University of Huddersfield

Doctoral Thesis

Comparative Genomics and the Evolution of Transposable Elements in Unicellular Eukaryotes

---

Author:
Jade Southworth

Supervisors:
Dr. Martin Carr
Dr. Jarek Bryk

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

Department of Biological and Geographical Sciences
School of Applied Sciences

University of HUDDERSFIELD
Inspiring tomorrow's professionals

Copyright Statement

Declaration of Authorship

I, Jade Southworth, declare that this thesis titled, “Comparative Genomics and the Evolution of Transposable Elements in Unicellular Eukaryotes” and the work presented in it are my own. I confirm that:

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

The work outlined in this thesis has contributed to two publications as listed below:

- Southworth, J., Armitage, P., Fallon, B., Dawson, H., Bryk, J. and Carr, M. (2018), ‘Patterns of ancestral animal codon usage bias revealed through holozoan protists’, *Molecular Biology and Evolution* **35** (10), 2499 – 2511.
- Southworth, J., Grace, C.A., Marron, A.O, Fatima, N. and Carr. M (2019). ‘A genomic survey of transposable elements in the choanoflagellate *Salpingoeca rosetta* reveals selection on codon usage’, *Mobile DNA* **10** (44), 1 – 19.

Thesis analysis detailed in Chapter 4 is described to differ from the Molecular Biology and Evolution publication, with reference to additional colleagues who contributed method support. The author contributed to analysis and writing of the manuscript for this publication, with PhD supervisor, Dr. Martin Carr. Additional analysis and comparative study to the publication was performed, as part of the thesis research, as detailed in Chapter 4.

For the work now published in Mobile DNA (Southworth et al., 2019), detailed in Chapter 5, the author devised the research, undertook analysis and writing of the manuscript, with Dr. Martin Carr. Both open access publications copyright ownership is held with the authors.

Signed: *J. Southworth*

Date: 03.06.2020

"*The great tragedy of science - the slaying of a beautiful hypothesis by an ugly fact."*

Thomas Henry Huxley

Abstract

Background

Transposable elements are mobile DNA sequences, which are ubiquitous in the majority of eukaryotic genomes. Unicellular eukaryotes have limited research on transposable elements and therefore the picture of evolution is far from conclusive. Similarly, codon usage bias, the frequency of synonymous codons present in a host species coding DNA, has been focused on multicellular organisms, with no clear explanation of the evolutionary pressures that drive bias in unicellular eukaryotic species.

Methods

Eight *Kazachstania* budding yeast species, and choanoflagellate species, *Salpingoeca rosetta*, were screened for the presence of mobile elements, with use of homology based methods. Protein and nucleotide phylogenies were constructed to review ancestral patterns and similarity across superfamilies. Codon usage statistics were employed to review patterns of bias in the host genes and mobile elements of the *Kazachstania* species, and *S.rosetta*, as well as two additional holozoan species, *Monosiga brevicollis* and *Capsaspora owczarzaki*.

Results

A diverse repetoire of transposable element families were uncovered in the species reviewed. A complete absence of DNA transposons was found in the *Kazachstania* species, however both classes of elements were uncovered in *S. rosetta*. Element phylogenies indicated vertical transfer for the majority of families, with the exception of one family in *S. rosetta*, which suggested acquisition by horizontal transfer. Patterns of codon usage were revealed in the genus *Kazachstania* and conservation was seen in the three holozoan species, with similar trends observed in the majority of host species mobile elements.

Conclusions

The known diversity of TE families for the yeast superfamily, and Choanoflagellatea has increased as a result of the study presented here. Codon usage bias for host genes and mobile elements provided evidence of selection, as well as mutational bias, suggesting that models of evolutionary pressures are more complex in unicellular eukaryotes.

Acknowledgements

I have so many people I would like to share thanks over the duration of my research at the University of Huddersfield, including my fellow Postgraduate Researchers, who have been the best councillors, experts and friends. I would like to thank my thesis supervisor Dr. Martin Carr, of the School of Applied Sciences at the University of Huddersfield. His breadth of knowledge, expertise and support have been continual, and allowed the work to evolve in a direction we hadn't originally anticipated! An additional thanks goes to my secondary supervisor, Dr. Jarek Bryk, for his infinite bioinformatic expertise. I would also like to thank my brother, Richard Southworth, for his critique, enthusiasm and brilliant mind, and my dear friend, Holly Dawson, who without her support and lab skills, I would not have been able to complete the project, which seemed like an impossible task towards the end of my final year. A further acknowledgement is to Dr. Cooper Grace, of the School of Applied Sciences at the University of Huddersfield. For every bioinformatic query, transposable element hypothesis, writing suggestion and general office companion, I thank him greatly. I would also like to thank my wonderful Mum for all the inspirational quotes and encouragement, and my partner, Adam, for convincing me to move to the North East and his unwavering support. Finally, I would like to thank the three golden boys - Finlay, Stevie and Aero, who have been the best distraction when I have needed to be separated from my research for a long walk.

Contents

List of Figures

List of Abbreviations

A	Adenine
AA	amino acid
BI	Bayesian Inference; phylogenetic method
biPP	Bayesian Inference Posterior Probability
BLAST	Basic Local Alignment Search Tool
BLASTn	nucleotide BLAST
BLASTp	Protein BLAST
BP	bootstrap
C	Cytosine
cDNA	complementary DNA
CDS	a CoDing Sequence
DNA	deoxyribonucleic acid
EF1A	Elongation factor 1-alpha
EFL	Elongation factor-like
FASTA	FAST-All, or Pearson; nucleotide or protein sequence format
FLE	full-length element
Fop	Frequency of optimal codons
G	Guanine
GAG; *gag*	capsid-like domain of transposable elements
GC3s	guanine+cytosine content at synonymous third codon positions
gDNA	genomic DNA
GIRI	Genetic Information Research Institute
HGT	horizontal gene transfer
HT	horizontal transfer
HTT	horizontal transfer (of) transposable elements
IN	integrase; enzyme catalysing the integration of DNA into a genome

ITR	Inverted terminal repeat
LINE	Long interspersed nuclear elements
LTR	Long terminal repeat
MAFFT	Multiple Alignment using Fast Fourier Transform
ML	Maximum Likelihood; phylogenetic inference method
mRNA	messenger RNA
Nc	Effective number of codons
NCBI	National Centre for Biotechnology Information
Ne	Effective population size
NEXUS	file format
NGS	Next Generation Sequencing
NT	nucleotide
ORF	open reading frame
PCR	polymerase chain reaction
PLE	Penelope-like elements
Pol; *pol*	polyprotein domain of transposable elements
Pol II or III	RNA Polymerase II or III
poly(A)	polyadenylate
PP	posterior probability
PR	protease
RAxML	Randomised Axelerated Maximum Likelihood; phylogenetic inference program
RH	ribonuclease H
RIP	repeat-induced point mutations
RNA	ribonucleic acid
RNAi	RNA interference
RT	reverse transcriptase; enzyme catalysing the synthesis of cDNA from RNA template
SINE	short interspersed nuclear repeat
SRA	Sequence Read Archive
T	Thymine
tBLASTn	translated nucleotide BLAST search of protein database
TBP	TATA-binding protein; transcription factor

TE	transposable element
tRNA	transfer RNA
Ty	Transposon (in) yeast; family names assigned in *Saccharomyces cerevisiae*
TSD	Target site duplication
U	Uracil
UTR	untranslated region
WGD	whole genome duplication
WGS	whole genome sequencing

Chapter 1

Introduction

1.1 Transposable Elements

1.1.1 Background

Transposable elements (TEs) are repetitive DNA sequences, that have been described as universal components of eukaryotic genomes (Kidwell, 2002). They have the ability to move position in the genome by the method of transposition, which can cause genomic rearrangements (McClintock, 1984). Advances in bioinformatics and genome sequencing has enabled exhaustive research into TE abundance, highlighting the importance and influence the repetitive DNA can have on host fitness (Gorinsek et al., 2004).

Discovered and published by geneticist Barbara McClintock in the late 1940's, the first TEs were found through the investigation of maize by the observed changes in kernel colour patterns (McClintock, 1950). It was at this point that mobile elements were theorised to move around the plant genome due to the proposed effects on gene expression, with TEs initially termed controlling elements (McClintock, 1950). From the initial research of TEs by McClintock, further work suggested that TEs were in fact "junk" DNA, with little to no impact on gene expression (Doolittle and Sapienza, 1980). This categorisation of the mobile elements was the predominant view from the 1970s, for several years, which contradicted McClitock's findings (Doolittle and Sapienza, 1980; McClintock, 1950). Extensive genetic analyses have since supported McClintock's work, and TEs are now accepted to have a major role in shaping genome evolution - an idea proposed by McClintock, who later won a Nobel prize in 1984 (Biemont, 2010; McClintock, 1984).

Upon acceptance of the mutational capacity of mobile elements, TEs were described as parasitic, "selfish" DNA (McClintock, 1950; Doolittle and Sapienza, 1980). Exhaustive research supported that TE invasion was solely detrimental to host fitness, causing mutagenesis, with the genomic

impact being predominantly deleterious (Doolittle and Sapienza, 1980; Orgel and Crick, 1980). This has since been challenged, suggesting that TEs can have a mutual relationship with their host, or with particular insertions having advantageous implications to host gene evolution, and potentially a driver for speciation (Fedoroff, 2012; Joly-Lopez et al., 2012). It is with this that a spectrum of host-element interactions has been considered, detailing evidence of both parasitism and symbiosis (Orgel and Crick, 1980; Hess et al., 2014). Genomic influence is heavily determined by the status of the TE family, however both autonomous and non-autonomous elements can shape genome evolution.

An autonomous element is defined as a full length element (FLE) with functional transcriptional machinery that enables mobility and therefore proliferation in the host genome (Feschotte and Pritham, 2007). From this, it is derived that a non-autonomous element is unable to move around the genome independently, and either depends on an autonomous element to enable transposition, or is stationary in the genome, and therefore defined as inactive (Feschotte and Pritham, 2007; Slotkin and Martienssen, 2007). Non-autonomous elements are typically mobilised by neighbouring autonomous elements, via the acquisition of enzymatic domains that allows the element to be cleaved and integrated at a new site in the genome by the process of hitchhiking (Feschotte and Pritham, 2007). The enzymatic domains encoded by a mobile element is the basis of transposable element classification, allowing elements to be assigned to one of two classes (Figure 1.1).

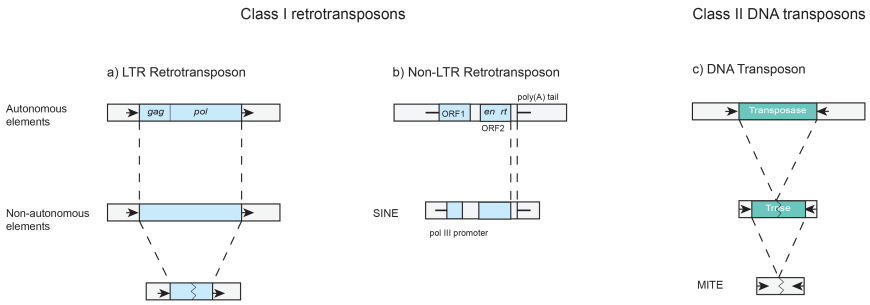


Figure 1.1: **Structural composition and classification of transposable elements.** Mobile elements are divided into two classes; Class I retrotransposons (a+b), and class II DNA transposons (c). (a) Class I elements are further categorised based on the presence or absence of long terminal repeats flanking the element; LTR and non-LTR retrotransposons. Autonomous elements transpose independently, encoding tranpositional machinery (*gag, pol*, ORF; open reading frame, en; endonuclease and rt; reverse transcriptase). Non autonomous elements do not have protein coding potential, but posssess cis elements which allow for transposition. Figure is based on Wessler (2006).

1.1.2 Transposition in DNA transposons

Element classification was established based on transposition mechanisms and are categorised into Class I and Class II elements (Figure 1.2) (Finnegan, 1989). Class II elements, DNA transposons, transpose directly as DNA, and do not require a reverse transcription step for mobility (McClintock, 1950). Transposition is enabled by Transposase, an enzymatic domain encoded by autonomous DNA transposons, which identifies inverted terminal repeats (ITRs) that are found at the terminal ends of the element to allow for element excision (Figure 1.2) (Slotkin and Martienssen, 2007). The Transposase enables the element to be cleaved from its original position, and reintegrated at another site of the genome (Slotkin and Martienssen, 2007). In contrast, *Helitron* DNA transposons do not possess ITRs or Transposase, and transpose via a rolling-circle mechanism that is catalysed by a DNA helicase protein (Slotkin and Martienssen, 2007). Diversity of both classes of TEs is host species dependent, with variance observed across the majority of eukaryotes (Carr, Nelson, Leadbeater and Baldauf, 2008; Carr et al., 2012; Lee and Kim, 2014; Elliott and Gregory, 2015).

1.1.3 Transposition in Retrotransposons

Class I elements, retrotransposons, transpose by an RNA copy, with the requirement of a reverse transcription step; the method is described as duplicative transposition, as the original copy is replicated and inserted into a new position in the genome (Slotkin and Martienssen, 2007). The duplicative nature of this class commonly results in an overall increase of family copy number (Cordaux and Batzer, 2009).

1.1.4 Retrotransposons

Retrotransposons are further classified into five orders, which are determined by the presence or absence of identical direct DNA repeats at the 5' and 3' end of the element, known as long terminal repeats (LTRs), as well as additional groupings known as *Penelope-like* elements and tyrosine-recombinase retrotransposons (Slotkin and Martienssen, 2007) (Figure 1.3). The orders themselves are further categorised into superfamilies.

Non- LTR retrotransposons

Non- LTR retrotransposons are comprised of two different elements; short interspersed elements (SINEs) or long interspersed elements (LINEs), which structurally differ to the LTR retrotransposon

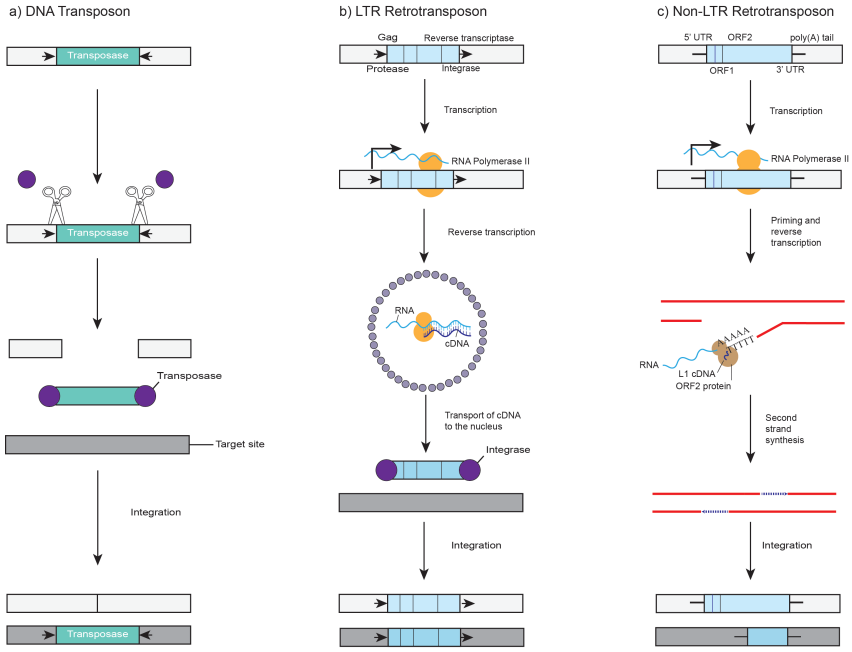


Figure 1.2: **Diversification of mechanisms of transposition.** Diverse mechanisms across all classes of transposable element. a) DNA transposons encode Transposase enzyme which allows for mobilisation, and are flanked by inverted terminal repeats (ITRs), illustrated by black arrowheads, which allow the element to be excised from its original position (scissors). Transposase (dark purple circles) binds to element ITRs allowing for integration into new genomic position (dark grey bar). b) LTR retrotransposons are mobilised via a "copy and paste" mechanism with the requirement of an RNA intermediate. LTR retrotransposons are flanked by long terminal repeats (LTRs), illustrated by black arrowheads, and encode *gag* and *pol* genes for tranposition machinery, including Protease, Reverse Transcriptase and Integrase. The promoter in the 5' LTR is recognised by RNA polymerase II, synthesising the element mRNA. Gag proteins (grey circles) assemble in a virus-like arrangement, encompassing the element mRNA, RT and IN. Here the RT transcribes the mRNA to cDNA, and IT (purple circles) integrates the element into its new genomic position (dark grey bar). c) Non-LTR retrotransposons are flanked by a 5' untranslated terminal repeat (UTR) and 3' polyA tail. An element encoded endonuclease cleaves a small target region of cDNA for integration, and second strand synthesis. Based on illustration from Levin and Moran (2011).

grouping (Eickbush and Jamburuthugoda, 2008) (Figure 1.1). Non- LTRs possess a 5' untranslated region, and a 3' polyA tail (Han, 2010). Similarity is drawn with LTR retrotransposons, in that groups typically possess two open reading frames (ORFs) (Eickbush and Jamburuthugoda, 2008) (Figure 1.2).

LTR retrotransposons

In contrast, the LTR retrotransposon classification system comprises of *Metaviridae* (*Ty3/gypsy*), and *Pseudoviridae* (*Ty1/copia*), BEL groups, tyrosine recombinase families and retroviruses (Figure 1.3) (Havecker et al., 2004). Structurally, LTR retrotransposons reflect an arrangement homologous to retroviruses (Xiong and Eickbush, 1990). However, tyrosine recombinase elements are flanked by ITRs, like DNA transposons (Piednoël et al., 2011). Like retroviruses, retrotransposons harbour

LTRs, and *gag* and *pol* genes, but lack an envelope (*env*) gene (Xiong and Eickbush, 1990) Three hypotheses have been suggested to shed light over the origin of the retroelements; (i) being that retroviruses originated from LTR retrotransposons, acquiring an *env*-like gene, (ii) that retrotransposons evolved from retroviruses, losing the env-like gene through selection, and (iii) that Retroviridae and Metaviridae are in fact sister groups, originating from the same common ancestor respectively (Hayward, 2017). Hypothesis one is predominantly accepted, but findings are still inconclusive (Hayward, 2017). Also, eukaryotic LTRs are frequently distinguished by conserved terminal dinucleotides T-G and C-A at the 5' and 3' end of the element (Freund and Meselson, 1984).

The typical structural arrangement of LTR retrotransposons is consisted of *gag* and *pol* genes that code for Gag-Pol proteins homologous to retroviral genes (Xiong and Eickbush, 1990). Gag plays a structural role, encoding for virus like particles that allow for reverse transcription, whereas *pol* encodes for several enzymatic domains to form a polyprotein of transposition machinery. The enzymatic capabilities that allow for transposition include; Protease (PR); Integrase (IN); Reverse Transcriptase (RT) and Ribonuclease-H (RNaseH). The classification of retrotransposons is determined by the structural composition of the polyprotein (Eickbush and Jamburuthugoda, 2008) (Figure 1.3).

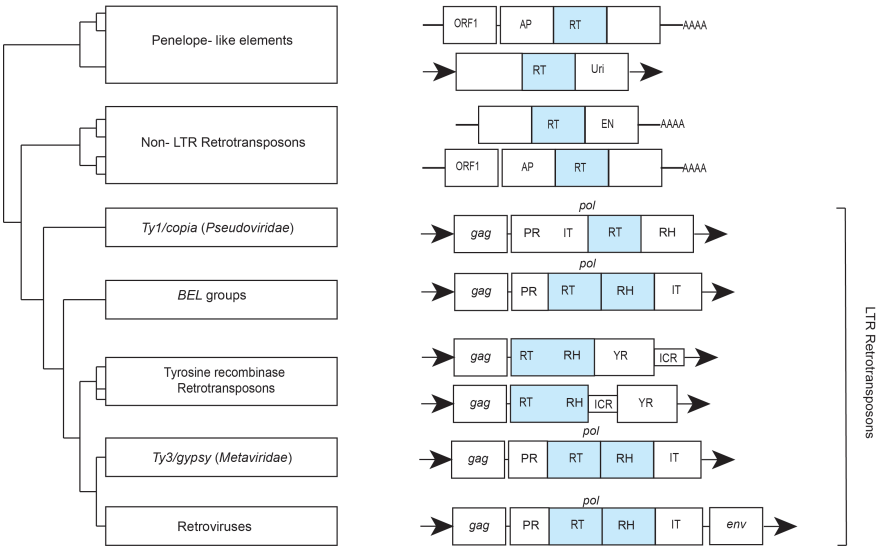


Figure 1.3: **Phylogenetic relationship and structure of class I retrotransposon families.** Left: RT based phylogeny of the main retrotransposon groups in cladogram format. Right: commonly accepted structure of mobile elements in each family. ORFs are represented by horizontal boxes, with those similar to *gag*, *pol* and *env* genes labelled consecutively. PR: protease; RT: reverse transcriptase; RH: RNase H domain; IN: integrase; EN: endonuclease domain; Uri: endonuclease-*like* domain found in some class I element introns; YR: tyrosine recombinase- *like*; ICR: internal complimentary repeats. Horizontal boxes with shaded arrowheads represent long terminal repeats (LTRs) and AAA: poly(A)tail. The blue boxes highlight RT and RH domains in the different families. The diagram is based on Eickbush and Jamburuthugoda (2008).

LTRs and their role in transposition

The presence of LTRs enable successful transposition to take place for LTR retrotransposons. The LTRs are comprised of three distinct structures; two unique regions (U3 and U5) flanking either side of a repeated region (R) (reviewed in Benachenhou et al. (2013)) (Figure 1.4). Transposition is initiated by the transcription of elements by the RNA intermediate to mRNA, originating at the 5' R region of the LTR, to the R region of the 3' LTR (Boeke and Corces, 1989; Zhang et al., 2014). The RNA intermediate is used as a copy for transposition by the process of reverse transcription to cDNA by Reverse Transcriptase. RNaseH is utilised in the degradation of the original RNA template, and the flanking U regions of 5' and 3' LTRs are replaced before integration back into the host genome by integrase (Zhang et al., 2014). The reintegration forms unique target site duplications due to the staggered breaks of host DNA upon transposition.

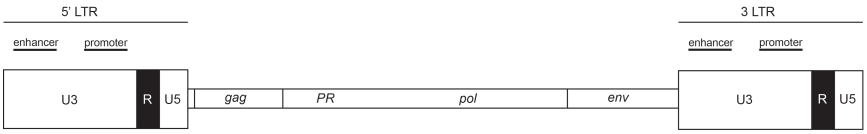


Figure 1.4: **Genomic organisation of long terminal repeats (LTRs) in retroelements**. LTRs are divided to three distinct regions; a repeated region, which is flanked by two unique regions - U3 and U5. Enhancer and promoter sequences that enable reverse transcription are found in U3. The polyA signal, and 5' capping sequences are encoded by the R region. Diagram is based on (Zhang et al., 2014).

1.1.5 Transposable element acquisition

The acquisition of TEs in eukaryotic species can occur by either vertical transfer, the inheritance of genetic material to daughter elements, or by horizontal transfer (Fortune et al., 2008). Horizontal gene transfer (HGT) is the mechanism in which genetic material is transferred from a donor species to a recipient, through interspecific mating or asexual processes(Marinoni et al., 1999; Bock, 2010). Due to the nature of the acquisition, the gene transfer is not confined to species barriers, thus recipient organisms can acquire genetic material from diverse donor species, of distant ancestral origin (Bock, 2010). Evidence has been found to support HGT events in eukaryotes, at both unicellular and multicellular levels (Fitzpatrick, 2012; Gilbert and Cordaux, 2013; Tucker, 2013; Walsh et al., 2013; El Baidouri et al., 2014). The processes in which HGT occurs in eukaryotes is unclear, however close relationships between species may facilitate the transfer of genetic material (Wang and Liu, 2016).

As described by Bock (2010), the facilitation of horizontal transfer of transposable elements (HTT) was thought to be predominantly autonomous, active elements. However, research has found that non-autonomous elements can also be acquired via horizontal transmission (Bock, 2010; Marsit et al., 2015; Legras et al., 2018). It is through stochastic loss, and natural selection, that element individual copies should be eradicated from host genomes until extinction, however this is challenged by the known proliferation in both prokaryotes and eukaryotes (Bartolome et al., 2009). With this, it is proposed that HTT can be defined as a catalyst to rearrangements in the genome, due to the process increasing the abundance of TE invasions between non-mating organisms (Schaack et al., 2010).

Evidence for HTT between two species is inferred when reviewing nucleotide diversity between elements proposed to be asexually acquired, when compared to vertically transmitted DNA, providing that the TEs are under the same evolutionary pressures to the host species (Figure 1.5) (Bartolome et al., 2009). Horizontally transferred DNA would be expected to have a lower level of diversity between two species, in contrast to sequences which are vertically inherited (Bartolome et al., 2009). Activity can also be recognised in a genome by the identification of DNA in daughter species, which is found to be absent in the parent (Huang et al., 2012). Transferred TEs have been predominantly identified as retrotransposons in several taxonomic groups, and transfer is proposed to commonly be facilitated between sequences of high similarity (Silva et al., 2005). In budding yeast, several HGT events have been detected (Liti et al., 2005; Carr et al., 2012; Fitzpatrick,

2012). Horizontal transfer of TEs has been supported in *S. cerevisiae*, with the acquisition of two *Ty* elements (*Ty2* and *Ty3*) from other species within the genus; *S. mikatae* and *S. paradoxus* (Liti et al., 2005; Carr et al., 2012). Carr et al. (2012) provided evidence to support that the two *Ty* elements uncovered in *S. cerevisiae* have been acquired via horizontal transmission between sister species with the genus.

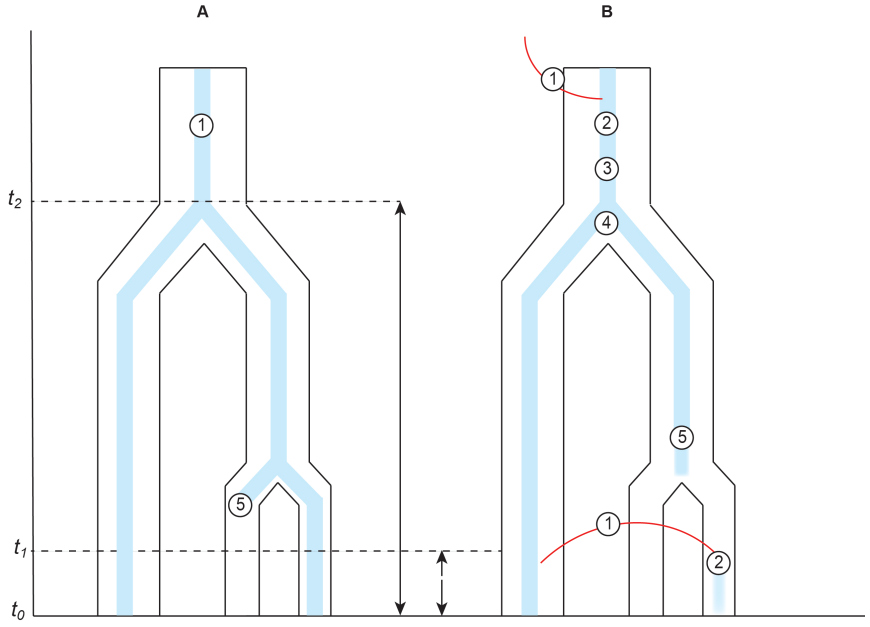


Figure 1.5: **Hypothesised transposable element acquisition and their host species**. (A) illustrates an example of vertical transfer (VT). The TE evolutionary pathway (in blue) corresponds to that of the host. TE eradication is common speciation events, eventually resulting in element extinction in the host species lineage, as well as random stochastic loss (5). (B) illustrates the mechanism of horizontal transfer (HT). The red line represents the recipient genome acquiring mobile DNA via horizontal transmission (1), and therefore facilitating TE persistence. The TE invasion causes proliferation in the recipient genome (2), until an equilibrium is released between selection pressures and transposition. The diversity between TEs upon insertion (4), can reverse previously described stochastic loss (5). Presuming that TEs are subjected to similar selection pressures as host genes, the synonymous divergence of TEs should reflect that of the host nuclear genes, as equal time has elapsed since inheritance (t0 – t2). In contrast, horizontally transmitted elements will have been subjected to fewer mutations as less time has elapsed since insertion (t0 – t1), thus a decreased level of divergence is expected when compared to host nuclear genes. Adapted from (Bartolome et al., 2009).

1.1.6 Advantages and disadvantages of TE insertions

Oliver and Greene (2012) successfully provided data that supports TEs having a major evolutionary impact, contributing significantly to genome evolution. The review detailed evidence of TEs being ”helpful”, in contrast to the detrimental view that was previously adopted (Doolittle and Sapienza, 1980; Oliver and Greene, 2012). Adaptation of the host genome has been facilitated by mobile elements, where the elements have been drivers of species diversity (Oliver and Greene, 2009; Warren et al., 2015). An extensive review of the specific characteristics that TEs possess highlighted the features that enable suitability for several influential events in the genome; agents of lineage and genome evolution, mutagenic agents to manipulate the rate that these processes occur and transposition due to interspersed insertions within the genome (Oliver and Greene, 2009).

The impact of TEs has been extensively reviewed over the past three decades (Doolittle and Sapienza, 1980; Orgel and Crick, 1980; Bartolome et al., 2009; Reilly et al., 2013; deHaro et al., 2014; Van't Hof et al., 2016). There are several forms of genomic rearrangements that can occur by transposition, which can arise with the integration of full length elements (FLE) or partial (truncated) elements (Goodier and Kazazian, 2008). These include inversions and deletions, transduction and recombination (Goodier and Kazazian, 2008; Lee et al., 2015). The insertions into new locations of the genome can have both beneficial and detrimental influence on the host genomes, causing insertional mutagenesis, deletions, splicing and transcript pausing/termination (Goodier and Kazazian, 2008).

Both autonomous and non-autonomous elements have been documented to insert into specific regulatory or coding regions within the host, reviewed by Kazazian (2004). Insertion patterns that accommodate transposition close to coding regions enable TE involvement in transcription, and thus influencing gene expression within the genome (Belyayev, 2014). Recombination has been found to be a strong driver of genomic rearrangement, specifically ectopic recombination (Kupiec and Petes, 1988; Charlesworth et al., 1997). The recombination between two regions of similarity (two LTR sequences) results in the removal of internal retrotransposon DNA and one LTR, with a solo LTR remaining (Kupiec and Petes, 1988). Therefore, the presence of solo LTRs in the host genome is a marker of ectopic recombination (Shirasu et al., 2000). The detrimental effects of TE insertions have been studied at length, however the genomic influence of the insertion is dependent on the target site, and the resulting mutagenesis (Beauregard et al., 2008; Levin and Moran, 2011).

In contrast, beneficial insertions have also been documented in several eukaryotic species, which has resulted in an increase in host fitness (Joly-Lopez et al., 2012, 2016; Van't Hof et al., 2016). Advantageous insertions of TEs have been acknowledged by the process of molecular domestication, where the elements have evolved to adapt to new molecular processes which are beneficial to the fitness of the host (Joly-Lopez et al., 2012; Jangam et al., 2017). Host organisms adapt transposable elements to their own advantage, for several purposes in the genome (Kaneko and Banno, 1991). Examples of domesticated elements include; *recombination activating genes* (*RAG1* and *RAG2*), and Telomerase Reverse Transcriptase, which is propsed to have evolved from an L1 relic (Kaneko and Banno, 1991; Miller et al., 1999; Curcio M. and Belfort, 2007).

Furthermore, TEs have been utilised through the application of transgenics (Wilson et al., 2007). The non-viral autonomy of TEs could provide a safe vector for gene delivery and thus employed in gene therapy (Wilson et al., 2007). *The Sleeping Beauty* (SB) transposon system was the first TE based gene therapy that was introduced, and employs the *Tc1*-like DNA transposon from the *mariner* superfamily (Ivics et al., 1997). Further transposon systems have been established, including *Tol2* DNA transposons from the *hAT* superfamily, and *piggybac/PB* transposons, which have been successful catalysts for efficient transposition in varied vertebrate models (Grabundzija et al., 2010). The application of TE domesication has been utilised in several genetic diseases, specifically cystic fibrosis, of which *piggybac* has been selected as a therapeutic agent (Cooney et al., 2015).

Although the presence of TE control mechanisms is inconsistent, it has been found that specific hosts, including *S. cerevisiae* can control TE mobility, as well as target site specificity (Curcio et al., 2015), and TE function by domestication (Sinzelle et al., 2008). An example of domestication include RAG proteins previously mentioned, which enable variable, diversity and joining (V(D)J) recombination (Melek et al., 1998). The process of V(D)J recombination catalyses the production of several antigen receptors in humans (Bassing et al., 2002). RAG proteins are homologous to DNA transposons, and are proposed to be derived from the class II elements (Melek et al., 1998).

1.1.7 TE elimination

Significant differences are seen in TEs across eukaryotic species, in relation to copy number, relative abundance and genomic influence. New TE insertions are commonly eradicated from the host genome by purifying selection, or genetic drift, resulting in TE relics throughout host

genomes (Carr et al., 2012). Mechanisms have evolved in eukaryotes in an attempt to control TE proliferation. The extensively documented mechanism of TE regulation is DNA methylation, predominantly seen in plants, mammalian and fungal species (Martienssen and Colot, 2001). In retrotransposons, cytosine methylation silences transposition by blocking the transcription of the RNA intermediate required for successful transposition (O'Donnell and Burns, 2010). However, several suppression mechanisms have been documented, such as RNA interference (RNAi) and host proteins causing TE inactiviation (Muñoz-López and García-Pérez, 2010).

The elimination process of RNAi, which has remained conserved from an ancestral eukaryotic species throughout the fungal kingdom, is documented as lost in several budding yeast species, including *S. cervisiae* (Drinnenberg et al., 2009). The mechanism involves RNA being silenced by an RNA degradation process, where double stranded RNA molecules destroy mRNA molecules (target transcripts) and therefore causing the termination of transcription (Agrawal et al., 2003). However, the yeast species lack homologues to RNAi host genes that enable cleavage of target transcripts (Drinnenberg et al., 2009). Observations of RNAi are described in the majority of eukaryotic taxa, including metozoans and protistan species which are part of supergroup, Opisthokonta (Agrawal et al., 2003).

1.2 Opisthokonta

Opisthokonts are a supergrouping of Fungi, Metazoa, *Incerta sedis* species, and three major protistan groups – Choanoflagellata, the nucleariids and Ichthyosporeans (Del Campo et al., 2015). The exclusive grouping was first proposed based on the basis of the posterior single flagellum in the 1940s (Visher, 1945), which was further supported by Cavalier-Smith (1987). From this initial classification, multigene phylogenies were produced to validate the grouping, including small subunit ribosomal (SSU) RNA and several host genes (Baldauf and Palmer, 1993), and subsequently it was defined as "Opisthokonta" (Cavalier-Smith and Chao, 1995). The supergroup further diversifies into two main lineages; Holozoa and Holomycota (Lang et al., 2002; Liu et al., 2009). Holomycota includes Fungi, and unicellular relatives nucleariids and *Fonticula alba* (Lang et al., 2002; Brown et al., 2009); The group Holozoa contains Metazoa, ichthyosporeans, choanoflagellata and filasterea (Shalchian-Tabrizi et al., 2008) (Figure 1.6).

Opisthiskonts have been of importance in the field of phylogenetics, with regards to predicting ancestral pathways, and the origins of multicellularity (Cavalier-Smith, 2017). An increase in

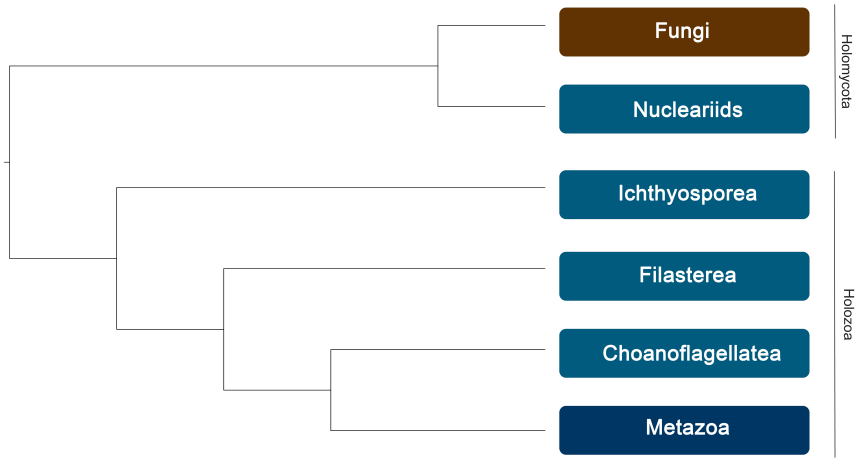


Figure 1.6: **Consensus cladogram of evolutionary relationships among Opisthokonta**. Opisthokonta are a supergrouping of two main lineages; Holozoa, which contain Metazoa, Choanoflagellates, Filasterea and Ichthyosporea; and Holomycota, containing Fungi and nucleariids. the cladogram was constructed in Newick format, and based on the relationship seen in Shalchian-Tabrizi et al. (2008).

genome sequencing across several eukaryotic species, has provided ample opportunity for extensive genomic analyses, and with this large quantities of TE data (Pritham, 2009). TE research in species of Opisthokonta has predominantly focused on multicellular organisms, such as fungal and metazoan species, with the majority of unicellular eukaryotes being some what over looked. Although in minority, TE research in single celled organisms has been based upon those found in the reference strain of the holomycota budding yeast species, *Saccharomyces cerevisiae* (Kim et al., 1998; Carr et al., 2012).

1.2.1 Yeast: a model organism

Yeast are eukaryotic, unicellular organisms, which have evolved from ancestors of multicellularity (Dickinson, 2005). The fungal microorganisms are not monophyletic, and have evolved from two separate phyla; Ascomycota and Basidiomycota (James et al., 2006). Within Ascomycota, and subphylum Saccharomycotina, budding yeasts are placed in the order Saccharomycetales, which is further categorised into superfamilies, including Saccharomycetaceae - which in itself contains 20 genera (Figure 1.7) (James et al., 2006).

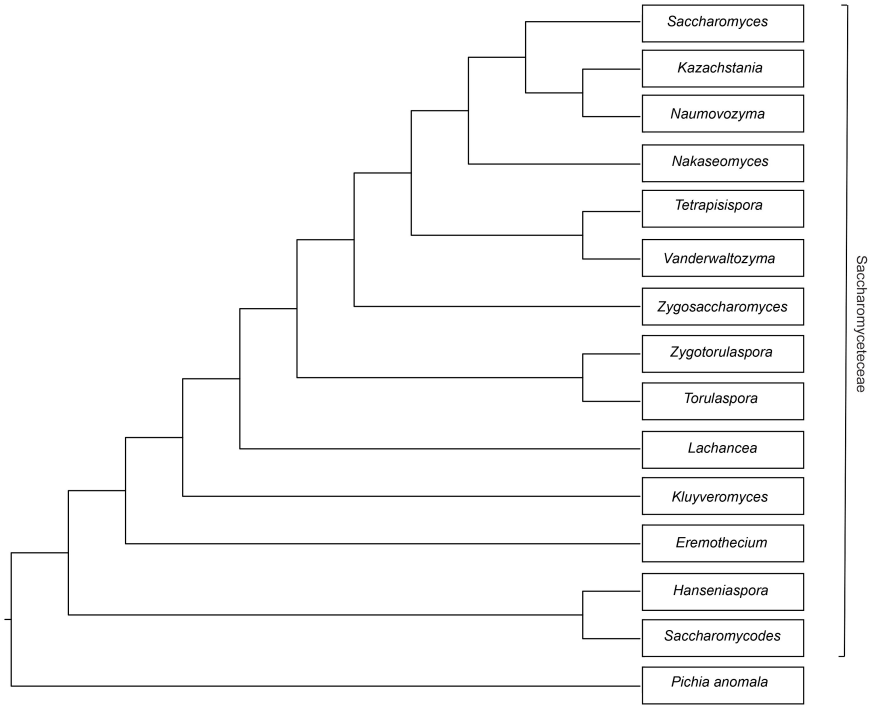


Figure 1.7: **Cladogram representation of species in the Saccharomyceteceae superfamily**. The 14 clades of Saccharomyceteceae are annotated at the terminal node respectively. *Pichia anomala* is used as the outgroup species. Species relationships are based on Kurtzman (2003).The cladogram was produced in Newick format.

S.cerevisiae is part of the superfamily Saccharomycetaceae ((Kurtzman, 2003; Wolfe et al., 2015)); the family was subjected to whole genome duplication event (WGD) approximately 100 million years ago (Wolfe and Shields, 1997; Chapman et al., 2004). This event caused the definition of a clade 'post-WGD species', of which the genomes possess features to support this shared event, whereas species that diverged from *S. cerevisiae* lineage prior to the WGD are classed as non-WGD species (Wolfe et al., 2015). The root of the WGD, the ancestral organism that was subjected to the genome duplication, originally possessed 5000 genes and post-WGD increased to 10000 genes; however most new copies seemed to be 'lost' (Wolfe et al., 2015). Post-WGD species usually include 500 pairs of genes in the 5500 genes documented, that were created by the WGD. The remaining 4500 loci were not conserved in duplicate form and only one copy of the gene survives (Wolfe et al., 2015).

The 12.2Mb draft genome of *S. cerevisiae* was the first eukaryotic species to be sequenced and has given great insight into yeast and TE evolution in unicellular eukaryotes (Goffeau et al., 1996; Kim et al., 1998; Carr et al., 2012; Bleykasten-Grosshans et al., 2013), being defined as a model organism (Botstein et al., 1997). The first valuable application of yeast as a model system was by the mapping of disease mediated genes in humans (Botstein et al., 1997). The disease

mediated gene of interest was identified in the host at a cellular level, and homology found in additional organisms, such as *S. cerevisiae* (Botstein et al., 1997). Building on this simplistic view, the application of the experimental organism has pioneered work in functional genomics (Botstein and Fink, 2011).

Ty elements of *Saccharomyces cerevisiae*

The *S. cerevisiae* reference genome (S288c) is made up of approximately 3% TEs (Carr et al., 2012). LTR retrotransposons have predominantly been found in *S. cerevisiae* with the majority of insertions being partial length elements (solo LTRs or truncated elements) (Kim et al., 1998). To date, only three *S. cerevisiae* strains have been found to have copies of a DNA transposons (Sarilar et al., 2015; Legras et al., 2018). A *hAT* transposon was the first to be discovered in a wild strain of *S. cerevisiae* (AWRI1631) (Sarilar et al., 2015). The *hAT* elements have been uncovered in several species of Ascomycota, and classified as presumably active, with the presence of multiple copies per genome (Sarilar et al., 2015). However, the element was only found as a single copy in *S. cerevisiae*, with limited synteny to copies found in the sister species to *Saccharomyces*, *Naumovozyma dairenensis* (Sarilar et al., 2015).

The yeast genome has been found to harbour six LTR retrotransposon families, *Ty1* – *Ty5* (Kim et al., 1998) and *Ty3p* (Carr et al., 2012). The majority of TE insertions are active families *Ty1* and *Ty2*, which comprise >75% of the genome TE content in the reference strain, S288c (Kim et al., 1998). Although similar, the families can be distinguished by diversity in both *gag* and *pol* ORFs (Kim et al., 1998). However, it is the high level of LTR identity between families, as well as breakpoints in the *pol* gene, that allows for intra-species recombination events, resulting in the presence of hybrid *Ty1/2* LTRs in the host genome (Jordan and McDonald, 1999; Carr et al., 2012). Both *Ty3* and *Ty4* are found as intact FLEs in the genome, however *Ty4* is presumably the least active of the two families, as there are fewer copies uncovered to support recent transposition events observed for this family (Carr et al., 2012). *Ty5* has also been defined as inactive, with only truncated fragments annotated to date (Kim et al., 1998; Zhu et al., 1999; Carr et al., 2012).

Due to the extensive study of mobile elements in *Saccharomyces cerevisiae*, the *Ty* families are the reference of all yeast TEs as a basis of comparison (Table 1.1) (Curcio et al., 2015; Neuvéglise et al., 2002). The five families range from 2600 - 5500 base pairs (bp) in length,

Table 1.1: **Genomic Organisation of *Ty* elements in *Saccharomyces cerevisiae***. *Ty* elements consist of flanking LTRs, which are positioned either side of the *TYA* and *TYB* ORF. The size of LTRs and Gag/Pol ORFs are included. Adapted from (Jordan and McDonald, 1998).

	Size (bp)		
Family	**LTR**	**ORF**	**Group**
Ty1	334	5250	*copia*
Ty2	332	5300	*copia*
Ty3	340	4730	*gypsy*
Ty4	371	5480	*copia*
Ty5	251	2640	*copia*

with LTR size of 250-370 bp. *Ty1/2* are the largest of the families, with *Ty5* described as the smallest. *Ty1* LTRs possess the dinucleotide inverted repeat 5'-TG and CA-3' at each terminal end, and are documented to encompass three domains; R, U3 and U5 that are consistent in all LTR retrotransposons (see Section 1.4) (Sandmeyer et al., 2015). Furthermore, the domains are characterised by their position in the major sense strand that is expressed from the DNA of *Ty1* – U5 and U3 are specific to the 5' and 3' end of *Ty1* RNA, whereas the R domain is replicated at both ends of the *Ty1* transcript (Clare et al., 1988). *Gag* (TYA) and *pol* (TYB) open reading frames overlap in *Ty1* -the Pol ORF overlaps the end 38 bp of the Gag protein, and encodes PR, IN, RT and RNaseH (Figure 1.8). (Curcio et al., 2015).

The *Ty1* replication cycle results in the parental element, and a copy of the retrotransposon in the host genome – the process is intracellular, mostly occurring in the cytoplasm of the cell (Curcio et al., 2015). RNA polymerase II transcribes *Ty1* elements, and *Ty1* RNA Gag-Pol and Gag are assembled to form nucleocapsids (virus-like particles). In the nucleocapsid, cleavage of Gag and Pol proteins are catalysed by a protease, which creates mature enzymatic domains (Curcio et al., 2015). Post maturation, the *Ty1* RNA is reverse transcribed, by RT, to form double stranded, linear DNA which is then transported into the nucleus of the cell – integrase targets host proteins to enable *Ty1* cDNA insertion in target areas of the host genome (Curcio et al., 2015).

Similarly, full length *Ty3* elements are comprised of two LTRs that flank the element which possess Gag3 and Pol3 ORF that also overlap (Figure 1.8) (Sandmeyer et al., 2015). Transcription of *Ty3* results in genomic RNA, which is then translated; Gag3 and Gag3-Pol3 are produced with the same polyprotein components of *Ty1*, however the catalytic domain order of Pol differs in *Ty3* (PR;RT; IN) (Sandmeyer et al., 2015). Genomic RNA, Gag3 and Gag3-Pol3 accumulate to

form virus like particles in the cytoplasm where *Ty3* RT reverse transcription generates cDNA from genomic RNA, which is integrated into the transcription start site of RNA polymerase II transcribed genes (Sandmeyer et al., 2015). The success of TE transposition highlights the influence on host genome characteristics.

