detection, bias, and perspectives. *Genome Biol Evol*, 4(8):801–811, January 2012.

[274] Gabriel Luz Wallau, Cristina Vieira, and Elgion Lucio Silva Loreto. Genetic exchange in eukaryotes through horizontal transfer: connected by the mobilome. *Mob DNA*, 9:6, January 2018.

[275] Qi-Ming Wang, Wan-Qiu Liu, Gianni Liti, Shi-An Wang, and Feng-Yan Bai. Surprisingly diverged populations of Saccharomyces cerevisiae in natural environments remote from human activity. *Mol Ecol*, 21(22):5404–5417, November 2012.

[276] Shi-An Wang and Feng-Yan Bai. Saccharomyces arboricolus sp. nov., a yeast species from tree bark. *International Journal of Systematic and Evolutionary Microbiology,*, 58(2):510–514, 2008.

[277] Lukas M. Weber, Wouter Saelens, Robrecht Cannoodt, Charlotte Soneson, Alexander Hapfelmeier, Paul P. Gardner, Anne-Laure Boulesteix, Yvan Saeys, and Mark D. Robinson. Essential guidelines for computational method benchmarking. *Genome Biol*, 20(1):125, June 2019.

[278] Lukas Weilguny and Robert Kofler. DeviaTE: Assembly-free analysis and visualization of mobile genetic element composition. *Mol Ecol Resour*, 19(5):1346–1354, 2019.

[279] Jonathan N. Wells and Cédric Feschotte. A field guide to eukaryotic transposable elements. *Annu Rev Genet*, 54:539–561, November 2020.

[280] Thomas Wicker, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, Philippe Leroy, Michele Morgante, Olivier Panaud, Etienne Paux, Phillip SanMiguel, and Alan H. Schulman. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*, 8(12):973–982, December 2007.

[281] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[282] C. M. Wilke and J. Adams. Fitness Effects of Ty Transposition in Saccharomyces cerevisiae. *Genetics*, 131(1):31–42, May 1992.

[283] F. Winston, D. T. Chaleff, B. Valent, and G. R. Fink. Mutations affecting Ty-mediated expression of the HIS4 gene of Saccharomyces cerevisiae. *Genetics*, 107(2):179–197, June 1984.

[284] Wenjing Xia, Lou Nielly-Thibault, Guillaume Charron, Christian R. Landry, Dahlia Kasimer, James B. Anderson, and Linda M. Kohn. Population genomics reveals structure at the individual, host-tree scale and persistence of genotypic variants of the undomesticated yeast Saccharomyces paradoxus in a natural woodland. *Mol Ecol*, 26(4):995–1007, February 2017.

[285] Weiwu Xie, Xiaowu Gai, Yunxia Zhu, David C Zappulla, Rolf Sternglanz, and Daniel F Voytas. Targeting of the yeast Ty5 retrotransposon to silent chromatin is mediated by interactions between integrase and Sir4p. *Molecular and cellular biology*, 21(19):6606–14, October 2001.

[286] Jinchuan Xing, David J. Witherspoon, and Lynn B. Jorde. Mobile element biology: new possibilities with high-throughput sequencing. *Trends in Genetics*, 29(5):280–289, May 2013.

[287] Haidong Yan, David C. Haak, Song Li, Linkai Huang, and Aureliano Bombarely. Exploring transposable element-based markers to identify allelic variations underlying agronomic traits in rice. *Plant Communications*, 3(3):100270, May 2022.

[288] Guangchuang Yu, David K. Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol*, 8(1):28–36, January 2017.

[289] Tianxiong Yu, Xiao Huang, Shengqian Dou, Xiaolu Tang, Shiqi Luo, William E Theurkauf, Jian Lu, and Zhiping Weng. A benchmark and an algorithm for detecting

germline transposon insertions and measuring de novo transposon insertion frequencies. *Nucleic Acids Res*, 49(8):e44, May 2021.

[290] Weichang Yu, Jonathan C. Lamb, Fangpu Han, and James A. Birchler. Cytological visualization of DNA transposons and their transposition pattern in somatic cells of maize. *Genetics*, 175(1):31–39, January 2007.

[291] Jia-Xing Yue, Jing Li, Louise Aigrain, Johan Hallin, Karl Persson, Karen Oliver, Anders Bergstrom, Paul Coupland, Jonas Warringer, Marco Cosentino Lagomarsino, Gilles Fischer, Richard Durbin, and Gianni Liti. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nat Genet*, 49(6):913–924, June 2017.

[292] Jia-Xing Yue and Gianni Liti. Long-read sequencing data analysis for yeasts. *Nat Protoc*, 13(6):1213–1231, 2018.

[293] Xiuwen Zheng, David Levine, Jess Shen, Stephanie M. Gogarten, Cathy Laurie, and Bruce S. Weir. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, 28(24):3326–3328, December 2012.

[294] Yuan O. Zhu, Gavin Sherlock, and Dmitri A. Petrov. Whole genome analysis of 132 clinical Saccharomyces cerevisiae strains reveals extensive ploidy variation. *G3*, 6(8):2421–2434, 2016.

[295] Jiali Zhuang, Jie Wang, William Theurkauf, and Zhiping Weng. TEMP: a computational method for analyzing transposable element polymorphism in populations. *Nucleic Acids Res*, 42(11):6826–6838, June 2014.

[296] S. Zou, N. Ke, J. M. Kim, and D. F. Voytas. The Saccharomyces retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci. *Genes Dev*, 10(5):634–645, March 1996.

[297] S. Zou, D. A. Wright, and D. F. Voytas. 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 USA*, 92(3):920–924, January 1995.