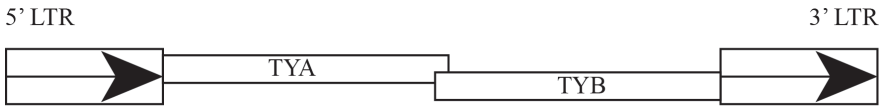


Figure 1.8: **Genomic Organisation of *Ty* elements in *Saccharomyces cerevisiae***. The consensus structural organisation of the LTR retrotransposon is comprised of two ORFs, TYA and TYB which overlap, and flanked by LTRs at the 5' and 3' end of the element, illustrated by black arrow heads. Diagram based on Jordan and McDonald (1998).

Ty insertion patterns

As outlined, TEs have been found to control target site specificity (van Luenen and Plasterk, 1994). The integration site heavily influences the success of the mobile element, which is defined by the elements ability to transpose and proliferate in the host species genome. With this, insertion into non-coding DNA should be under decreased selection pressure, allowing families to thrive. In *S.cerevisiae*, both *Ty1* and *Ty2 copia* elements have been documented to integrate upstream of Polymerase- III (Pol III) transcribed genes, usually 80bp upstream of the coding sequence (Kim et al., 1998). tRNA genes predominantly possess the majority of Pol III transcribed genes, and have limited coding DNA (Figure 1.9) (Boeke and Devine, 1998). Due to this, the genomic effect that *Ty1* has on gene expression is reduced (Boeke and Devine, 1998). This targeted integration is suggested to aid in the facilitation of double strand break repair by *Ty1* recombination, which in turn would improve host fitness (Cheng et al., 2012). Similarly, *Ty3* has been found to target upstream of tRNA genes (Figure 1.9) (Boeke and Devine, 1998).

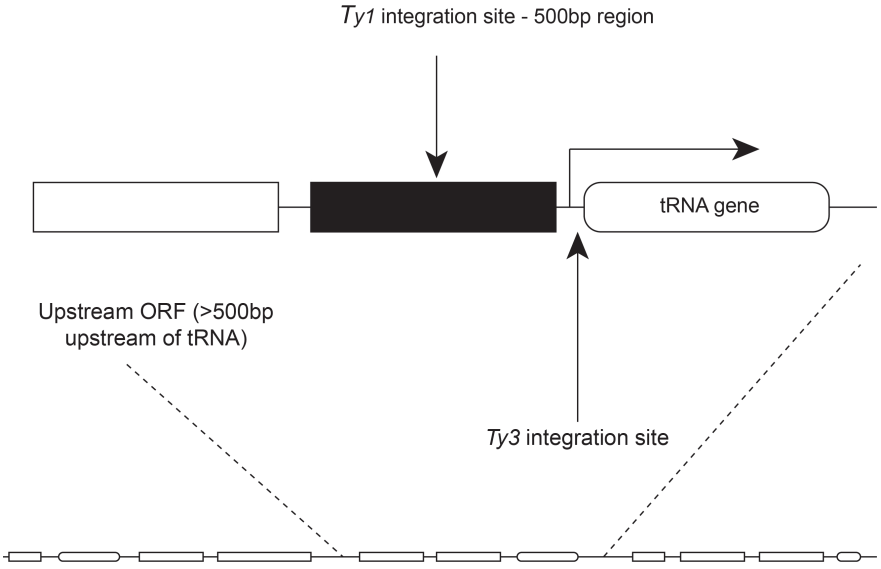


Figure 1.9: **Target integration site patterns of LTR retrotransposons; *Ty1* and *Ty3*, upstream of Pol III**. *Ty1* and *Ty3* are found to target upstream of pol III, by independent insertion mechanisms.The yeast chromosome, represented by the lower line, is annotated with boxes to denote protein coding regions. tRNA gene is shown in an oval shape, downstream to the hypothesised integration sites. Based upon Boeke and Devine (1998).

TE in abundance in the Saccharomycetaceae superfamily

Additional fungal species have been reviewed for TE content since the annotation of *S. cerevisiae*; many of which are from the Saccharomycetaceae superfamily (Neuvéglise et al., 2002; Novikova and Blinov, 2009; Bleykasten-Grosshans et al., 2011; Muszewska et al., 2011; Carr et al., 2012; Bleykasten-Grosshans et al., 2013). At present, *S. cerevisiae* has the highest percentage of TE genome content (Carr et al., 2012). However, other yeast species reflect similar characteristics, with several possessing *Ty-like* elements (Neuvéglise et al., 2002; Muszewska et al., 2011). The work outlined by Muszewska et al. (2011) detected LTR retrotransposons in 58/59 analysed fungal genomes, the majority of which were classified as *Ty3/gypsy* but *Ty1/copia* were also identified (Muszewska et al., 2011). Liti et al. (2009) found that lab strains of yeast species have an increased abundance of TEs when compared to wild populations, with the isolated conditions enabling the mobile DNA to proliferate and be subjected to limited selection pressures. Wild yeast strains have been found to present decreased TE copy numbers, supporting effective elimination of TEs by natural selection via several elimination mechanisms employed by the host (Drinnenberg et al., 2009) (see Section 1.1.7).

Muszewska et al. (2011) reviewed LTR retrotransposons in fungi by analysing all 59 previously published fungal genomes that had documented TE content – this included full, functional elements,

partial elements (detected by known enzymatic domains) and solo elements (Muszewska et al., 2011). 66,000 TEs were identified, all of which were categorised as either *Ty3/gypsy* or *Ty1/copia* retrotranposon superfamilies (Muszewska et al., 2011). The majority of the elements represented Chromoviridae (an LTR retrotransposon family) characterised by the presence of a chromodomain in the *pol* ORF (Muszewska et al., 2011). The data was analysed by documenting the LTR TE content in various types of individual yeast species, recording the highest copy element from each of the genomes (Muszewska et al., 2011). This review gave great insight into the variety of TE content that spans across a taxonomic group, as element abundance fluctuated from 50 to 8000 copies (Muszewska et al., 2011). These observed TE expansions seen in fungi were found to correlate with an increase in copy number for the number of types of elements, as well as individual elements, with *Ty3/gypsy* representing the highest copy number in all yeast genomes reviewed (Muszewska et al., 2011). Further phylogenetic analysis of the TEs supported the theory that the TE expansions appeared independently in more distant yeast genomes, in various taxonomic groups (Muszewska et al., 2011).

Carr et al. (2012) took the opportunity to further investigate *Ty* evolution in *S. cerevisiae* by systematic re-annotation of *Ty* elements in the reference genome, S288c. The genome annotation was initially documented in a paper by Kim et al. (1998), who gave insight into the abundance of *Ty* elements within *S. cerevisiae*, including copy number and distribution of *Ty1-5* families. The re-annotation of S288c in Carr et al. (2012), saw an increase in copy number compared to Kim et al. (1998), as well as identification of a new family of *Ty3*- like elements, *Ty3p* (Kim et al., 1998; Carr et al., 2012). This is a prime example as to how advances in bioinformatics can spur re-evaluation of the phylogenetic relationships of TEs, create opportunity for further research to be employed and excitement of what is yet to be discovered. Carr et al. (2012) also successfully provided evidence to support that the *Ty2* family had arose in *S. cerevisiae* genome by the process of horizontal gene transfer (HGT), which is another theory to explain TE evolution (Carr et al., 2012; Tucker, 2013).

Furthermore, Bleykasten-Grosshans et al. (2013) examined complete genome sequences taken from 41 different strains of *S. cerevisiae*, including the reference genome S288c used in Carr et al. (2012). This study successfully shed light over *Ty*- related polymorphism (Bleykasten-Grosshans et al., 2013). The 41 strains were both geographically and ecologically diverse, reducing the limitation of sampling bias, often apparent in phylogenetics (Bleykasten-Grosshans et al., 2013). The results showed the differences that exist within strains in reference to number of full length

Ty elements, as well as evidence to support insertion polymorphism (Bleykasten-Grosshans et al., 2013). These findings have fuelled further investigation into the impact of such polymorphism on the phenotypic characteristics of these strains and emphasised the variability that can exist within species, genera and taxa (Bleykasten-Grosshans et al., 2013).

It is evident to see the influence that full genome sequencing of *S.cerevisiae* has had on predicting TE evolution in budding yeast, but many other members of Saccharomycetaceae are yet to be sequenced or investigated. The genus *Kazachstania*, closely related to *Saccharomyces*, is yet to be sequenced in full (Wolfe et al., 2015).

1.2.2 *Kazachstania* species

The genus *Kazachstania* is also from the Saccharomycetaceae superfamily (Kurtzman, 2003), and only has four species sequenced at whole genome level; the available species included *Kazachstania africana* CBS 2517, *Kazachstania naganishii* CBS 8797, *Kazachstania saulgeensis* CLIB 1764 and *Kazachstania servazzii* SRCM102023/CBA6004 (Sayers et al., 2009). Although documented as morphologically identical, *Kazachstania* species variability can be seen at DNA level (Figure 1.10). As a close relative to *S. cerevisiae* (see Section 1.2.1; Figure 1.10) (Kurtzman, 2003), the limited data availability for this genus leaves a niche left to be explored. The *Kazachstania* work will primarily explore TE content of four novel species from *Kazachstania*, that have been selected for greatest phylogenetic diversity based on multigene phylogenies (*K. bovina*, *K. exigua*, *K. lodderae* and *K. viticola*) (Table 1.2).

Table 1.2: **Characteristics of four novel *Kazachstania* species.**

	NCYC Number	Country of Origin	Origin
K. bovina	526	Unknown	Caecum of cow
K. exigua	814	Unknown	Fermenting cucumber brine
K. lodderae	1417	South Africa	Soil
K. viticola	2701	Kazakhstan	Fermenting grapes

Previous research into *Kazachstania* indicated it can be isolated from various habitats, such as soil, animal associated sampling and food products (Table 1.2). Partial sequencing of several species in the genera has shown phylogenetic diversity that exists across the species that are publicly available (Suh and Zhou, 2011). Kurtzman (2003) reclassified the superfamily, which reassigned several previously defined *Saccharomyces* species, to the new clade, *Kazachstania*.

Nisiotou and Nychas (2008) depicted the relationship between four yeast strains isolated from Botrytis affected fermenting grape juice, where phylogenetic analysis showed that the four strains represented a distinct species within the genus (Nisiotou and Nychas, 2008). Although successful in providing support of the affiliation of the four strains, which were given the name *Kazachstania hellenica*, and representing its relationship in reference to other *Kazachstania* species, the dataset was limited and therefore has given rise to further investigation into species within this genus (Nisiotou and Nychas, 2008).

Similarly, Suh and Zhou (2011) isolated three yeast strains from the gut of passalid beetle, *Odontotaenius disjunctus* (Suh and Zhou, 2011). These strains were also identified as species within *Kazachstania* – given the name Kazachstania intestinalis, and through sequence similarity searches; molecular phylogeny represented a basal lineage of a clade including other documented *Kazachstania spp*., none of which were a close sister to *K. intestinalis* (Suh and Zhou, 2011). This paper has effectively portrayed complexity of phylogenetic relationships within this genus, and supported that further investigation should be employed to determine further conclusions (Suh and Zhou, 2011).

Regarding TE content, Neuvéglise et al. (2002) constructed a study to investigate genomic evolution of LTR retrotransposons that reviewed 49,199 random sequence tags (RSTs) from 13 species of hemiascomycetous yeasts; two of the *Saccharomyces* species reviewed in this paper have been reclassified to the genus *Kazachstania* since the research publication (Neuvéglise et al., 2002; Kurtzman, 2003; Genolevures et al., 2009). LTRs were identified in all hemiascomycetous yeasts; 17 distinct families of full length elements were identified as well as five families of solo LTRs (Neuvéglise et al., 2002). In the *Kazachstania* species studied (*Kazachstania exigua* and *Kazachstania servazzii*), RSTs were matched to *Ty* proteins from *Saccharomyces* (Table 1.3). The newly identified elements were named according to the species name and *Ty* number element similarity (Neuvéglise et al., 2002). This finding supported that different elements exist in single hosts within this genus, and that a close relationship exists with *Saccharomyces* which would facilitate vertical inheritance and potentially horizontal transposable element transfer (Neuvéglise et al., 2002). The evolutionary model constructed for *Ty1/copia* in this review, can be expanded upon in future work, as well as the production of a similar evolutionary model to represent *Ty3/gypsy* (Neuvéglise et al., 2002). The presence of TE element copies in these two *Kazachstania* species supports further work to expand upon evolutionary genomics in this genus, drawing comparison

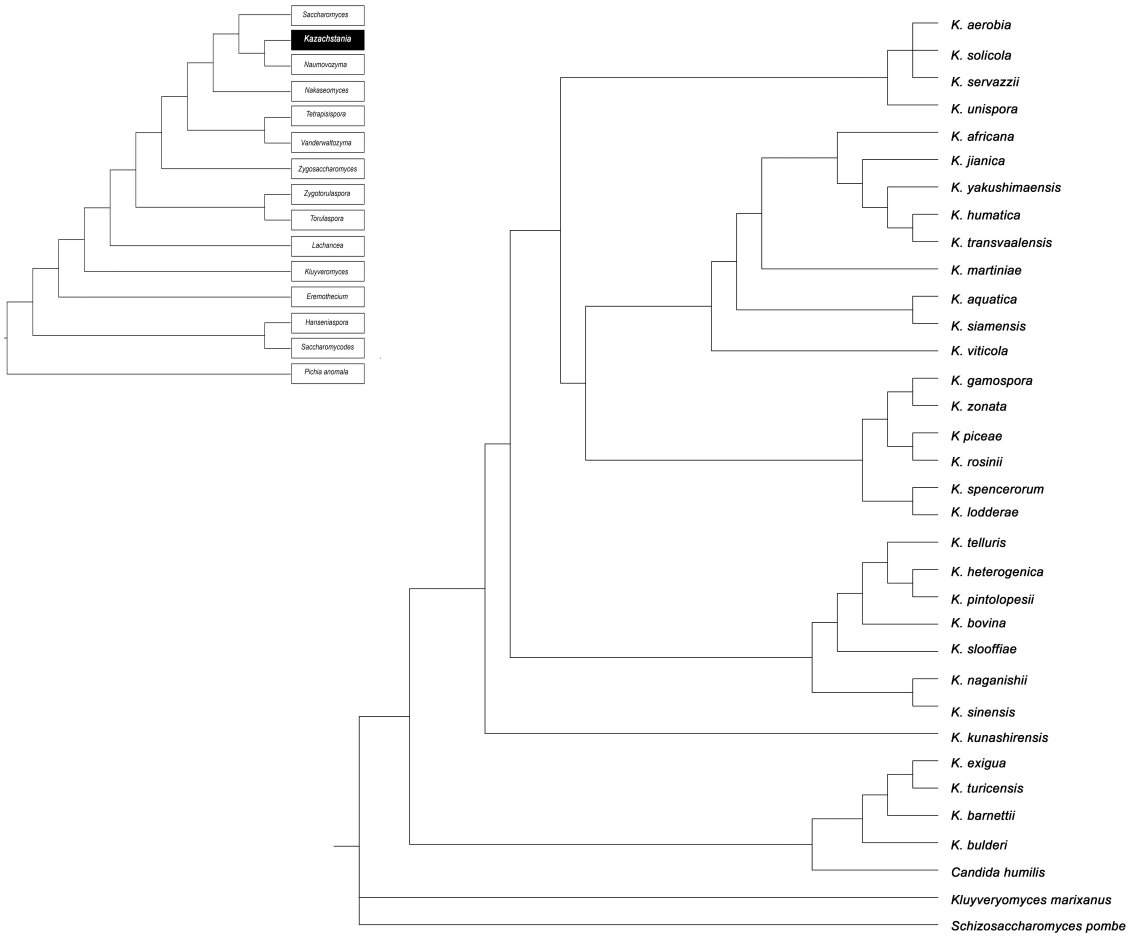


Figure 1.10: **Cladogram representation of *Kazachstania* spp.** The 31 species of *Kazachstania* are annotated at the terminal node respectively. *Kluyveryomyces marixanus* and *Schizosaccharomyces pombe* are used as the outgroup species. Species relationships are based on Vaughan-Martini et al. (2011). The cladogram was produced in Newick format. A cladogram of the Saccharomycetaceae phylogeny is annotated to show species relationships within the superfamily.

with *S. cerevisiae* and other species of the Saccharomyceteceae superfamily.

Table 1.3: **Homology between RSTs and Ty families in Saccharomyces.** Random sequence tags (RSTs) were matched with all *Ty* families found in *Saccharomyces*, and homologous copies were found in *K. exigua* and *K. servazzii*. Data was proposed by Neuvéglise et al. (2002).

Yeast species	***Ty1/2***	***Ty3***	***Ty4***	***Ty5***
K. exigua	19 (Tse1)	15 (Tse3)	0	10 (Tse5.1, Tse5.2)
K servazzii	1 (Tss1)	3 (Tss3)	0	0

1.2.3 Choanoflagellates

Choanoflagellates are a second line of unicellular eukaryotes, a lineage of Opisthokonta (Figure 1.6), and the closest living relative to metazoans (Lang et al., 2002). Discovered by James-Clark (1868), a likeness between the morphology of choanoflagellates and the collared cells of Porifera (sponges) was described, leading to the confirmed relationship between the choanoflagellates and Metazoa (Lang et al., 2002). These protists are found in both freshwater and marine environments, often existing as single-celled species, but occasionally as ephemeral multicellular colonies (King, 2005).

Choanoflagellates are characterised by a distinct feeding collar of microvilli, and apical flagellum (Figure 1.11) (Hoffmeyer and Burkhardt, 2016). The actin-filled microvilli feeding collar allows for effective prey capture of bacteria, and aquatic detritus, through the process of phagocytosis (King et al., 2009).

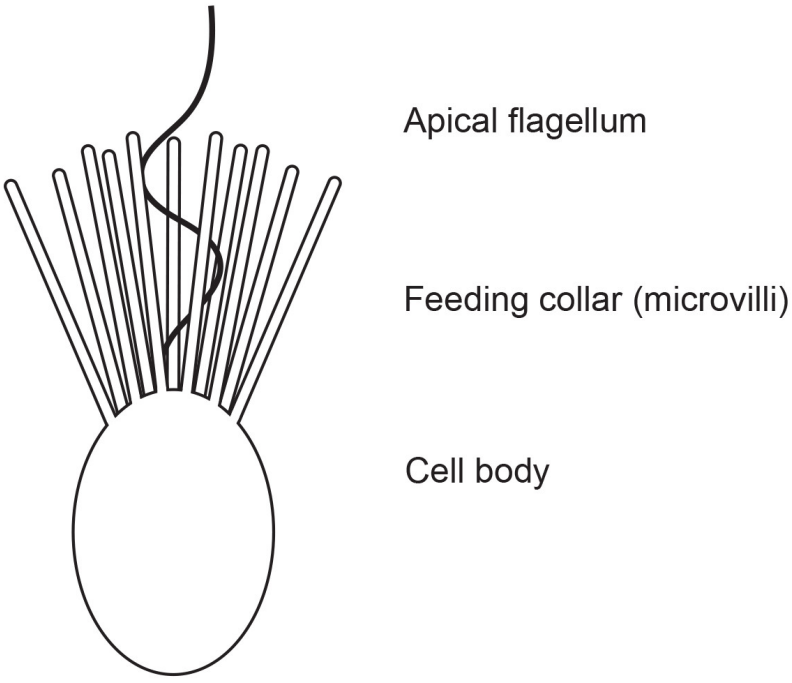


Figure 1.11: **Simplified choanoflagellate morphylogy.** Choanoflagellates are comprised of a cell body, microvilli feeding collar and apical flagellum, as annotated. Image is based on (Hoffmeyer and Burkhardt, 2016).

Although the choanoflagellate cell morphology is conserved (Figure 1.11), the external covering (periplast) is highly variable across species; this initially lead to species classification into distinct families; Salpingoecidae, Acanthoecida, and Codonosigida (Norris, 1965). The morphology of Salpingoecidae presents a rigid theca, whereas Codonosigida instead possess a mucilaginous

cover, otherwise named a glycocalyx (Carr, Leadbeater, Nelson and Baldauf, 2008). However, molecular phylogenies show that the two families did not descend from the same common ancestor, and thus were not monophyletic as first hypothesised (Carr et al., 2017). The remaining species, within the order Acanthoecida, are characterised by “cage-like” silica baskets (Carr et al., 2017). Phylogenetic analysis supported that Codonsigidae evolved from within the Salpingoecidae lineage, and thus formed a major clade, Craspedida, which is subdivided into three groupings.

King et al. (2008) outlined that >125 species of choanoflagellate species are identified, with several more presumably undiscovered (King et al., 2008). The first choanoflagellate to provide genome availability, *Monosiga brevicollis*, was sequenced by the Joint Genome Institute (JGI), and is a marine choanoflagellate with a globally wide distribution (King et al., 2008). The genome is 41.7Mb in size, with >9000 genes annotated. Although a small genome, when compared to the majority of metazoan species, *M. brevicollis* genes are significantly intron-rich, with an average of 6.6 introns per gene (King et al., 2008).

The second choanoflagellate with genome availability was *Salpingoeca rosetta*, previously known as *Proterospongia* sp., sequenced by the Broad Institute (BI) (Sayers et al., 2009; Fairclough et al., 2010). *S. rosetta* was predicted to have >11000 gene models, with a greater genome size of 55.4Mb, compared to *M. brevicollis* (Sayers et al., 2009). In contrast to *M. brevicollis*, *S. rosetta* is a colony forming species, by a process that mirrors the early phases of marine invertebrate embryogenesis (Fairclough et al., 2010). The possible evolutionary relationship between metazoan embryogenesis and colony development outlines the importance of further study for the unicellular species, at both molecular and cellular level. This ancestry allows an insight into animal origin and early evolution of metazoan species. Theoretically, this will enable pathways to be predicted regarding gene loss and gain, and TE acquisition. Furthermore, as the unicelluar eukaryotes are typically prey for several aquatic multicellular organisms, choanoflagellates create opportunity for HGT to take place including the transfer of mobile elements. As phagotrophs, the species also harbour the ability of gene acquisition from their host species (Yue et al., 2013). With this, Choanoflagellatea are an additional class to be revised for TE abundance (Carr, Nelson, Leadbeater and Baldauf, 2008).

1.2.4 Filasterea

As a member of the Opisthkonts, *Capsaspora owczarzaki* is the only representative of the genus, *Capsaspora*, assigned to the holozoan clade, Filasterea (Shalchian-Tabrizi et al., 2008) (Figure 1.6). Similarly, to choanoflagellates, *C. owczarzaki* is a second close relative to Metazoa, and therefore provides further insight to ancestral predictions of multicellularity in the animal lineage.

1.2.5 Transposable elements in protists

The globally distributed choanoflagellate, *Monosiga brevicollis*, is the only species of choanoflagellate which has been annotated for mobile elements, and has been found to possess three families of LTR retrotransposons only (Carr, Nelson, Leadbeater and Baldauf, 2008). Further insight into the evolution of TEs in opisthokont protists was provided by the annotation of TEs in the filasteran protist, *Capsaspora owczarzaki* (Carr and Suga, 2014). The draft genome presented 23 families from Class I and Class II TEs. The TEs identified were shown to have orthologous TE families in Metazoa, which supports the hypothesis that the common ancestor of the metazoans had a diverse repertoire of mobile elements (Carr and Suga, 2014).

The two marine choanoflagellates, *Monosiga brevicollis* and *Salpingoeca rosetta* are both placed in the major clade of choanoflagellates, Craspedida (Jeuck et al., 2014). *S. rosetta* is yet to be studied regarding TE content, allowing for a comparative genomic study to unfold with the previously reviewed *M. brevicollis*. With genome availability, and the known close relationship to metazoans, *M. brevicollis* and *S. rosetta* are found to be presenting as promising model organisms to predict evolutionary origin of multicellularity (Hoffmeyer and Burkhardt, 2016). In addition to *C. owczarzaki*, the three holozoan species are key subjects of investigation to aid understanding of ancestral pathways that present in animals in present day. Exhaustive research has detailed evolutionary traits in multicellular species, including trends in codon usage bias. With extensive genomic research focused on metazoan species, the availability of the protistan species allowed for the reconstruction of several ancestral traits, including codon usage (Southworth et al., 2018).

1.3 Codon usage

1.3.1 Background

The study of choanoflagellates, as the closest known relatives to Metazoa, has aided depiction of evolutionary pathways of the Opisthokonts (Carr, Nelson, Leadbeater and Baldauf, 2008; Carr et al., 2010; Suga et al., 2013; Tucker, 2013; Carr et al., 2017). Recent research has determined the genome complexity of the last common ancestor of Holozoa, in that the majority of host genes are homologous to Metazoa, which were previously determined to be kingdom specific (Hehenberger et al., 2017). With this, traits uncovered in both choanoflagellates and filastereans are ancestral to Metazoa (Figure 1.12).

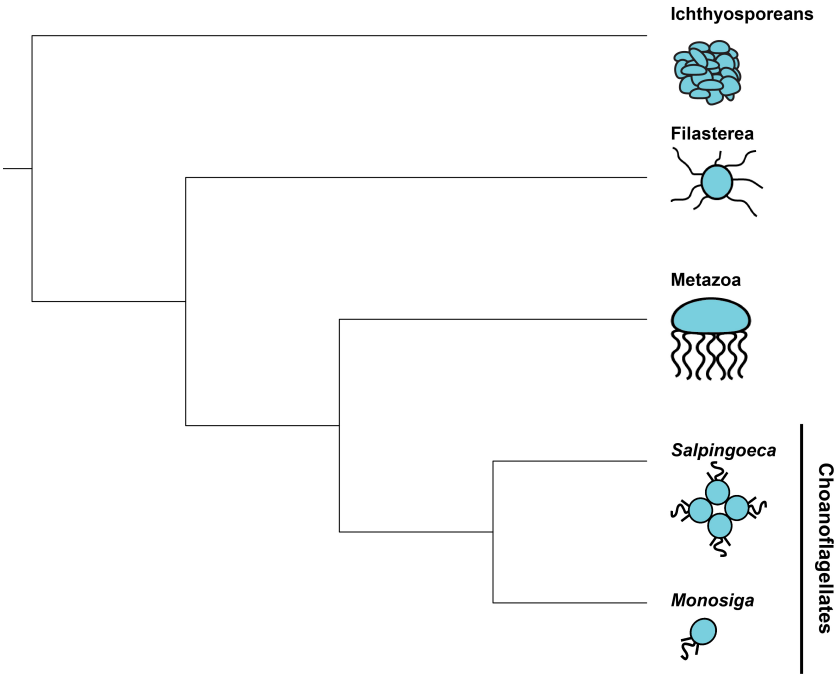


Figure 1.12: **Simplified phylogenetic representation of Holozoa.** The cladogram was produced in Newick format, and based upon the holozoan phylogeny outlined in Shalchian-Tabrizi et al. (2008).

Ancestral gene pathways have been hypothesised in Holozoa, with the genome availability of *Monosiga brevicollis* (King et al., 2008), *Salpingoeca rosetta* (Suga et al., 2013), and filasterean, *Capsaspora owczarzaki* (Suga et al., 2013). Research has predominantly focused on molecular processes, rather than analysis of population genetics, which has left a niche to be explored. Carr et al. (2017) study focused on the transcriptomes of 19 choanoflagellate species (Figure 1.13), and the impact of natural selection on two elongation factor genes; EF1-A and EFL.The research

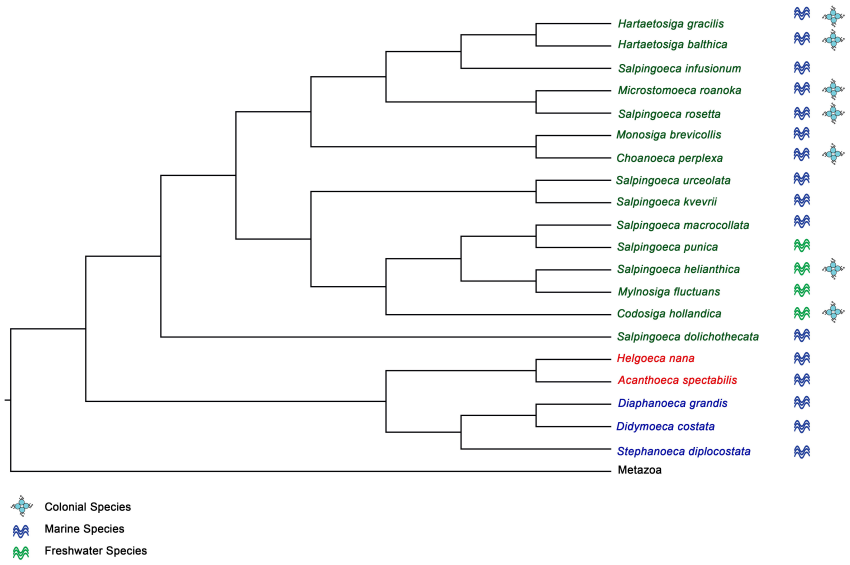


Figure 1.13: **Ecological and morphological characteristics of 19 choanoflagellate species.** Phylogenetic representation of 19 choanoflagellate species with transcriptome availability. The first column represents the species habitat, and the second column highlights the species with colonial traits. Cladogram is produced in Newick format and based on Carr et al. (2017).

focused on population genetics, including codon usage analyses to measure selective constraint among the species.

Individual codons are utilised for accuracy and efficiency, and variation is seen across all genomes, with patterns of codon preference documented (Ikemura, 1981; Ehrenberg and Kurland, 1984). The genetic code allows for degeneracy, with 18 amino acids encoded by two or more different synonymous codons, with only methionine and tryptophan encoded by one codon respectively (Hinegardner and Engelberg, 1963). This leaves 59 possible codon combinations that code for the remaining 18 degenerate amino acids (Hinegardner and Engelberg, 1963). Grantham et al. (1980) showed that codon selection bias is not random, and that genes are biased to codons ending in GC or AT.

Codon usage bias can operate through the process of three mechanisms; natural selection, mutation pressure and genetic drift (Bulmer, 1991). The three mechanisms do not always operate exclusively, and it may be a combination of pressures which contribute to codon selection in a host species.

In the study of codon usage bias, the level of bias can be measured by the effective number of codons, which is denoted *Nc*, which range from 20 to 61 (Wright, 1990). Values of 20 signifies genes which have amino acids that are coded by a single codon, and 61 is when all codon degenerates are equally utilised for each amino acid. Evidence to support that codon usage is

driven by selection, suggests that these species possess codons defined as optimal, for translational accuracy.

Optimal codons were originally proposed by Ikemura (1981) are those codons that are found to be complementary to the major tRNA genes within a host species. This concept was redefined by Lloyd and Sharp (1992), who stated that optimal codons are as those that are utilised more frequently in the first 5% highly expressed genes, in comparison to the 5% lowest expressed genes. The abundance of optimal codons within a gene is described as the Frequency of Optimal Codons (Fop), and can be calculated when the total number of codons in a gene is divided by the number of optimal codons (Ikemura, 1981). The selectively optimal codons commonly complement the highly expressed tRNAs (Kanaya et al., 2001). Optimal codons have been shown to vary across organisms. Leucine, with six-fold degeneracy, is coded by TTG in *S. cerevisiae* and in contrast, by CTG in *Drosophila melanogaster* (Sharp et al., 1988).

Carr et al. (2017) was the first study to comment on codon usage in choanoflagellates, which found that the highly expressed elongation factor genes, showed to have low levels of *Nc*, and thus strong codon usage bias. It was also found that EF1-A were also shown to be evolving under weaker constraint, and showed less biased codon usage in species where the gene was co-expressed with EFL (Carr et al., 2017). This finding supported consistency with natural selection, rather than bias driven by mutation pressure (Carr et al., 2017). The correlation observed prompted further investigation into codon usage in holozoan species, to determine the main influence on shaping codon usage. With this in mind, Southworth et al. (2018) aimed to determine the main driving forces of codon usage in *S. rosetta*, *M. brevicollis* and *C. owczarzaki*.

1.4 Project aims

Although the majority of eukaryotic species are found to be unicellular, evolutionary genomic research has predominantly focused on multicellular host species. With this, it is presumed that the overview of evolutionary pathways in eukaryotes is truncated based on multicellular bias, and limited information of unicellular species. The limited data for unicellular eukaryotes is the basis of this project, to develop a more robust picture of evolutionary trends across diverse eukaryotic species.

The bias in data includes transposable element evolution in multicellular hosts and outlines the necessity of research centred on unicellular species. The investigation of TEs in unicellular eukaryotes, will aid the determination of several TE families' origin. The project presented here will aim to review TEs in novel species of yeast, and holozoan species, as well as comparative genomics between species as well as superfamilies.The project have been divided into the following outcomes:

- Genomic survey of novel *Kazachstania* species, including codon usage bias of host genes
- TE review in four novel *Kazachstania* species
- TE review in choanoflagellate, *Salpingoeca rosetta*, with comparative review to *M. brevicollis*
- Codon usage review of three holozoan species; *S. rosetta*, *M. brevicollis* and *C. owczarzaki*
- Codon usage review of TEs found in *S. rosetta*, *M. brevicollis* and *C. owczarzaki*

At present, only one choanoflagellate species has been analysed (Carr, Nelson, Leadbeater and Baldauf, 2008), and as the metazoan sister group, the review of mobile element content of *S. rosetta* would provide new insight into the evolutionary origin of animal TEs, as well as an avenue for comparative genomics with *M. brevicollis*. As stated, comprehensive work has been employed on TEs in yeast species to date, creating a strong basis to build upon. The wide genome availability allows for exhaustive analysis, which is not obtainable in other species. Furthermore, an increase in whole genome sequences would further enable the study of TE origin, function and influence across genera of the unicellular eukaryotes. The extensively studied yeast *S. cerevisiae* has

caused interest in TE inheritance across closely related species, leaving *Kazachstania* the ideal genus for continued research.

Fungi are an ideal candidate for genomics due to their small genome size (Gladieux et al., 2014). The genetic variability seen across the kingdom allows for comparative study to take place, and to what variables influence the measured differences seen in species synteny, genomic size and TE content.

With this, TEs would seem to be a crucial aspect of genome analysis, with the annotation and characterisation of the repetitive DNA allowing for a greater overview of genomic content. Despite known abundance, TE annotation is still poorly practiced in whole genome analyses, with no elements annotated in publicly available WGS for the choanoflagellate or *Kazachstania* species (Sayers et al., 2009). Here, a homology based method is employed for TE detection, RepeatMasker (*RepeatMasker*, 1996; Huda and Jordan, 2009). The program searches for areas of high conservation and similarity between DNA sequences, against a database of consensus TE sequences from an employed library (Tempel, 2012). Two types of TE libraries were ran to allow for comparison in detection and annotation; Repbase (*GIRI*, 2016) and two custom libraries. A Saccharomycetaceae custom library of TEs annotated in species within the superfamily, and a choanoflagellate library of TEs found in available species to date.

The codon usage review will allow for comparative genomics within the *Kazachstania* genus, as well as patterns of bias across the yeast superfamily, Saccharomycetaceae. Furthermore, the research of codon usage bias for TEs is limited, with very few evidence found to show that elements codon usage is driven by selection bias, rather than mutation pressure (Lerat et al., 2002; Jia and Xue, 2009; Jiang et al., 2006). With this, a review of codon usage in mobile elements that are uncovered will reveal signatures of bias in previously unsequenced species. In addition to this, the work published by Southworth et al. (2018) detailed that codon usage bias was found to be conserved across three holozoan species, *S. rosetta*, *M. brevicollis* and *C. owczarzaki*. The work presented here will replicate the research outlined, as well as reciprocal analyses on smaller bias categories. The codon usage of TEs in the three holozoan species will also be explored, to determine if selection bias is the main driver of codon usage in the mobile elements, as seen in the host genes (Southworth et al., 2018).

Chapter 2

A genomic survey of novel species of the genus *Kazachstania*

2.1 *Kazachstania*; a relative to *Saccharomyces*

2.1.1 Introduction

Species within the genus *Kazachstania* were formerly attributed to *Saccharomyces*, and only in 2003 reclassified into their own genus on the basis of homology (Kurtzman, 2003; Kurtzman and Robnett, 2003). Initially described as a single species, *Kazachstania viticola*, the original isolation originated in fermenting grapes in Kazakhstan (Zubkova, 1971). It was later found through the analysis of multigene phylogenies that several yeast species, originally categorised to other genera in the superfamily, were reclassified to *Kazachstania* (Kurtzman, 2003, 2011). The genus contains 32 accepted species, which are described as the sister group to *Saccharomyces* and *Naumovozyma* within the Saccharomycetaceae superfamily (Vaughan-Martini et al., 2011) (Figure 2.1). *Kazachstania* has little published literature in relation to its TE content (Neuvéglise et al., 2002).

Kazachstania species

The genus *Kazachstania* is also from Saccharomycetaceae (Kurtzman, 2003), and only has four species sequenced at whole genome level; *Kazachstania africana* (CBS 2517), *Kazachstania naganishii* (CBS 8797), *Kazachstania saulgeensis* (CLIB 1764) and two strains of *Kazachstania servazzii* (NJIJ01 /PTQT01) (Sayers et al., 2009). Although morphologically identical, *Kazachstania* species variability can be seen at the DNA level (Figure 2.1). As a close relative to *S. cerevisiae* (Kurtzman, 2003), the limited data availability for this genus leaves a niche left to be explored. The

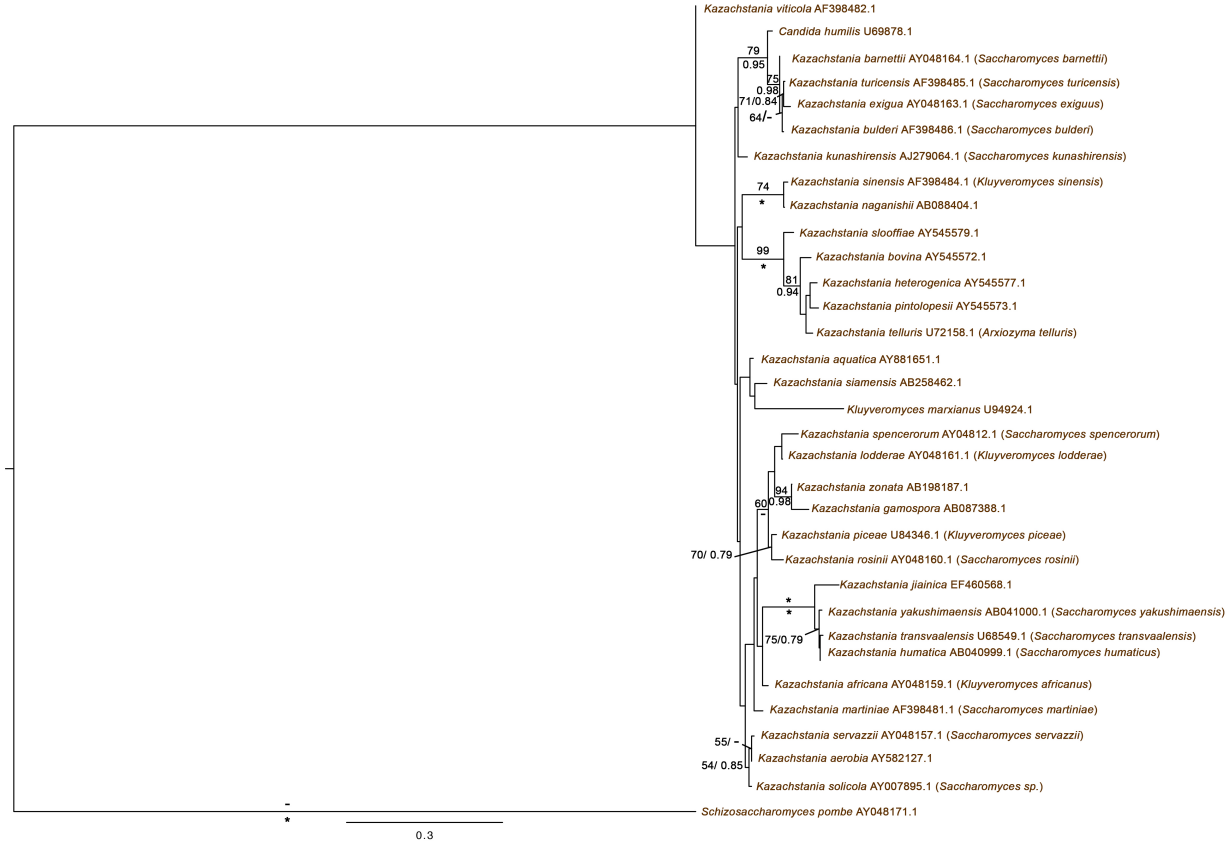


Figure 2.1: **Maximum likelihood phylogeny of species from the genus *Kazachstania* using partial or full 26S rRNA sequences** Species sent for whole genome sequencing are included in the phylogeny (*Kazachstania bovina* (AY545572); *Kazachstania exigua* (AY048163); *Kazachastania lodderae* (AY048161); and *Kazachstania viticola* (AF398482)). National Center for Biotechnology Information (NCBI) Accession Numbers are labelled with the corresponding species name (Sayers et al., 2009). Species names listed on NCBI are annotated in brackets, next to species name within *Kazachstania* post reclassification (Kurtzman, 2003). Species names annotated are that represented on NCBI; all species are from *Kazachstania* genera. The phylogeny was produced using raxmlGUI 1.5 beta via python and constructed by an alignment of 578 nucleotide constructs with the employment of raxmlGUI using the GTRCAT model (Silvestro and Michalak, 2011). ML and biPP values are labelled above and below corresponding branches. Maximum support is annotated by '*' (100ML/1.0biPP) and low support (<50 ML/ <0.70 biPP) are annotated by '-'. The scale bar signifies the number of nucleotide substitutions per site. The tree is based on maximum parsimony analysis (Vaughan-Martini et al., 2011).

yeast review in this project primarily explored the TE content of the four species publicly available and four novel species from *Kazachstania*, that have been selected for greatest phylogenetic diversity based on rRNA sequences *(K. bovina, K. exigua, K. lodderae, K. viticola*).

Table 2.1: **Genome statistics of four publicly available *Kazachstania* species.** For each *Kazachstania* species, the species strain, genome size, average G+C content and total number of coding genes are annotated. Chromosome or CDS data was not available for either strain of *K. servazzii*.

Species	Strain	No. of chromosomes	Genome size (Mb)	Average G + C content (%)	Total no. of CDS
K. naganishii	CBS8797	13	10.85	45.9	5321
K. africana	CBS2517	12	11.13	36.3	5378
K. saulgeensis	CLIB 1764	16	12.94	32.2	5869
K. servazzii	SRCM102023	-	12.84	34.90	-
K. servazzii	CBA6004	-	12.50	34.40	-

Kazachstania africana *Kazachstania africana* was classified as *Kluyveromyces africanus* initially (Kurtzman, 2003), and its genome sequence is publicly available on NCBI (Kurtzman, 2003; Sayers et al., 2009). *K. africana* (CBS 2517) has a genome size of 11.1 Mb and 12 chromosomes (Wolfe et al., 2015).

Kazachstania naganishii *Kazachstania naganishii* was formerly classified as *Saccharomyces naganishii* (Mikata et al., 2001) and the whole genome sequence (WGS) is publicly available (Sayers et al., 2009). *K. naganishii* (CBS 8797) boasts 13 chromosomes with a genome size of 10.8 Mb (Wolfe et al., 2015).

Kazachstania saulgeensis The species *Kazachstania saulgeensis* was isolated from sour dough (Sarilar et al., 2017) and whole genome is publicly available on NCBI (Sayers et al., 2009). Of the four *Kazachstania* species sequenced, *K. saulgeensis* is the largest, with a genome size of 12.94Mb, and approximately 16 chromosomes (Sarilar et al., 2017).

Kazachstania servazzii *Kazachstania servazzii* was formerly classified as *Saccharomyces servazzii* (Kurtzman, 2003), with two strains publicly available on NCBI (Sayers et al., 2009).

Chromosome number is unknown, but estimated at <12, with genome size ranging from 12.50-12.84 Mb (Sayers et al., 2009). Before genome availability, the species was investigated for TE content using random sequence tags (RST), and was found to possess both *gypsy* and *copia* elements (Neuvéglise et al., 2002).

Kazachstania exigua *Kazachstania exigua* has been partially sequenced, and TE content has been investigated (Neuvéglise et al., 2002). Formerly classified as *Saccharomyces exiguus* (Kurtzman, 2003), *K. exigua* has been documented to possess *copia* elements previously named *Transposon Saccharomyces exiguus -1 (Tse1)*, *Transposon Saccharomyces exiguus-5 (Tse5)*, and *Transposon Saccharomyces exiguus-3 (Tse3) gypsy* elements, similar to *Saccharomyces Ty* sequences (Neuvéglise et al., 2002).

2.1.2 Experiment overview

The investigation of the novel *Kazachstania* yeast species, allowed for trends across the genus to be drawn between genomic characteristics, as well as within the superfamily. Very little is known regarding the genus *Kazachstania*, and the review was constructed to allow for comparative analysis for the novel species. Similarly, the review of TEs in species, will aid the determination of several TE families' origin. Codon usage analyses were performed on the host genes and ORFs of all LTR retrotransposons identified in the *Kazachstania* species, in order to determine if selection was operating at either the level of translational accuracy, efficiency or if other evolutionary pressures were affecting bias. The work outlined would also be comparable to patterns of codon usage bias of other host genes and TEs in unicellular organisms. (Lerat et al., 2002, 2003; Jia and Xue, 2009).

2.2 Methods

2.2.1 Transposable element content and sequence similarity

The whole genomes of *Kazachstania naganishii* (CBS 8797), *Kazachstania africana* (CBS 2517), *Kazachstania saulgeensis* (FXLY01000001-17) and *Kazachstania servazzii* (SRCM102023 and CBA6004) were downloaded from NCBI (Sayers et al., 2009) in FASTA format and ran through RepeatMasker version open -4.0.6 (Smit, 1996) to review TE content. Parameters for RepeatMasker are listed in Appendix E. A custom TE library was created and employed to annotate TEs (both internal and LTR DNA) from the Saccharomycetaceae superfamily (Cooper Grace, Aug 2016 per comms) within the *Kazachstania* genomes. Additional Saccharomycetaceae species were ran via RepeatMasker to review TE content (genome accession numbers are listed in Appendix E). *Kazachstania* TEs identified from initial RepeatMasker analysis were added to the library, for a more accurate representation of TE content. LTRs that were found in close proximity on a chromosome were located in the whole genome reads, and putatively internal DNA translated using ExPASy on the SIB server (Artimo et al., 2012), and ran through the Basic Local Alignment Search Tool (BLAST) protein database (BLASTp) and NCBI conserved domain tool to search for conserved enzymatic regions (Altschul et al., 1990). The same protocol was repeated for the four novel *Kazachstania* species upon genome availability.

Elements uncovered from the *Kazachstania* species were annotated and were added to the *Kazachstania* custom library dataset for future *Kazachstania* species annotation. The same *Kazachstania* TE specific library was ran against all *Kazachstania* species to validate annotation and to detail copy number for each TE family. The newly identified *gypsy* elements were added to previously constructed *Ty3*-like fungal dataset and phylogeny was created with the employment of raxmlGUI 1.5 beta (Silvestro and Michalak, 2011) and Mr Bayes 3.2.6 XSEDE (Ronquist et al., 2012) via the server based platform, Cipres Science Gateway (Miller et al., 2010). Default parameters for Maximum Likelihood and Bayesian Inference trees are listed in Appendix E. Alignments were created using MAFFT on the EMBL-EBI server with default settings (Katoh, 2002). An additional custom library was created of *Kazachstania* TEs only and employed through RepeatMasker against all available *Kazachstania* genomes, to accurately determine element copy number. This included the internal DNA and LTR sequences (where applicable) for elements found in the eight *Kazachstania* species reviewed.

2.2.2 Chromoviral work

Similarity searches and domain prediction

Sequences from *Ty3*-like fungal dataset were reviewed for the presence of a chromodomain. The conserved amino acid sequences downstream from Integrase domain found in *Saccharomyces* species were employed as a query sequence and ran using BLASTp and tBLASTn on NCBI Blast (Sayers et al., 2009) as a reciprocal blast. The query sequence to search for additional species containing the predicted chromodomain was from *S. cerevisiae* Pol3 (AAA98435.1) with a length of 60bp. Hits were downloaded in FASTA format and aligned using MAFFT as default (Katoh, 2002). Text files in FASTA format were uploaded onto CLUSTALX 2.1 to view amino acid conservation (Thompson et al., 2002).

Protein modelling

A crystallised structure of a known chromodomain from *S. cerevisiae* was downloaded from the crystallographic database; Protein Data Bank (Berman et al., 2000). The crystal structure of the chromodomain-ATPase portion of the yeast Chd1 chromatin remodeler (3MWY) was uploaded to the database by Hauk et al. (2010). 3MWY protein was uploaded to SWISS-MODEL (Biasini et al., 2014) as a template model in pdb file format, and the potential CD domain as the target sequence in FASTA format. Default settings were employed to build the model. The mapped model result file was downloaded in pdb file format. The mapped model was viewed using Py-MOL version 1.74 (Schrodinger, 2017). The predicted chromodomain was mapped onto 3MWY template protein. Server based PSIPRED version 3.3 (Jones, 1999; Buchan et al., 2013) was employed to predict the secondary structure of the predicted chromodomain from *S. cerevisiae*.

2.2.3 Yeast husbandry

Yeast species were purchased from the National Collection of Yeast Cultures Catalogue (NCYC) as live cultures (NCYC, 2016). The cultures were stored at 4˚c with no light exposure. All yeast cultures were grown in YPD liquid medium (Amberg et al., 2005), with glucose carbon source, as applicable to each species (NCYC, 2016). Yeast Extract-Peptone-Dextrose (YPD) liquid media using 50g of YPD broth per litre of deionised H_2O. The media was then autoclaved for 15 minutes at 121˚c (Amberg et al., 2005). YPD Broth constituents are listed in Appendix A. 1ml/L of Ampicillin

was added to the media. Yeast stocks of each *Kazachstania* species were made and stored at -80˚c in cryotubes, in a glycerol media (*Fisher*, 2017).

Yeast cultures were also grown on YPD agar plates. YPD agar was made using 65g agar per litre of deionised water, and then autoclaved for 15 minutes at 121˚c (Amberg et al., 2005). 1ml/L of Ampicillin was added to the agar when the media was cooled in a 50˚c water bath. The plates were then poured aseptically and left to set at room temperature. Yeast cultures were streaked on YPD agar plates and incubated at 25˚c for 48 hours. Stocks were kept at 4˚c and inoculated every 4 weeks.

The live yeast cultures were suspended to YPD liquid medium aseptically, using a sterile loop. *Kazachstania* species were incubated at 25˚c for a minimum of three days (Vaughan-Martini et al., 2011). The four *Kazachstania* species were stained using methylene blue to test sample purity, and to assess morphology within the genus. Samples were aseptically mounted on a glass side, and dried. The samples were needed flooded with methylene blue for 1 minute. The dye was washed off the slide using sterile water, and blotted dry, to allow for observation under a microscope.

2.2.4 RNA extraction

Two methods of RNA extraction were employed; Trizol extraction (Qiagen, 2017) and RNAswift (Nwokeoji et al., 2016). The results were comparatively assessed for quantity, absorbance ratios and yield. Recipes for constituents of the RNA extraction for both methods are listed in Appendix A. The protocol for Trizol extraction and RNASwift was performed by the manufacturer's instructions, and are listed in Appendix A.

DNase reaction

Post RNA extraction, a DNase reaction was ran on samples to remove any gDNA. For every 500ng of RNA, 0.5ul of DNase buffer and 0.5ul DNase was added to the sample. The sample was then incubated at 37˚c using a heat block for 60 minutes. Following incubation, 0.5ul of EDTA was added, and the sample was incubated for a further 10 minutes at 65˚c.

2.2.5 DNA extraction

Two methods of DNA extraction were employed; DNeasy blood and tissue extraction kit (Qiagen, 2017) and LiOAC-SDS method (Looke et al., 2011). The results were comparatively assessed

for quantity, absorbance ratios and yield. DNeasy blood and tissue extraction kit was used and protocol employed from the Qiagen DNeasy manual. The protocol was specifically designed for the purification of DNA from ≤ 5 x 107 yeast cells (Qiagen, 2017) and is listed in Appendix A. The DNA extraction protocol using lithium acetate (LiOAc), 1% SDS solution (LiOAc-SDS) was modified from the procedure detailed in Looke et al. (2011). The employed protocol is listed in Appendix A.

2.2.6 Whole genome sequencing of four novel *Kazachstania* species

Due to time limitations, the four species which were cultivated in the laboratory were sent to Macrogen, where DNA was extracted for whole genome sequencing.

DNA extraction and sequencing

DNA was extracted using AMPure purification, performed according to manufacturers instructions. Following extraction, the samples were ran through quality control (QC) and library preparation (RSII SMRT Library [20kb]) to allow for PacBio RSII SMRT Cell (700Mb output) whole genome sequencing to be performed (Figure 2.2).

Figure 2.2: **Method for sequencing and analysis workflow for WGS and genome assembly and annotation, based on procedure by Macrogen**.

Preprocessing

The method of QC used involved a library size check, as well as library quantity check, to ensure the highest quality of data on the PacBio sequencing platforms. The size of the DNA fragments incorporated from each species were ran on a Agilent Technologies 2100 Bioanalyzer with a DNA 12000 chip. DNA bioanalyser results can be found in Appendix B. The requested SMRTbell Library with 20kb SMRTbell templates, for the PacBio sequencing platform, required a concentration of >15ng/ul, with a size of >10000bp. For the generation of a standard curve of fluorescence readings, and for library sample concentration calculations, Macrogen employed Roche's Rapid library standard quantification solution and calculator (Roche, 2018).

For the preprocessing, a sequence of nucleotides from each species sample was incorporated with use of a circular SMRT bell template, and DNA polymerase. Following sequencing, Macrogen performed genome assembly, as well as genome prediction and annotation.

Table 2.2: **Macrogen library preparation results for the four novel species of *Kazachstania*.**

#	Library Name	Library Type	Conc. (ng/ul)	Size (bp)	Result
1	*K. bovina*	20kb SMRTbell Templates	41.6	20000	Pass
2	*K. exigua*	20kb SMRTbell Templates	33	20000	Pass
3	*K. lodderae*	20kb SMRTbell Templates	34.7	20000	Pass
4	*K. viticola*	20kb SMRTbell Templates	40.3	20000	Pass

Analysis

Macrogen performed a pre-assembly step, which mapped single reads to seed reads, which showed the longest section of the distribution of read lengths. This led to the generation of a consensus sequence for all the mapped reads, producing large, highly resolved reads for the species genome. This process was repeated for each *Kazachstania* genome. The *de novo* assembly was performed using FALCON (Chin et al., 2016), and the read filtering using Quiver (Chin et al., 2013). Details for each contig per genome are listed in the Appendix A. The reads were then filtered to ensure duplicate and too low/ high overlapping reads were removed prior to genome construction. Based on the data which overlapped, contigs were constructed for each genomes, which varied in quantity per species. Post genome assembly, protein coding sequences were located and genes identified. Maker v2.31.8 (Cantarel et al., 2007) was used to predict the

location of the genes, and Protein BLAST+ v2.6.0 was performed to identify the genes. The best hits of the NCBI BLAST against NCBI database were annotated (Altschul et al., 1990).

2.2.7 Codon usage and tRNA genes

Codon usage bias was investigated in all *Kazachstania* species except *K. saulgeensis* and *K. servazzii* due to no transcripts available for these two species. Transcript files for *K. africana* and *K. naganishii* were downloaded from NCBI (Sayers et al., 2009). Codon usage statistics of the complete annotated transcriptome sequences for six of the yeast species were analysed using CodonW (Peden, 1999). Optimal codons for each species were generated by correspondence analysis (COA), using relative synonymous codon usage (RSCU), with employment of default parameters. The optimal codons identified were then assessed comparatively to the anticodons of the tRNA genes for each yeast species to determine if the optimal codons were found to be complementary to the tRNA anticodons. For each TE family in the *Kazachstania* genomes, values of *Nc*, Fop and GC3s were calculated for all coding sequences using CodonW 1.4.4 (Peden, 1999). The fop.coa file generated for each species was employed to the TEs on a species specific basis. For elements where the *gag* and *pol* ORFs are separated (*Tkn3, Tse1, Tkl3* and *Tkv3*), the sequences were concatenated to provide comparable values with the other LTR retrotransposon families, where Gag and Pol were transcribed in the same ORF.

2.2.8 Major tRNA gene screening

Each annotated genome for the six *Kazachstania* species were ran for major tRNA genes. The two publicly available species (*K.africana*) and *K. naganishii*) were downloaded from NCBI. The program tRNAscan-SE 2.0 Lowe and Chan (1997) was employed to identify major tRNA genes using default settings, by postgraduate student, Holly Dawson.

2.2.9 *K. exigua* gene annotation

K. exigua cds files were ran using eggNOG version 4.5 via a server to assess gene function through orthology assignment (Huerta-Cepas et al., 2016, 2017).

2.2.10 Species synteny

To assess synteny across the *Kazachstania* genomes, syntenic blocks were produced using SyMAP v4.2 (Soderlund et al., 2011, 2006). For *K. africana* and *K. naganishii*, the genome annotation gff file were downloaded from NCBI for genome annotation. For each species in SyMAP, gff genomic data was uploaded as the annotation file, and WGS uploaded as sequences reference in FASTA format. Synteny was then reviewed between *Kazachstania* species through gene alignment, where the collinearity of gene order was analysed at whole chromosome level.

2.3 Results

2.3.1 Characteristics of *Kazachstania* genomes

The publicly available *Kazachstania* species (*K. africana* CBS 2517, *K. naganishii* CBS8797, *K. saulgeensis* CLIB1764 and *K. servazzii* SRCM102023/CBA6004) were downloaded from NCBI (Sayers et al., 2009), and characteristics compared to that of other budding yeast species from the superfamily, Saccharomycetaceae (Genolevures et al., 2009) (Table 2.3). TE content varied across the species, ranging from 0.06% in *Eremothecium gossypii* (ATCC10895) and 3.4% in *S. cerevisiae* (S288c) draft genome (Table 2.3). The re-annotation of *Kazachstania* genomes with the custom library increased the whole genome TE content for both species, from the original results when employed with species parameters specified, rather than a custom library (*GIRI*, 2016). *K. africana* increased from 0.15% to 0.18%, and *K. naganishii* increased from 0.21% to 0.63%. *K. saulgeensis* has a TE content of 0.29% when RepeatMasker was employed with the default Repbase library (*GIRI*, 2016), with an increase to 0.38% with the use of the custom Saccharomycetaceae family. The two strains of *K. servazzii* varied slightly regarding TE content, with an increase from 0.35% in SRCM102023 strain, to 0.50% in CBA6004.

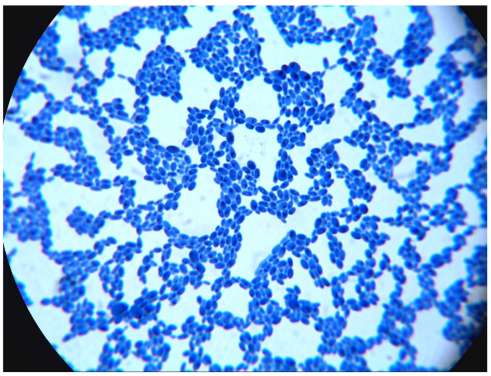
Relationships were reviewed between TE content and other species characteristics, to identify any trends between genomic data and host mobile elements. No correlation was seen between genome size, and TE content (R^2 = 0.099) (Appendix B). Furthermore, no correlation was present between TE content and Average GC content of the yeast species (R^2 = 0.001) (data not shown) and between TE content and number of coding genes (R^2 = 0.262)(Appendix B). Previously, a negative correlation has been observed between TE content and G+C content in other eukaryotic species, including metazoans and plant species (Shen et al., 2013; Dhillon and Goodwin, 2014).

Table 2.3: **Summary of annotated characteristics in eleven yeast species from Saccharomycetaceae** *Kazachstania* species were added to the tabulated data set from (Genolevures et al., 2009), with TE content added from RepeatMasker results (*RepeatMasker*, 1996). RNA data was removed as transcriptome data is unavailable for the *Kazachstania* species on NCBI (Sayers et al., 2009). *Kazachstania* species are written in bold font.

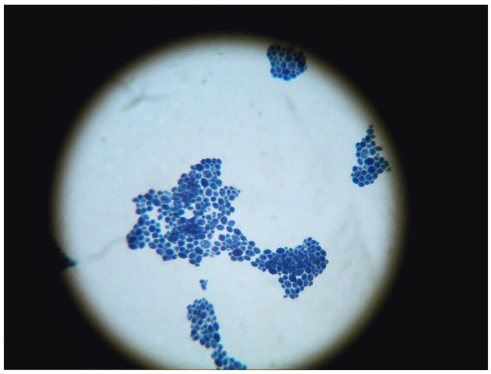
Species	Strain	No. of chromosomes	Genome size (Mb)	Average G + C content (%)	Total no. of coding genes	TE Content (%)
K. naganishii	CBS8797	13	10.85	45.9	5321	0.63
K. africana	CBS2517	12	11.13	36.3	5378	0.18
K.saulgeensis	CLIB 1764	16	12.94	32.2	5869	0.38
K. servazzii	SRCM102023	-	12.84	34.9	-	0.35
K. servazzii	CBA6004	-	12.5	34.4	-	0.5
S. cerevisiae	S288c	16	12.1	38.3	5769	3.4
Candida glabarata	CDS138	13	12.3	38.8	5204	0.17
Zygosaccharomyces rouxii	CBS732	7	9.8	39.1	4992	0.07
Kluyveromyces thermotolerans	CBS6340	8	10.4	47.3	5094	0.29
Lachancea kluyveri	CBS3082	8	11.3	41.5	5320	1.05
Kluyveromyces lactis	CBS2359	6	10.7	38.8	5076	0.26
Eremothecium gossypii	ATCC10895	7	8.7	52	4715	0.06
Debaryomyces hansenii	CBS767	7	12.2	36.3	6395	1.13
Yarrowia lipolytica	CBS7504	6	20.5	49	6580	1.32

Morphological characteristics of the four novel species of *Kazachstania*

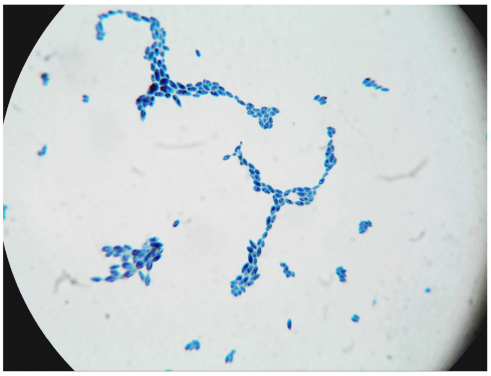
The four species of yeast which were cultured and sent for whole genome sequencing were viewed to observe any morphological differences that may exist between the species and to ensure sample purity prior to sending for sequencing. At 100x magnification, the four species presented similarly, as each budding yeast cell had a spherical to ellipsoidal shape. With this, the importance of RNA extraction for species identification was supported, as morphological differences were limited within the genus (Figure 2.3).



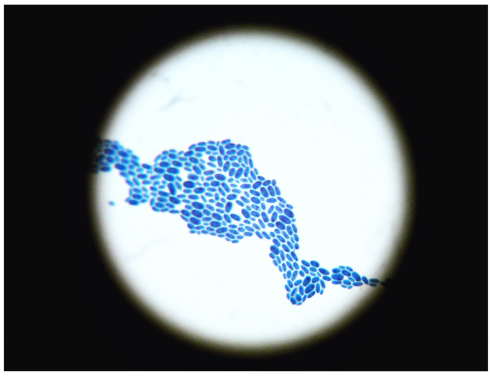
K. exigua 100x



K. lodderae 100x



K. bovina 100x



K. viticola 100x

Figure 2.3: **Four novel *Kazachstania* species stained with methylene blue at 100x magnification**. The four yeast species were cultured from a live state at 25˚C in YPD broth for 72 hours.

Genome characteristics of four novel species of *Kazachstania*

Four novel *Kazachstania* species, *K. bovina*, *K. exigua*, *K. lodderae* and *K. viticola*, were sequenced, and assembled allowing for genome characteristics to be identified, and comparative analysis to be reviewed across the genus. With the exception of *K. exigua*, genome size for the three species ranged from 11.4Mb - 12.4Mb. This was expected, as previously annotated *Kazachstania* species genome size varied between 10.85Mb and 12.84Mb. In contrast, *K. exigua* was found to have a much larger genome of 24.8Mb (Table 2.4). The genomic size exceeds all other yeast species within Saccharomycetaceae, and the total number of coding genes were almost double the other species investigated (Table 2.3).

Gene annotation of the host genes of *K. exigua*

The whole genome sequencing of four novel *Kazachstania* species, revealed an unexpected finding that *K. exigua* was found to have a genome size which was twice as big as other *Kazachstania* species, and possessed over 9000 host genes (Table 2.4). To investigate this, the genes of *K. exigua* were ran using eggNOG version 4.5 via a server to assess gene function through orthology assignment (Huerta-Cepas et al., 2016, 2017). Upon gene annotation, 611 genes were found to be undetermined. The remaining 8901 were assigned to 23 different categories (Table 2.5 and Figure 2.4).

The greatest proportion of genes (1249 genes; 14%) in *K. exigua* were assigned to category 'Unknown Function (S)' (Figure 2.4). The next highest proportion of genes were assigned to the following categories; Intracellular Trafficking, Secretion and Vesicular Transport (U), Transcription (K), Translation (J) and Posttranslational Modification, Secretion and Vesicular Transport (O). With the exception of (S), the majority of genes were assigned to the three main categories of annotation relatively evenly; Cellular Processes and Signalling (25.65%), Information Storage and Processing (35.33%); Metabolism (24.99%). It was found that 86% of the genes have been annotated with known function, and therefore are likely to be functional in the genome, rather than misannotation.

Table 2.4: **Summary of *Kazachstania* species genome characteristics, including four novel *Kazachstania* species.** Kazachstania species were sequenced and assembled, with key genome characteristics identified, including genome size, GC content and number of coding genes. TE content added from RepeatMasker results (*RepeatMasker*, 1996).

Species name	Strain	No.of contigs	Genome size (Mb)	Average G + C content (%)	Total no. of coding genes	TE Content (%)
K. africana	CBS2517	12	11.13	36.3	5378	0.18
K. bovina	NCYC 526	24	11.4	28.27	5920	2.37
K. exigua	NCYC 814	32	24.8	32.5	9964	2.21
K. lodderae	NCYC 1417	17	12.4	33.84	5555	0.67
K. naganishii	CBS8797	13	10.85	45.9	5321	0.63
K. saulgeensis	CLIB 1764	16	12.94	32.2	5869	0.38
K. servazzii	SRCM102023	-	12.84	34.9	-	0.35
K. servazzii	CBA6004	-	12.5	34.4	-	0.5
K. viticola	NCYC 2701	30	11.5	32.71	5524	2.91

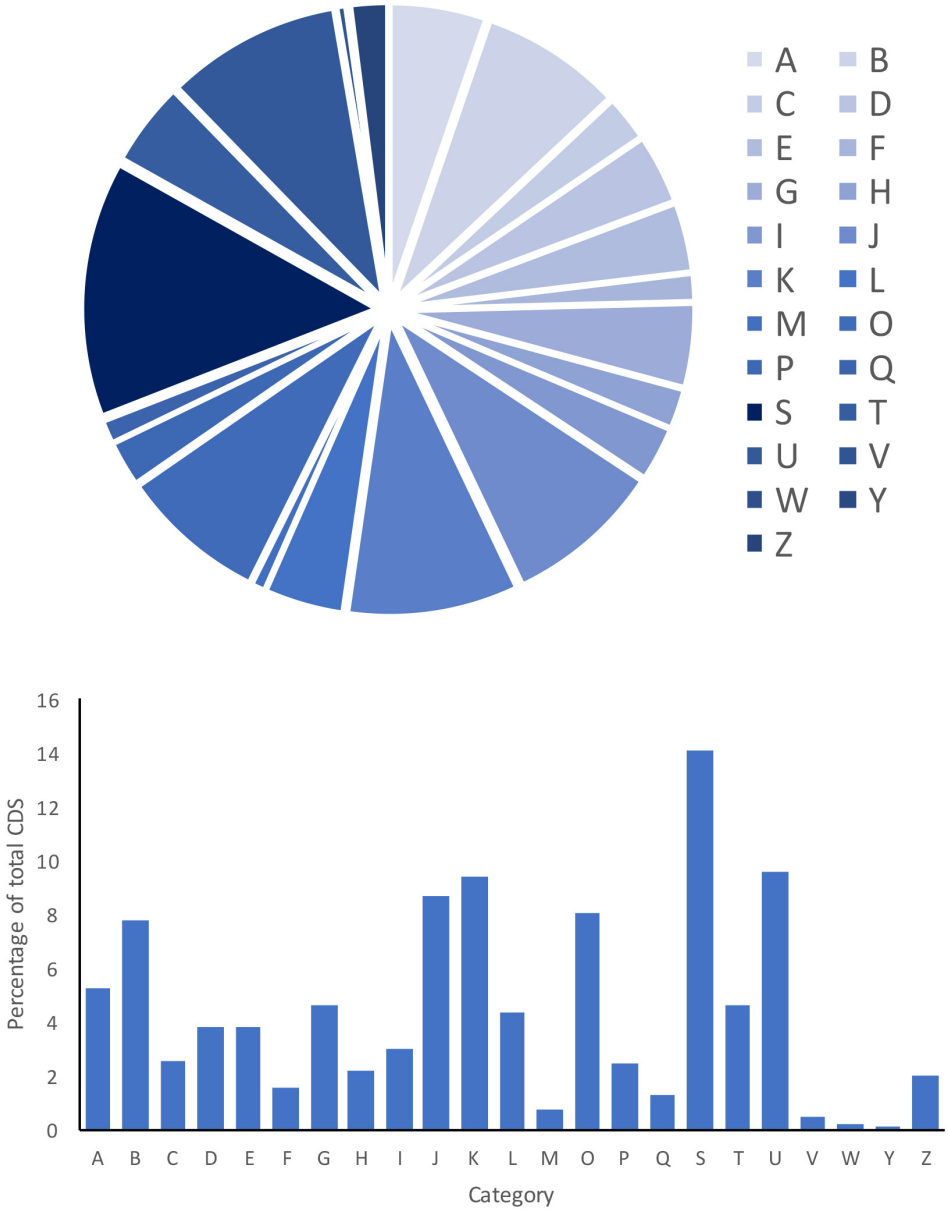


Figure 2.4: **Gene groupings and categories the host genes of *K. exigua* based on eggNOG annotation.** Each gene category was ran through eggNOG annotation to investigate gene assignment across the genome.

Table 2.5: **Gene ontology percentage distribution for the genes annotated in *K. exigua*.**

KOG Category	Percentage Distribution
Cellular Processes and Signalling	
(M) Cell wall/membrane/envelope biogenesis	0.74
(N) Cell motility	0.00
(O) Posttranslational modification, protein turnover, chaperones	8.02
(T) Signal transduction mechanisms	4.61
(U) Intracellular trafficking, secretion and vesicular transport	9.58
(V) Defence mechanisms	0.47
(W) Extracellular structures	0.17
(Y) Nuclear structure	0.06
(Z) Cytoskeleton	2.00
Information Storage and Processing	
(A) RNA processing and modification	5.26
(B) Chromatic structure and dynamics	7.75
(J) Translation, ribosomal structure and biogenesis	8.64
(K) Transcription	9.37
(L) Replication, recombination and repair	4.31
Metabolism	
(C) Energy production and conversion	2.49
(D) Cell cycle control, cell division, chromosome partitioning	3.80
(E) Amino acid transport and metabolism	3.79
(F) Nucleotide transport and metabolism	1.51
(G) Carbohydrate transport and metabolism	4.58
(H) Coenzyme transport and metabolism	2.17
(I) Lipid transport and metabolism	2.94
(P) Inorganic ion transport and metabolism	2.46
(Q) Secondary metabolites biosynthesis, transport and catabolism	1.25
Poorly Characterised	
(R) General function prediction only	0.00
(S) Function Unknown	14.03

Synteny across the *Kazachstania* species

Synteny between *K. exigua* and the remaining *Kazachstania* species was assessed to visualise conservation between species. The method was employed to investigate the striking contrast seen between species regarding genomic size and number of genes. It was proposed that evidence of a potential duplication event in this species would be visualised using a syntenic program, such as SyMAP (Soderlund et al., 2006, 2011). The synteny analysis revealed a total of 227 syntenic blocks between *K. exigua* abd *K. viticola* (Figure 2.5 and 2.6). Syntenic mapping was performed for all combinations of *K. exigua* and the remaining *Kazachstania* species with transcript availability, and similar levels of syteny were seen for each analysis (Appendix B). Additional *Kazachstania* species were compared for synteny to review if patterns of conservation were similar between species of similar genome size and number of genes (Appendix B).

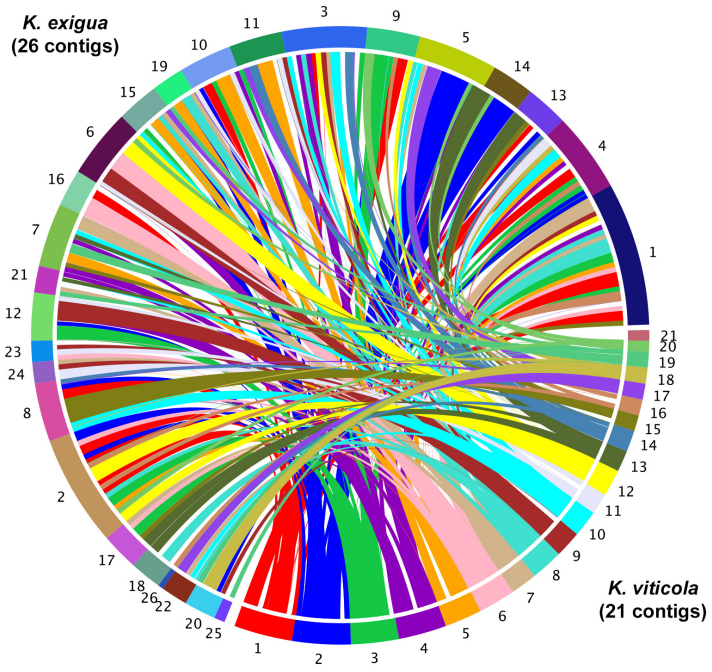


Figure 2.5: **SyMap synteny analyses between *K. exigua* and *K. viticola*.** Sytenic mapping between the contigs of *K. exigua* to the contigs of *K. viticola*. Syntenic blocks were viewed using a circular view, with scaling based on genome size.

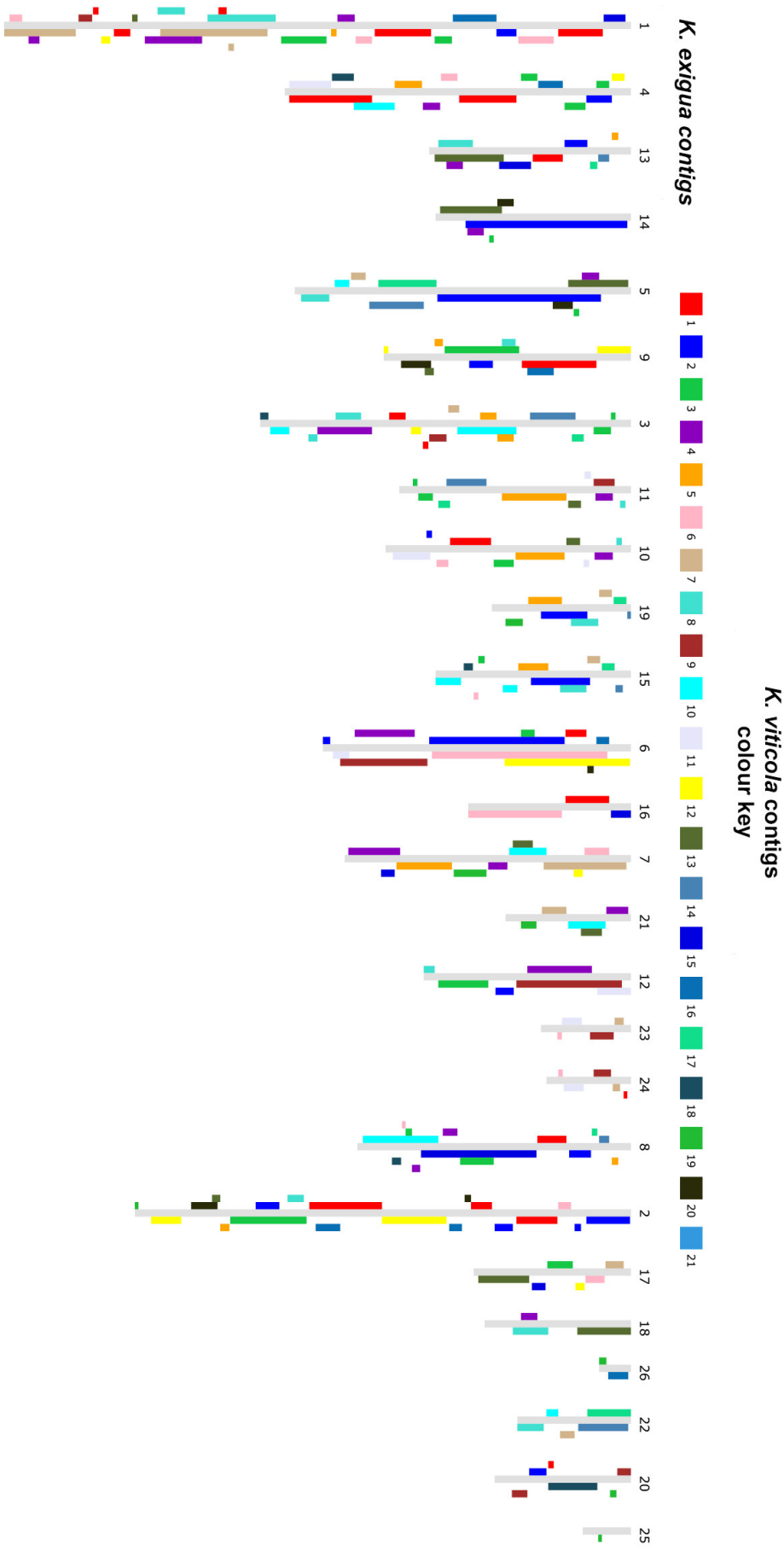


Figure 2.6: **SyMap syntenic block analyses between *K. exigua* and *K. viticola*.** Sytenic mapping of *K. viticola* contigs of *K. exigua* contigs.

Transposable element annotation of publicly available *Kazachstania* genomes

Only LTR retrotransposon TE families were detected in all *Kazachstania* genomes, which were annotated and added to the custom library dataset. *K. africana* revealed relics of *copia-like* elements similar to *Ty1, Ty3* and *Ty5* from *S. cerevisiae*. However, only one partial element was found and annotated as *Transposon Kazachstania africana-3* (*Tka3*) (Table 2.6). *Tka3* was found with no LTRs flanking the element, however all conserved enzymatic domains were intact (Figure 2.7).

K. naganishii held a more diverse selection of families possessing both *gypsy-like* and *copia-like* elements. Three TE families were uncovered in *K. naganishii*, named *Transposon Kazachstania naganishii-1* (*Tkn1*), *Transposon Kazachstania-3* (*Tkn3*) and *Transposon Kazachstania naganishii-5* (*Tkn5*) (Table 2.6). The *Tkn5* element, which had a length of 4643 bp, was found to potentially encode an additional domain, compared to the expected *Ty5*-like pol protein structure (Figure 2.7). The FLE also had flanking 100% identical LTRs at 5' and 3' end, however the 5' LTR was in reverse compliment. The putative domain inserted in *Tkn5* was annotated in BLASTp as Herpes virus major outer envelope glycoprotein (pfam05109). Reciprocal similarity searches showed that the protein had no homologues in other species, and therefore is presumed to be incorrectly annotated (Sayers et al., 2009). The inserted domain is therefore of unknown function, and may be typical of *Tkn5*, however as only one full length copy is present in the genome, this cannot be determined. The putative domain was not found during conserved domain searches of additional *Ty5*-like elements in other *Kazachstania* species. *Tkn3* was found to encode *gag* and *pol* in two separate ORFs (Figure 2.7).

Similarly to *K. naganishii*, screening of the recently published *K. saulgeensis* genome revealed two FLE from *gypsy-like* and *copia-like* families, named *Transposon Kazachstania saulgeensis-1* (*Tks1*) and *Transposon Kazachstania saulgeensis-3* (*Tks3*) (Table 2.6 and Figure 2.7). Both FLEs were similar to *Ty1* and *Ty3* from *S. cerevisiae* (Kim et al., 1998). LTRs identified for each element presented conserved dinucleotides of T..G and C..A at terminal ends of the repetitive sequences (Table 2.6).

K. servazzii was found to possess only one mobile element, in multicopy. The *copia* element, *Transposon Kazachstania servazzii-5* (*Tkn5*) was intact with LTRs flanking the 5' and 3' end of the element.

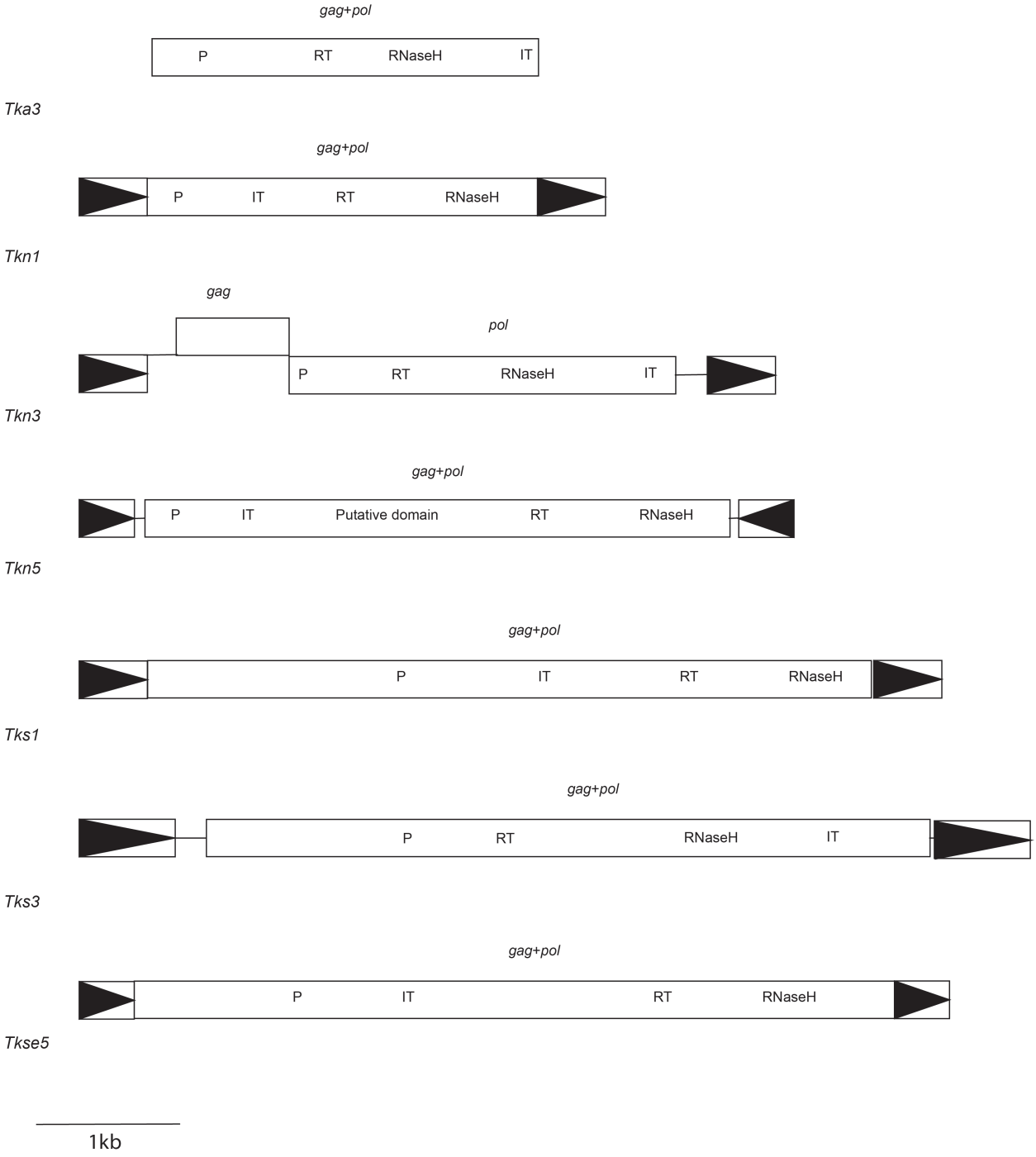


Figure 2.7: **Genomic organisation of the four LTR retrotransposon families characteristed in the four publicly available *Kazachstania* species.** Horizontal boxes represent *gag* and *pol* open-reading frames (ORFs). Protein coding domains are indicated as follows: P, protease; RT, reverse transcriptase; RNaseH, ribonuclease H; IT, integrase. Boxes with black arrowheads represent long terminal repeat sequences.

2.3.2 TE annotation of four novel species

The project involved the sequencing of four novel *Kazachstania* species, *K. bovina, K. lodderae, K. exigua* and *K.viticola*. The species had previously not been sequenced, and left a niche to be explored for both genome characteristics, as well as TE content. Similarly to the four publicly available *Kazachstania* species, the TEs annotated in the novel species were also class I elements, from either *copia* or *gypsy* families (Figure 2.8). As with most yeast species within the superfamily, Saccharomycetaceae, elements were initially identified with likeness to the *Ty*-elements of *S. cerevisiae* (Kim et al., 1998; Carr et al., 2012).

All novel *Kazachstania* species, except *K. lodderae* were found to possess both *copia* and *gypsy* elements, similar to *Ty1*, *Ty4*, *Ty5* and *Ty3* (Figure 2.8). In contrast, *K. lodderae* only possessed a *gypsy*-like element. Elements were named based on the species which the element was identified, as well as similarity to the publicly known *Ty* elements (Kim et al., 1998).

Although TE data has been publicly available (Neuvéglise et al., 2002; Sayers et al., 2009) for *K. exigua* (formerly classified as *Saccharomyces exiguus*), the whole genome sequence was not available for this species. Therefore, the sequencing in this project has allowed for a review of genome characteristics, as well as further evidence to support previously published TE data, as well as the addition of TE genome content, and family copy number. All three TEs previously annotated were identified in the whole genome of *K. exigua* as FLEs with multiple copies (Table 2.6).

Similarly, *K. bovina* revealed one *copia*-like element, Transposon *Kazachstania bovina*- 5 (*Tkb5*), and one *gypsy*-like element, Transposon *Kazachstania bovina*-3 (*Tkb3*) (Figure 2.8). *K . lodderae* and *K. viticola* also possessed *gypsy*-like elements similar to *Ty3*; Transposon *Kazachstania lodderae* -3 (*Tkl3*) and Transposon *Kazachstania viticola* -3 (*Tkv3*) (Figure 2.8). *K. viticola* was the only species that possessed a copia element similar to *Ty4*, which was named Transposon *Kazachstania viticola* -4 (Figure 2.8). The element was originally categorised as *Ty1*-like, however in the *copia* phylogeny, *Tkv4* was found to form a clade with *Ty4*, separate to *Ty1* and *Ty5* clades, with maximum support (100ML/1.0biPP) (See Phylogenetic analyses of transposable element families in *Kazachstania* species).

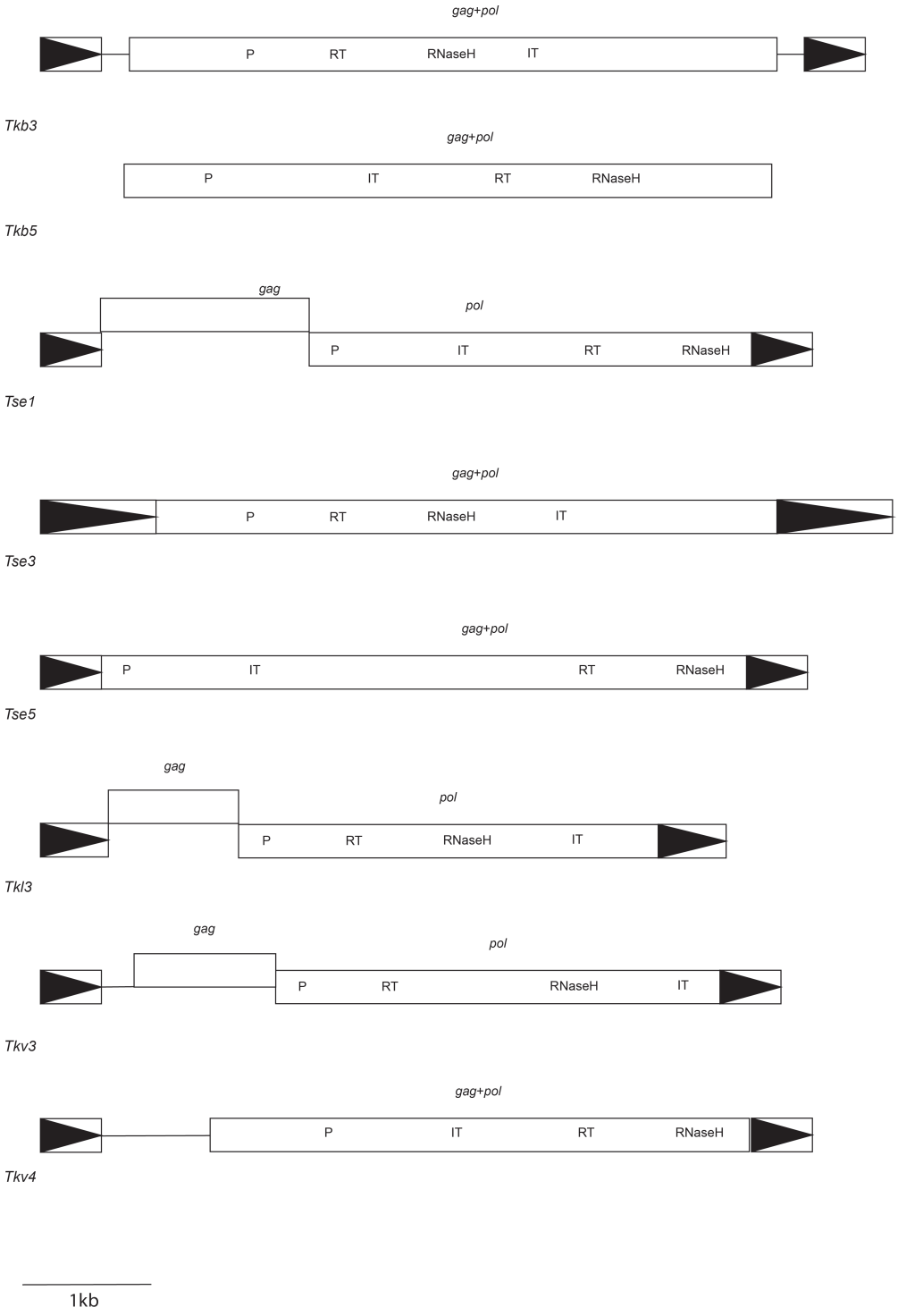


Figure 2.8: **Genomic organisation of the one *gypsy*-like family characteristed in the *K. bovina* genome.** Format is stated in Figure 2.7.

In reference to element length, the TEs in the *Kazachstania* species typically ranged between 3Kb and 5Kb. *Tka3* was a smaller size of 2.5Kb, however was not classified as a full length element, as LTRs were not identified. The element was only found as one copy, and therefore there was no evidence of proliferation in the *K. africana* genome. In contrast, both *gypsy* elements revealed in *K. bovina* and *K. exigua* were both above 6Kb in length (Table 2.6).

With the exception of *Tka3*, the elements uncovered in the *Kazachstania* species were found in multicopy. Copy number ranged from 2 copies of *Tkn1* and *Tks3*, and 30 copies of *Tkv4*. The presence of multiple copies of elements supported transposition in the majority of *Kazachstania* species.

As previously detailed, LTR retrotransposons are flanked by long terminal repeats at the 5' and 3' ends of the elements. LTR size ranged from 200 - 600bp in length, which is typically seen for elements in Saccharomycetaceae yeast species (Neuvéglise et al., 2002). In contrast, *Tse3* was found to have LTRs which were 962bp in length. All LTRs were found to have conserved terminal dinucleotides T-G and C-A flanking each element, which is frequently used to identify eukaryotic LTR retrotransposons (Freund and Meselson, 1984).

The initial *Kazachstania* species analysed for TE content were found to have a low percentage of the genome possessed by mobile elements, with values <0.70% for *K. africana*, *K. naganishii*, *K. saulgeensis* and *K. servazzii*. However, with the addition of the four novel species, TE content was found to be higher, with percentage values of over 2% for three of the four species (Table 2.4). These values are similar to that documented in *S. cerevisiae*, which was found to have a genomic TE content of approximately 3.3% (Kim et al., 1998; Carr et al., 2012).

Table 2.6: **Characteristics of novel LTR retrotransposons in *Kazachstania* species from *copia*-like and *gypsy*-like superfamilies.** Similar elements are annotated with species name, with corresponding new element name represented in bold. The size of each element are given in bp, and the current status of the element detected. LTR characteristics are also tabulated, including size (bp) and conserved 5' and 3' terminal ends. Table structure was based on (Neuvéglise et al., 2002).

					LTR	
Species	**Name**	**Size (bp)**	**Copy Number**	**Status**	**Size (bp)**	**Conserved flanking nucleotides**
Ty1 - like						
K. exigua	***Tse1***	5791	27	Full	431	TG…CA
K. naganishii	***Tkn1***	3455	2	Full	356	TGT…CAACA
K. saulgeensis	***Tks1***	4960	5	Full	432	TG..CA
Ty3- like						
K. africana	***Tka3***	2537	1	No LTR		
K. bovina	***Tkb3***	6055	23	Full	462	TGT…TTACA
K. exigua	***Tse3***	6601	28	Full	962	TGT…TTACA
K. lodderae	***Tkl3***	5216	6	Full	354	TGT…TACA
K. naganishii	***Tkn3***	4501	4	Full	358	TGT…CAACA
K. saulgeensis	***Tks3***	3824	2	Full	759	TGT…TTACA
K. viticola	***Tkv3***	5164	14	Full	342	TGT…TTACA
Ty4 - like						
K. viticola	***Tkv4***	5745	30	Full	466	TG…CA
Ty5 - like						
K. bovina	***Tkb5***	4673	14	No LTR		
K. exigua	***Tse5***	5378	5	Full	376	TGT…CAACA
K. naganishii	***Tkn5***	4643	9	Full	245	TGT…CAACA
K. servazzii	***Tkse5***	5474	5	Full	229	TGT…CAACA

2.3.3 Phylogenetic analyses of transposable elements families in *Kazachstania* species

Protein phylogenies were created for all the TE families present in the *Kazachstania* genomes. Phylogenies for *Kazachstania* elements were created to compare to species phylogeny, to explore evidence of horizontal transfer, or if elements were vertically inherited. All LTR retrotransposon phylogenies were moderately to highly resolved, with support values found to range from 89-100% for maximum likelihood (ML) and 1.0 bayesian inference posterior probabilities (biPP) (100%ML/ 1.0biPP). For the *gypsy*-like phylogeny of *Kazachstania* families, the element were split to two distinct clades. The first clade (*TY3A*) included elements from *K. viticola*, *K. africana*, *K. lodderae* and *K. bovina*. The second clade was comprised of the three remaining *gypsy*-like families found in *K. saulgeensis*, *K. exigua* and *K. naganishii*. The two clades were separated with maximum support, on long branches, which supported that divergence within the *gypsy* family was an ancient occurrence (Figure 2.9). This positioning which supported two independent, vertical inheritance events of *Ty3*-like elements has been documented within the Saccharomycetaceae superfamily

(Wolfe et al., 2015). To investigate the position of the additional *Kazachstania* species within the superfamily, the *gypsy*-like sequences were added to a *Ty3*- like Pol dataset (Figure 2.10).

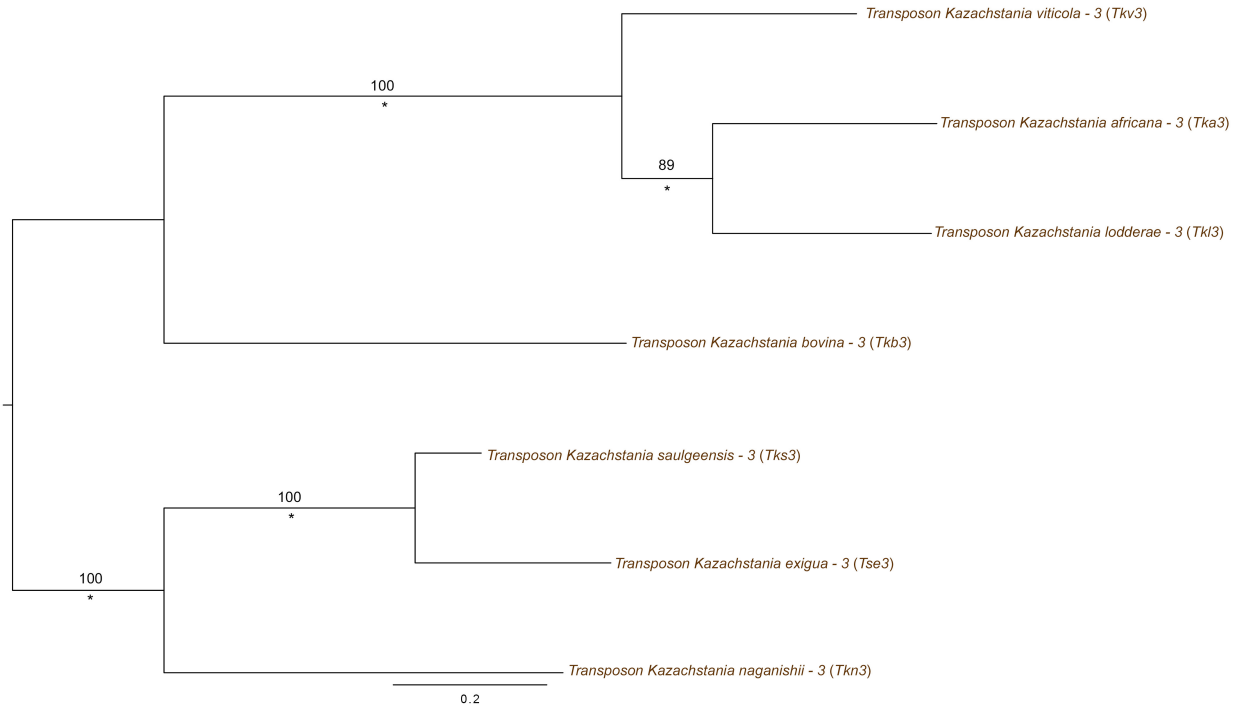


Figure 2.9: **Maximum likelihood phylogeny of chromoviral *Ty3*- like Pol amino acid sequences from *Kazachstania* species.** The phylogeny was created with raxmlGUI 1.5 beta via python with the employment of the PROTCAT model and estimated amino acid frequences with the RTREV matrix (Silvestro and Michalak, 2011) from 518 amino acid positions. The scale bar signifies the number of amino acid substitutions per site. All enzymatic domains from Pol were included. Formatting is stated in Figure 2.1.

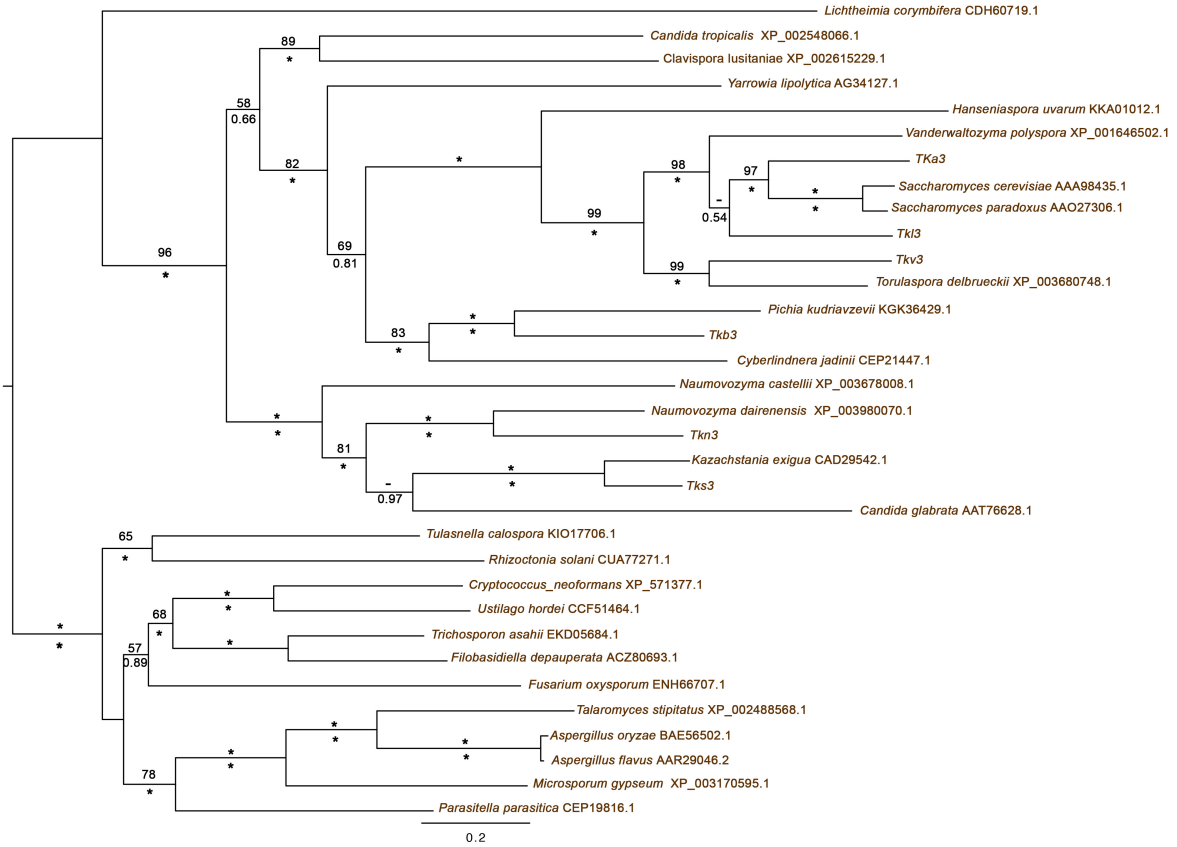


Figure 2.10: **Maximum likelihood phylogeny of chromoviral *Ty3*- like Pol amino acid sequences from Saccharomycetaceae.**The phylogeny was created with raxmlGUI 1.5 beta via python with the employment of the PROTCAT model (Silvestro and Michalak, 2011) with RTREV substituition matrix from 725 amino acid positions. An outgroup of non-saccharomycetaceae TE sequences are included to root the tree. Formatting is stated in Figure 2.1.

The yeast phylogeny of *Ty3*-like elements showed maximum support for biPP and ML (100/1.00) for the elements from species of Saccharomycetaceae to be positioned in a separate clade to other elements of diverse fungal species, in concordance with species phylogeny (Wolfe et al., 2015) (Figure 2.10). Within the Saccharomycetaceae grouping, the divergence of two types of *gypsy* element was highly supported (96%ML/1.00biPP). The positioning of *Kazachstania* species seen in Figure 2.9, has been further supported, with one group clustering *K. saulgeensis*, *K. exigua* and *K. naganishii* elements, with *gypsy* elements from *Naumovozyma* species. The remaining *Kazachstania* species are nested in a group with *Saccharomyces* species (Figure 2.10). The two types of *gypsy* element were grouped with maximum support (100%/1.0biPP). The distribution of *gypsy*-like elements did not reflect expected species phylogeny, which provided support for ancestral diversity or horizontal transposable element transfer (HTT) within the superfamily (Figure 2.11 and Figure 2.1).

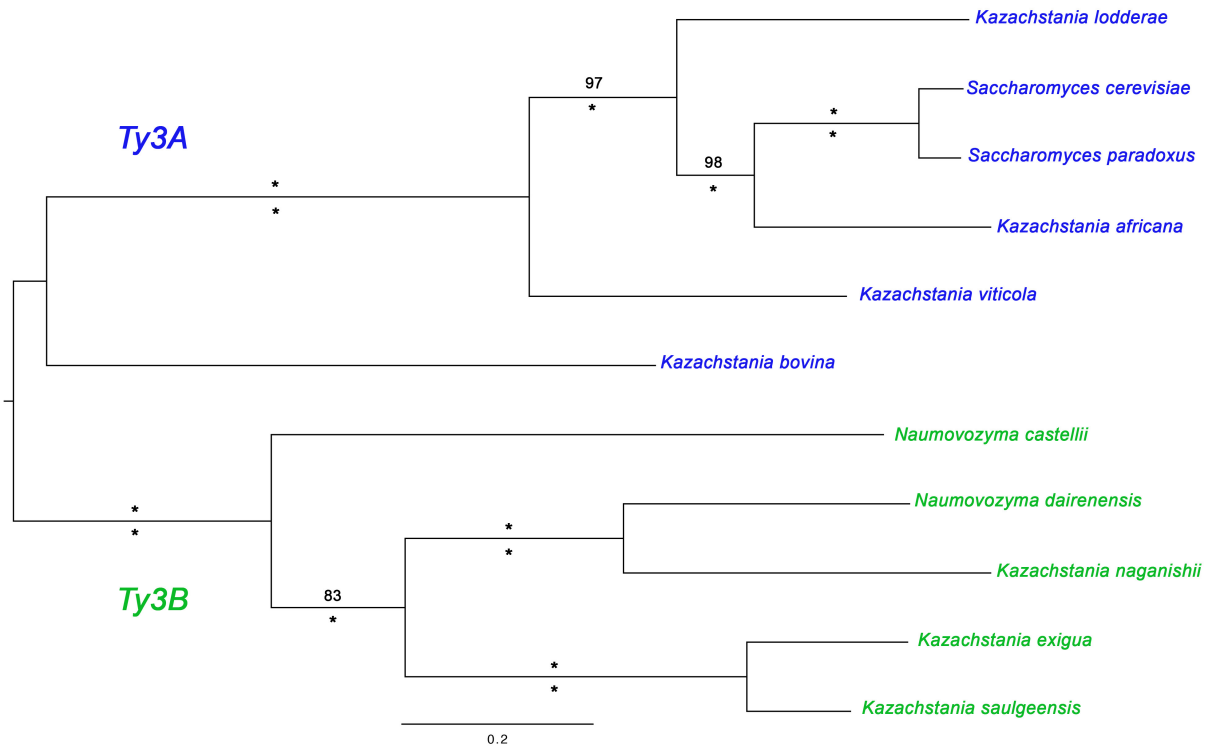


Figure 2.11: **Maximum likelihood phylogeny of chromoviral *Ty3*- like Pol amino acid sequences from 11 members of Saccharomycetaceae superfamily.**The phylogeny was created with raxmlGUI 1.5 beta via python with the employment ofthe PROTCAT model (Silvestro and Michalak, 2011) from 557 amino acid positions . The scale bar signifies the number of amino acid substitutions per site. All enzymatic domains from Pol were included. All elements grouped to *Ty3A* are coded in blue, and sister elements of *Ty3B* are coded in green. Annotated species names are based upon the species origin that the *gypsy* element was uncovered.

Similarly to the *gypsy*-like phylogeny for the elements of *Kazachstania* species, phylogenetic analysis was ran to review the *copia*-like elements identified by RepeatMasker (*RepeatMasker*, 1996) and reciprocal BLAST (Altschul et al., 1990). The *copia* phylogeny was moderately resolved with high support (81-100%ML/1.00biPP) and the elements were positioned as two distinct clades; *Ty1*-like and *Ty5*-like (Figure 2.12). As seen in the *gypsy* phylogeny, the element from *K. bovina* was placed as a sister group to the other *Kazachstania* elements, *Tse5, Tkn5* and *Tkse5*. Within the *Ty1*-like clade, *Tkv1* is placed on a long branch, separate to *Tkn1, Tse1* and *Tks1* (Figure 2.12). With this, a second phylogeny was ran with known *copia-like* elements from species within Saccharomycetaceae to further support sequence similarity of elements to *Ty* families, as outlined by the original RepeatMasker results (*RepeatMasker*, 1996).

As shown in Figure 2.13, three main clades were found with high resolution (100%ML/1.0biPP). *Tkv1* was grouped with *Ty4* from *S. cerevisiae*, as a sister group to *Ty1*-like elements. The *copia* element from *K. viticola* was renamed to *Tkv4*, due to the high similarity with *Ty4*. As seen in the *gypsy* phylogeny, further evidence of ancestral divergence is supported here, with the mobile elements distribution not reflecting species phylogeny (Figure 2.13). Strong evidence to support this is within the *Ty5*-like clade, where *Tkn5* and *Tkse5* are positioned together with high support (98%ML/1.00biPP). This is in contrast to species phylogeny, where *K. servazzii* and *K. naganishii* are distant relative within the *Kazachstania* genus. As horizontal transfer of TEs has been hypothesised in *S. cerevisiae*, with the acquisiton of two *Ty* elements (*Ty2* and *Ty3*) from other species within the genus; *S. mikatae* and *S. paradoxus* (Carr et al., 2012), it was considered that a similar transfer event could have occurred between species of *Kazachstania*. Transferred TEs from both classes have been documented, and facilitated between sequences of high similarity (Walsh et al., 2013), which outlines that unicellular organisms are commonly susceptible to gene transfer (Fitzpatrick, 2012). However, as the phylogeny indicates 40% divergence, it is plausible for the positioning to be due to ancestral diversity, or HTT.

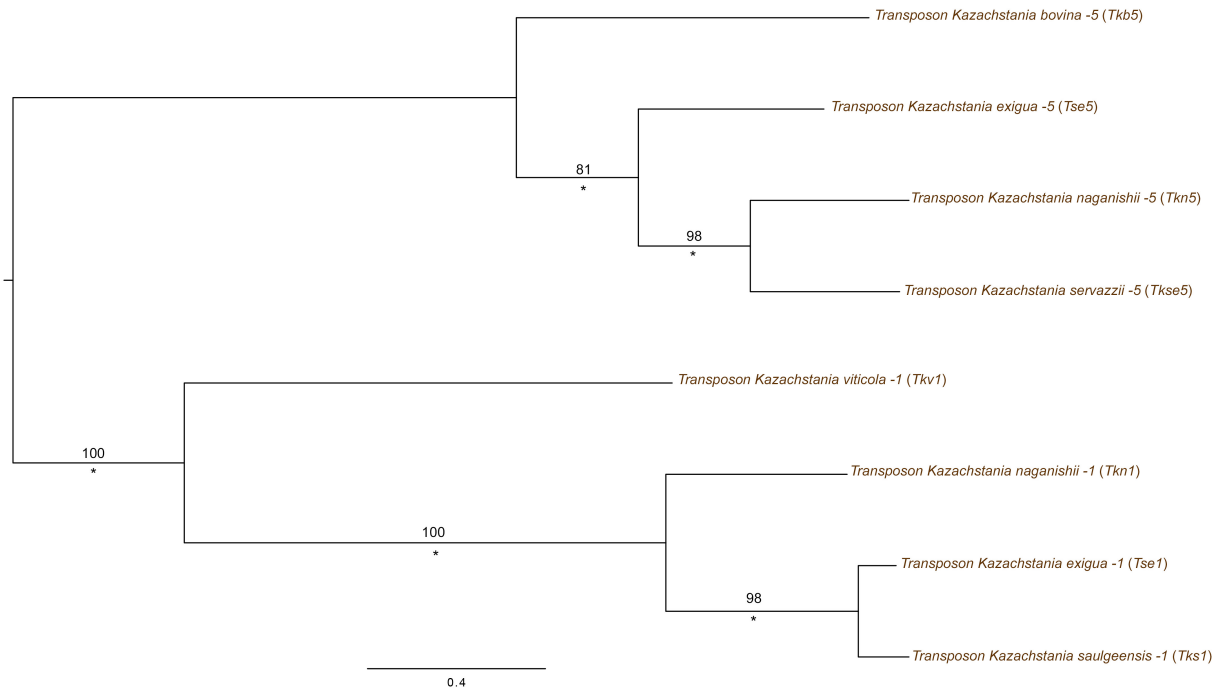


Figure 2.12: **Maximum likelihood phylogeny of *copia*- like Pol amino acid sequences from *Kazachstania* species.** The phylogeny was created with raxmlGUI 1.5 beta via python with the employment of the PROTCAT model (Silvestro and Michalak, 2011) from 600 amino acid positions. The scale bar signifies the number of amino acid substitutions per site. All enzymatic domains from Pol were included. Formatting is stated in Figure 2.1.

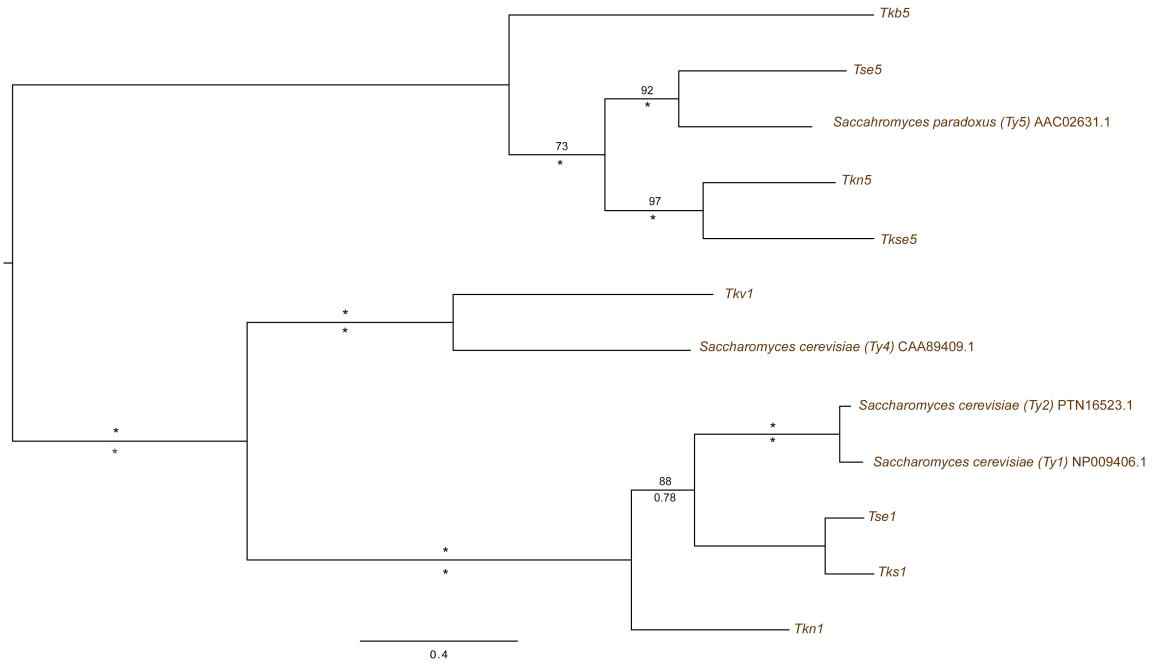


Figure 2.13: **Maximum likelihood phylogeny of *copia* Pol amino acid sequences from Saccharomycetaceae**The phylogeny was created with raxmlGUI 1.5 beta via python with the employment of the PROTCAT model (Silvestro and Michalak, 2011) from 622 amino acid positions . The scale bar signifies the number of amino acid substitutions per site. All enzymatic domains from Pol were included. Formatting is stated in Figure 2.1.

Chromodomain annotation

During phylogenetic analyses, the *gypsy*-like elements uncovered in budding yeast species of Saccharomycetaceae were reviewed for the presence of a chromodomain, to determine if *gypsy* elements were likely to be chromoviruses. No putative chromodomains were uncovered in the eight *Kazachstania* species. Details of the predicted chromodomain uncovered in *S. cerevisiae* are outlined in Appendix B.

2.3.4 Identification of major tRNA genes and optimal codons

Optimal codons were determined for the four novel *Kazachstania* species, and publicly available *Kazahstania* species, *K. africana* and *K. naganishii* using CodonW (Peden, 1999), and major tRNA genes using tRNAscan-SE 2.0 (Lowe and Chan, 1997). In line with the AT-bias observed in five of the six *Kazachstania* species, 98 out of 151 optimal codons ended in adenine or uracil (Table 2.7).

It is acknowledged that the CodonW correspondence analysis can give false positives regarding optimal codons, and therefore the defined codons shown here may not be accurate (Table 2.7). However, due to lack of expression data for the *Kazachstania* species, optimal codons could not be calculated based on expression. The codon usage of known highly expressed gene, EF1A was run using CodonW, to assess bias and optimal codons calculated for the known highly expressed gene for each *Kazachstania* species. It was found that the codons identified by CodonW for EF1A are supported by the COA output, with the majority of optimal codons found for EF1A to mirror those identified for the *Kazachstania* host genes, however it cannot be considered as definitive proof that the optimal codons are genuine (Appendix B).

Major tRNA genes were identified in each species, in order to compare the host species tRNA genes to the optimal codons identified in CodonW (Peden, 1999). The tRNA genes identified in the six species showed some similarities, with the exception of *K. exigua*. The *Kazachstania* species were found to harbour 160 - 310 tRNA genes, with *K. exigua* which was found to possess an increased number of 384 tRNA genes, which would be expected with the increase genomic size in comparison to the other species (Appendix B). In two-fold degenerate amino acids, there was a complete absence of tRNA genes with adenine at the first position of the anticodon (Appendix B). However, the other amino acids show complementary optimal codons and major tRNA genes, with many of the higher degeneracy amino acids showing evidence of deamination. Signatures for deamination is seen in tRNA anticodons which have adenine at the wobble 1st position, that would

Table 2.7: **Optimal codons assigned for the *Kazachstania* species based on CodonW analysis.** An asterisk by a codon denotes that the codon would be optimal to the major tRNA gene if the adenosine at the 3rd wobble position underwent deamination modification. Optimal codons which complement the major tRNA genes are written in bold.

Amino acid	*K. africana*	*K.bovina*	*K. exigua*	*K. lodderae*	*K. naganishii*	*K. viticola*
Phe	**UUC**	**UUC**	**UUC**	**UUC**	**UCC**	**UUC**
Leu	**UUA**	**UUA**	**UUA**, CUA	**UUA**	**UUA**, UUG, CUU, **CUA**	**UUA**
Ile	**AAU**, AUC*	**AUU**, AUC*	**AUU**, AUC*	**AUU**, AUC*	**AUU**	**AUU**, AUC*
Val	**GUU**, GUC*	**GUU**, GUC*	**GUU**, GUC*	**GUU**, GUC*	**GUU**, GUC*, GUA	**GUU**, GUC*
Ser	**UCU**, UCC*	**UCU**, UCA	**UCU**, UCC*	**UCU**, UCC*	**UCU**, UCC*, UCA	**UCU**, UCC*
Pro	**CCA**	**CCA**	**CCA**	**CCA**	**CCU, CCA**	**CCA**
Thr	**ACU**, ACC*	**ACU**	**ACU**, ACC*	**ACU**, ACC*	**ACU**, ACC*	**ACU**, ACC*
Ala	**GCU**, GCC*	GCU	**GCU**, GCC*	**GCU**, GCC*	**GCU**, GCC*	**GCU**, GCC*
Tyr	**UAC**	UAU	**UAC**	**UAC**	UAU	**UAC**
His	**CAC**	CAU	**CAC**	**CAC**	CAU	**CAC**
Gln	**CAA**	**CAA**	**CAA**	**CAA**	**CAA**	**CAA**
Asn	**AAC**	AAU	**AAC**	**AAC**	AAU	**AAC**
Lys	**AAG**	AAA	**AAG**	**AAG**	**AAA**	**AAG**
Asp	**GAC**	GAU	**GAC**	**GAC**	GAU	**GAC**
Glu	**GAA**	**GAA**	**GAA**	**GAA**	**GAA**	**GAA**
Cys	UGU	UGU	UGU	UGU	UGU	UGU
Arg	**AGA**	**AGA**	**CGU, AGA**	**CGU, AGA**	**CGU, AGA**	**CGU, AGA**
Gly	GGU	GGU	GGU	GGU	GGU	GGU
Stop	UAA	UAA	UAA	UAA	UAA	UAA

allow complementary pairing to codons with cytosine, guanine and uracil, if adenine is deaminated to inosine.

In the six *Kazachstania* species there is evidence of deamination in amino acids with three-fold or greater degeneracy (Table 2.7 and Appendix B). For amino acids; Ile, Leu, Ser, Thr and Val, evidence was found to support the deamination of their major tRNA genes in all six *Kazachstania* species (Appendix B). The majority of tRNA anticodons identified for the listed amino acids were found to have adenine at the first wobble position. Evidence for deamination was also found for Alanine in the yeast species, with the exception of *K. bovina*. With this, the accompanying optimal codons identified by CodonW (Peden, 1999), are predominantly C-ending, which would suggest that the adenine base in the anticodon is deaminated to inosine, which would allow complementary base pairing to the cytosine nucleotides of optimal codons (Table 2.7). In contrast, data would suggest that Arg, Gly, Leu and Pro do not deaminate, as they do not possess tRNA genes with adenosine at the wobble position (Appendix B). However, with the exception of Gly, the remaining three amino acids had optimal codons which were complementary to the major tRNA genes with standard Watson-Crick nucleotide pairing (Table 2.7). No optimal codons were found to complement tRNA genes for amino acids; Cys and Gly in any of the six *Kazachstania* species.

Conservation was seen across the yeast species in optimal codons for many amino acids. Firstly, all six species have CAA as an optimal codon for glutamine (Gln), with the complementary anticodon TTG with the highest frequency for Gln tRNA genes. Conserved optimal codons were also seen for glutamic acid (Glu) , with GAA as the preferred codon, and have the complementary TTC anticodon in their most abundant Glu tRNA genes (Table 2.7). Optimal codons were found to complement major tRNA genes for the majority of *Kazachstania* species, with the exception of *K. bovina* and *K. naganishii* (Table 2.7).

2.3.5 Codon usage of host genes in novel *Kazachstania* species

Southworth et al. (2018) found that host gene codon usage bias in unicellular protists is driven by natural selection at the level of translational accuracy and efficiency. Codon usage statistics via CodonW (Peden, 1999) were employed for six *Kazachstania* species, upon transcriptome availability. Mean values of *Nc* varied from 31.60 - 51.03 (Table 2.8 and Figure 2.14). The strongest level of bias based on *Nc* was seen in *K. bovina* (31.60±4.57), and the least bias was seen for host genes of *K. naganishii* (51.03±6.90). The direction of bias was also reviewed based on the value of GC at synonymous third positions (GC3s), which showed that host genes of all six species favoured AT ending codons (Table 2.8), with all GC3s mean values being <0.32, with the exception of *K. naganishii* which was found to have a GC3s value of 0.569 (Table 2.8).

Table 2.8: **Mean codon usage statistics in the transcriptomes of six *Kazachstania* species.** Data included the average value for effective number of codons (*Nc*), GC3s and frequency of optimal codons (Fop) for each species. Standard deviation was calculated for each data set (±s.d).

Species	GC3s	Nc	Fop
K. africana	0.315±0.054	46.07±6.69	0.506±0.113
K. bovina	0.113±0.052	31.60±4.57	0.700±0.075
K. exigua	0.216±0.069	37.66±6.36	0.548±0.114
K. lodderae	0.260±0.053	42.30±6.93	0.517±0.119
K. naganishii	0.569±0.116	51.03±6.90	0.463±0.106
K. viticola	0.243±0.057	41.39±6.20	0.509±0.110

Trends were observed for the *Nc* plot of the six *Kazachstania* species. Similar patterns were seen between the *Kazachstania* species, with the exception of *K. bovina* and *K. naganishii* (Figure 2.15). A positive correlation was seen between GC3s and *Nc* for *K. bovina*, with the majority of genes representing a GC3s value of <0.3 (R^2= 0.528) (Figure 2.15 and Appendix B). In contrast, *K. naganishii* was found to show contrasting patterns of codon usage, with the highly biased genes to favour GC ending codons (Figure 2.15). The remaining four species host genes were found to favour AT codons, with a weak positive relationship between *Nc* and GC3s value (Figure 2.15).

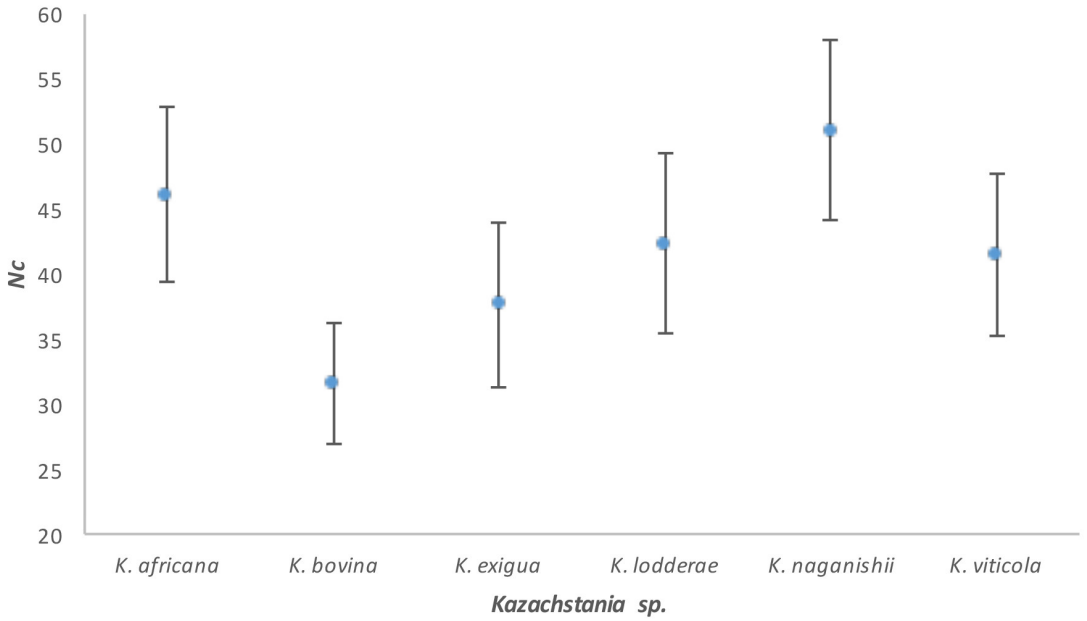


Figure 2.14: **Average *Nc* values for host genes in six *Kazachstania* species.** Standard deviation error bars for each species are shown.

2.3.6 The role of selection as a driver for codon usage in the *Kazachstania* species

One possible explanation of the AT-preference observed in the majority of *Kazachstania* species, is that AT-ending codons are being selected for translational efficiency. In order to determine if selection is driving the AT bias in the budding yeast species, rather than mutation pressure, bias categories were calculated for each *Kazachstania* species, based on *Nc*. Each 5% bias category for high, mid and low biased genes were reviewed for GC3s and Fop, and comparatively assessed within the genus (Figure 2.16).

Figure 2.16 shows that each species Fop value decreases from the highly biased to medium biased and least biased genes, except for *K. naganishii*, where the low biased genes were found to be significantly different to the medium biased genes, with a greater value of Fop. The relationship between Fop and bias categories provided support for selection being a driver of codon usage, with the employment of optimal codons being higher in genes which are presumably more highly expressed, and therefore require efficient translation.

In contrast, GC3s showed a less consistent pattern for the six species (Figure 2.17). As the yeast species were found to have an AT-preference for direction of bias, it would be expected that

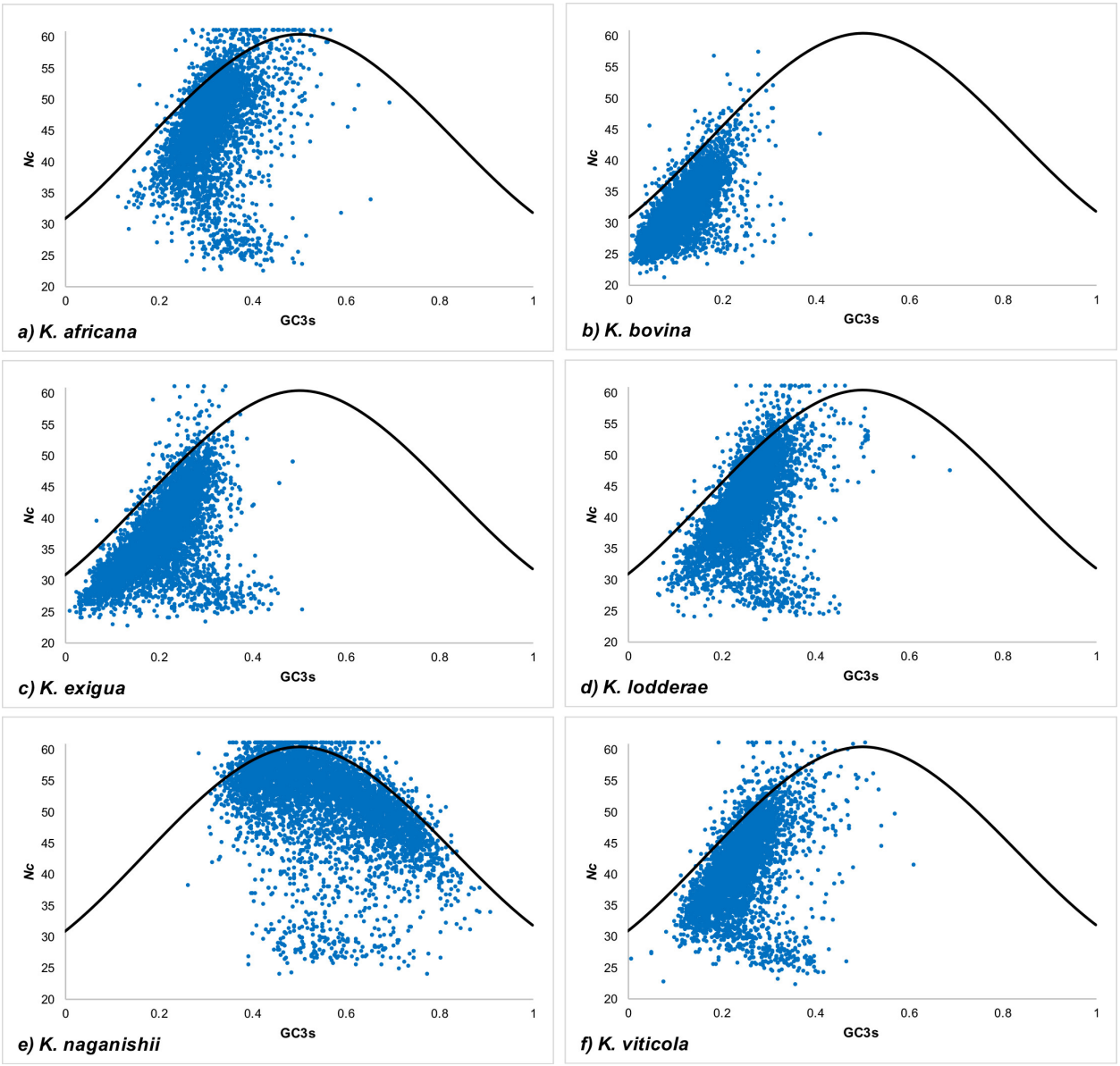


Figure 2.15: **Nc plot against GC3s for the genes of the *Kazachstania* species.** Nc values were plotted against GC3s for a) *K. africana* ; b) *K. bovina*; c) *K. exigua*; d) *K. lodderae*; e) *K. naganishii* and f) *K. viticola*. The modified equation, *Nc*=2+S+29/[S2+(1-S)2], from Wright (1990), with S=GC3s, was used to create the parabolic curve on each *Nc* plot.

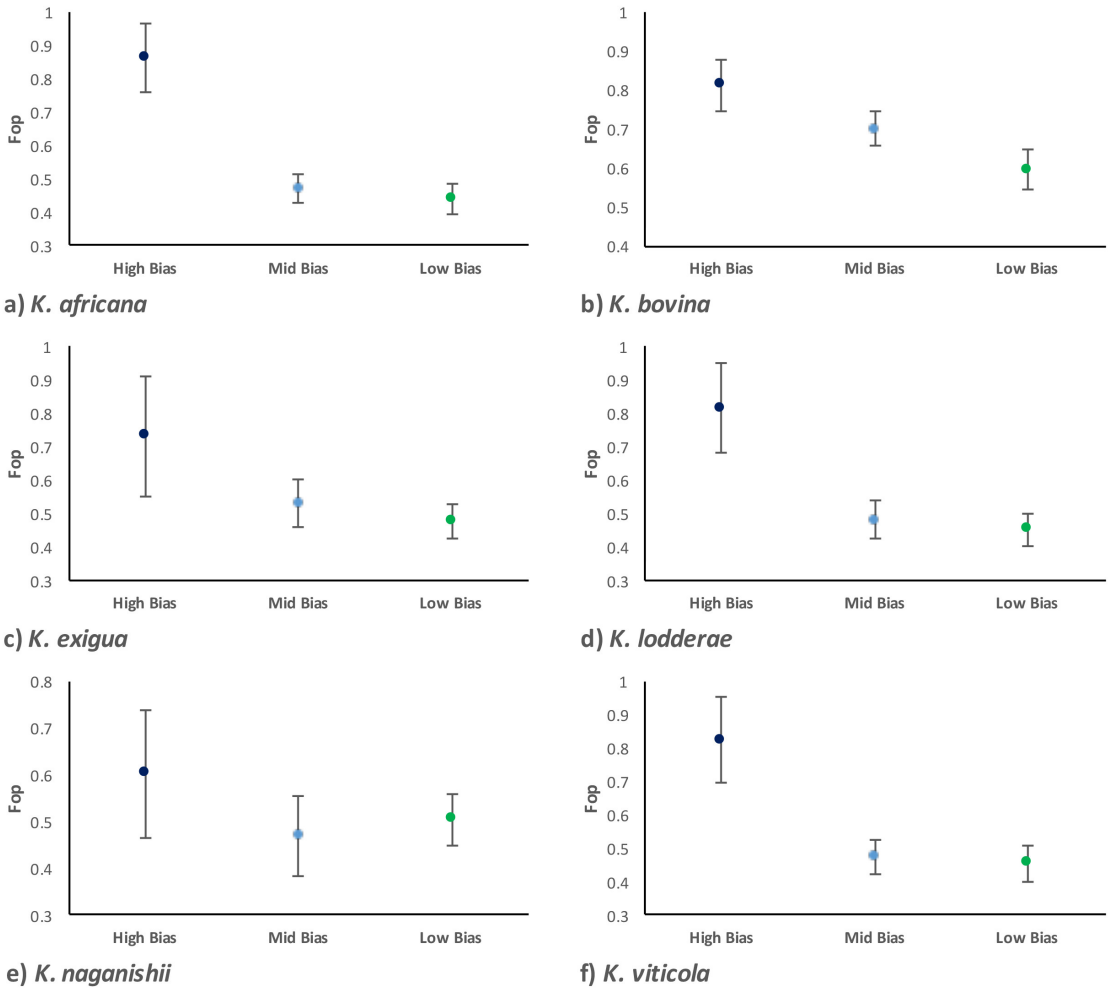


Figure 2.16: **Average Fop value for 5% bias categories for the six yeast species**. Error bars were included for each bias category to show values of standard deviation per dataset.

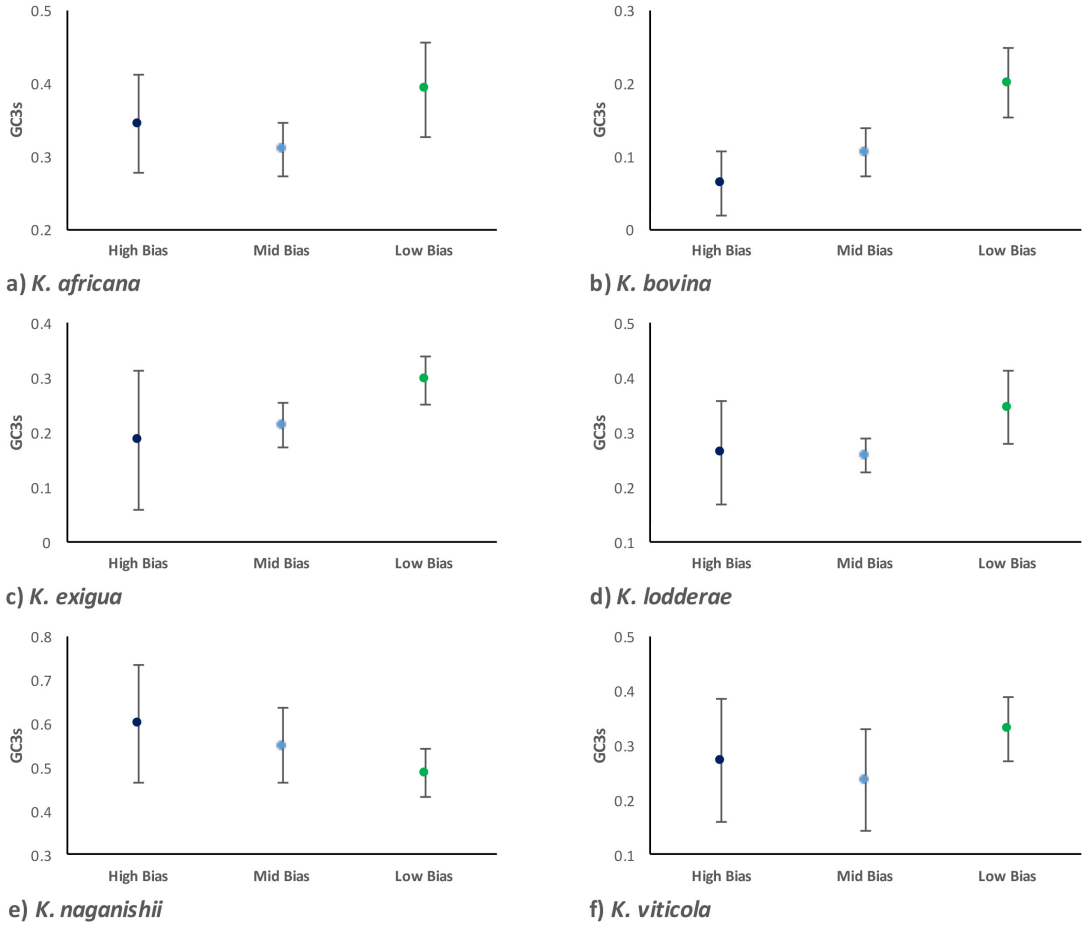


Figure 2.17: **Average GC3s value for 5% bias categories for the six yeast species**. Error bars were included for each bias category to show values of standard deviation per dataset.

the value of GC3s would increase as the level of bias decreased, with AT-ending codons being more prevalent in highly expressed genes. As shown in Figure 2.17, this pattern was seen in the majority of *Kazachstania* species, with five of the six species showing diversity difference between GC3s when comparing the highly bias categories to the low bias categories, with a far greater value of GC3s in less biased genes (Figure 2.17). As expected from the atypical results found based on GC3s data and *Nc* plot, *K. naganishii* presented with a contrasting pattern, where GC3s was found to decrease as the bias category decreased. With this, it is supported that *K. naganishiii* has a preference for GC-ending codons, which are employed at a higher frequency in highly biased genes.

2.3.7 Codon usage of transposable element families in novel *Kazachstania* species

For the TE codon usage analysis, a contrasting association between GC3s and *Nc* was observed for the TE ORFs of all *Kazachstania* species (Figure 2.18). For the TEs found in *Kazachstania* species, a strong positive correlation was seen between GC3s and *Nc* (R^2=0.878) (Figure 2.19). This supported that codon usage bias is seen to decrease with the greater abundance of guanine or cytosine at the third position. With this, it was found that the majority of TE families all exhibited an excess of AT-ending codons (Table 2.9).

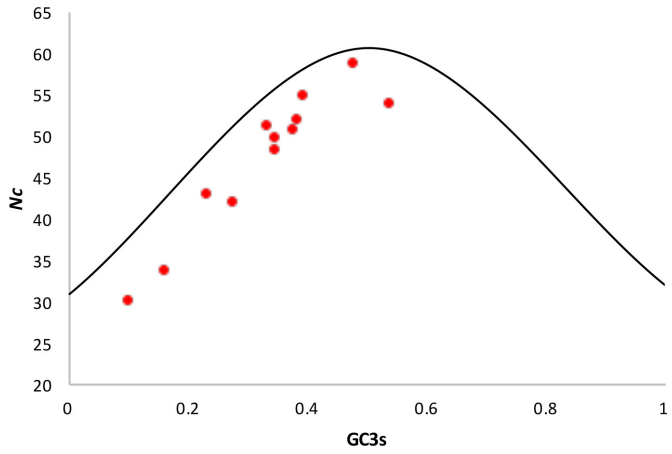


Figure 2.18: ***Nc* plot against GC3s for the TEs uncovered in *Kazachstania* species** Nc values were plotted against GC3s for the TEs uncovered in the eight *Kazachstania* species. The modified equation, *Nc*=2+S+29/[S2+(1-S)2], from Wright (1990),with S=GC3s, was used to create the parabolic curve on each *Nc* plot (Southworth et al., 2018).

The strongest GC3s bias is observed in the TE families of *K. naganishii*, with GC3s values ranging from 0.38-0.54, mirroring the host species GC preference (Table 2.9). In contrast, a very weak GC3s is observed in the mobile elements for *K. bovina*, with both elements GC3s value scoring <0.17 (Table 2.8). GC-bias in synonymous 3rd positions is mirrored in each species TE codon usage bias, with conservation observed between elements and the host genes. A strong positive correlation was seen between GC3s and Nc of the TEs in the *Kazachstania* species (Figure 2.19). Specifically, the elements of *K. bovina* were found to show the strongest codon usage bias with a mean *Nc* value of 31.82, which was similar to the host genes bias of 31.60 (Table 2.8).

Table 2.9: **Codon usage statistics for the LTR retrotransposon families from the genus *Kazachstania***. Mean values for GC3s, Nc and Fop are listed for each TE family and host species are annotated.

Family	GC3s	Nc	Fop
K. africana			
Tka3	0.395	54.84	0.512
K. naganishii			
Tkn1	0.377	50.65	0.571
Tkn3	0.479	58.79	0.520
Tkn5	0.54	53.95	0.468
K. bovina			
Tkb3	0.102	30.00	0.738
Tkb5	0.163	33.64	0.708
K. exigua			
Tse1	0.348	49.7	0.440
Tse3	0.333	51.13	0.496
Tse5	0.275	42.08	0.577
K. lodderae			
Tkl3	0.384	52.02	0.519
K.viticola			
Tkv3	0.345	48.23	0.652
Tkv4	0.232	43.04	0.696
	Mean=0.33±0.12	**Mean=47.34±8.63**	**Mean=0.57±0.10**

However, the mean *Nc* for all TEs of *Kazachstania* is much higher than values seen for *Tkb3* and *Tkb5* (47.34±8.63), showing less codon usage bias for the elements collectively and the highly biased elements of *K. bovina* being inconsistent when compared to the other mobile elements uncovered in the remaining *Kazachstania* species (Table 2.8, and Figure 2.20).

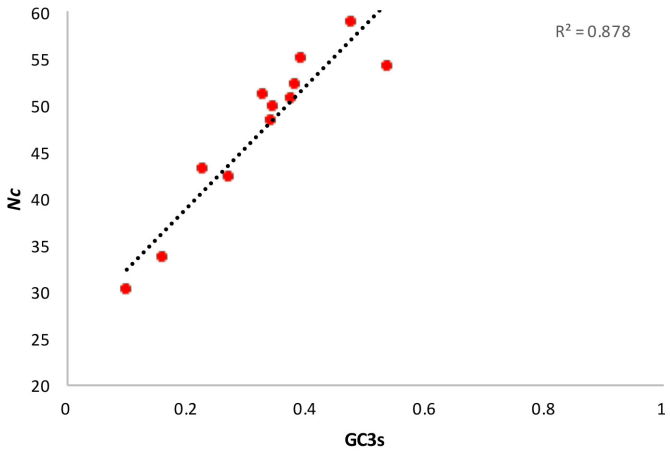


Figure 2.19: **Relationship between GC3s and *Nc* for the 12 TE families uncovered in the *Kazachstania* species.** A linear trendline, with R^2 value, was added to the graph to assess the strength of the positive relationship.

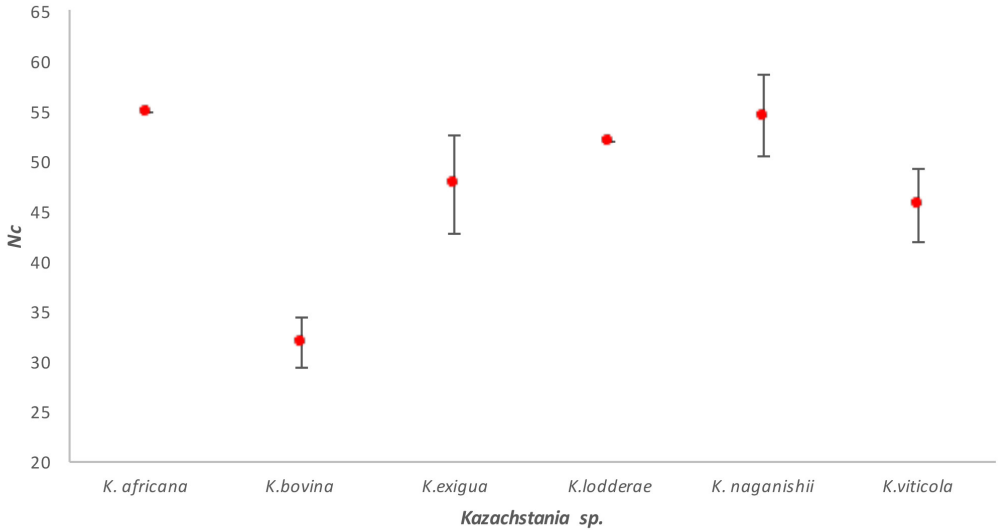


Figure 2.20: **Average Nc values for TE families in eight *Kazachstania* species.** Standard deviation error bars for each species are shown.

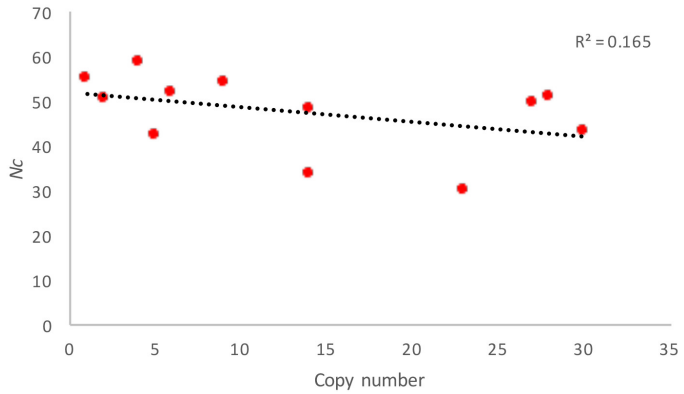


Figure 2.21: **Relationship between copy number of TE families and effective number of codons (*Nc*).** Copy number for each TE family was plotted against *Nc* for each of the *Kazachstania* species investigated.

Trends between strength of codon usage bias (*Nc* value), and copy number were reviewed for the TE families. It was found that high copy number families seemed to show a stronger codon usage bias, in comparison to families with lower copy number seen in the host species, however it was a very weak association (R^2=0.165) (Figure 2.21). Similarly, no trend was seen between copy number and Fop (R^2= 0.06) (Figure 2.22). However, the two data points which influence the association are *Tse1* and *Tse3* of *K. exigua*. Both TE families are present with a high number and copies and seem to present with a lower value of Fop, when compared to the remaining TE families, the majority of which have a value of Fop >0.50 (Table 2.8). As expected, a negative relationship was seen between *Nc* and Fop, supporting that the frequency of optimal codons is higher in more highly biased genes (Figure 2.23).

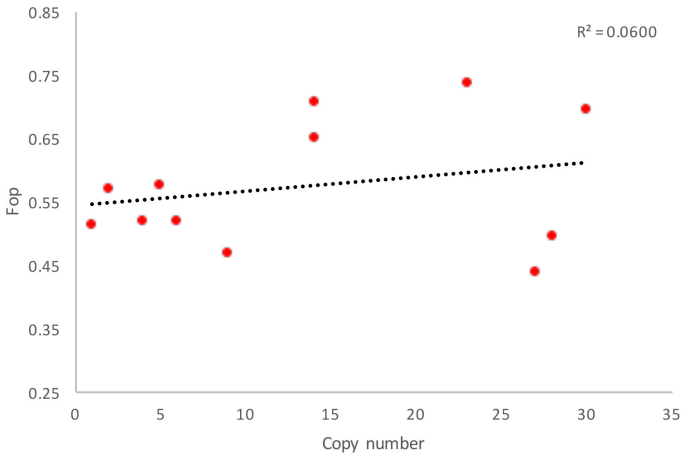


Figure 2.22: **Relationship between copy number of TE families and frequency of optimal codons (Fop).** Copy number for each TE family was plotted against Fop for each of the *Kazachstania* species investigated.

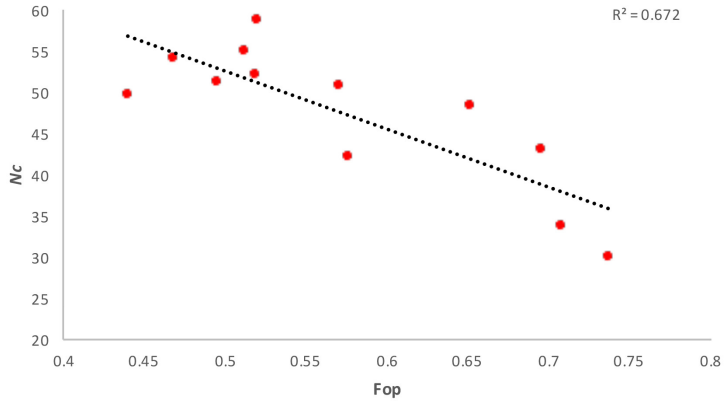


Figure 2.23: **Relationship between frequency of optimal codons (Fop) and strength of codon usage bias (*Nc*)**.

2.3.8 The influence of selection on codon usage bias in the TE families of the *Kazachstania species*

The most abundant codon for each amino acid was calculated using the Pol domains for the LTR retrotransposons uncovered in the *Kazachstania* species (Table 2.10). Abundant codons were then reviewed in relation to the major tRNA genes identified for each host species, as well as the optimal codons determined per amino acid for each *Kazachstania* species. Several of the abundant codons identified for the TE families were complementary to the tRNA anticodons for the

host species, providing evidence that preferred codon employment for the mobile elements should result in efficient translation, similar to the highly biased host genes. The highest complement was found between the optimal codons of TEs uncovered in *K. bovina* and the host species major tRNA genes (Table 2.10).

Tse1 and *Tkn5* were found to not have a majority of amino acids where the abundant codons were complementary to the major tRNA genes of the host. However, with the exception of *Tkn5*, the optimal codon CAA for Gln remained conserved in the TE families, as well as GAA which was found to be the most abundant codon in all TE families for amino acid, Glu, as identified for the host genes. Evidence supported that TE families are under the same evolutionary pressure as host genes, with the suggestion that selection is driving codon usage bias in both host genes, and uncovered TE families, with the exception of *K. naganishii* which may be evolving under different pressures to explain the contrasting GC preference for this species.

Table 2.10: **Abundant codons assigned for the TE families of the *Kazachstania* species based on CodonW analysis.** Optimal codons which complement the major tRNA genes are written in bold.

Amino acid	*Tka3*	*Tkb3*	*Tkb5*	*Tse1*	*Tse3*	*Tse5*	*Tkl3*	*Tkn1*	*Tkn3*	*Tkn5*	*Tkv3*	*Tkv4*
Phe	UUU	UUU	UUU	UUU	UUU	UUU	UUU	UUU	UUC	UUC	UUU	UUU
Leu	**UUA**	**UUA**	**UUA**	CUG	CUG	**UUA**	**UUA**, CUU	**UUA**	**UUA**	**CUU**	**UUA**	**UUA**
Ile	**AUU**	**AUU**	**AUU**	AUA	**AUC**	**AUU**	**AUU**	AUA	**AUU**	**AUC**	**AUU**	**AUU**
Val	**GUU**	**GUU**	**GUU**	GUA	**GUU**	**GUU**	**GUU**	GUG	**GUU**	**GUC**	**GUU**	**GUU**
Ser	UCA	**UCA**	**UCA**	UCA	UCA	UCA	UCA	**UCU**	**UCA**	AGC	**UCC**	**UCC**
Pro	**CCA**	CCU	CCU	**CCA**	**CCA**	**CCA**	**CCA**	**CCA**	CCG	CCC, **CCA**, CCG	**CCA**	**CCA**
Thr	**ACU**, ACA	**ACU**	**ACU**	ACA	ACA	**ACU**	ACA	ACA	**ACC**	ACG	**ACU**	ACA
Ala	**GCC**	**GCU**	**GCU**	GCA	GCA	**GCU**	**GCU**	**GCU**	**GCU**	**GCC**	**GCU**	**GCU**
Tyr	**UAC**	**UAU**	**UAU**	UAU	**UAC**	UAU	**UAC**	UAC	UAC	UAC	**UAC**	UAU
His	**CAC**	**CAU**	**CAU**	CAU	**CAC**	CAU	CAU	**CAU**	CAC	CAC	**CAC**	CAU
Gln	**CAA**	**CAA**	**CAA**	**CAA**	**CAA**	**CAA**	**CAA**	**CAA**	**CAA**	CAG	**CAA**	**CAA**
Asn	**AAC**	**AAU**	**AAU**	AAU	**AAC**	AAU	**AAC**	AAC	AAC	**AAU**	AAU	AAU
Lys	AAA	**AAA**	**AAA**	AAA	AAA	AAA	AAA, **AAG**	**AAA**	**AAA**	AAG	AAA	AAA
Asp	GAU	**GAU**	**GAU**	GAU	GAU	GAU	GAU	GAC	GAC	**GAU**	GAU	GAU
Glu	**GAA**	**GAA**	**GAA**	**GAA**	**GAA**	**GAA**	**GAA**	**GAA**	**GAA**	GAA, GAG	**GAA**	**GAA**
Cys	UGC	**UGU**	**UGU**	UGC	**UGU**, UGC	**UGU**	UGU	UGC	**UGU**	**UGU**, UGC	UGC	**UGU**
Arg	CGA	**AGA**	**AGA**	**AGA**	**AGA**	**CGU**	**CGU**	**AGA**	**AGA**	**AGA**	AGA	AGA
Gly	**GGU**	**GGU**	**GGU**	**GGU**	**GGU**	**GGU**	**GGU**	**GGU**, GGA	**GGU**	GGA	**GGU**	**GGU**
tRNA	12/19	16/18	16/18	5/18	11/19	11/18	12/19	10/19	12/18	10/18	12/18	10/18
tRNA & Optimal Codons	12/19	16/18	17/18	5/18	11/19	11/18	13/19	10/19	13/18	10/18	13/18	11/18

2.4 Discussion and concluding remarks

The genus *Kazachstania* had limited genome sequence availability and therefore very little data regarding TE content and genome characteristics. Following screening of four publicly available *Kazachstania* genomes (*K. africana, K. naganishii, K. saulgeensis* and *K. servazzii*), as well as the addition of the four novel *Kazachstania* species that were sequenced for this project, similarities and differences were drawn between the novel species and the *S. cerevisiae* reference genome, S288c (Carr et al., 2012), as well as divergence within the *Kazachstania* genus itself. The more robust dataset allowed for unprecedented trends to be drawn for the genus, including TE abundance, or if variation exists across species as seen in other genera within Saccharomycetaceae (Neuvéglise et al., 2002).

As well as TE content, the genome availability allowed for comparative genomics within the superfamily and genus, specifically the first investigation of codon usage bias in the *Kazachstania* species with transcript availability. With this, codon usage of the host genes allowed for comparison with TE families, to determine if the mobile elements seemed to be under the same evolutionary pressures that drive codon usage bias in the host.

2.4.1 Genome characteristics and TE review in *Kazachstania* species

Across the *Kazachstania* genomes studied, the major difference between species regarding genome characteristics was observed for the size and gene quantity of *K. exigua*. Typically *Kazachstania* species genomic size ranged from 10Mb - 13Mb, with a total cds between 5300 and 6000 genes. Striking contrast was seen for *K. exigua*, which was found to have a genomic size of 24.8Mb, and a total number of 9964 coding genes. The genes were run through a gene orthology annotation program to depict the function and categorisation of each gene, to assess whether the original annotation was valid and the genes could be defined as functional. Results showed that 86% of the genes (8569) were annotated with known function. However the greatest proportion of gene allocation was found for category S, of ”Unknown Function”.

If the unidentified genes were misannotated, *K. exigua* would still be found to have 8500 coding genes, which is far greater than any other *Kazachstania* species, or budding yeast within the Saccharomycetaceae superfamily (Genolevures et al., 2009). This finding suggests that 14% of the coding genes annotated in the *K. exigua* genome may be false positives, and annotation may be inaccurate. However, as the genome availability is limited for the genus *Kazachstania*, the

unknown categorisation may be due to low homology, and that the subset of genes may be left undetected due to species divergence (Casaregola et al., 2000; Gaillardin et al., 2000). Sequence divergence is a well known limitation of homology-based analysis, with the methodology being reliant on the assumption that all genes possess homology and therefore can be assigned to a specific category or function that has been previously identified (Gaillardin et al., 2000). However, with the extensive study in genome characteristics within Saccharomycetaceae (Casaregola et al., 2000; Gaillardin et al., 2000; Kurtzman, 2003; Dujon et al., 2004; Dujon and Louis, 2017), the presence of a novel gene orthologue that is found to possess low similarity to the other host genes of closely related yeast species is unlikely.

TE annotation in *Kazachstania* species

An initial similarity between the *Kazachstania* species and *S. cerevisiae* is the absence of DNA transposons. The *S. cerevisiae* reference genome has been documented to lack DNA transposons, and possess LTR retrotransposons only (Kim et al., 1998; Carr et al., 2012). From the investigation of eight species within the genus *Kazachstania*, it can be said that these species possess the same trait, only harbouring families classified as LTR retrotransposons. *Hat* DNA transposons have been found in a wild strain of *S. cerevisiae*; however, they have not spread throughout the global population (Sarilar et al., 2015). The absence of DNA transposons in the budding yeast species is not a finding that is universal within the Saccharomycetaceae superfamily. Although typically absent from *Saccharomyces*, *Kazachstania* and sister group *Naumovozyma*, DNA transposons have been uncovered in other genera within the superfamily (Sarilar et al., 2015). The absence highlighted in the species studied is likely to be caused by stochastic loss in an ancestral strain prior to divergence. Due to the nature of class II elements, it is known that active transposition is required for proliferation in the genome via a ”cut and paste” method, rather than duplicative transposition. With this, TE acquisition to the host genome is crucial to TE family success, and the loss in the last common ancestor (LCA) of the clade, would eliminate DNA transposons from the lineage, as no copies would be passed to daughter elements by vertical transmission. In contrast, evidence has shown that LTR retrotransposons have remained in the Saccharomycetaceae lineage with greater success, with documented abundance in the majority of species within the superfamily (Kim et al., 1998; Neuvéglise et al., 2002; Bleykasten-Grosshans et al., 2011; Carr et al., 2012; Bleykasten-Grosshans et al., 2013; Wolfe et al., 2015).

The immediate difference between *Kazachstania* and *S. cerevisiae* is the overall TE content percentage of the whole genome. *S. cerevisiae* TE families are known to represent >3% of the reference strain (Carr et al., 2012), whereas the TE families in the majority of *Kazachstania* species only constitute to 0.18 -0.67% of the genomes. In contrast, *K. bovina, K. exigua* and *K. viticola* were found to have a genome TE content of >2% and elements were found in multicopy; a value with greater likeness to the greater abundance seen in *S. cerevisiae*. The poor representation seen in five of the eight species cannot be generalised to the genus due to sampling limitations, and therefore would require further investigation in additional *Kazachstania* species. However, TE content diversity variation is common within Saccharomycetaceae (Neuvéglise et al., 2002). Comparative genomics of several members of the Saccharomycetaceae family with whole genome availability were ran with RepeatMasker and found that TE content varied from 3% in *S. cerevisiae* to 0.06% in *Eremothecum gossypii*. No correlation could be drawn between TE content and yeast species as variation existed within genera and clades, as well as superfamilies (Table 2.3).

From the phylogeny created with *Ty3*-like gypsy elements, *Tka3* of *K. africana* clustered with the *gypsy* elements uncovered in *S. cerevisiae* and *S. paradoxus* with high support from ML and biPP (97%ML/1.00biPP) (Figure 2.10). The elements uncovered in *K. lodderae*, *K. viticola* and *K. bovina* were also positioned within this clade with moderate support (69%ML/0.81biPP) (Figure 2.10).

The phylogeny is consistent with potential vertical inheritance for the two clades that separate the *Ty3*-like elements uncovered in the *Kazachstania* species (Figure 2.10 and 2.11). The *gypsy* sequence found in *K. naganishii* showed closest similarity with the gypsy element in *Nauvomozyma dairenensis*, in a clade with *K. saulgeensis, K. exigua, N. castellii* and *Candida glabrata* (100%ML/ 1.0 biPP). This finding provided strong support that the *gypsy* elements in *K. naganishii* are more similar to *Tse3* in *K. exigua* and *K. saulgeensis*, rather than the *Ty3* elements found in *Saccharomyces* species. A theory to rationalise this finding is referenced and annotated in Figure 2.11 which was a simplified Saccharomycetaceae *gypsy* phylogeny to outline the divergence of two types of *gypsy* element (*TY3A* and *TY3B*). The positioning of the two clades within the yeast sequences showed maximum support values, with the addition of the elements uncovered in the novel *Kazachstania* species (100%ML/1.00 biPP). The theory to explain the two types of *Ty3*; *Ty3A* and *Ty3B*, has not been outlined in previous literature (Figure 2.11), but a plausible explanation to explore.

It is considered that two ancestral copies may have existed in a common ancestor and species

in both lineages lost one type of *Ty3*, and therefore only harboured one single type of *Ty3/gypsy*, generating the branch positioning we see here (Figure 2.11). However, this theory is not conclusive, as there is little evidence to support the existence of both types of *Ty3* in a common ancestor of the two lineages, other than the phylogenetic relationships reflected in the TE phylogeny (Figure 2.11). Another explanation is that several horizontal gene transfer (HGT) events have occurred across the superfamily, resulting in many forms of the gypsy element; however this is a less probable theory, as the order of HGT events would have to mirror the divergence of the yeast species.

The reannotation of *Kazachstania* species presented here, highlights that the employment of RepeatMasker, with a Repbase library is not as comprehensive as thought, underestimating TE content for all *Kazachstania* species reviewed (*RepeatMasker*, 1996; *GIRI*, 2016). The limitation expressed here is that the use of Repbase library is a reliable program if TEs uncovered have previously been annotated in an alternative genome, and thus added to the default library. Novel *Kazachstania* elements were detected by regions of high similarity with conserved TE enzymatic domains across the eight species reviewed. With the employment of a refined custom library, including the newly annotated elements found in the initial *Kazachstania* species which were publicly available, a more accurate representation of genome content was provided.

Additional findings theorised included that of the potential existence of a chromodomain in *S. cerevisiae*, which has previously been documented as ancestrally lost or specialised (Malik and Eickbush, 1999; Marin and Llore´ns, 2000). Malik and Eickbush (1999) found that the conserved chromodomain present at the C-terminal of Integrase was apparent in several *gypsy*-like LTR retrotransposons, but absent in *S. cerevisiae* (Malik and Eickbush, 1999). However, at the same position of the element, a domain of similar length was uncovered (Malik and Eickbush, 1999). It was speculated that the chromodomain had been replaced or specialised in *Ty3* of *S. cerevisiae* (Malik and Eickbush, 1999), and the work seen here supports the latter. The investigation into the module found in *S. cerevisiae* has drawn several characteristics of the conserved chromodomain observed in many *gypsy*-like elements. Both domains are approximately 50 - 60 base pairs in length, in the same position in Pol, and from further analysis, have similar secondary structure (Kordis, 2005). From this evidence, it suggests that the domain has become specialised in *S. cerevisiae*, and other budding yeast species, rather than an ancestral loss across the entire superfamily. Chromodomains were found to be absent within the *gypsy* elements uncovered in *Kazachstania* species (data not shown).

2.4.2 Codon Usage across the genus, *Kazachstania*

Closely related species, *S. cerevisiae* has been extensively studied regarding codon usage (Sharp and Cowe, 1991; Lerat et al., 2002). Sharp and Cowe (1991) detailed selection as the main driver of codon usage in the model yeast species, with great emphasis on highly expressed genes observed to have the highest codon usage bias. It was also outlined that closely related species commonly show similar codon usage patterns, and therefore it was expected a similar trend would be seen across all *Kazachstania* species studied.

Five of the six *Kazachstania* species studied showed a similar direction of bias towards AT-ending codons, and optimal codons, which predominantly ended in adenine or thymine. For three-fold, four-fold and six-fold degenerate amino acids, there is a defined preference to adenine or uracil at the synonymous third position. It was found that optimal codons were shared across the species, for each amino acid, and only seven amino acids were found to show diverse variation in optimal codon selection (Appendix B, Table 2.7). The findngs outlined provide evidence for codon usage conservation across the majority of *Kazachstania* species. Furthermore, conservation was seen between four of the *Kazachstania* species, and the expression-determined optimal codons of *S. cerevisiae* (Sharp et al., 1986). Conservation was seen between the optimal codons of *S. cerevisiae* and *K. africana*, *K. exigua*, *K. lodderae* and *K.viticola* for all amino acids, except Leu, where UUG was optimal for *S. cerevisiae*, whereas UUA was optimal for the *Kazachstania* species. Optimal codons for *S. cerevisiae* and the remaining two *Kazachstania* species were also found to be conserved for the majority of amino acids, with the exception of Tyr, His, Aln, Lys and Asp. The high conservation between yeast species provided further support that the optimal codon selection based on COA by CodonW was likely to be accurate, as the same codons were found to be selected for closely related species, based on expression data (Sharp et al., 1986).

K. naganishii differed from the other *Kazachstania* species, with the highest GC3s value, and *Nc* plot showing a bell curve distribution (Table 2.8 and Figure 2.15). The *Kazachstania* species was found to show a preference to GC-ending codons, in contrast to the AT-bias in the other species of the genus. The finding was further supported by patterns observed across bias categories for each of the *Kazachstania* species. For *K. naganishii*, GC3s value decreased as the bias category decreased, indicating that evolutionary pressures may be driving the use of GC at synonymous third positions in highly biased genes, compared to genes of low bias. In contrast, the remaining *Kazachstania* species were found to show an increase in GC3s value as the bias category decreased,

providing evidence that evolutionary pressure is driving the employment of AT-ending codons in highly biased genes. However, it can not be determined whether the bias seen across species is predominantly driven by selection, or mutation pressure.

The variance in codon usage between *K. naganishii* and the remaining *Kazachstania* species is unclear, with no clear distinctions observed between species when reviewing genomic characteristics. Typically, species within Saccharomycetaceae have been found to show an AT-preference, leaving the *K. naganishii* finding to be unexpected (Sharp et al., 1986, 1988; Sharp and Cowe, 1991; Lerat et al., 2002; Harrison and Charlesworth, 2011). Although the GC preference observed in *K. naganishii* is in contrast to the other *Kazachstania* species, it is of note that the average *Nc* value for the host genes is much higher than the five remaining species, as well as the lowest value of Fop. This suggested that the genes of *K. naganishii* employ optimal codons at a lower frequency compared to the other *Kazachstania* species, and genes overall are less biased in this species. It is plausible that *K. naganishii* is evolving under a different evolutionary pressure, therefore explaining the GC preference observed here.

2.4.3 Selection for Optimal Codons in *Kazachstania* species

For five of the six *Kazachstania* species, there was a significant increase in GC3s when comparing the bias categories, high-mid and mid-low, consistent with the employment of translationally optimal AT-ending codons in the highly biased genes. For *K. africana* and *K. viticola*, a lower GC3s value were observed for mid bias genes compared to high bias genes, however contrasting values were observed between the high bias genes, and low bias genes, with greater values of GC3s in the low biased gene categories, and an overall positive correlation observed (Figure 2.17). A pattern was also observed for all yeast species between Fop and bias categories, with greater values of Fop observed in highly biased genes, compared to mid and low bias, across all *Kazachstania* species reviewed. The lower enrichment of optimal codons for less biased genes is consistent with predictions for points to selection for translational efficiency, however mutation bias can still not be ruled out as an additional pressure.

2.4.4 Evidence for Deamination in *Kazachstania* species

The major tRNA genes of two-fold degenerate amino acids in the majority of *Kazachstania* species were found to be complementary to their optimal codons, except for cysteine in all species. A

contrast was seen for *K. bovina* which was found to only have three two-fold degenerate amino acids that showed complement to their optimal codons (Gln, Glu and Phe) (Appendix B).

Many higher degeneracy amino acids (three-fold to six-fold) were found to have tRNA genes which held adenosine at the first synonymous position of the anticodon (wobble position) in the *Kazachstania* species. Amino acids with evidence for deamination include; Ile, Ser, Thr and Val. Alanine was also found to possess adenosine in the first position in *K. exigua*, *K. lodderae* and *K. viticola*. In contrast, no evidence for deamination was seen for higher degeneracy amino acids, Arg, Gly, Leu and Pro in the yeast species reviewed. Although there is evidence that standard Watson-Crick base pairing between optimal codons and major tRNA genes for the amino acids with adenosine at the wobble position, the high degeneracy amino acids have multiple optimal codons per amino acid, which show cytosine at the degenerate position, with the exception of Ser and Thr in *K. bovina*. The presence of cytosine at the degenerate position provides evidence that the optimal codons do not always bind to tRNA genes by the standard base pairing expected, but also deaminate the adenosine to inosine, which would allow pairing to cytosine at the degenerate position. Southworth et al. (2018) provided evidence for deamination of adenosine in unicellular protist species, which found that high degeneracy amino acids relied on the deamination of adenosine to complement major tRNA genes across all three holozoans.

Evidence of deamination in eukaryotic taxa has been uncovered, with a greater abundance of tRNA modification in multicellular species, than unicellular organisms (Rafels-Ybern et al., 2017). However, evidence of deamination in unicellular species, such as protists, and the *Kazachstania* species reviewed here, provides support that tRNA modification and deamination evolved ancestrally prior to multicellularity. Furthermore, the data provided evidence that deamination is present in Holomycota, as well as premetazoans of Holozoa, as seen in Southworth et al. (2018).

2.4.5 Conservation of Codon Usage in TE families of host species

Previous literature of codon usage for TE families has shown a preference to AT-ending codons, which has been observed in for several TE ORF in diverse taxa (Lerat et al., 2002; Jia and Xue, 2009). A similar pattern was seen here, with the majority of TEs favouring AT-ending codons, mirroring the bias reported in the host species. The only exception were TE families of *K. naganishii*, which were found to favour GC-ending codons, which was representative of the GC-preference

seen for the host species. TEs were typically AT-rich at synonymous third positions, with mean GC3s values of less than 0.40 for all families except elements uncovered in *K. naganishii*.

As seen for the host species, *K. bovina* TEs were found to show the highest level of bias (Mean *Nc*=31.82), and *K. naganishii* and *K. africana* elements had the lowest codon usage bias (Mean *Nc*=54.46;54.84). Furthermore, each TE family was found to favour optimal codons which were complementary to the major tRNA genes of the host species (Table 2.7 and 2.10). The findings supported that selection was also driving codon usage bias in the mobile elements, as well as the host genes. Although no significant relationship was observed between Fop or *Nc*, and copy number across all TE families, it is of note that the highest bias was observed for *Tkb3* and *Tkb5*, which are found to be present in multicopy within the *K. bovina* genome (14 - 24 copies). High values of Fop (>0.65) were also found for the TE families of *K. viticola*, which are also found in multicopy (14 - 30). The observed poor relationship is likely to be due to the TE families of *K. exigua*, which although found in multicopy with the genome, were found to have lower Fop values (<0.58). Removing the *K. exigua* TE families from the dataset was found to significantly change the relationship observed between Fop and copy number, with an significant increase in R^2 value from R^2=0.060 to R^2=0.631.

Evidence has been found to support selection as a driver of codon usage in the host organisms, with translationally optimal codons found to abundant in TE families of the *Kazachstania* species. With codon usage appearing to conserved between host genes, and TE families, it is theorised that the employment of optimal codons within the mobile elements, would facilitate efficient transposition in the genome, compared to families which favour codons which do not complement major tRNA genes of the host species, therefore being beneficial to TE proliferation in the host genome. However, although evidence has been uncovered that points towards selection, it is of note that mutation bias can not be ruled out. Although the complementary major tRNA genes and the abundant codons in EF1A match the optimal codons defined by CodonW for each of the host species, the evidence for selection does not eliminate the potential for mutation bias, as no signatures against mutation bias have been uncovered for the yeast species. The work has introduced an intriguing picture of codon usage bias within the *Kazachstania* genus, and an area of study which should be developed upon, to determine if selection is a driver of codon usage in the yeast species, as evidence would point towards here.

2.4.6 Concluding remarks

The review of the novel *Kazachstania* species has provided an unprecedented insight into genome characteristics for the genus, as well as TE data and codon usage bias for each of the budding yeast species. Typically, the findings revealed showed similarity amongst other yeast species within the Saccharomycetaceae superfamily, and conservation was evident for the majority of *Kazachstania* species regarding genome size, TE content and codon usage bias. However, striking results were uncovered, including the increased genome size of *K. exigua*, and varied codon usage bias observed for the host genes of *K. naganishii*, which hopefully will inspire further investigation with an increase in genome availability for the genus, to see if the unexpected trends were apparent in any other species yet to be analysed. The comparative genomics analyses detailed here is the first review of the *Kazachstania* genus, including four novel yeast species, and has opened an exciting avenue for continued study.

Chapter 3

A genomic survey of transposable elements in the choanoflagellate *Salpingoeca rosetta*

Aspects of the work described in this chapter was published in Southworth, J., Grace, C.A., Marron, A.O, Fatima, N. and Carr. M (2019). 'A genomic survey of transposable elements in the choanoflagellate *Salpingoeca rosetta* reveals selection on codon usage', *Mobile DNA* **10** (44), 1 – 19.

3.1 Introduction

Transposable elements have predominantly been investigated in multicellular organisms from major kingdoms such as plants, fungal species and metazoans, with very little research into unicellular eukaryotic organisms. This niche has been targeted to a degree with studies investigating TE evolution within species across known eukaryotic diversity, including opisthokonts, amoebozoans, alveolates and excavates (Silva et al., 2005; Carr, Nelson, Leadbeater and Baldauf, 2008; Elliott and Gregory, 2015).

Choanoflagellates and their TEs

Opithoskonta is the eukaryotic supergroup that includes Metazoa, Fungi, as well as unicellular groups nucleariod amoebae, Ichythosporea, Filasteria and Choanoflagellatea (Adl et al., 2018). Within Holozoa, choanoflagellates are the closest living relative to the metazoans, and give great insight into the origin of animals (Dayel et al., 2011). Found in marine and freshwater habitats, as well as hydrated soils, choanoflagellates are predominantly unicellular, but many species develop ephemeral multicellular colonies (Dayel et al., 2011; Dayel and King, 2014).

M. brevicollis was the first unicellular holozoan to have its genome sequenced (King et al., 2008), allowing a study of its TEs. Only three families were identified, all of which were LTR retrotransposons; *Mbcv* (*Monosiga brevicollis chromovirus*), *Mbpv1* (*Monosiga brevicollis pseudovirus -1*) and *Mbpv2* (*Monosiga brevicollis pseudovirus-2*) (Carr, Nelson, Leadbeater and Baldauf, 2008). The same study also screened available expressed sequence tags (EST) sequences from a second choanoflagellate, *Mylnosiga fluctuans* (incorrectly deposited in ATCC as *Monosiga ovata* (Carr et al., 2017), uncovering LTR and non-LTR retrotransposons in addition DNA transposons, suggesting that *M. brevicollis* may be unusual in having a limited diversity of TE families (Carr, Nelson, Leadbeater and Baldauf, 2008). From this, the evolution of TEs in opisthokont protists was further investigated by the annotation of TEs in the filasteran protist, *Capsaspora owczarzaki* (Carr and Suga, 2014). Screening of the draft genome uncovered a total of 23 families from both classes of TEs - a far greater repertoire of elements when compared to *M. brevicollis* (Carr, Nelson, Leadbeater and Baldauf, 2008; Carr and Suga, 2014). With this, the *C. owczarzaki* TE families identified were shown to have orthologues in other opisthokonts, predominantly in Metazoa and Fungi.

The draft genome of *Salpingoeca rosetta* was released in 2013, and sequenced by Broad Institute (Fairclough et al., 2010). Analysis of the *S. rosetta* genome for TEs, provides an ideal opportunity to determine if Carr, Nelson, Leadbeater and Baldauf (2008) were correct in suggesting that *M. brevicollis* may be atypical of choanoflagellates in having a limited diversity of TE families. Both *M. brevicollis* and *S. rosetta* fall into Clade 2 of Craspedida within the choanoflagellates (Nitsche et al., 2011); however they are not close relatives, with the *M. brevicollis* SSU gene showing 72.6% nucleotide identity to the *S. rosetta* orthologue (Carr et al., 2017).

The survey of *M. brevicollis* in Carr, Nelson, Leadbeater and Baldauf (2008) TEs could not determine their evolutionary origin, either through vertical or horizontal inheritance, due to the limited volume of whole genome sequences availability. The larger *C. owczarzaki* study (Carr and Suga, 2014) showed that all of the families present in this species appeared to have been vertically inherited during the opisthokont radiation, highlighting the long-term co-existence of TEs and their hosts within this lineage (Carr and Suga, 2014). A disparity between *M. brevicollis* and *C. owczarzaki* was the finding that all families in the former are active, whereas the latter contains families that are no longer functional (Carr, Nelson, Leadbeater and Baldauf, 2008; Carr and Suga, 2014).

3.1.1 Experiment overview

This project involved bioinformatic screening of the *S. rosetta* genome for TE content. A minimum of 22 TE families; including both retrotransposons and DNA transposons, were uncovered in the genome. Overall, seven families are part of the *Ty3/gypsy* family of LTR retrotransposons, five being placed in chromovirus clade and 2 were non-chromoviral *gypsy* elements. A further six families of LTR retrotransposons were uncovered from the *Ty1/copia* group. Seven families were classified as DNA transposons. Similarity searches uncovered two families of non-LTR retrotransposons, however due to poor sequencing coverage full-length consensus sequences could not be reconstructed.

3.2 Methods

3.2.1 Identification of TE families in the *S. rosetta* genome

The annotation of elements were produced by Dr. Martin Carr. Genomic supercontigs of *S. rosetta* were downloaded from the Origins of Multicellularity Project at the Broad Institute (https://www.broadinstitute.org/scientific-community/data/origins-multicellularity). Two methodologies were employed by Dr. Martin Carr for TE annotation. Firstly, the supercontigs were ran using Protein Based RepeatMasker server, available by the Institute for Systems Biology(http://www.repeatmasker.org/cgi-bin/RepeatProtein MaskRequest). Similiarity hits were deemed potential candidates, with an e value of <e-05, with hits found in LTR retrotransposons, Non-LTR retrotransposons and DNA transposons. Successful nucleotide hits were downloaded, and translated to amino acid sequences using ExPASy (Artimo et al., 2012). The proteins were then subjected to a reciprocal blast, using the BLAST protein database (BLASTp) (Altschul et al., 1990), to determine if the candidates were genuine TEs.

Secondly, TE query sequences and methodology were taken from Carr, Nelson, Leadbeater and Baldauf (2008), with the employment of translate nucleotide BLAST database (tBLASTn). The query sequences were constructed using Pol and Transposase amino acid sequences from a wide range of eukaryotic species, selected based on phylogentic diversity to ensure for broad sampling. Pol and Transposase were selected as query sequences due to high conservation across eukaryotic taxa (Table 3.1).

The RepeatMasker and BLAST hits failed to recover full-length TE sequences (Altschul et al., 1990). Carr, Nelson, Leadbeater and Baldauf (2008) employed a method to increase consensus sequence coverage using overlapping sequencing reads from the NCBI Trace Archive with organism specified - "*Proterospongia* sp. ATCC 50818". The sequencing reads allowed the generation of full-length consensus sequences for all families, with the exception of two non-LTR retrotransposons.

Table 3.1: **Query sequences employed for BLAST search of the *Salpingoeca rosetta* genome.** The methology and query sequences were duplicated from Carr, Nelson, Leadbeater and Baldauf (2008).

Family	Query sequence (accession number)
Ty1/copia	*1731* (CAA30503) *Athila* (AAF19227), *Evelknievel* (AAC02669) *Melmoth* (Y12321) *SIRE-1* (AAC64917) *Tca5* (AAC24836) *Yokozuna* (BAA74713)
***DIRS1*- like families**	*DIRS1* (AAA33195)
Ty3/gypsy	*gypsy* (AAC82604) *mdg1* (AAD14015), *mdg3* (CAA65152) *Maggy* (L35053) *marY1* (AB028236) *opus* (Q8I7P9) *Osvaldo* (AAC60519) *skippy* (S60179) *Tirant* (AAX28844)
non-LTR retrotransposon	*Bari-1* (CAA47913) *Bilbo* (AAB92394) *Doc* (CAA35587) *F-element* (AAA28508) *L1* (P11369) *marY2N* (BAB32469) *R1* (CAA36227) *R2* (CAA36225) *R5* (AAP69990) *Rte-1* (AF054983) *TART* (CAD92793) *TvsL10* (AJ850265)
***Pao*-like family**	*roo* (AAN87269)
***Penelope*-like family**	*Penelope* (AAX11377)
Retrovirus	*PERV* (AAT77167)
DNA transposon	*flipper* (AAB63315) *Foldback* (CAA23501) *Harbinger/PIF* (ABB83644) *Hermes* (AAC37217) *Hobo* (P12258) *Hop* (AAP31248) *hupfer* (DQ074974) *Impala* (AAB33090) *IS1* (AP_004308) *IS600* (AAK18596) *mariner* (DQ197023) *maT* (CAD31217) *P-element* (CAA43305) *piggyBac* (ABC88680) *pogo* (Q80TC5) *pokey* (AAM76341) *punt* (AF181822) *restless* (AAK16925) *S-element* (AAC47095) *tigger* (NP_997161) *Tn3* (YP_665994)

3.2.2 Phylogenetic analyses

Superfamily phylogenies were created for all of the *S. rosetta* families by using amino acid query sequences of Transposase for DNA transposons, and Pol for LTR retrotransposons. The query sequences for each novel family are listed in the Appendix. Sequence similarity searches of whole genome shotgun-contigs (wgs) were performed using translated nucleotide BLAST datbase (tBLASTn), and non-redundant protein sequences (nr/nt) database for BLASTp on National Centre for Biological Information (NCBI) (Altschul et al., 1990), to identify closely related TE families in diverse taxonomic group of eukaroytic species (Table 3.2).

Table 3.2: **Taxonomic groups employed for BLAST searches with the query sequences of the *Salpingoeca rosetta* genome.**

Kingdom	Superphylum
Metazoa	*Deuterostomia* *Gnathostomulida* *Platyhelminthes* *Protostomia* *Cnidaria* *Ctenophora* *Mesozoa* *Placozoa* *Porifera*
Fungi	*Blastocladiomycota* *Chytridiomycota* *Cryptomycota* *Ascomycota* *Basidiomycota* *Entomophthoromycota* *Glomeromycota* *Microsporida* *Neocallimastigomycota*
Protistan groups	*Alveolata* *Amoebozoa* *Apusozoa* *Breviatea* *Centroheliozoa* *Cryptophyta* *Rhodophyta* *Stramenopiles*
Plant	*Chlorophyta* *Mesostigmata*
Excavata	*Euglenazoa* *Parabasalida*
Rhizaria	*Cercozoa*

Nucleotide sequences of high identity and similarity were translated using EMBOSS Transeq on EMBI EBI (Li et al., 2015). Amino acid sequences were aligned using Multiple Aligment Fast Fourier Transform (MAFFT) on EMBL-EBI server with default parameters (Katoh et al., 2002). Amendments were made to the alignments by eye to reduce conserved indel regions. Problematic and unconserved regions were removed from alignments. Bayesian inference support values were constructed using a mixed amino acid model via MrBayes 3.2.6 on XSEDE (Ronquist and Huelsenbeck, 2003; Ronquist et al., 2012). The analyses consisted of 500000 generations, a sampling frequency of 1000, with a burnin value of 1250. Mr Bayes was accessed via CIPRES Science Gateway, a server based platform that allows for phylogenetic analyses (Miller et al., 2010). Maximum likelihood phylogenies were ran using raxmlGUI 1.5 beta (Silvestro and Michalak, 2011). The ML and thorough bootstrap analysis were performed with 1000 replicates and 100 runs using the PROTCAT model for amino acid sequences. The ML amino acid substitution model used for each family was determined from the output of the mixed model analysis from MrBayes (Figure 3.1).

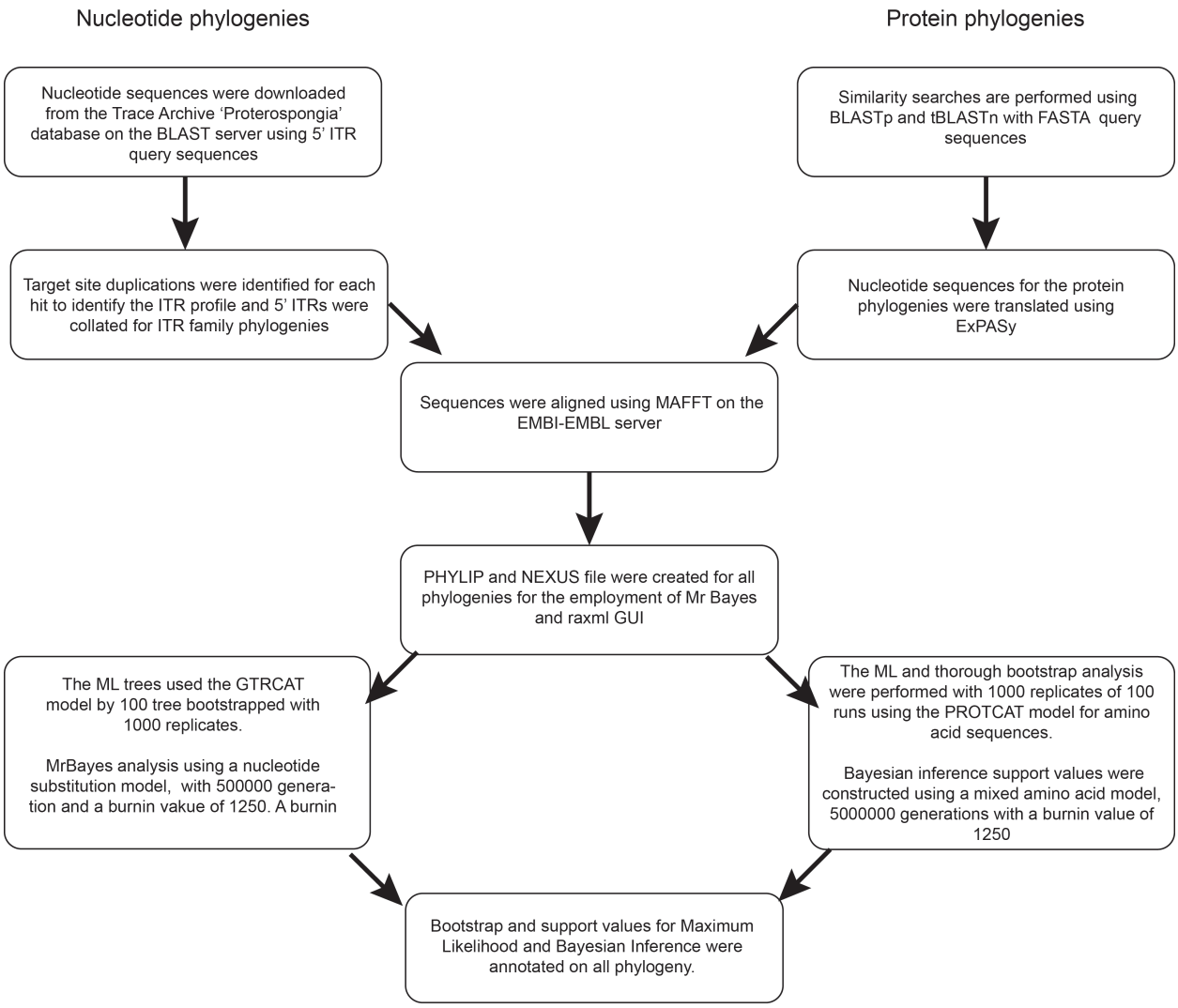


Figure 3.1: **Schematic diagram to represent the bioinformatic protocol employed for the construction of nucleotide and protein phylogenies in the genome of *S. rosetta*.**

LTR and ITR phylogenies for each TE family were generated using *S. rosetta* 5' query sequences to construct phylogenies of BLAST hits with high similarity to the consensus sequences. Due to the short length of 5' ITRs, the 5' UTR was also included in phylogeny construction. Terminal sequences were downloaded from NCBI Trace Archive against '*Proterospongia*' with a threshold of e-05 (Figure 3.1). The TSD could be identified for each individual insert to distinguish the LTR/ITR profile. If identical TSDs from the Trace Archive were present, the two termini of the same element could be identified. Partially sequenced LTRs/ITRs were excluded from further analyses, as the element status could not be determined. Copy number estimates are therefore likely to be lower than the actual copy number for each family. ITR phylogenies could not be created for *SrosT2* and *SrosT3* due to low copy number, and *SrosH* as the element lacked ITRs.

All trees used Maximum Likelihood (ML) with the employment of raxmlGUI and Bayesian Inference via MrBayes. The ML trees were generated using from 100 starting parsimony trees, using the GTRCAT model and supported with 1000 bootstrap replicates. MrBayes analysis was employed using the protocol stated for the amino acid datasets, although the GTR+I+Γ nucleotide substitution model was used. Support value analyses were based on thresholds outlined in Table 3.3 documented by Hillis and Bull (1993) and Rannala and Yang (1996).

Table 3.3: **Support value thresholds for Maximum Likelihood (ML) and Bayesian Inference Posterior Probabilities (biPP) phylogenies**. The level of support is annotated with both tree methodologies respectively.Values were proposed in Rannala and Yang (1996); Hillis and Bull (1993).

	Maximum Likelihood %(ML)	**Bayesian Inference Posterior Probabilities (biPP)**
High Support	≥70	≥0.97
Moderate Support	50 - 69	0.70 - 0.96
Low Support	<50	<0.70

3.2.3 TSD preference patterns and nucleotide diversity

TSD nucleotide preference patterns were reviewed using WebLogo version 2.8 via University of Berkeley, California server (Crooks et al., 2004). TSDs for each TE family were uploaded, and default parameters set. The Y axis height was varied per family depending on the output file and nucleotide value per TSD position. A graphical representation of nucleotide conservation was exported for each TE family, and conserved bases analysed to observe trends between superfamilies and classes.

Levels of nucleotide diversity for each TE family were calculated, using values of π (Nei and Kumar, 2000), with DnaSP version 5. 10. 01 (Rozas et al., 2003). Values of π were calculated for DNA transposon families by the analysis of 5' ITR alignments. Values of π for LTR retrotransposon families were determined with individual alignments containing solo elements, individual full length elements (FLE) and combined FLE, partial and solo elements to produce ALL, FLE and SOLO phylip input files.

3.2.4 Determining TE family expression levels

Raw Illumina RNASeq transcriptome reads (SRA files SRX042046-SRX042054) were downloaded from NCBI and mapped to the TE family sequences through SMALT v. 0.2.6 (Ponstingl, 2014). The total number of reads for each family was calculated from the SMALT output SAM files in Tablet v.1.17.08.06 (Milne et al., 2013).

3.3 Results

3.3.1 *S. rosetta* harbours a higher diversity of TE families than *M. brevicollis*

The genomic survey of *M. brevicollis* by Carr, Nelson, Leadbeater and Baldauf (2008) identified three families of LTR retrotransposon, but a lack of non-LTR retroelements and DNA transposons (Carr, Nelson, Leadbeater and Baldauf, 2008). In contrast, through both RepeatMasker and BLAST similarity searches, the *S. rosetta* genome (ATCC 50818) was found to have a minimum of 20 TE families. Both methodologies described identified the same families within the choanoflagellate genome, supporting the validity of the identification. The *S. rosetta* TEs were classified in to 10 superfamilies (Table 3.4 and 3.5). The LTR retrotransposons were named *Salpingoeca rosetta chromovirus-1* to *Salpingoeca rosetta-5* (chromoviruses), *Salpingoeca rosetta gypsy-like element-1* and *Salpingoeca rosetta gypsy-like element-2* (non-chromoviral gypsy-like families) and *Salpingoeca rosetta pseudovirus-1* to *Salpingoeca rosetta pseudovirus-6* (*copia-like* families). The DNA transposon families were *Salpingoeca rosetta MULE-like* element, *Salpingoeca rosetta Helitron*, *Salpingoeca rosetta Tigger-1* and *Salpingoeca rosetta Tigger-2*, as well as three uncategorised transposon elements. Partial pol sequences from two putative families of non-LTR retrotransposon were also uncovered, however the complete full-length sequences could not be reconstructed from Trace Archive sequencing reads due to poor coverage. As the two families were unable to be fully sequenced, they have not been considered in the remainder of the project.

Table 3.4: **Characterisation of the seven identified families of DNA transposons in the genome of *Salpingoeca rosetta*.**

Family	Length	ITR Size	TSD Length	Copy Number (5' ITR/3' ITR)	No. of RNASeq Reads	ITR Nucleotide Diversity (π)
SrosH	3614	-	-	6-9 (6/3)	108,920	-
SrosM	8324	28	9	24 (14/16)	873,989	0.017
SrosT1	2071	28	-	8-14 (6/8)	9,715	0.083
SrosT2	3270	32	4	2 (2/2)	77,180	-
SrosT3	3112	27	-	1-2 (1/1)	140,210	-
SrosTig1	2122	22	2	7-11 (7/4)	174,979	0.03
SrosTig2	2165	23	2	4-7 (4/3)	28,176	0.050

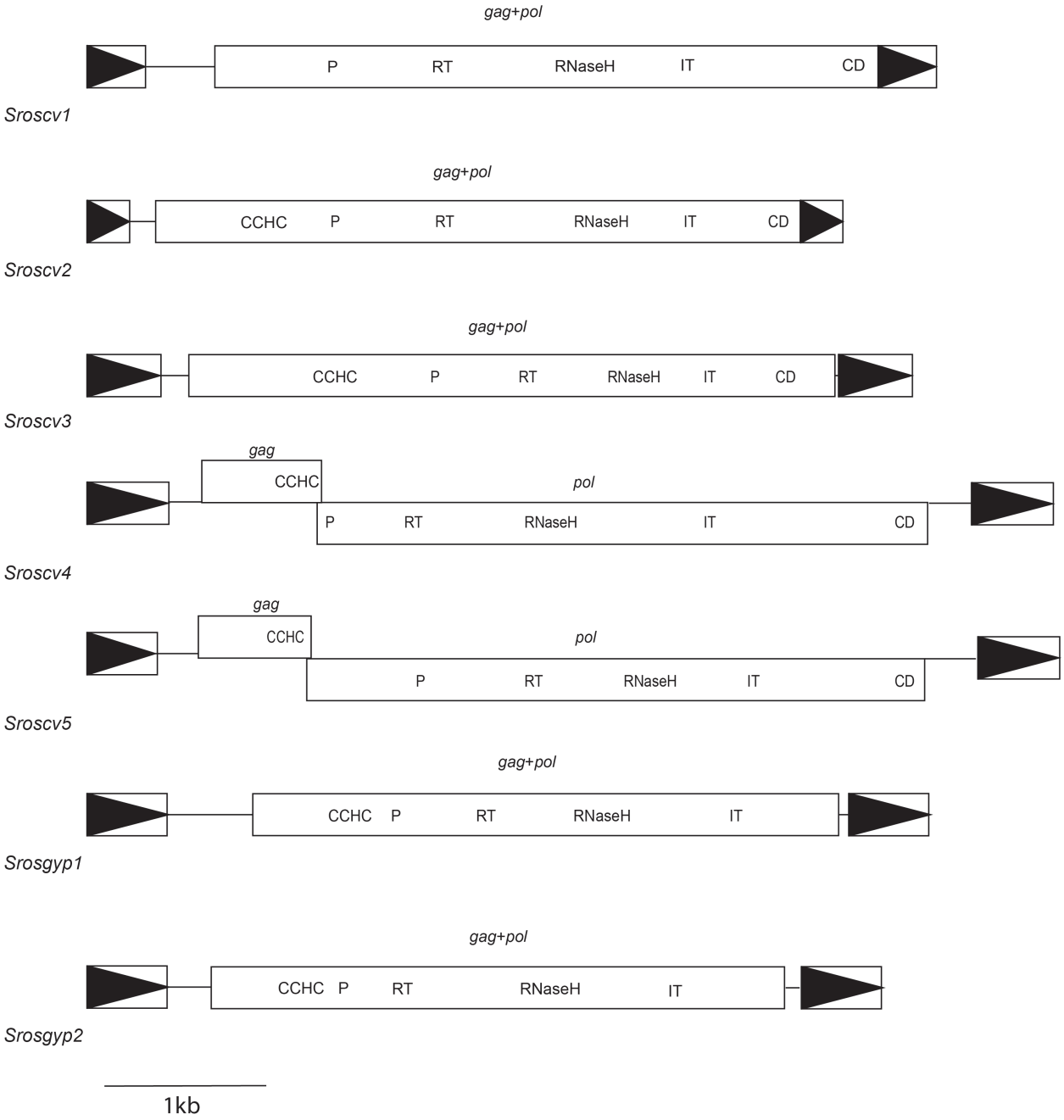
Copy numbers for full-length elements (FLE), solo LTR and truncated elements were calculated per LTR retrotransposon family, by identifying the number of unique target site duplication (TSD) sequences flanking element termini sequences (Table 3.4 and 3.5). Although *S. rosetta* possesses a far larger number of TE families than *M. brevicollis*, the copy numbers for each family are similar to those previously observed (Carr, Nelson, Leadbeater and Baldauf, 2008). Copy number ranged from single copy elements to >100 copies across all families identified (Table 3.4 and 3.5). The family with the highest copy number was found to be *Srospv3*, and the single copy families were *Srospv6* and *SrosT3*.

Table 3.5: **Characterisation of the thirteen identified families of LTR retrotransposon in the genome of *Salpingoeca rosetta*.**

Family	Length	LTR Size	Copy Number (FLE/Solo LTR/ Truncated)	No. of RNASeq Reads	No. of Identical Paralogous Copies	Intra-element LTR Identity (%)	LTR Nucleotide Diversity (π - Total/FLE/Solo)
Sroscv1	5290	243	18 (8/5/3)	923	6	99-100	0.020/0.006/0.048
Sroscv2	4813	190	10 (4/1/5)	36190	4	100	0.177/0.106/-
Sroscv3	5165	394	25 (10/10/5)	1617	11	99.24-100	0.045/0.006/0.051
Sroscv4	5806	421	15 (12/1/2)	44506	2	99.1	0.050/0.048/-
Sroscv5	6071	408	20 (8/3/9)	17425	4	99.51	0.092/0.037/0.075
Srosgyp1	5460	373	15 (6/4/5)	170	8	100	0.034/0.003/0.039
Srosgyp2	5170	385	9 (5/1/3)	390	6	100	0.020/0.000/-
Srospv1	4943	387	17 (12/0/5)	517	0	99.74 -100	0.015/0.015/-
Srospv2	5452	554	71 (40/4/27)	4369	5	99.30 - 100	0.047/0.043/0.077
Srospv3	5681	445	121 (59/14/48)	35094	22	87.0 - 100	0.098/0.068/0.066
Srospv4	5150	359	4 (1/3/0)	574	2	99.17	0.054/0.000/0.002
Srospv5	5086	362	3 (3/0/0)	1679	2	100	0.006/0.000/-
Srospv6	5948	168	1 (1/0/0)	57997	0	100	-

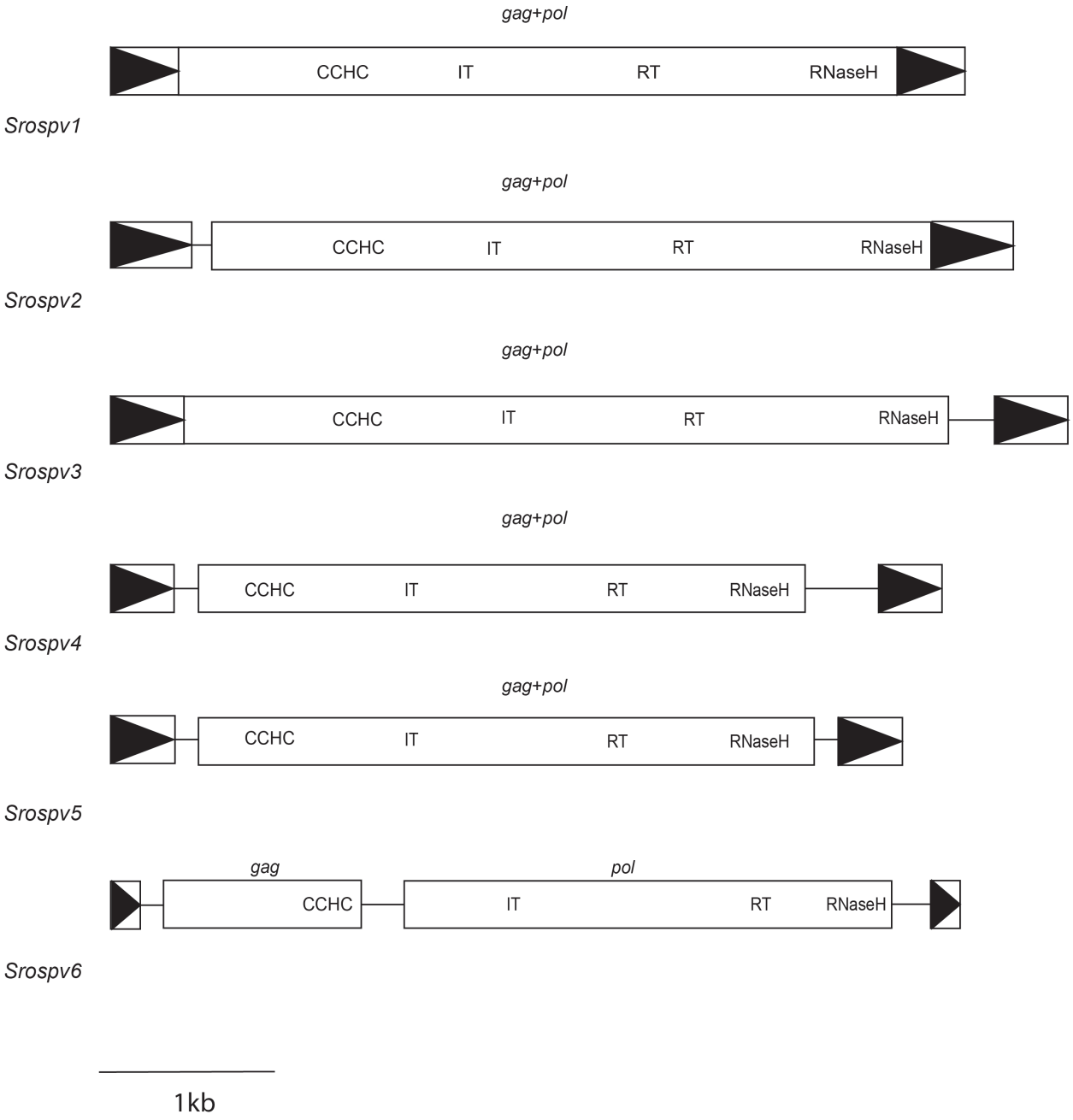
The predicted full-length *gypsy-like* elements ranged from 4.8-6.1 kb in length (Figure 3.2). *Sroscv-1,2,3* and *Srosgyp-1,2* encode the *gag* and *pol* open reading frames in the same frame, whereas *Sroscv-4,5* encodes *gag* and *pol* in separate frames. Similarly, the *copia-like* elements all encoded gag and pol within the same reading frame, with the exception of *Srospv6*, which presents two separate reading frames (Figure 3.3). All *copia*-like families had a similar length between 4.9-5.9 kb. The length of the LTRs for each retrotransposon family varied considerably with a range from 168bp - 554bp, with the greatest diversity documented in the *pseudovirus* superfamily; *Srospv6* and *Srospv2* (Table 3.5). In *M. brevicollis*, LTR size ranged from 265-668bp (Carr, Nelson, Leadbeater and Baldauf, 2008). *Mbcv* possessed the greatest LTR size of the three LTR retrotransposon families uncovered in *M. brevicollis*, however the same pattern was not found here, with the pseudoviruses having the largest mean LTR size (Table 3.5). The genomic organisation of LTR retrotransposons and enzymatic domains are detailed in Figure 3.2 and 3.3.

In contrast, the DNA transposon elements varied considerably in length, with a range from 2.1-8.3 kb in length (Figure 3.4). The DNA transposon families each possessed a single ORF, which encoded a putative Transposase protein. Five of the seven ORFs harboured introns, with only *SrosT1* and *SrosTig1* not possessing introns (Figure 3.4). ITR size was similar for all class II elements, ranging from 22-32bp in length (Table 3.4). The genomic organisation of DNA transposons and enzymatic domains are detailed in Figure 3.4.



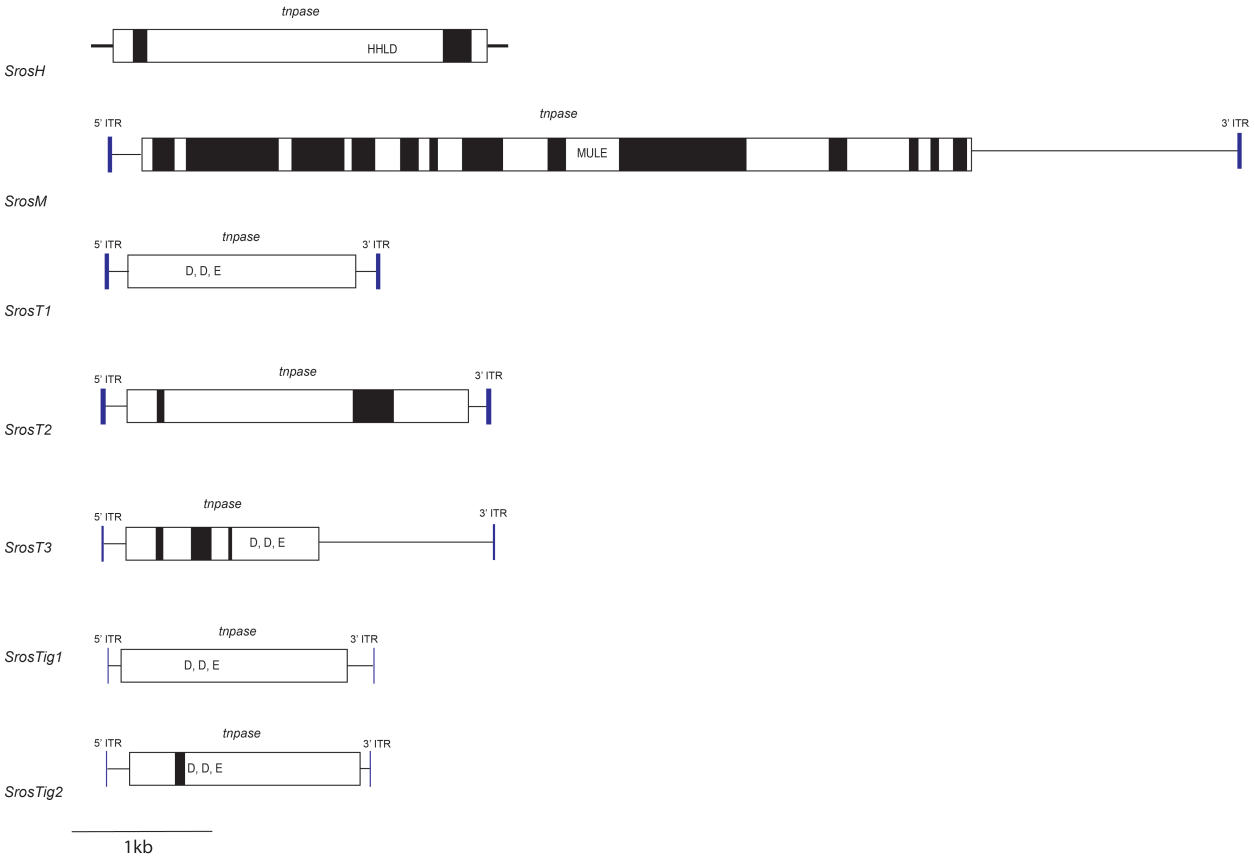
a) *gypsy-like* families

Figure 3.2: **Genomic organisation of the 7 *gypsy*-like families characterised in the *S. rosetta* genome.** *gypsy*- like LTR retrotransposons: boxes with black arrowheads represent long terminal repeat sequences, horizontal boxes represent gag and pol open-reading frames (ORFs). Protein coding domains are indicated as follows: CCHC, RNA binding motif; CD, chromodomain; IT, integrase; P, protease; RT, reverse transcriptase.



b) *copia-like* families

Figure 3.3: **Genomic organisation of the 6 *copia*-like families characterised in the *S. rosetta* genome**. *copia-like* LTR retrotransposons: The format follows that of Figure 3.2.



c) Transposon families

Figure 3.4: **Genomic organisation of the 7 DNA transposon families characterised in the *S. rosetta* genome.** DNA transposons: blue boxes represent inverted terminal repeat sequences, white boxes represent tnpase exon sequences, and black boxes represent tnpase intron sequences. Protein coding domains are indicated as follows: D,D,E, aspartic acid and glutamic acid catalytic domain; HHLD, helix-turn-helix like domain; MULE, Mutator-like element transposase domain. Non-coding regions are indicated as follows: ITR, inverted terminal repeat.

3.3.2 Transposable element genome content in *S. rosetta* and *M. brevicollis*

Both choanoflagellates *S. rosetta* and *M. brevicollis* whole genome contigs were both screened for TEs using RepeatMasker and BLAST searches to uncover novel families (*RepeatMasker*, 1996; Sayers et al., 2009). *M. brevicollis* showed to possess 0.57% of TE content, whereas *S. rosetta* had a TE genome content of 18.50%. *M. brevicollis* has previously been documented to have a TE genome content of 1%, which appears to be an over representation of TE content based upon cruder methodology, if the genome assembly was accurate (Carr, Nelson, Leadbeater and Baldauf, 2008). Although *S. rosetta* was found to have a far greater TE family diversity than *M. brevicollis*, similarity was drawn between copy number for the families uncovered in both choanoflagellate

species (Table 3.4 and 3.5). The majority of families were found to have less than 30 copies, with the exception of two *copia-like* families (*Srospv2* and *Srospv3*). Carr and Suga (2014) found that the families uncovered in *Capsaspora owczarzaki* also presented with low copy number, which supported that a small number of copies for each TE family could be an expected feature of unicellular protists.

3.3.3 Phylogenetic analyses of *S. rosetta* TE Families

Protein phylogenies were created for the all TE superfamilies present in the *S. rosetta* genome (Figure 3.6, 3.7 3.8, 3.9, 3.10 and Appendix F) with the use of conserved ORFs from each TE family; Transposase for DNA transposons and Pol for LTR retrotransposons. The backbone for all LTR retrotransposon phylogenies ranged in resolution, however, some branches were positioned with high support values of 98-100% maximum likelihood (ML) and 1.00 bayesian inference posterior probabilities (biPP) (Figure 3.6, 3.8, 3.7 and Appendix F). Both the *pseudovirus* and non-chromoviral *gypsy* phyogenies clustered the *S. rosetta* families together with high support (96-100 ML/1.00 biPP) (Figure 3.7 and 3.8).

The chromoviral phylogeny positioned the *Sroscv* sequences in to one clade; *Sroscv1-3* cluster together with strong support (95% ML/ 1.0 biPP), however *Sroscv4-5* were placed as sister groups to the main *S. rosetta* cluster with weak support (Figure 3.6). *Monosiga brevicolls chromovirus* (*Mbcv*) is placed in the clade with low/moderate support (<50% ML/ 0.78 biPP), which rejected the monophyly of the *S. rosetta* families, although not highly supported. However, the clustering of families within choanoflagelatte species supported that the families have an ancient history within the choanoflagellate lineage. Furthermore, no strong support is seen between the choanoflagellates and the main opisthokont group of chromoviruses (Figure 3.6). The chromoviral distribution across taxonomic groups was expected, based on documented chromoviral phylogenies (Figure 3.5) (*Gypsy Database 2.0*, 2010; Llorens et al., 2010). The documented inferred chromoviral relationships and abundance is predominantly in plant and fungal species, with few metazoans and one amoebozoan representative (Llorens et al., 2010). A similar pattern was observed in the chromoviral phylogeny, with the majority of elements from plant and fungal host species, and the minority found in vertebrates. As seen in Carr and Suga (2014), the metazoan and fungal Pols cluster together with 97%ML/ 0.99biPP, but the nested relationship had no support. The tree is consistent with the metazoan and fungal Pols being sister groups (Figure 3.6).

As with the chromoviral phylogeny, the non-chromoviral gypsy-like phylogeny was not robustly resolved, however, the choanoflagellate species were found to be monophyletic (Figure 3.7). A similar pattern was seen in the *Srosgyp* phylogeny, with the majority of elements uncovered in metazoan species. The *gypsy* sequences in *S. rosetta* were placed nested within the metazoan sequences, however with low support (<50%/<0.70biPP). *Gypsy-like* sequences were uncovered in several metazoan taxonomic groups, including arthropods, fish, insects and reptiles (Figure 3.8).

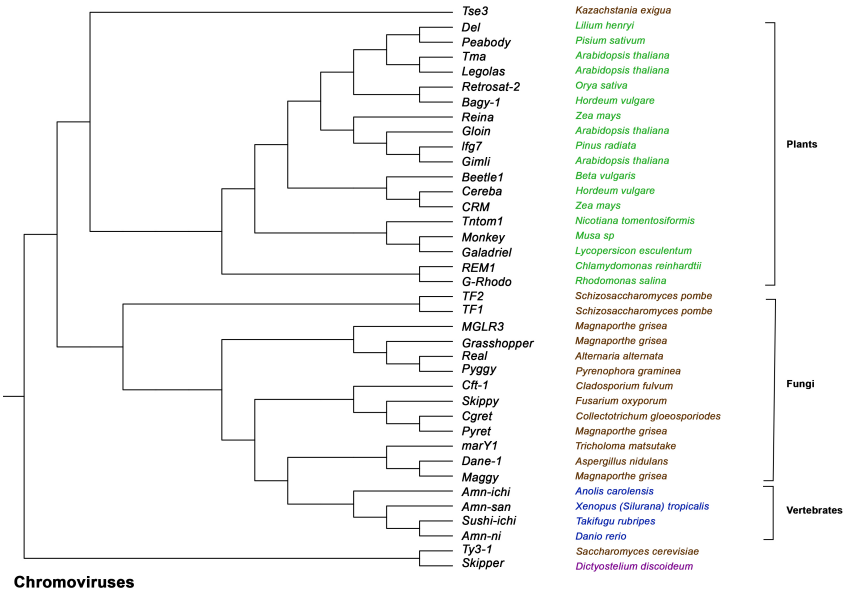


Figure 3.5: **Phylogenetic reconstruction of chromoviral elements among taxonomic groups**. The inferred relationships are based on integrase domain, from *Gypsy Database 2.0* (2010); Llorens et al. (2010). The cladogram was produced in Newick format.

In contrast, the *copia-like* phylogeny was unrooted, clustering all *copia-like* elements from both choanoflagellate species in one clade with maximum support (100% ML/ 1.00 biPP). The grouping provided support that the superfamily may have evolutionary history within the choanoflagellate lineage long term. However, choanoflagellate Pol sequences were found to be positioned separate to the other elements of Opisthokont species (Figure 3.8). Within the choanoflagellate clade, the majority of *S. rosetta* pol sequences cluster together with medium/high support (68%ML/0.98biPP). However, monophyly was not apparent for the *S. rosetta* famiies, as *Srospv3* was found to cluster with the *M. brevicollis* families, although with low support (<50%ML/0.86biPP). The two *M. brevicollis* families were found to form a strongly supported group with maximum support (100%ML/1.00 biPP). The main opisthokont group also included three Pol sequences from the stramenopiles; *Phytophora infestans*, *Klebsormidium nitens* and *Nannochloropis gaditana*. A higher number of *copia-like* elements were uncovered in metazoan species, when compared to the loss seen in

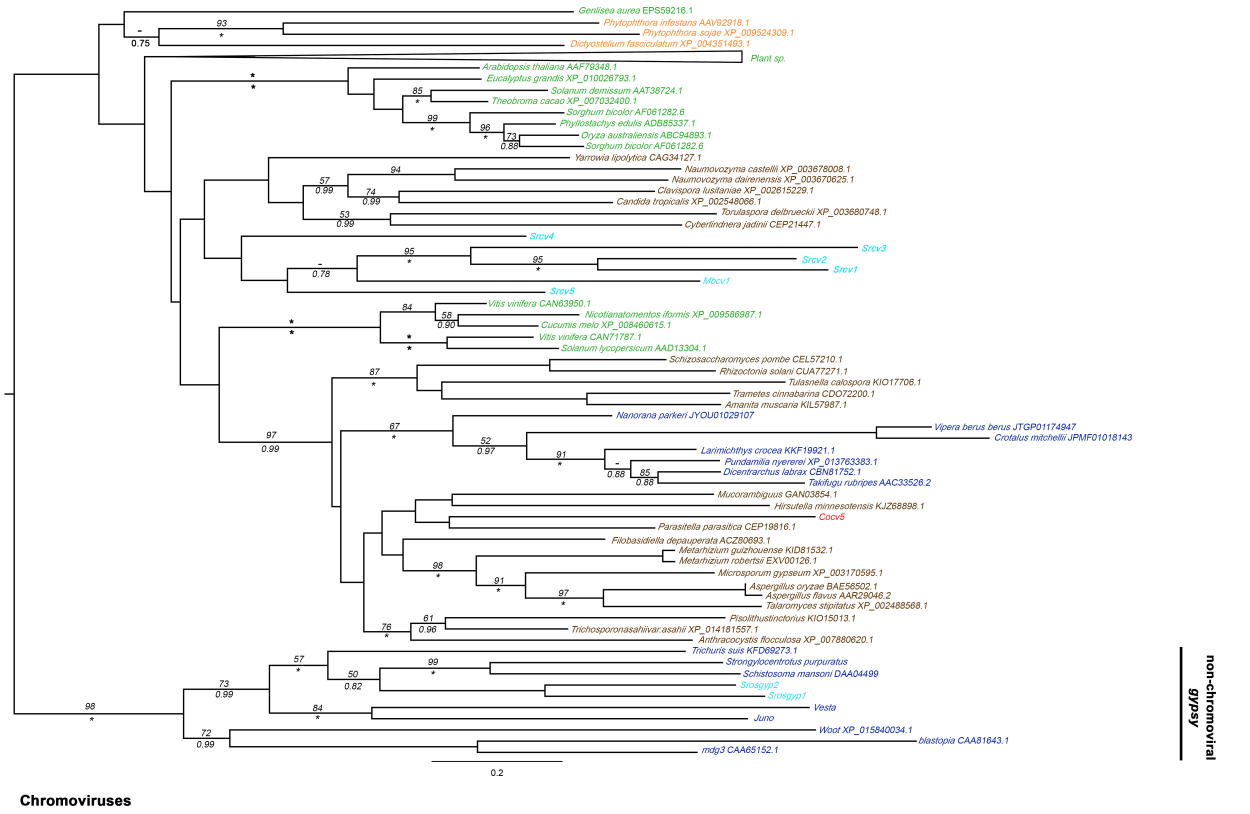


Figure 3.6: **Maximum Likelihood phylogeny of chromoviral amino acid sequences**. The phylogeny was constructed by an alignment of 406 amino acid constructs with the employment of raxmlGUI using the PROTCAT model and estimated amino acid frequencies with RTREV substitution matrix. ML and biPP values are labelled above and below corresponding branches. Maximum support is annotated by '*' (100ML/1.0 biPP) and low support (<50 ML/ <0.70 biPP) are annotated '-'. The scale bar signifies the number of amino acid substitution per amino acid site. Metazoan proteins are in dark blue, choanoflagellate proteins are in light blue, fungal sequences are written in brown, plants are in green, stramenopiles are in orange and Capsaspora in red. The outgroup is constructed of non-chromoviral *gypsy* elements from previous literature (Carr et al., 2008b), and *S. rosetta gypsy-like* elements uncovered in the RepeatMasker analysis.

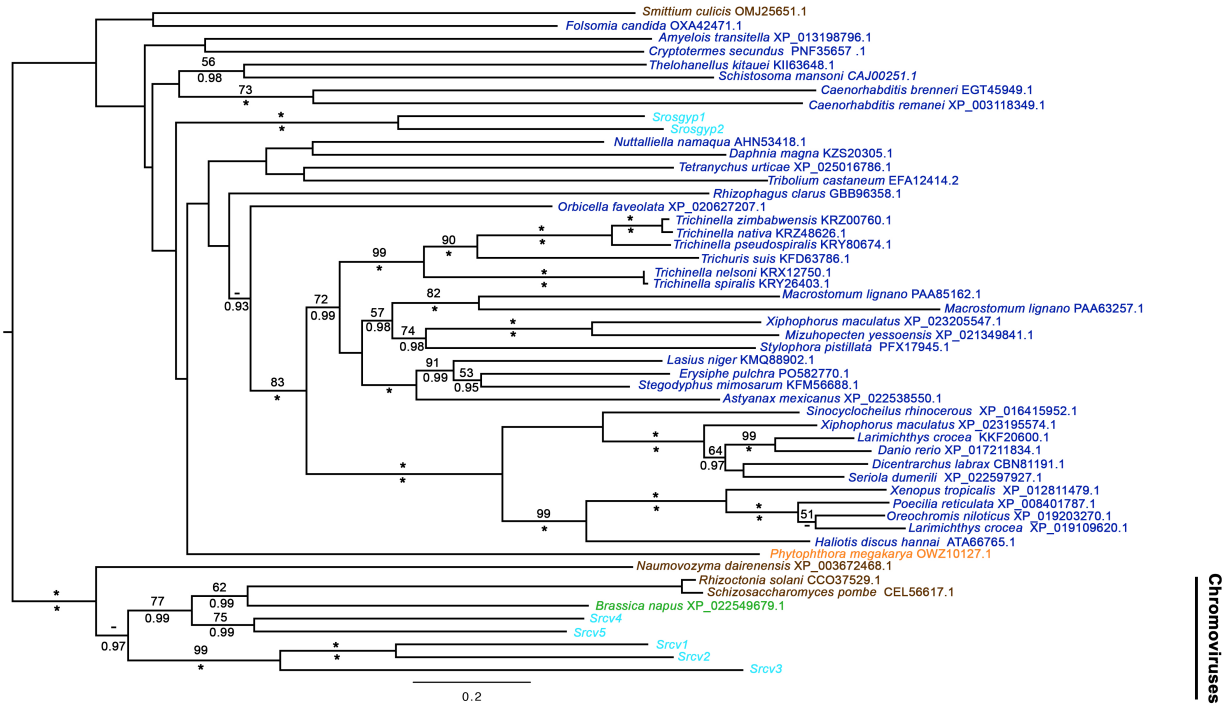


Figure 3.7: **Maximum Likelihood phylogeny of non-chromoviral *gypsy* amino acid sequences**. The phylogeny was constructed by an alignment of 743 amino acid constructs with the employment of raxmlGUI using the the PROTCAT model and estimated amino acid frequencies with the RTREV substitution matrix. ML and biPP values are labelled above and below corresponding branches. Maximum support is annotated by '*' (100ML/1.0 biPP) and low support (<50 ML/ <0.70 biPP) are annotated '-'. The scale bar signifies the number of amino acid substitution per amino acid site. Format is described in Figure 3.6. The outgroup is constructed of *S. rosetta* chromoviral elements uncovered in the RepeatMasker analysis.

chromoviruses (Figure 3.6 and 3.8).

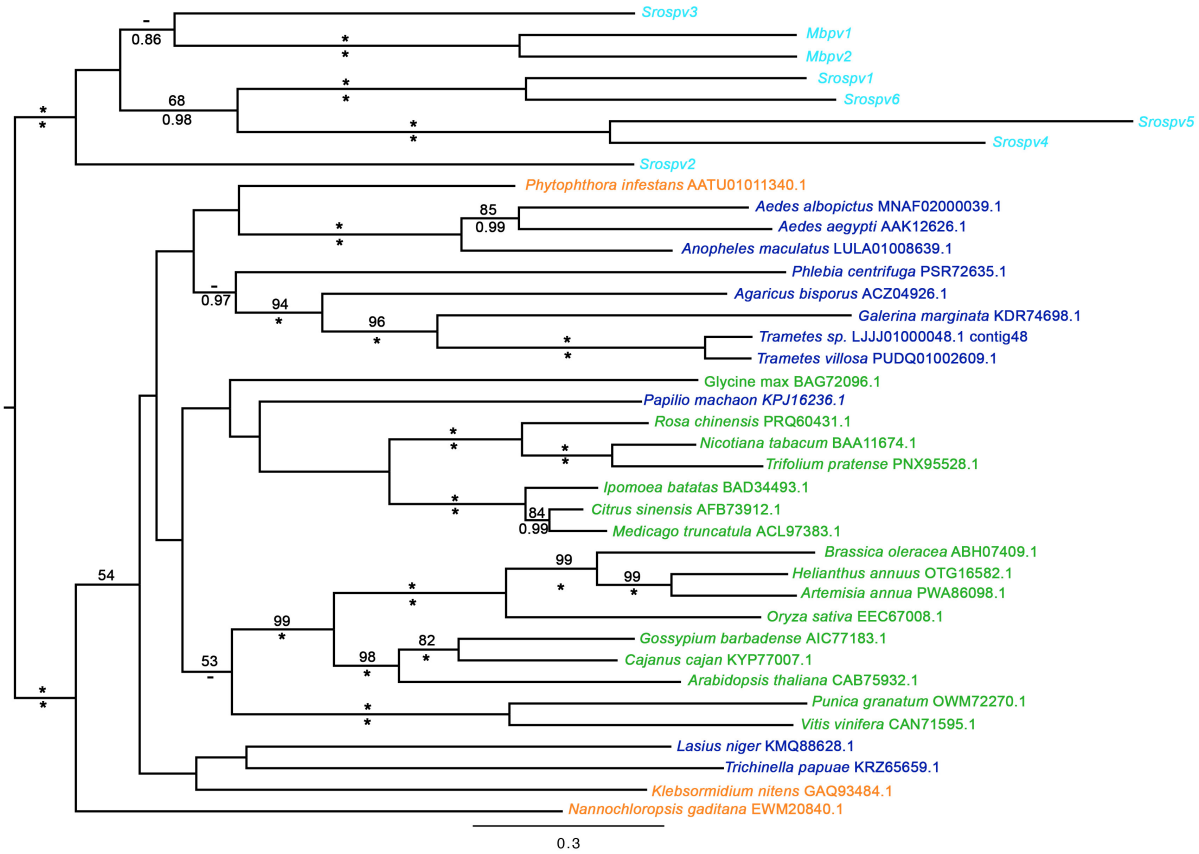


Figure 3.8: **Maximum Likelihood phylogeny of *copia-like* amino acid sequences.** The phylogeny was constructed by an alignment of 711 amino acid constructs with the employment of raxmlGUI using the PROTCAT model and estimated amino acid frequencies with RTREV substitution matrix. ML and biPP values are labelled above and below corresponding branches. Maximum support is annotated by '*' (100ML/1.0 biPP) and low support (<50 ML/ <0.70 biPP) are annotated '-'. The scale bar signifies the number of amino acid substitution per amino acid site. Format is described in Figure 3.6.

In contrast to the *gypsy-like* phylogeny, the phylogeny for *SrosT1* did not position the *S. rosetta* Transposase sequence with the other opisthokonts Transposase elements (Figure 3.9). Similarity searches with *SrosT1* query unveiled an abundance of stramenopile Transposase proteins, as well as diverse metazoan elements, and a minority found in plant species (Figure 3.9). The *SrosT1* element was found nested within the stramenopile proteins with high support (83%ML/0.99biPP). Furthermore, the BLAST searches with *SrosT1* found previously unannotated elements in the choanoflagellate species, *M. brevicollis*. The unicellular organism has previously been described to possess LTR retrotransposons only; *Mbcv1*, *Mbpv1* and *Mbpv2* (Carr, Nelson, Leadbeater and Baldauf, 2008). This review has since been challenged, with the unveiling of two additional predicted proteins found; *M. brevicollis Tigger-1* (*MbTig1*) and *M. brevicollis Transposon-1* (*MbT1*). All *M. brevicollis* predicted transposons are included in superfamily phylogenies and described in Section 3.4.8 (Figure 3.9 and Appendix F). *SrosT1* is grouped with a predicted transposon element found in the choanoflagellate, *M. brevicollis* (XP_001743358.1).

The nested grouping described is not consistent with the vertical inheritance of the transposon families since choanoflagellates and stramenopiles last shared a common ancestor. In contrast, the phylogeny indicates an ancient horizontal transfer of a stramenopile transposon into a common ancestor of both *M. brevicollis* and *S. rosetta*. As phagotrophs, choanoflagellates ingest unicellular prey and the acquisition of genes by choanoflagellates from a variety of prey species is already well documented (Tucker et al., 2015). The positioning seen here is consistent with horizontal transfer, from a stramenopile donor, to a choanoflagellate host.

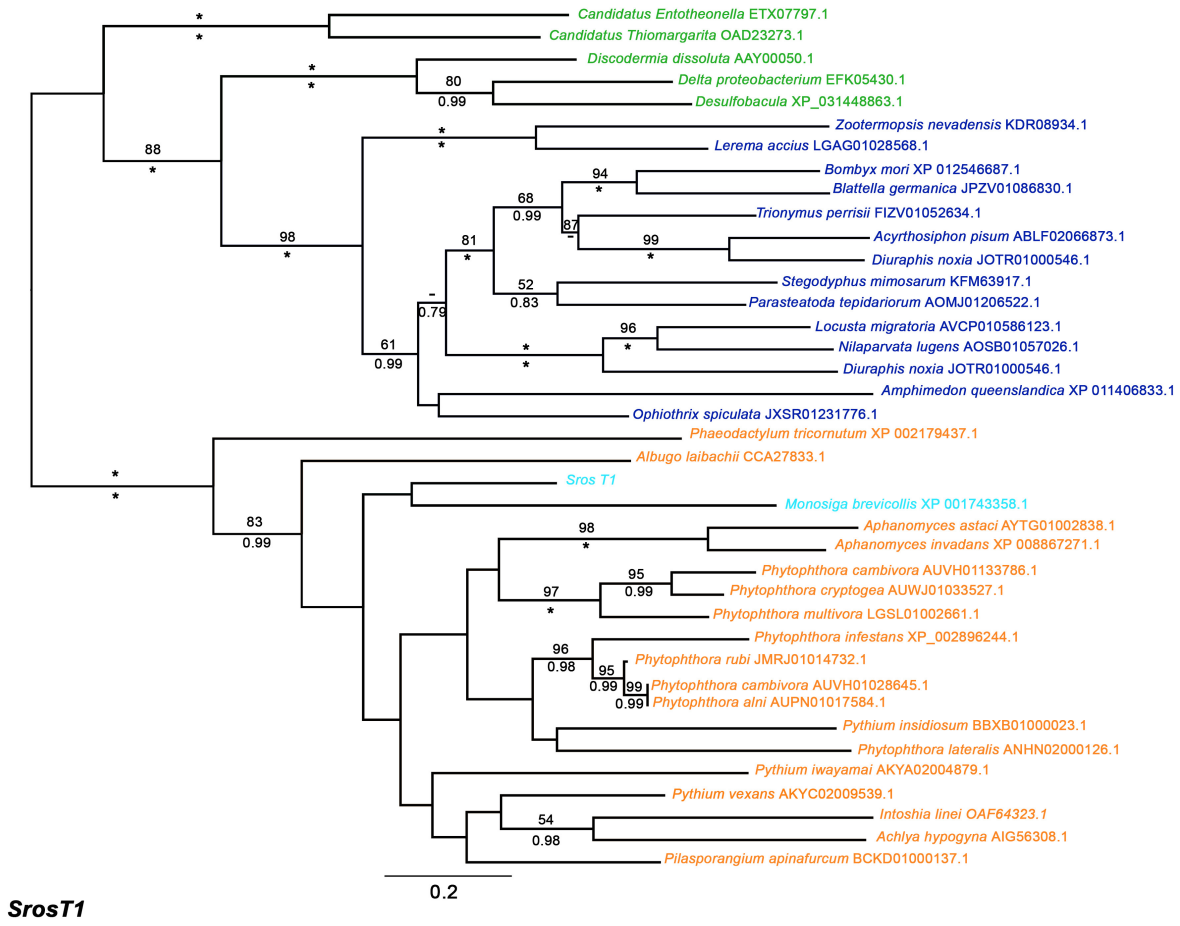


Figure 3.9: **Maximum Likelihood phylogeny of *T1* amino acid sequences across eukaryotic supergroups**. The phylogeny was constructed by an alignment of 253 amino acid constructs with the employment of raxmlGUI using the the PROTCAT model and estimated amino acid frequencies with the WAG substitution matrix. Format is described in Figure 3.6.

With the exception of *SrosT1*, the transposon family phylogenies were not robustly resolved (Figure 3.9, 3.10 and Appendix). The elements uncovered in the *MULE* similarity search uncovered full length tranposons in predominantly metazoan species. With this, a non-*MULE* outgroup was added to the phylogeny; the outgroup included *MuDR*, *Jittery* and *Hop* transposon families (Carr and Suga, 2014). *SrosMule* is placed with *Mule* tranposon in plant species, *Chlorella variabilis* (green algae), with moderate support (60%ML/0.71biPP), as a sister group to metazoan species, *Orechromis niloticus* (Nile Tilapia) and *Maylandia zebra* (Zebra Fish) with moderate support (58%ML/ 0.84biPP) (Figure 3.10). The positioning seen for the *Mule* phylogeny is likely to be poorly supported vertical inheritance, as the choanoflagellate transposases are nested with the metazoan species, which would not allow for a plausible explanation of HTT. Protein phylogenies for *Helitron*, *T2*, *T3*, *Tigger-1* and *Tigger-2* are shown in the Appendix F. The remaining phylogenies for the transposon families were inconclusive, with no clear explanation to the choanoflagellate positioning within the phylogenies. The Transposase proteins of *SrosH, SrosT2, SrosT3, SrosTig1* and *SrosTig2* were found to cluster with other opisthokont proteins with weak support (<0.50%ML/<0.70biPP). The Transposase of *SrosTig1* was found to cluster with stramenopile sequences as a sister group to the main opisthokont elements. However, the position was not well supported (<50%ML/<0.70biPP) and therefore it is unclear whether inheritance is vertical or horizontal.

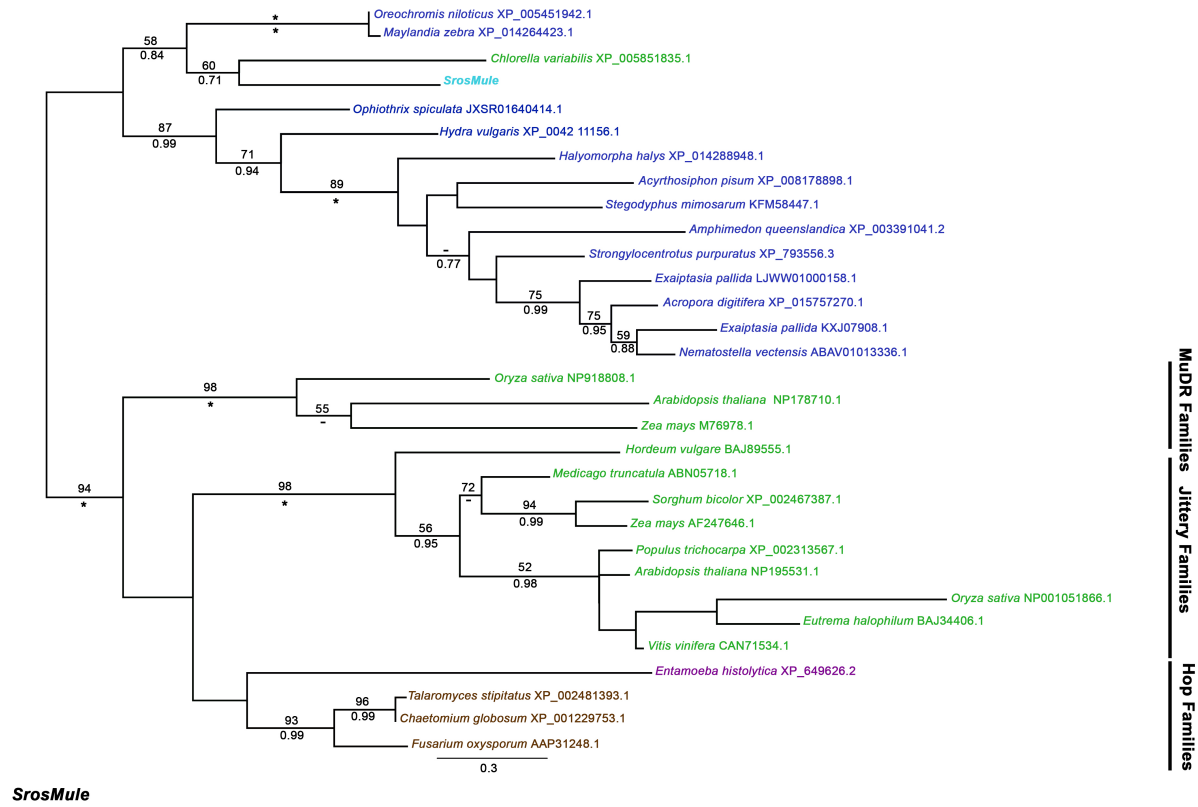
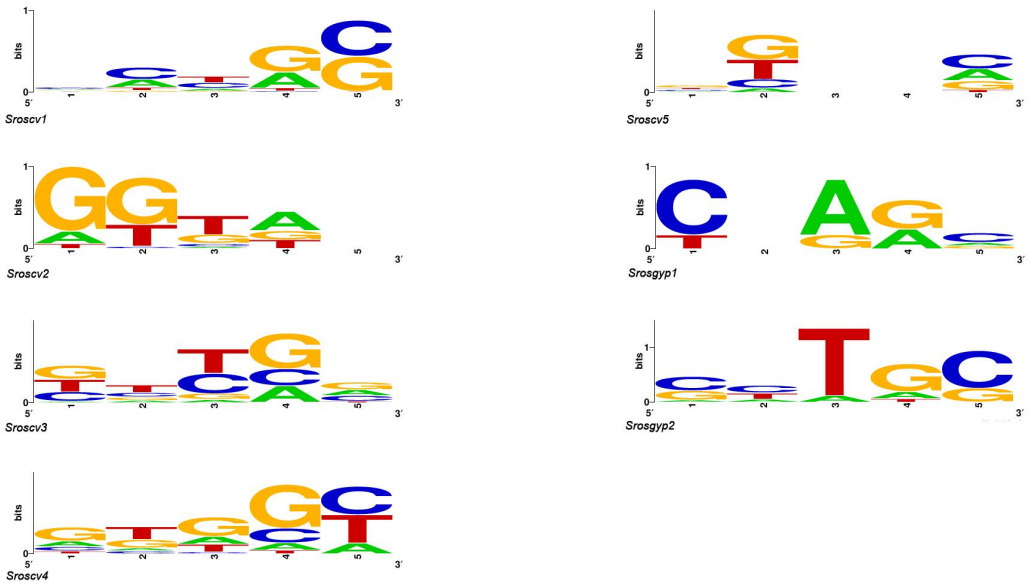


Figure 3.10: **Maximum Likelihood phylogeny of *Mule* transposase amino acid sequences across eukaryotic supergroup**. The phylogeny was constructed by an alignment of 92 amino acid constructs with the employment of raxmlGUI using the PROTCAT model and estimated amino acid frequencies with the BLOSUM substitution matrix. Format is described in Figure 3.6, except amoebozoa proteins which are written in purple.The outgroup is constructed of *MuDR*, *Jittery* and *Hop* transposon subfamilies of *Mule* from Carr and Suga (2014).

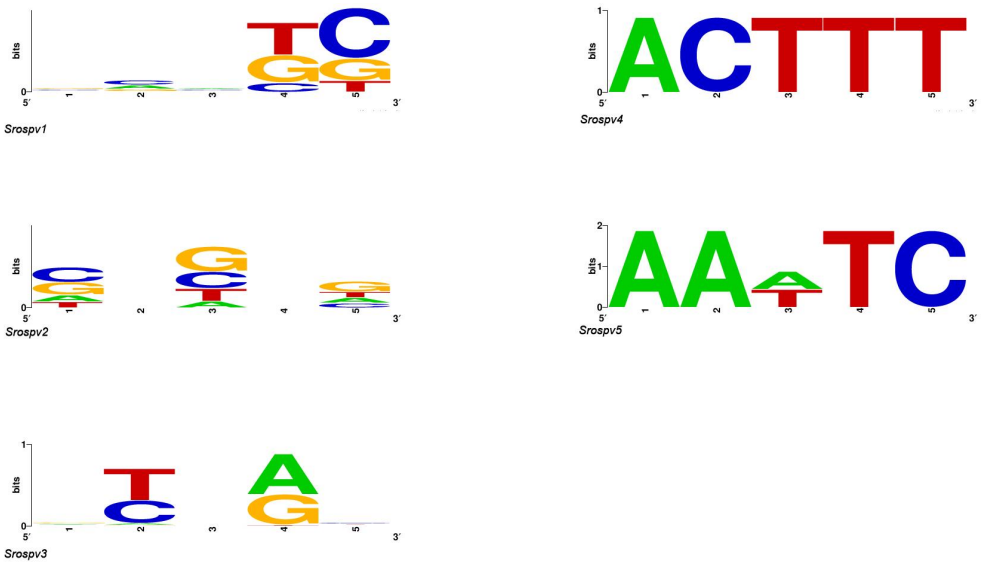
3.3.4 Target site insertion patterns of *S. rosetta* transposable elements

A moderate level of length and base conservation is seen in TE target site duplications across various species and genera (Lee and Harshey, 2003). This finding was also evident in the work with *S. rosetta* TEs for the length of the TSDs. Following insertion, the LTR retrotransposon families had target site duplications (TSDs) of 5 base pairs in length. The highest level of base conservation was observed for *gypsy-like* families with the majority of families favouring cytosine or guanine at the 5' and 3' termini of TSDs, however this is inconclusive and could still be described as random (Figure 3.11). Although high conservation in target site duplications was not seen for the *copia-like* families, a preference for guanine or cytosine (GC) at the 5' and 3' terminal positon of the TSD was still observed, except *Srospv4* and *Srospv5* which showed strong conservation of adenine at the 5' terminal end of the TSD, and *Srospv4* also showed preference to thymine at the 3' termini (Figure 3.12). However, it is of note, that *Srospv4* and *Srospv5* represent a very small percentage of overall genomic TE content, with only 3-4 copies identified per family. With this, the documented TSD are based on a smaller dataset, when compared to more active families that have proliferated in the genome (Table 3.5).



a) *gypsy-like* families

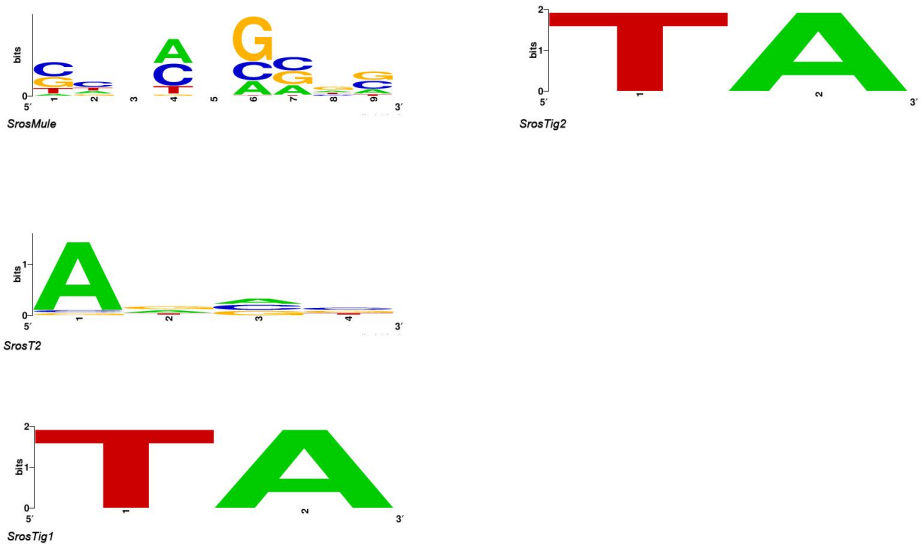
Figure 3.11: **Conserved base composition of target site motigs for *gypsy*-like families in the *S.rosetta* genome.** Conserved target site duplications for each transposable element family in the genome of *S. rosetta* were reviewed using WebLogo version 2.8 via University of Berkeley, California server (Crooks et al., 2004).



b) *copia-like* families

Figure 3.12: **Conserved base composition of target site motigs for *copia*-like families in the *S.rosetta* genome.** The methodology for TSD review is detailed in 3.11

In contrast, DNA transposon families had TSDs of varied length, ranging from 2 – 9 bp and showed higher conservation (Figure 3.13). All DNA transposon families, except *SrosMule*, preferred Adenine or Thymine at the 5' terminal end of the TSD, with the majority preferring Adenine at the 3' end (Figure3.13). The favoured base motifs of adenine and thymine is well documented target sites for DNA transposons (Geurts et al., 2006; Carr and Suga, 2014; Eide and Anderson, 1988). As seen in Carr and Suga (2014) for the *Mule* families uncovered in *Capsaspora owczarzaki*, *SrosMule* had GC conserved at the 5'/3' termini, with a longer TSD of 9bp in length. ITR target site insertions could not be reviewed in *SrosHel*, *SrosT1* and *SrosT3*, as the length of TSD could not be determined.



b) Transposon families

Figure 3.13: **Conserved base composition of target site motigs for DNA transposon families in the *S.rosetta* genome.** The methodology for TSD review is detailed in Figure 3.11.

3.3.5 Recent TE activity in the *S. rosetta* genome

Evidence for persistence of active TEs has been reviewed in several eukaryotic genomes, including *M. brevicollis*, *C. owczarzaki* and *S. cerevisiae* (Kim et al., 1998; Carr, Nelson, Leadbeater and Baldauf, 2008; Carr et al., 2012; Carr and Suga, 2014). Several characteristics were considered when reviewing TE activity including; terminal branch length, identical paralogous copies, expression and nucleotide diversity. Copy number is not an accurate representation of TE activity alone, as the host genome may have a high copy number, relating to several TE copies that are inactive, partial/truncated elements. The 20 families annotated in the *S. rosetta* genome ranged from multiple copies (161 copies of *Srospv2*) to single copy elements (*Srospv6*) (Table 3.5). The host genome contained higher copies of LTR retrotransposons than DNA transposons. The terminal sequences from NCBI Trace Archive were employed to investigate several methods of population genomics to determine TE activity. Nucleotide phylogenies for all TE families represented FLE, presumably young elements, on short branches, with few ancient copies (defined in Carr and Suga (2014) as having a terminal branch length <0.05 substitutions per site) present in the genome.

In the *S. rosetta* genome, the majority of LTR retrotransposon families were dominated by younger elements, with average terminal branch length of <0.05 per family (Figure 3.14). From the terminal branch length values, *Sroscv2* is defined as the retrotransposon family with the oldest copies, with the greatest range of length, and range exceeding 0.05 (Figure 3.14 and 3.17). This supported persistence in the *S. rosetta* genome for this family, which is further supported by family characteristics with truncated element copies uncovered in the genome (Figure 3.14 and Table 3.5). The youngest LTR retrotransposon families, based on terminal branch length, are *Srosgyp2*, *Srospv4* and *Srospv5*. When reviewing family characteristics, the three families have limited copies uncovered in the genome, with copy numbers ranging from 3-9 (Table 3.5). Recombination events are presumed to have occurred in all LTR retrotransposons families, as solo elements were identified in each LTR phylogeny, except *Srospv5* and *Srospv6* (Figure 3.21, 3.17, 3.22; Appendix F).

In contrast, the majority of DNA transposon families seem to possess older insertions, with an overall average terminal branch lengths of >0.05. *SrosTig1* was found to have the youngest copies, with all elements represented on short branches (Figure 3.14). Based on terminal branch lengths, the copies of *SrosT1* seem to be the oldest in the genome, with a range of up to 0.9. Reviewing the ITR phylogeny for this family indicated that one long branch is present, caused by a truncated ITR; the majority of copies are presumably young elements (Figure 3.16). The truncated ITR indicates family persistance, with an accumulation in mutations over time, increasing nucleotide diversity when compared to the *S. rosetta* FLEs. Alternatively, the element may be an ancient relic of an ancestral TE family that is no longer active in the host population, but is uncovered as the element shows similarity with the *T1* transposon family copies.

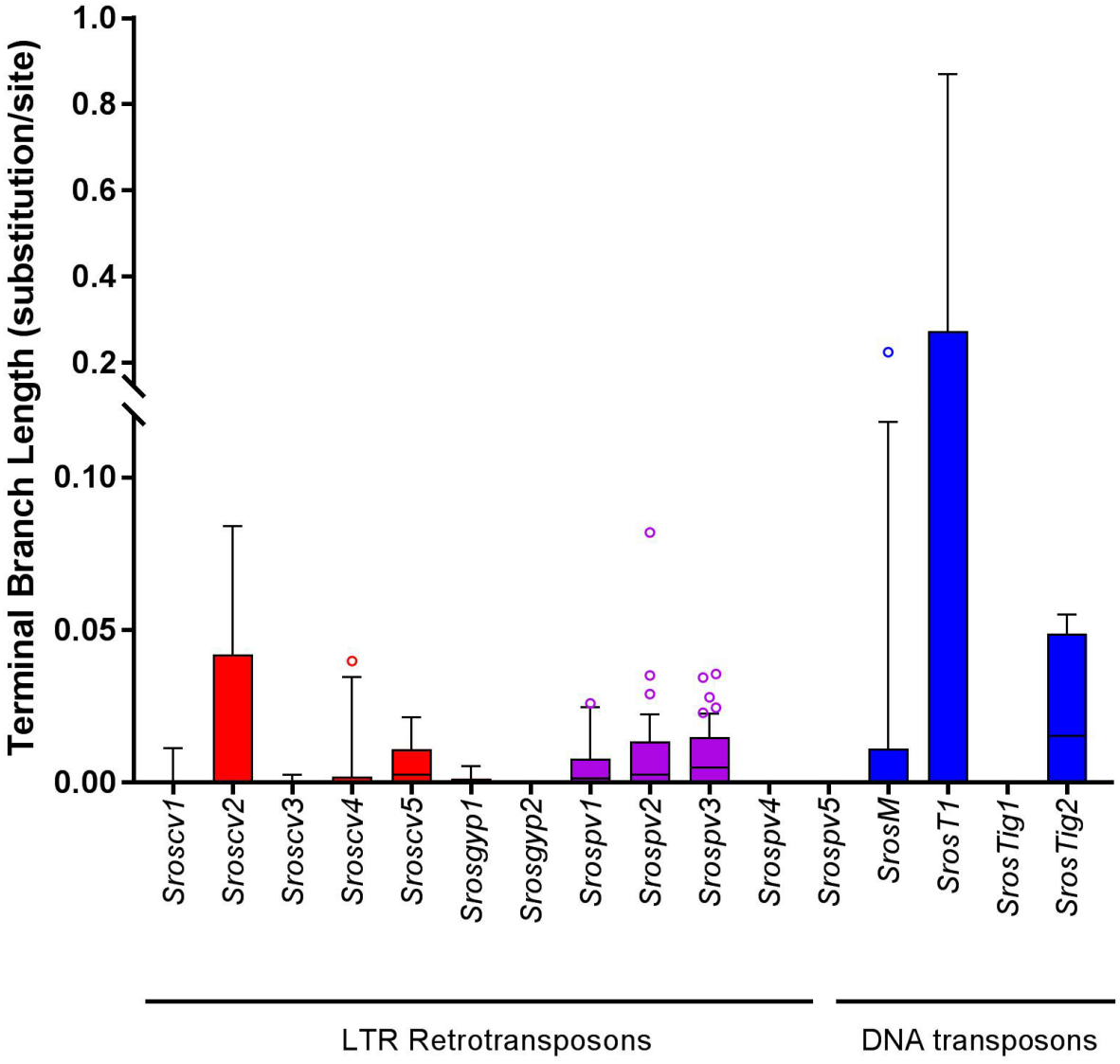


Figure 3.14: **Box and Whisker diagram to show terminal branch length of 20 TE families in the *S. rosetta* genome**. LTR retrotransposon and DNA transposon families are represented by red, purple, and blue boxes, respectively, which *gypsy* families in red, and *copia* families in purple. Branch lengths for full length LTR retrotransposons were taken from the 5′ LTR when this was present in the phylogeny; in its absence, the 3′ LTR was used. The filled boxes denote the 10 - 90 percentile range and the horizontal dark line represents the median branch length. The whiskers the percentile range from the median and the circles represent branch lengths outside this range.

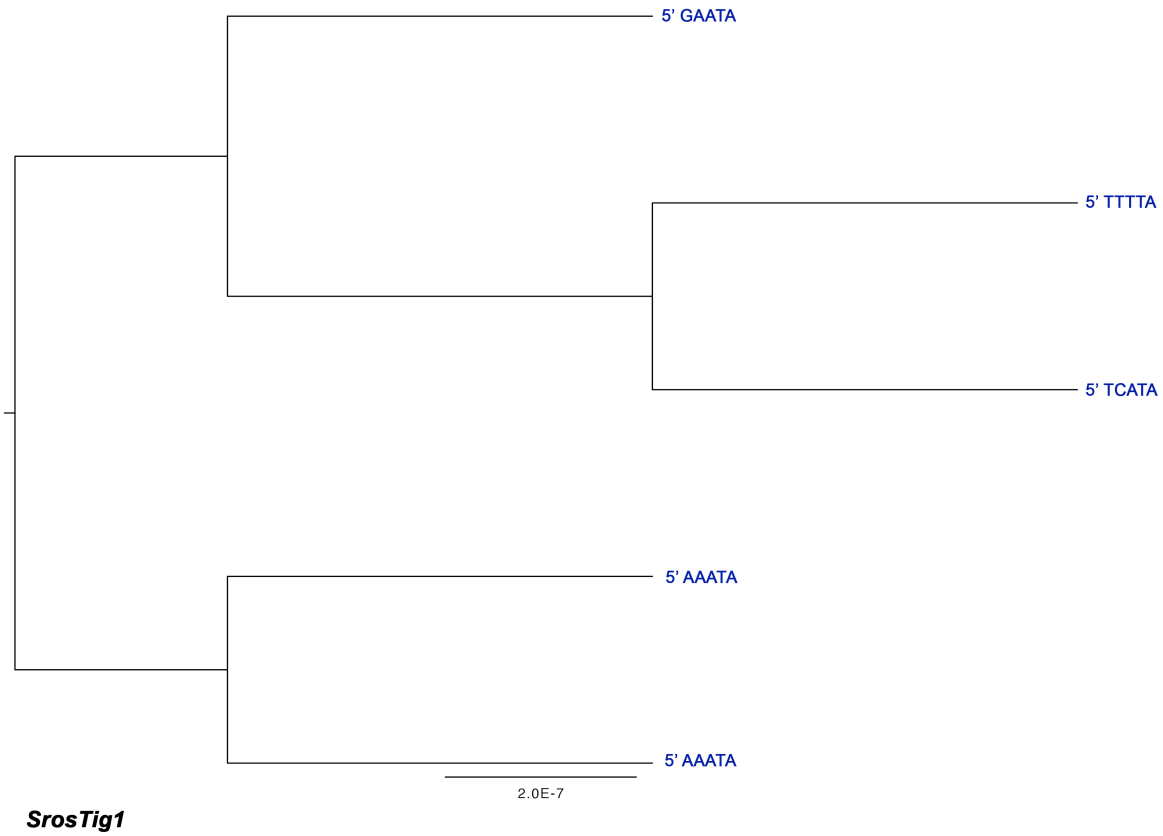


Figure 3.15: **Maximum Likelihood phylogeny of individual element copies of *SrosTig1*.** The phylogeny was constructed by an alignment of 393 nucleotide constructs with the employment of raxmlGUI using the GTRCAT model. ML and biPP values are labelled above and below corresponding branches. Maximum support is annotated by '*' (100ML/1.0 biPP) and low support (<50 ML/ <0.70 biPP) are annotated '-'. The scale bar signifies the number of nucleotide substitutions per amino acid site. 5' and 3' ITR sequences are written in blue, with individual TSDs annotated on terminal branches respectively.

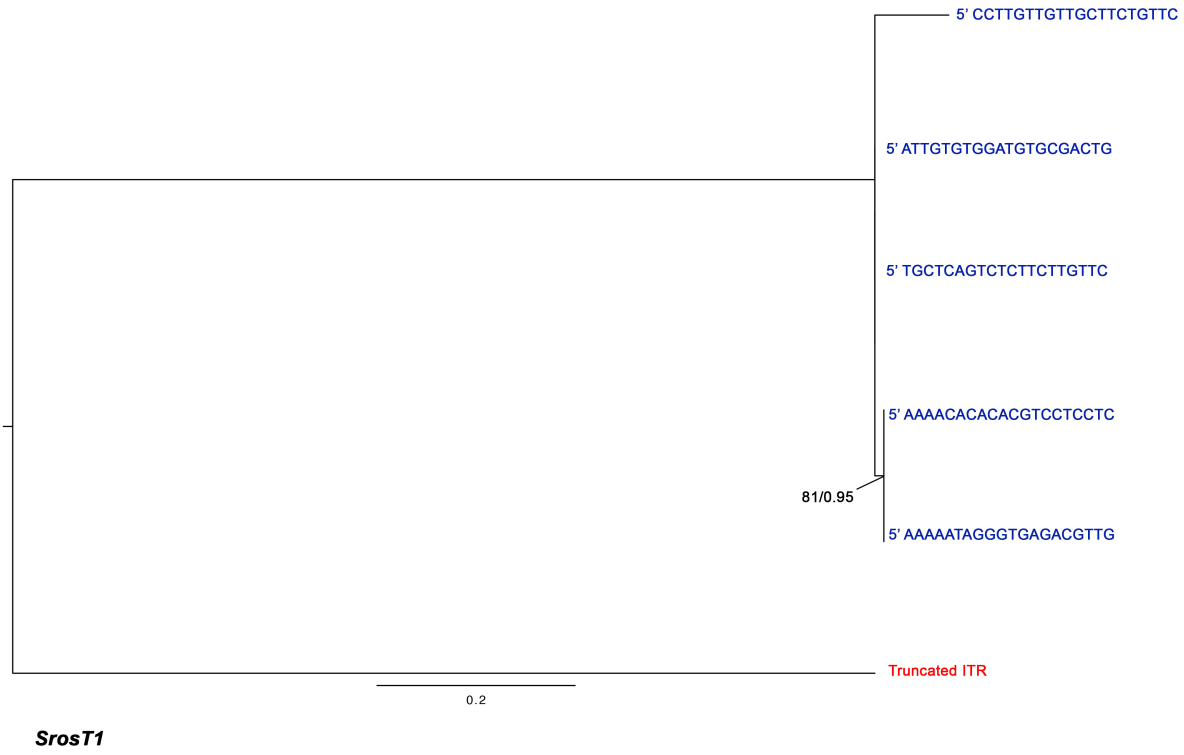


Figure 3.16: **Maximum Likelihood phylogeny of individual element copies of *SrosT1*.** The phylogeny was constructed by an alignment of 87 nucleotide constructs with the employment of raxmlGUI using the GTRCAT model. The tree format is stated in Figure 3.15.

3.3.6 TE expression in *S. rosetta* genome

Transcription is a fundamental stage of transposition for TEs, to produce RNA daughter elements in retrotransposons, Tnpase in DNA transposons, and to synthesis catalysts for transposition to occur (Slotkin and Martienssen, 2007). The RNASeq reads generated for the *S. rosetta* transcriptome project showed to include TE sequences. The total RNASeq dataset for *S. rosetta* contained 316,464,000 kb of reads, of which 1,616,278 were TE families (0.5%) (Table 3.5 and 3.4). This was similar to RNASeq read values seen in the filasterean species, *C. owczarzaki*, which was found to be 0.3% TEs (Carr and Suga, 2014). RNA expression varied dramatically across the TE families, with the lowest RNA expression seen for *Srosgyp1*, showing only 170 RNASeq reads during analysis (Table 3.5). The transcripts for *SrosMule* showed the highest levels of expression, with 54% of the TE reads being from this DNA transposon family (Table 3.4). *SrosMule* is found to have the highest number of copies of the DNA transposon families, as well as expression. The family is also the greatest in length, and predominantly only young copies are found in the *S. rosetta* genome. The family characteristics are consistent with the high level of expression calculated

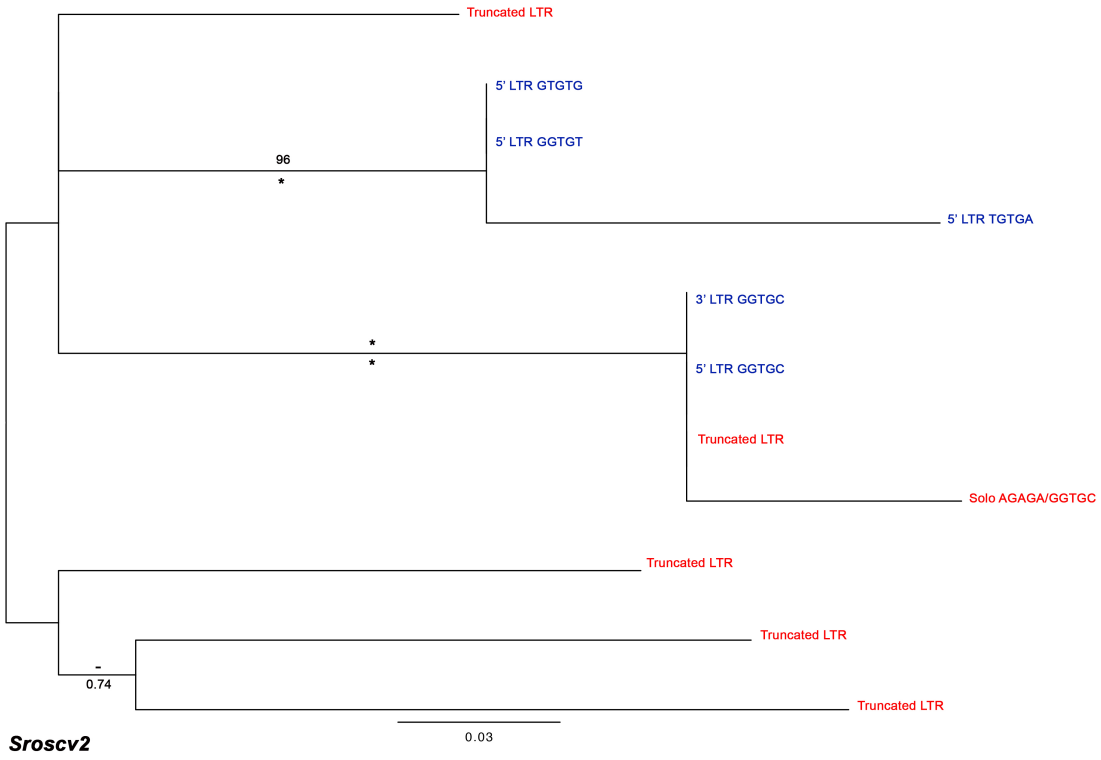


Figure 3.17: **Maximum Likelihood phylogeny of individual element copies of *Sroscv2*.** The phylogeny was constructed by an alignment of 214 nucleotide constructs with the employment of raxmlGUI using the GTRCAT model. The tree format is stated in Figure 3.17

here. Previously, trends have been seen between expression level and family characteristics, so this was investigated here (Carr and Suga, 2014). A strong positive correlation was seen between expression and copy number for DNA transposon families (R^2=0.837), however no relationship was seen for LTR retrotransposons (R^2=0.011) (Figure 3.18). No relationship was observed between RNASeq reads and identical paralogous copies as previously seen in protist *Capsaspora owczarzaki* (data not shown) (Carr and Suga, 2014).

Although from the same TE superfamily, *SrosTig1* and *SrosTig2* expression varied considerably within the *S. rosetta* genome. RNASeq reads for *SrosTig1* differed by over an order of magnitude when compared to *SrosTig2*. Both families have provided evidence for recent transposition in the genome, representing only FLE copies with low diversity in ITR copies seen (Figure 3.15 and 3.19; Table 3.4). However, *SrosTig2* has lower levels of expression (28,176), compared to *SrosTig1* value of 174,979 (Table 3.4). Although identical nucleotide diversity is seen here, *SrosTig2* shows greater terminal branch lengths greater than *SrosTig1*. The variance between the two families expression could also be explained by differences in copy number. As *SrosTig1* has a higher copy number, the elements are therefore more likely to be expressed.

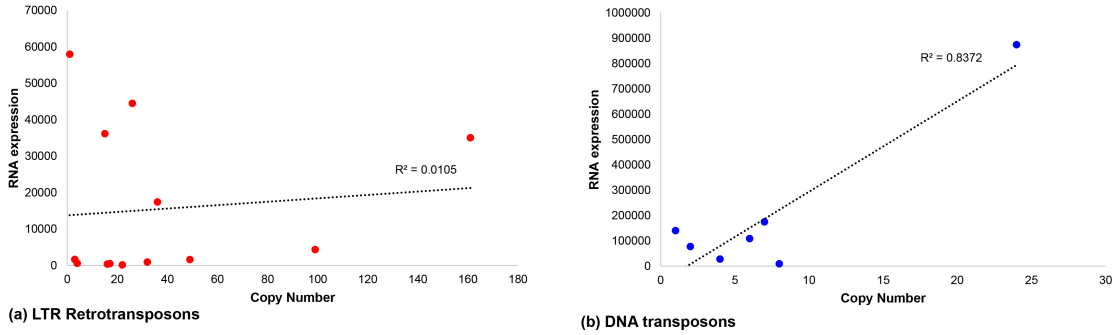


Figure 3.18: **Relationship between copy number and expression for TE families in *S. rosetta*.** The relationship between copy number and expression was reviewed in (a) LTR retrotransposon families, and (b) DNA transposons families. The R^2 value is annotated on each trendline.

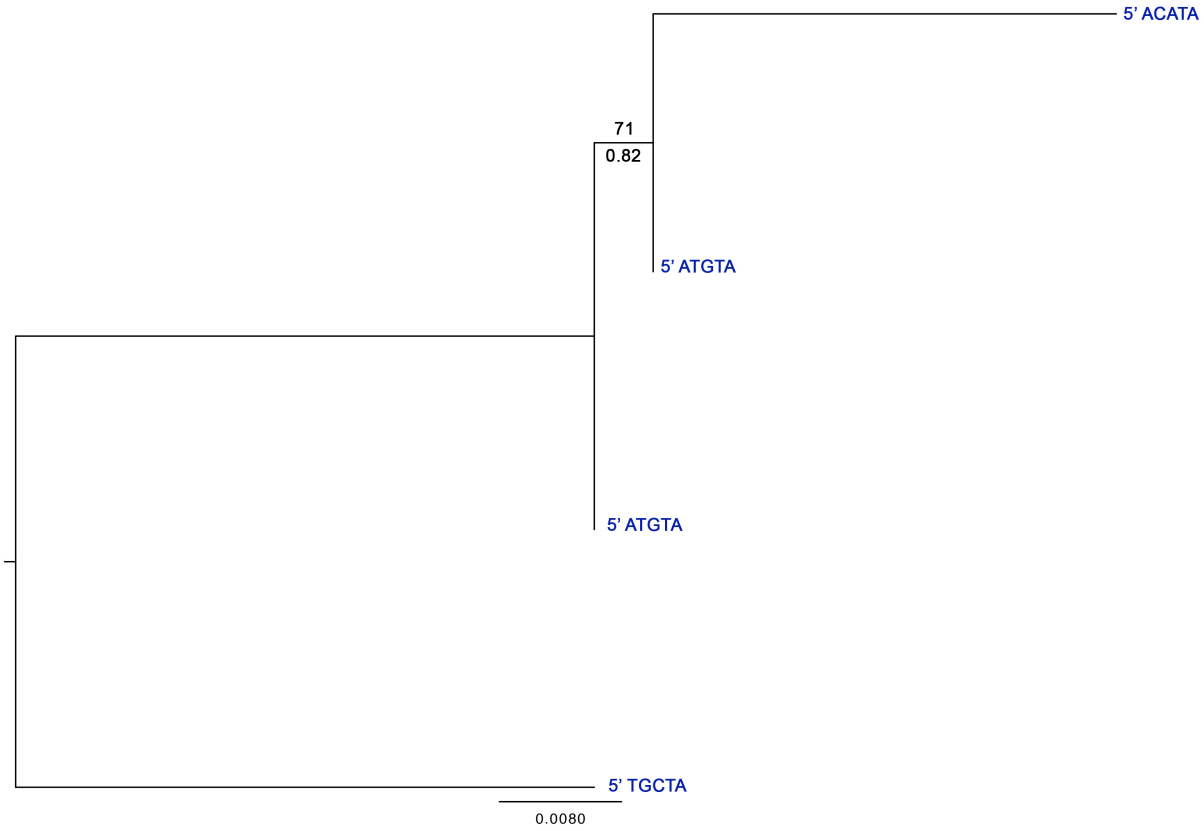


Figure 3.19: **Maximum Likelihood phylogeny of individual element copies of *SrosTig2*.** The phylogeny was constructed by an alignment of 276 nucleotide constructs with the employment of raxmlGUI using the GTRCAT model. The tree format is stated in Figure 3.15.

However, with copy number diversity across the 20 TE families uncovered in the *S. rosetta* genome, the expression data was normalised in order for relative expression to be reviewed for

the families (Figure 3.20). Only FLE copy numbers were included for the analysis. Overall, DNA transposons showed higher levels of relative expression when reviewing the two classes of TE in the *S. rosetta* genome (Figure 3.20). The family found to show the highest relative expression level per copy was *SrosT3*, with an expression value of 140,210 per copy, as the family was only found to have one copy in the *S. rosetta* genome.

3.3.7 TE nucleotide diversity in *S. rosetta*

Nucleotide diversity was an additional line of evidence to review recent transposition for both class of TEs. LTR sequences flanking elements are identical upon insertion, and evolve respectively by several mutations that accumulate over time. 5' and 3' elements were determined by the unique TSDs flanking the LTR sequences in Class I elements and ITR sequences for Class II, across all 20 TE families and alignments created for both solo, and FLEs. Intraelement LTR identity ranged from 87% to 100% (Table 3.5), supporting recent transposition of the LTR retrotransposons. Values for LTR identity could not be determined for*Srosgyp1-2*. In LTR retrotransposons, diversity was reflective of variance in copy number, element status and population substructure. *Sroscv2* LTR nucleotide diversity differed from the remaining chromoviral families by over an order of magnitude (π = 0.1765) (Table 3.5). This is reflected in the topology of the *Sroscv2* phylogeny, reflecting subdivision of two active clades with high support (96 - 100% ML/ 1.0 biPP) (Figure 3.17). The abundance of identical paralogous copies in the LTR retrotransposon families is an additional method to review recent transposition activity within the *S. rosetta* genome, as seen here (Table 3.5). Older copies would harbour several mutations through genome persistence and represented on long branches. All LTR retrotransposon families contained ancient copies that have existed in the *S. rosetta* genome for presumably long periods of time. This is supported by the presence of long branched solo/truncated elements, which are products of mutation (figure 3.21 and 3.17). Recombination events are detected in all LTR retrotransposon families as solo elements were identified in each phylogeny excluding *Srospv5-6* (Appendix F). *Sroscv1* reflected a star-like phylogeny, supporting a recent common origin of all family copies in the *S. rosetta* genome. A similar topology was seen for *Sroscv4, Srosgyp1 and Srosgyp2* (Appendix F). ITR nucleotide diversity was calculated in the multicopy families; *SrosM*, *SrosT1*, *SrosTig1* and *SrosTig2*, and found that diversity was low for the transposon families with no values >0.05 (Table 3.4). *SrosHel* has no ITRs annotated and therefore could not be reviewed for diversity.

Fewer ancient element copies were present in the ITR phylogenies for DNA transposon families, but were present in all families except *SrosTig1* and *SrosTig2* (Figure 3.15 and 3.19). This was reflected in the phylogenies with very few truncated ITRs uncovered. *SrosMULE* elements showed evidence for recent transposition in all copies except one full length element which was represented on a long branch (Figure 3.23). The diversity between the majority of Mule elements and the outgroup sequence could be rationalised to be an ancient copy that has shown persistence in the

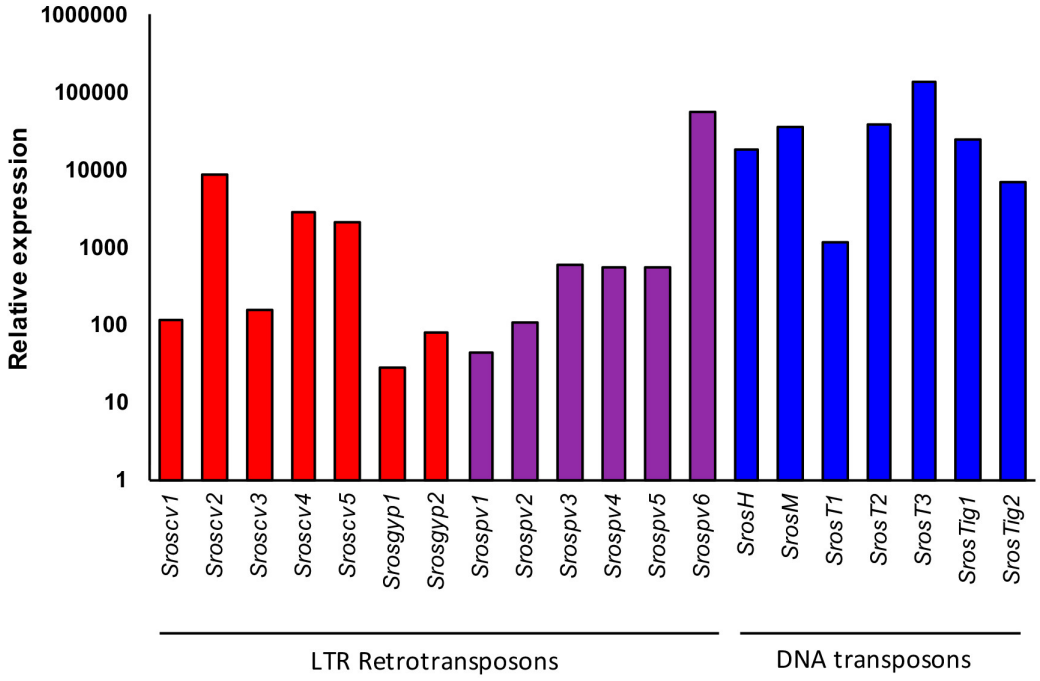


Figure 3.20: **Bar chart to show normalised expression data per element copy for 20 TE families in the *S. rosetta* genome.** The RNAseq raw data for each family was divided by family copy number to normalise the data for comparative analyses. Relative expression for each family is shown using a logarithmic scale.

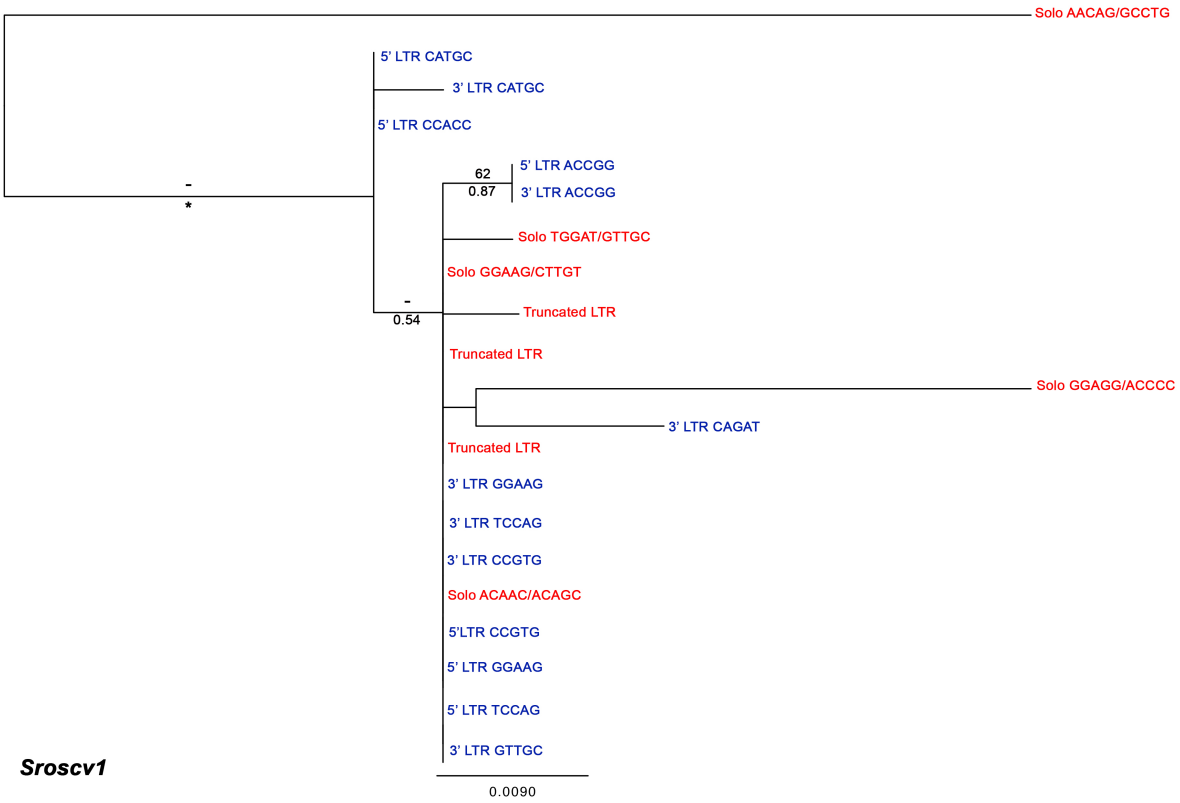


Figure 3.21: **Maximum Likelihood phylogeny of individual element copies of *Sroscv1*.** The phylogeny was constructed by an alignment of 252 nucleotide constructs with the employment of raxmlGUI using the PROTCAT model. ML and biPP values are labelled above and below corresponding branches. Maximum support is annotated by '*' (100ML/1.0 biPP) and low support (<50 ML/ <0.70 biPP) are annotated '-'. The scale bar signifies the number of nucleotide substitutions per amino acid site. 5' and 3' LTR sequences are written in blue, solo and truncated elements are in red, with individual TSDs annotated on terminal branches respectively.

genome, and been subjected to several mutations, causing a high value of nucleotide divergence. In contrast, the element may be a copy from a previously active DNA transposon family,that has shown similarity with *MULE* elements (Figure 3.23). The absence of ancient copies is common for Class I elements (Gorinsek et al., 2004; Carr, Nelson, Leadbeater and Baldauf, 2008; Carr and Suga, 2014), which is also apparent in *S. rosetta*.

Solo copies were detected in LTR retrotransposon families (Table 3.5 and 3.4). Previous literature has reported greater nucleotide diversity of solo LTRs and LTRs of FLEs (Carr, Nelson, Leadbeater and Baldauf, 2008). The same pattern was seen here when comparing nucletide diversity of solo LTRs to FLEs, except *Srospv3* (Table 3.5). This is reflected in the family LTR phylogenies, with levels of population substructure.The same is seen in *Srospv3*, with several truncated elements represented on long branches, however the solo copies show little sequence divergence to FLE, represented on short branches. This could be explained by recent ectopic

recombination upon insertion in the *S. rosetta* genome.

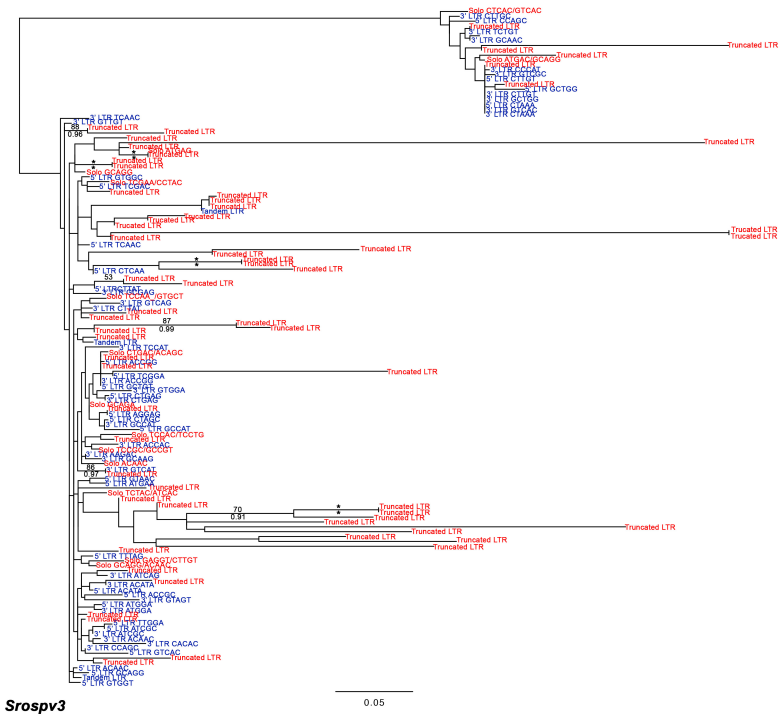


Figure 3.22: **Maximum Likelihood phylogeny of individual element copies of *Srospv3*.** The phylogeny was constructed by an alignment of 679 nucleotide constructs with the employment of raxmlGUI using the PROTCAT model. ML and biPP values are labelled above and below corresponding branches. Maximum support is annotated by '*' (100ML/1.0 biPP) and low support (<50 ML/ <0.70 biPP) are annotated '-'. The scale bar signifies the number of nucleotide substitutions per amino acid site. 5' and 3'. Format is stated in Figure 3.24

3.3.8 Evidence of recent transposition in the *S. rosetta* genome

Following acquisition, via vertical or horizontal transmission, TE families often multiply in the host genome. Persistence in the genome leads to individual elements acquiring mutations over time, resulting in inactive copies that are no longer able to transpose. The 20 TE families reviewed here varied considerably when comparing phylogenies, nucleotide diversity and RNA expression.

All LTR retrotransposon families contained presumably ancient elements,that were categorised as solo or truncated, which were represented on long branches (Figure 3.21, 3.17 3.22 and Appendix F). FLE in the families were predominantly short branched. *Sroscv1*, *Sroscv3*,*Srosgyp1* and *Srospv4* show similar phylogenies, with the majority of elements being short branched, presumably young LTRs (Figure 3.21, 3.24, 3.25 and Appendix F). The three *gypsy* families also show the highest number of identical paralogous copies within the LTR retrotransposons present in the *S. rosetta* genome (Table 3.5). This would suggest that the families are active within the choanoflagellate

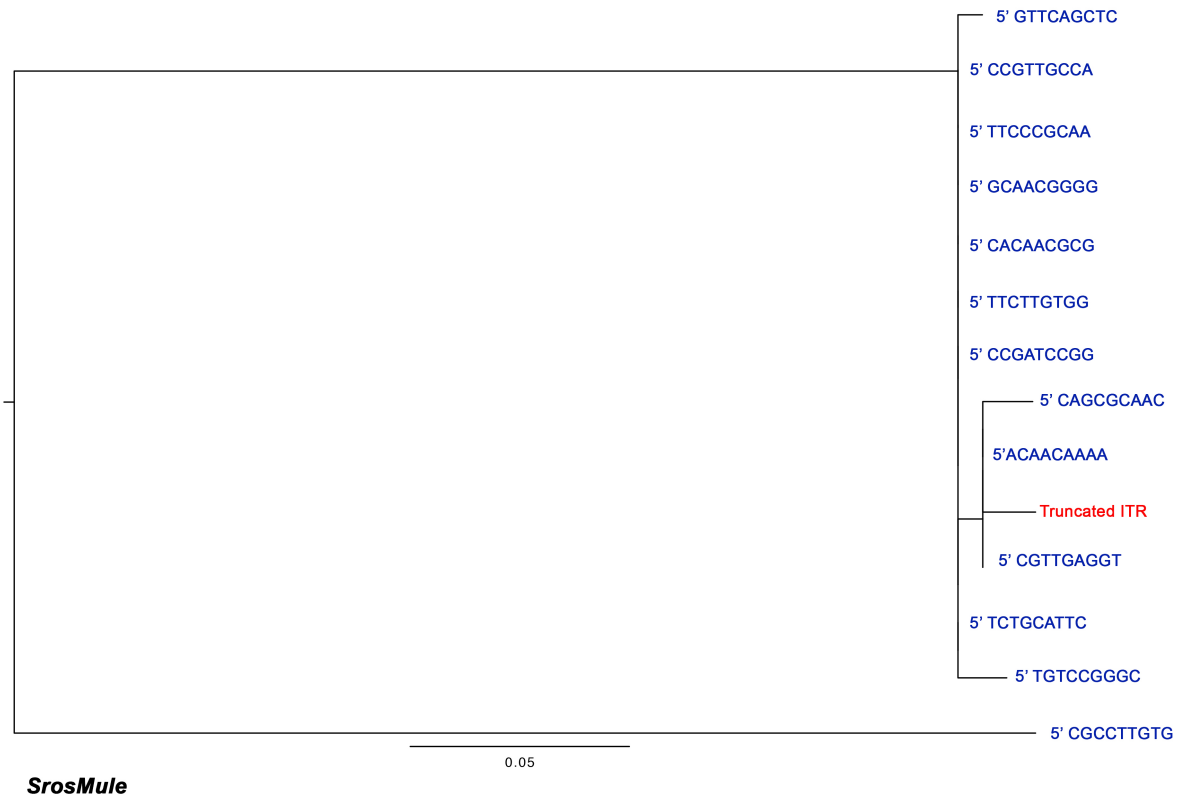


Figure 3.23: **Maximum Likelihood phylogeny of individual element copies of *SrosMule*.** The phylogeny was constructed by an alignment of 265 nucleotide constructs with the employment of raxmlGUI using the PROTCAT model. ML and biPP values are labelled above and below corresponding branches. Maximum support is annotated by '*' (100ML/1.0 biPP) and low support (<50 ML/ <0.70 biPP) are annotated '-'. The scale bar signifies the number of nucleotide substitutions per amino acid site. 5' and 3' ITR sequences are written in blue, solo and truncated elements are in red, with individual TSDs annotated on terminal branches respectively.

species. In contrast, no identical copies are seen in *Srospv1* and *Srospv6*, and the two *copia-like* families show LTR identity of 99.74 -100% (Table 3.5). This would suggest a recent transposition event for the two *copia-like* families. The majority of LTR retrotransposon families showed higher levels of nucleotide diversity in solo LTRs when compared to FLE LTRs within the same family, similar to *M. brevicollis* (Carr, Nelson, Leadbeater and Baldauf, 2008).

In contrast, DNA transposon families harboured fewer ancient copies, with *SrosTig2* consisting of FLE only (Figure 3.19). The majority of FLE in the DNA transposon families are represented on short branches, supporting recent transposition events within the host genome. The nucleotide diversity varies across the TE families by over an order of magnitude. This is respective of families showing either subdivision of diverse full length and truncated copies, as seen in *Sroscv2* (Figure 3.17), in contrast to young elements, which show decreased element subdivision, with limited diversity amongst family copies, as seen in both *Tigger* families (Figure 3.15 and 3.19).

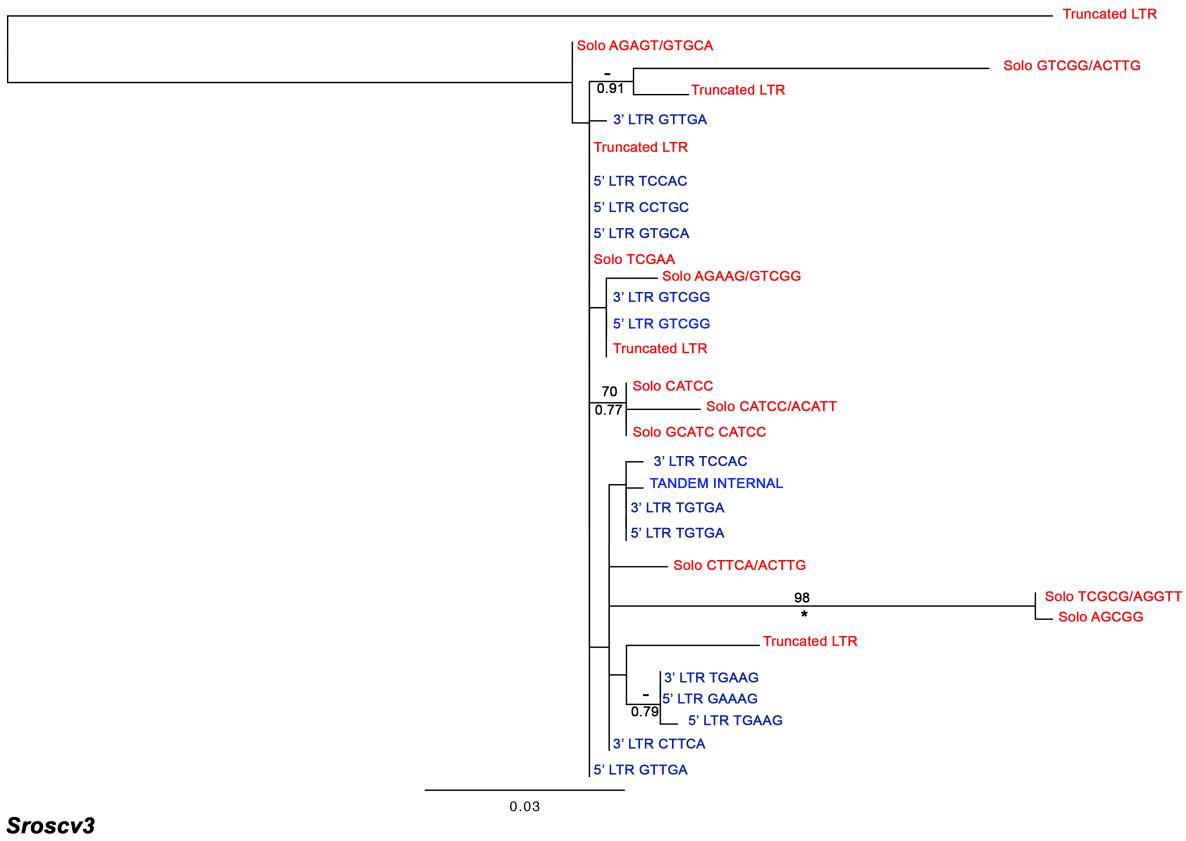


Figure 3.24: **Maximum Likelihood phylogeny of individual element copies of *Sroscv3*.** The phylogeny was constructed by an alignment of 418 nucleotide constructs with the employment of raxmlGUI using the PROTCAT model. ML and biPP values are labelled above and below corresponding branches. Maximum support is annotated by '*' (100ML/1.0 biPP) and low support (<50 ML/ <0.70 biPP) are annotated '-'. The scale bar signifies the number of nucleotide substitutions per site. 5' and 3' LTR sequences are written in blue, solo and truncated elements are in red, with individual TSDs annotated on terminal branches respectively.

3.3.9 Novel DNA transposons uncovered in choanoflagellate *M. brevicollis*

Novel DNA transposon elements were uncovered in the previously annotated choanoflagellate, *M. brevicollis*. Carr, Nelson, Leadbeater and Baldauf (2008) presented the initial TE annotation upon genome availability, and only LTR retrotransposons were uncovered. From BLAST similarity searches for protein superfamily phylogenies, two DNA tranposon families were found in the marine choanoflagellate, with similarity to *SrosT1* and *SrosTig2*. With this, the elements have been named *Monosiga brevicollis Transposon-1* (*MbT1*) and *Monosiga brevicollis Tigger-1* (*MbTig1*).The DNA transposon elements varied in length, with *MbT1* 4.7kb in length, and *MbTig1* a shorter 2.5kb (Figure 3.26). The DNA transposon families each possessed a single ORF, which encoded a putative *transposase* gene, and both ORFs harboured introns, (Figure 3.4). ITR size was similar

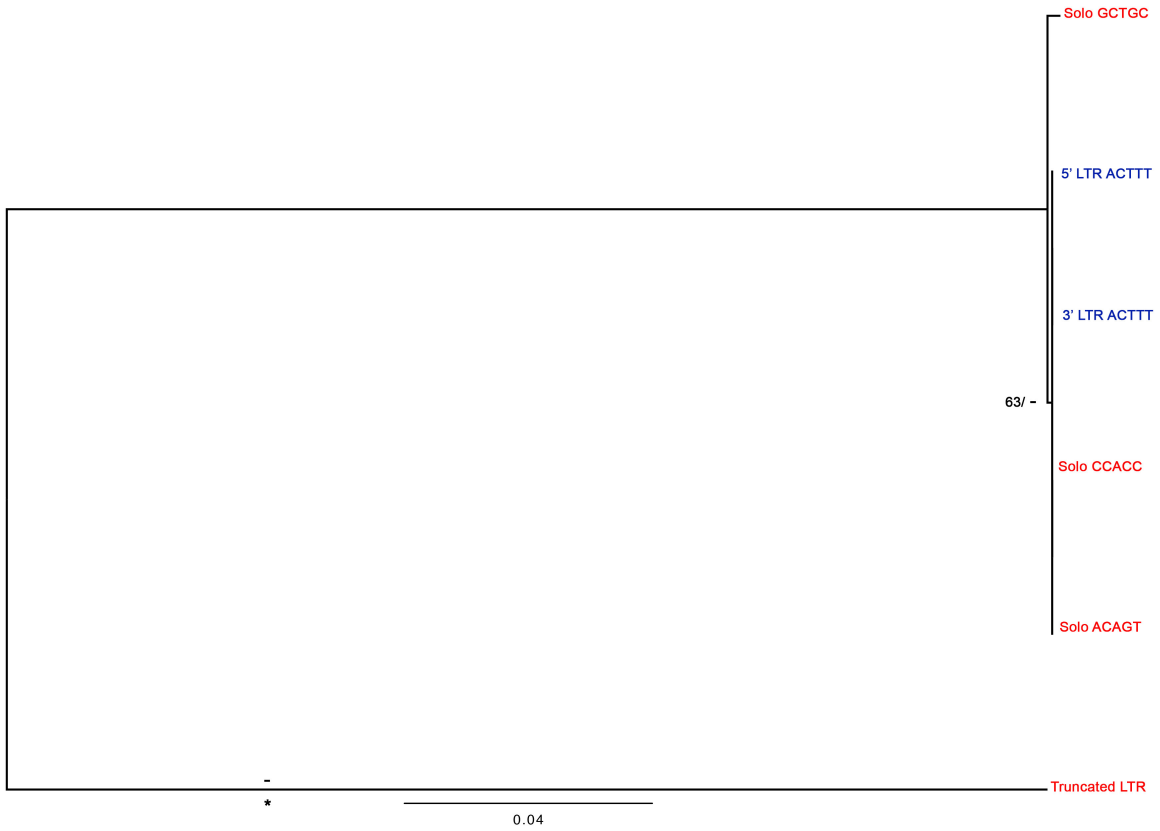
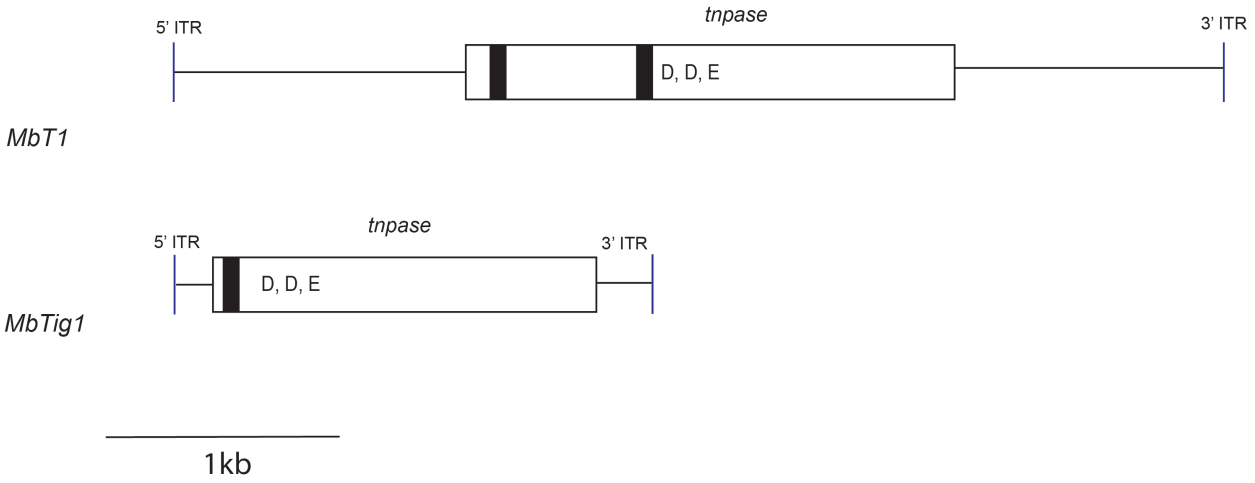


Figure 3.25: **Maximum Likelihood phylogeny of individual element copies of *Srospv4*.** The phylogeny was constructed by an alignment of 371 nucleotide constructs with the employment of raxmlGUI using the PROTCAT model. ML and biPP values are labelled above and below corresponding branches. Maximum support is annotated by '*' (100ML/1.0 biPP) and low support (<50 ML/ <0.70 biPP) are annotated '-'. The scale bar signifies the number of nucleotide substitutions per amino acid site. 5' and 3' LTR sequences are written in blue, solo and truncated elements are in red, with individual TSDs annotated on terminal branches respectively.

for both families, ranging from 10-11bp in length. The newly annotated elements were assessed comparatively with their similar copies in *S. rosetta*.

SrosT1 and *MbT1* differed considerably in length, with *SrosT1* being a shorter 2.0kb in length. Structurally, both elements encode Transposase, however *MbT1* possesses an intron, which are found to be absent in *SrosT1*. Amino acid sequences for the Transposase domains were compared using the BLAST2 function on NCBI (Sayers et al., 2009; Altschul et al., 1990). The sequences were found to have 56% identity. Conserved amino acid motifs were viewed in ClustalX (Thompson et al., 2002) (Figure 3.27). Amino acid conservation was seen between the elements, with largely conserved blocks throughout the alignment (Figure 3.27).



M. brevicollis transposon families

Figure 3.26: **Genomic organisation of the DNA transposon families characterised in the *M. brevicollis* genome.** DNA transposons: blue boxes represent inverted terminal repeat sequences, white boxes represent tnpase exon sequences, and black boxes represent tnpase intron sequences. Protein coding domains are indicated as follows: D,D,E, aspartic acid and glutamic acid catalytic domain. Non-coding regions are indicated as follows: ITR, inverted terminal repeat.

In addition, *SrosTig2* and *MbTig1* were similar in length, with *SrosTig2* being 2.1kb in length, compared to 2.5kb of *MbTig1*. Like all DNA transposons families annotated in the choanoflagellates, both elements encoded Transposase, and both *MbT1* and *SrosTig2* possessed introns. Despite the similarities detailed above, variation was seen when assessing the families at protein level; when compared BLAST2 function on NCBI Altschul et al. (1990); Sayers et al. (2009), the sequences were found to have a low percentage identity of 34%. With this, amino acid conservation was low for the *Tigger* elements, with few conserved blocks throughout the alignment (Figure 3.28).

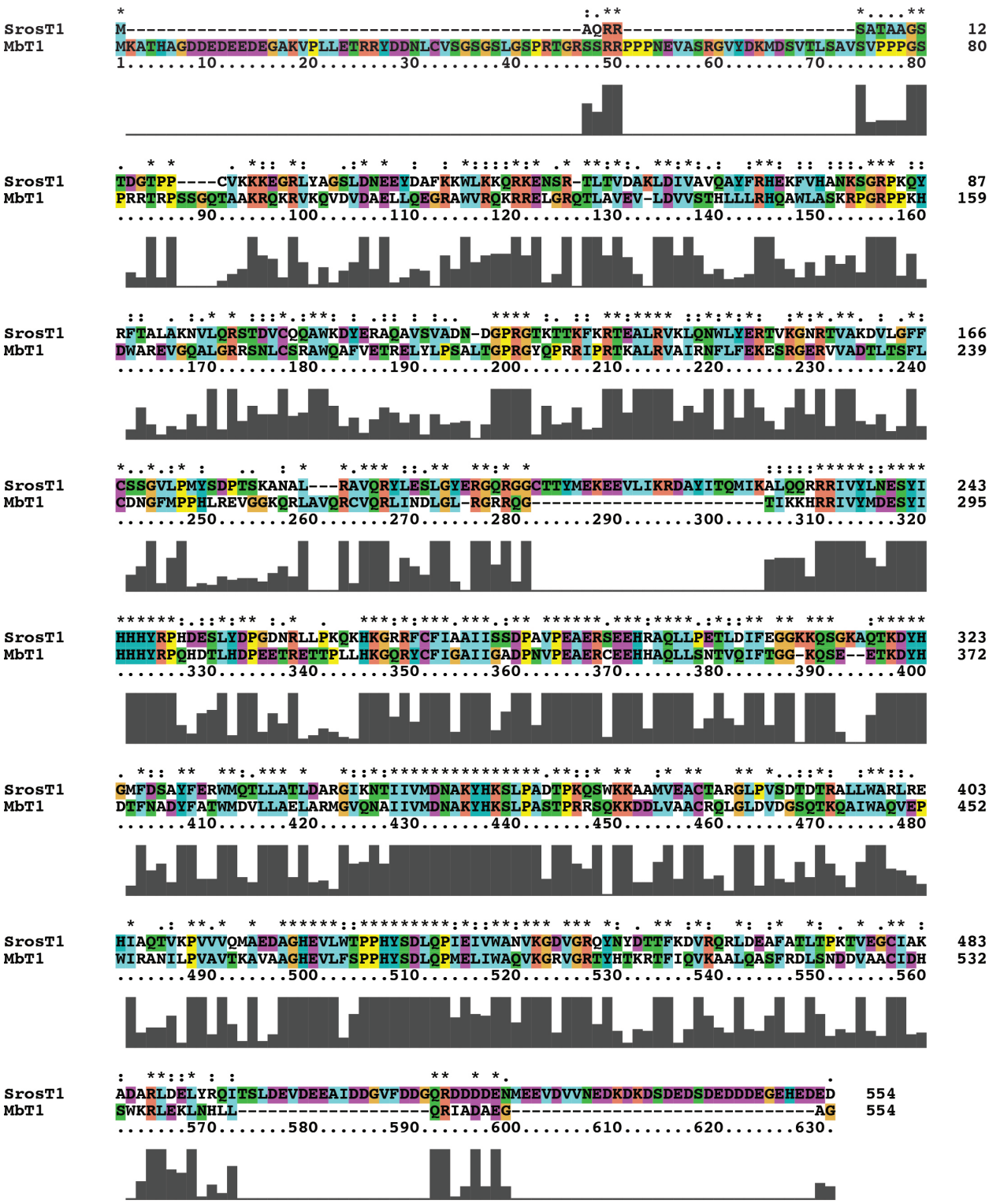


Figure 3.27: **A graphic representation of amino acid conservation for unclassified DNA tranpsoson family, *T1*, in choanoflagellate species, *S. rosetta* and *M. brevicollis*.** Sequences were aligned used MAFFT on the EMBL-EBI server, and viewed using ClustalX v2.1 (Thompson et al., 2002).

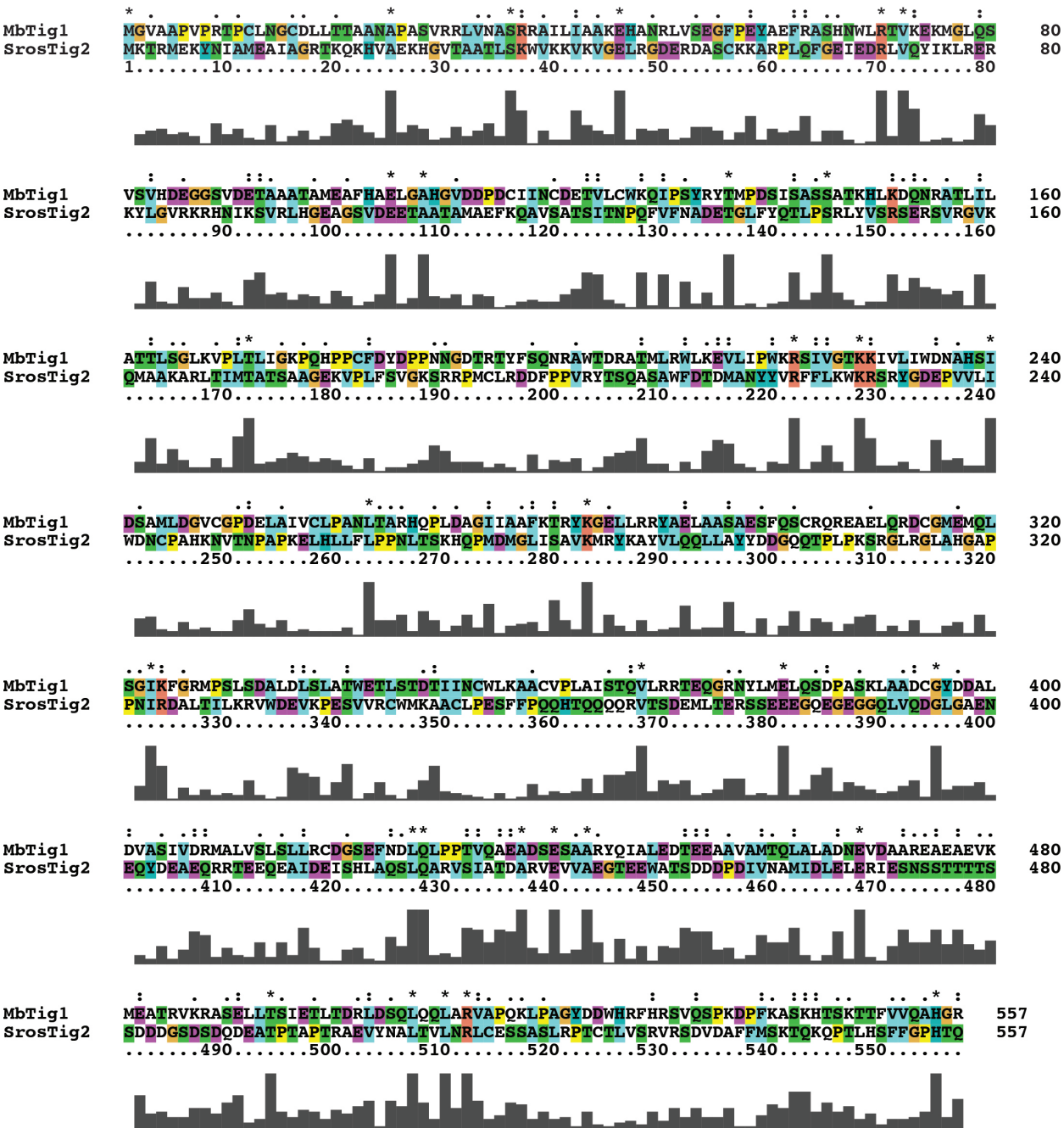


Figure 3.28: **A graphic representation of amino acid conservation for the DNA transposon family, *Tigger*, in choanoflagellate species, *S. rosetta* and *M. brevicollis*.** Format is detailed in Figure 3.27.

3.4 Discussion

3.4.1 TE family diversity in the *S. rosetta* genome

Salpingoeca rosetta is the second choanoflagellate species to be analysed for TE content. With this comparison can be drawn with the choanoflagellate species, *M. brevicollis*, as well as the filasterean species, *C. owczarzaki*, as both have been reviewed regarding TE content. *S. rosetta* was found to have a wider repertoire of elements when compared to *M. brevicollis*, harbouring multiple TE families from Class I and II, with a minimum of 20 TE families uncovered in this review. The abundance of TE families was a similar trait found in *C. owczaraki*, which harbours 23 families from both TE classes (Carr and Suga, 2014). This review has shed further light over TE evolution within the holozoans.

Comparatively, DNA transposon families were uncovered in *S. rosetta*, when Class II elements were documented as absent in the previously annotated choanoflagellate species, *M. brevicollis*. *S. rosetta* harboured 7 DNA transposon families; *MULE*, *Helitron*, *Transposon1-3* and *Tigger1-2*. However, similarity searches with the predicted transposons detected the presence of putative DNA transposons in *M. brevicollis*, challenging the initial TE review (Carr, Nelson, Leadbeater and Baldauf, 2008). Elements from superfamilies unclassified *Transposon-1* and *Tigger* were found in *M. brevicollis*. However, only with an increase in genome availability of choanoflagellates, can any ancestral inheritance be deduced for the families found. Unlike *C. owczarzaki*, non-LTR retrotransposons were found to be absent in *M. brevicollis*. Two non-LTR TE families were found in *S. rosetta*, however the full-length element could not be constructed, and therefore was not part of the TE analysis.

All superfamily phylogenies showed the elements to be placed within groups of other opisthokont families supporting inheritance by vertical transfer, with the exception of *SrosT1*. With this, protein phylogenies have provided evidence for the presence of TEs in the last common ancestor of the opisthokonts. The *copia-like* phylogeny showed moderate resolution, which the majority of branches representing low/ moderate support values. However, the *copia-like* sequences are clustered with *M. brevicollis pseudoviruses* to form one group, which supported that the superfamily may have had ancient homologues in the choanoflagellate lineage long term.

Chromoviruses were uncovered in three holozoan protists; *M. brevicollis*, *C. owczarzaki* and *S. rosetta*. This finding has provided further evidence that chromoviruses may have been present

in the last common ancestor of the holozoans (Carr, Nelson, Leadbeater and Baldauf, 2008). The chromoviral phylogeny detailed presented with moderate support, however the choanoflagellate copies are not a well supported grouping. However, the choanoflagellate clade is not positioned as a sister group to the metazoan copies, which would be expected with vertical inheritance. The metazoan chromoviruses form a well supported clade nested within fungal species, which has previously been observed in chromoviral phylogenies. This finding has been documented previously and is potential evidence for horizontal transmission of chromoviral elements from a fungal lineage to an ancestral vertebrate, as documented in (Carr and Suga, 2014). The transfer of mobile DNA could be facilitated between host species who have a shared habitat, or through predator-prey relationships, as seen in choanoflagellate species (Tucker, 2013). However, the findings are speculative, and further support would need to be provided to support the genetic transfer within the opisthokonts. An increased volume of protistan species would aid to resolve the evolutionary distances between mobile elements uncovered in the opisthokont phylogenies.

With most families appearing to be vertically inherited since the origin of the opisthokonts, there is evidence to support that *SrosT1* appears to be acquired by horizontal inheritance from a stramenopile donor. Horizontal transfer has been documented in choanoflagellates, with the transfer of genetic material from the prey species, including bacteria and other unicellular species (Tucker, 2013; Yue et al., 2013).The shared habitat of the marine species and phagocytosis of the stramenopiles would facilitate this transfer (Carr, Nelson, Leadbeater and Baldauf, 2008; Tucker, 2013). Due to the close proximity of the food vacuole and nucleus in the unicellular species, the transfer of genetic material may be facilitated between organelles of the cell. Furthermore, as an additional predicted Transposase protein was uncovered in *M. brevicollis*, this would support that the transfer was ancient in the choanoflagellate lineage. It would appear that the family remains active in *S. rosetta*, however no evidence to support recent transposition was uncovered in the *M. brevicollis* genome.

3.4.2 TE activity in the *S. rosetta* genome

Similarly to the three active families found in *M. brevicollis* (Carr, Nelson, Leadbeater and Baldauf, 2008), all TE families in the *S. rosetta* genome appear to be active, with the exception of *SrosT3* and *Srospv6*, which both have only one copy per family. It is known that inactive TE families are a common trait in several genomes of Opisthokonta (Carr et al., 2012; Carr and Suga, 2014). The

majority of families revealed elements in *S. rosetta* appear to be active, and also show persistence in the genome, except the two *Tigger* families. Both *tigger* families showed evidence of recent transposition, with no long branched, presumably ancient elements. The family has low levels of expression, as well as low nucleotide diversity. Similarly, *Srospv6* was found to have only a single copy in the genome, with 100% identical LTRs uncovered and therefore is suggested to have recently invaded the host genome, with potential to proliferate. However, this is not conclusive, as the decreased genome availability of choanoflagellates leaves the absence of potential donor species, and thus the evolutionary pathway is unknown. The putative donor would be a choanoflagellate, as in the *copia-like* phylogeny, the family appears to be monophyletic (Figure 3.8).

The TEs found in both *S. rosetta* and *M. brevicollis* (Carr, Nelson, Leadbeater and Baldauf, 2008) have low copy number, with only one family, *Srospv3*, found to harbour greater than 100 copies. Similar characteristics were also found in *C. owczarzaki* (Carr and Suga, 2014). This finding supports the hypothesis that low TE family copy number could be due to the large population size of protistan species, and therefore would allow for both efficient and effective elimination mechanisms to combat proliferation in the genome (Carr and Suga, 2014). A similar observation was found for the *Kazachstania* species, with TE copy number across the genus ranging from 1 - 30 (Chapter 2). As discussed in Carr, Nelson, Leadbeater and Baldauf (2008), it is proposed the protistan species are found to have large effective population sizes (Snoke et al., 2006), and therefore allow elimination mechanisms of individual TE insertions within a species. The same theory can be proposed for the yeast species, with presumed large effective population size (Tsai et al., 2008), who present with active elements in low copy.

Further similarity can be drawn with *M. brevicollis*, in that all FLE of LTR retrotransposons are presumably young elements (Carr, Nelson, Leadbeater and Baldauf, 2008). This finding supports the successful elimination of full length LTR retrotransposons, as very no ancient copies were uncovered and therefore elements are presumed to be eradicated prior to accumulating mutations. A similar trait is seen in the DNA transposon families, with very few ancient copies uncovered, and all copies presented on short branches, suggesting recent transposition events in the genome. The low number of families uncovered in *M. brevicollis* (*Mbpv1*, *Mbpv2* and *Mbcv*) would suggest a major TE loss has occurred in this species, however this may not be a representative trait of all choanoflagellate species (Carr, Nelson, Leadbeater and Baldauf, 2008). This finding is supported by the survey completed here, with increased diversity of TE superfamilies

in the *S. rosetta* genome, as well as the repertoire of elements uncovered in *M. fluctuans* by the screening of EST sequences (Carr, Nelson, Leadbeater and Baldauf, 2008). With this, the TE survey uncovered 20 novel TE families in the choanoflagellate, *S. rosetta*, and two potential TE families in *M. brevicollis* from a class that was previously defined as absent for the choanoflagellate species (Carr, Nelson, Leadbeater and Baldauf, 2008). The findings uncovered here will hopefully inspire further investigation of mobile elements in additional choanoflagellate species.

Chapter 4

Codon usage in three holozoan species

Aspects of the work described in this chapter was published in Southworth, J., Armitage, P., Fallon, B., Dawson, H., Bryk, J. and Carr, M. (2018), 'Patterns of ancestral animal codon usage bias revealed through holozoan protists', *Molecular Biology and Evolution* **35** (10), 2499 – 2511.

4.1 Introduction

The genetic code possesses 61 independent codons, which encode for 20 amino acids. From these 20,18 amino acids are encoded by more than one synonymous codon, (reviewed by Clark (1970)). The remaining 2 amino acids are methionine and tryptophan (Grantham et al., 1980). The degeneracy observed between codon families, has been an area of interest, with suggestion that codon usage is biased, resulting in specific codons within codon families being present at a greater frequency (Yannai et al., 2018). The theory of codon usage bias was first proposed by Clark (1970), and later supported by (Grantham et al., 1980) who showed that the non-random use of codons was found in both eukaryotes and prokaryotes.

Three hypotheses are found to drive codon usage bias; natural selection, mutational forces and genetic drift (Sharp et al., 1988; Bulmer, 1991; Smith and Eyre-Walker, 2001). Natural selection shapes codon usage to enable translational efficiency in the majority of tRNA genes within the host cell (Ikemura, 1981). As defined by Ikemura (1981), "optimal codons" were originally categorised as the codons which were found to be complementary to the host major tRNA genes, by standard Watson-Crick base pairing. The definition was adapted by Lloyd and Sharp (1992), who proposed that optimal codons were those found to be show greater abundancy in highly expressed genes in relation to other gene categories. The selective usage of optimal codons within a gene is defined as the Frequency of Optimal Codons (Fop), which is calculated by the total number of codons within a

gene, divided by the number of optimal codons found in the gene. Codons selected for this reason are defined as ”optimal” as they provide a selective advantage. The level of bias in a host gene is frequently assessed by the measure of the ’effective number of codons’ (*Nc*) (Wright, 1990). *Nc* can range from 20 to 61, with 20 showing the use of a single codon, and 61, where codons within families are used equally, with no bias observed (Wright, 1990). In contrast to natural selection, codon usage bias can also be shaped by mutation pressure, influencing nucleotide composition at neutral synonymous coding positions. If mutation pressure is found to drive codon usage, it would have influence on both coding and non-coding region respectively. Evidence for mutational forces influencing codon usage includes a correlation between GC-content at third positions (GC3s) and non-coding local base composition. With this, if no relationship is seen, this would indicate that mutation is not the main driver of codon usage, and therefore the bias would be caused by selection and/or genetic drift.

Codon usage of highly expressed genes is proposed to enable efficient translational efficiency, and therefore the use of ’optimal codons’ facilitating more rapid translation (Ikemura, 1981; Ehrenberg and Kurland, 1984). Efficiency is increased with enriched optimal codons as it is thought that the decoding of optimal codons is found to be faster by the ribosome in comparison to codons which are not determined as optimal (reviewed by Tuller et al. (2010)). A gene defined as highly expressed, with low translational efficiency, would lead to an increase ribosomal usage, and therefore the quantity of free ribosomes would be in decline, reducing availability for other host genes (Frumkin et al., 2018). Furthermore, a signature of selection for translational accuracy is that optimal codons are found in abundance within domain coding regions compared to non-domain coding regions within host genes (Akashi, 1994; Stoletzki and Eyre-Walker, 2007; Ran and Higgs, 2012).

In diverse taxonomic groups, evidence has shown that selection is predominantly the main driver of codon usage (Lerat et al., 2003; Yannai et al., 2018). The closest known relatives of Metazoa, are several lineages of unicellular eukaryotes; this grouping of metazoans and unicellular relatives is known as Holozoa (Shalchian-Tabrizi et al., 2008). Within Holozoa, the sister group to choanoflagellates are metazoans, and Filasterea as a more distally related lineage (Figure 4.1) (Parfrey et al., 2011). The study of the unicellular members of Holozoa has shown novel patterns of codon usage bias for the holozoan protists, as well as predictions of ancestral traits now seen in multicellular eukaryotes (Carr et al., 2010; Tucker, 2013; Carr et al., 2017). Evolutionary characteristics conserved in both the filastereans and choanoflagellates have been defined as

ancestral, and therefore ancestral to Metazoa, even if these traits has since been lost through multicellularity, or lost in premetazoan lineage prior to multicellularity. In the last decade, whole genome sequences had been made available of three holozoan species; two choanoflagellate species, *Salpingoeca rosetta* (Fairclough et al., 2013) and *Monosiga brevicollis* (King et al., 2008), and filasterean, *Capsaspora owczarzaki* (Suga et al., 2013).

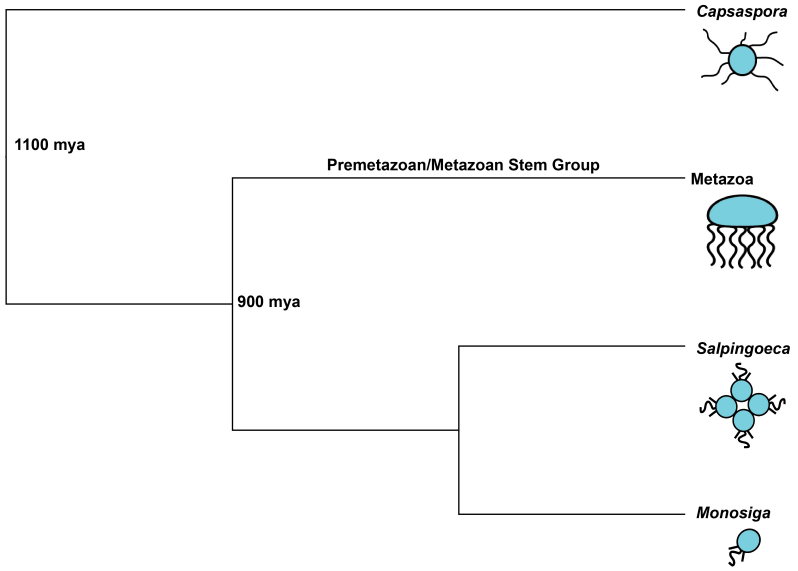


Figure 4.1: **Simplified phylogenetic representation of Holozoa** The cladogram outlines that choanoflagellate species (*Monosiga* and *Salpingoeca*) are the sister group to the metazoans. Approximate divergence dates are based on figures outlined in Parfrey et al. (2011).

Upon genome availability, *S. rosetta*, *M. brevicollis* and *C. owczarzaki* genomic characteristics have been studied (Carr, Nelson, Leadbeater and Baldauf, 2008; Carr, Leadbeater, Nelson and Baldauf, 2008; King et al., 2008; Suga et al., 2013; Carr and Suga, 2014). Previously, codon usage bias had not been reviewed in unicellular holozoan species, leaving a niche to be explored. Southworth et al. (2018) explored codon usage bias across the three holozoan species, to determine whether bias is more likely to be driven by natural selection and/ or mutation bias. The analysis in Southworth et al. (2018) was repeated here to ensure validity, and to assess whether narrowing the bias category margins from 5% to 1% reflects the same results when comparing bias gene statistics (Southworth et al., 2018). The larger cohort of researchers involved with the Southworth et al. (2018) publication allowed for greater sample sizing to be studied, whereas a smaller grouping was employed for independent research.

4.2 Methods

4.2.1 Codon usage analysis

Annotated transcript sequences were downloaded for each of the holozoan species; *S. rosetta*, *M. brevicollis* and *C. owczarzaki*. The *C. owczarzaki* transciptome was downloaded from the Ensembl Protists database (Capsaspora owczarzaki atcc 30864.C owczarzaki V2.cds). *S. rosetta* and *M. brevicollis* transcriptomes were downloaded from the Origins of Multicellularity Project at the Broad Institute (salpingoeca rosetta 1 transcripts and monosiga brevicollis mx1 1 transcripts). CodonW was used to determine codon usage statistics for each set of cds transcriptome sequences. Optimal codon files were initially generated by the CodonW correspondence analysis on relative synonymous codon usage (COA on RSCU) using default parameters. CodonW calculates optimal codons by COA, which is used to identify trends in datasets, creating a series of continuous axes which show variation between genes and codons. RSCU was calculated for each gene in CodonW to show deviation from even usage (Sharp et al., 1986; Peden, 1999). RSCU values indicate the number of times a codon is uncovered, in relation to the number of time it would be expected to be observed based on equal synonymous codon usage (Sharp et al., 1986). An RSCU of 1.00 would indicate that a codon would be employed with equal frequency based on random codon usage; RSCU >1.00 would indicated that a codon was selected more frequently than expected for random codon usage, and <1.00 would indicate that a codon is used at a lower frequency based on a non-biased random model.

Expression data was calculated using SMALT v. 0.2.6. (Ponstingl, 2014) to determine the RNA reads per host genes for both *S. rosetta* and *C. owczarzaki*. Genes with reads >100 were concatenated from the high and low 5% bias categories, and new datasets based on expression were ran using CodonW to determine optimal codons based on expression. The fop files for *S. rosetta* and *C. owczarzaki* were manually revised based upon expression data. Optimal codons based on expression data were compared to those generated by CodonW, and the fop file was amended to show the optimal codons for amino acids based on expression data. The fop file is a measure of codon bias for each codon, where all 64 codons must have a score; 1 indicates a rare codon, 3 for an optimal codon and 2 for other codons (standard usage) (Peden, 1999). Values of GC3s, F_{op} and *Nc* were calculated for each gene per species. F_{op} values were determined using the fop.coa file from CodonW for *M. brevicollis*, and an amended fop.coa file was used for *C.*

owczarzaki and *S. rosetta*. Mean and standard deviation values were determined for each codon usage statistic in each species.

4.2.2 Codon usage bias categories

For each holozoan species, two sets of three categories were created based upon effective number of codons, *Nc*. Initially, the high category represented the 5% most highly biased genes, the median 5% biased genes for the mid category, and the lowest 5% bias genes for the low category, as seen in Southworth et al. (2018). From this, the categories were decreased to represent 1% for each category. The categories were formulated to review codon usage characteristics across the transcriptomes, and amended to study if 1% effectively represented trends observed for 5% seen in Southworth et al. (2018), and to allow for repeated independent study, as the larger categories would not be feasible due to time constraints.

4.2.3 Determining GC content for intronic and flanking DNA

For each bias category, genes were reviewed using NCBI gene annotation to determine intronic DNA, which was ran for both 5% bias categories (Southworth et al., 2018) and 1% categories. With this, introns were excised for each gene, and GC content was calculated using CodonW (Peden, 1999). The intronic DNA dataset from Southworth et al. (2018) was used to extract the values for the 1% bias category sequences. For each gene in the 1% category, flanking DNA of 200bp was extracted upstream and downstream from the GenBank nucleotide sequence reads. The flanking DNA was truncated if the read did not cover the entirety of the flanking region, or if a neighbouring gene was found to overlap the flanking DNA at the 5' or 3' region of the gene of interest. The flanking DNA was concatenated, and GC content calculated using CodonW. Mean and standard deviation values were determined for each category for both intronic and flanking DNA in all species.

4.2.4 Major tRNA gene screening

Each annotated genome for the three holozoan protists were downloaded from NCBI. *S. rosetta* was made up from 154 scaffolds, the *M. brevicollis* dataset was made up of 219 scaffolds, and *C. owczarzaki* was comprised of 84 scaffolds. The program tRNAscan-SE 2.0 Lowe and Chan (1997)

was employed to identify major tRNA genes using default settings, by postgraduate student, Holly Dawson.

4.2.5 Gene expression in *C. owczarzaki* and *S. rosetta*

For *S. rosetta* and *C. owczarzaki*, gene expression was determined for all genes by the examination of sing SMALT v. 0.2.6. (Ponstingl, 2014). The bias categories were determined using *Nc* values as previously described, to assess trends between gene expression and codon usage bias. Transcriptome SRA files SRX042046-SRX042024 for *S. rosetta* were downloaded and SRX155789-SRX155797/ SRX1690425-SRX1690428 for *C. owczarzaki*. A total of 122.1 million reads were ran for *S. rosetta* and 665.2 million reads for *C. owczarzaki*. SMALT aligned the SRA reads to each gene sequence per bias category to review the expression level per gene. The reads per gene were visualised and calculated using Tablet v. 1.16.09.06 (Milne et al., 2013).

4.2.6 Codon usage analysis in domain and non-domain codons

As seen in Southworth et al. (2018), for the three bias categories per species, each gene was assessed and separated into functional domain and non-domain codons. Any genes that did not have a specified domain region were removed from the analysis. The domain and non-domain codons for each bias category were ran using CodonW to determine Fop values (Peden, 1999). The domain and non-domain codons were separated for each 1% bias category and compared to the results seen in Southworth et al. (2018) for the 5% bias category. Domain and non-domain codons for *S. rosetta* were excised by Paul Armitage, undergraduate project student, and sequences extracted based on the accession list for the 1% categories, and reran for Fop values using CodonW (Peden, 1999).

4.3 Results

4.3.1 Review of codon usage bias in three holozoan protists

Genome characteristics for each holozoan species are listed in Table 4.1, including genome size, GC-content and number of CDS sequences. As seen in Table 4.1, *C. owczarzaki* has the smallest genome, with a size of 28.0Mb. In contrast, *S. rosetta* has a much larger genome of 55.0Mb, and the *M.brevicollis* genome is 41.6Mb in size (Table 4.1). Besides *C. owczarzaki* having the smallest genome, it was found to possess over 10,000 coding genes, which is similar to the numbers seen in both choanoflagellate species (Table 4.1). GC-content was found to be conserved across the three holozoan species, with values ranging from 54% - 57% (Table 4.1).

Table 4.1: **Whole genome characteristics of three holozoan protists**. Data included genome size (Mb), number of CDS sequences, and overall GC-content.

Species	Genome Size (Mb)	CDS total	GC-content
M.brevicollis	41.6	9171	55%
S. rosetta	55.0	11736	57%
C. owczarzaki	28.0	10123	54%

The direction of codon usage was initially determined for the transcriptomes of *C. owczarzaki*, *M. brevicollis* and *S. rosetta* with use of the effective number of codons value (*Nc*) calculated by CodonW. Values of *Nc* varied, ranging from 20, highly biased, to 61; mean scores for each transcriptome are shown in Table 4.2. The strongest level of bias based on *Nc* was seen in *S. rosetta* (average *Nc* = 44.79±5.37). *C. owczarzaki* showed the next strongest level of bias, with *M. brevicollis* being the least bias (average *Nc* = 47.60±6.45;48.05±5.62) (Table 4.2). The distribution of bias for the genes was observed in the three species. It was found that for *S. rosetta* the highest percentage of the transcriptome was found to show bias *Nc* values between 40-54.99, whereas for *M. brevicollis* and *C. owczarzaki*, the majority of genes were in the higher range of 50-54.99, showing weaker bias (Figure 4.2).

Following the methods from Wright (1990), the direction of bias can also be determined by the value of GC at synonymous third positions (GC3s), which reflects whether the host genes of the transcriptomes favour AT or GC ending codons. Figure 4.3 shows that all three species showed a bias towards GC-ending codons, with highly biased genes having the highest GC3s values. In line with *S. rosetta* showing the highest level of bias based on *Nc*, the mean GC3s value is highest for this species, followed by *C. owczarzaki*, and last *M. brevicollis* (Table 4.2). It was seen that very few genes in the three species are composed of AT-ending codons, with the minority of genes showing <0.5 GC3s value (0.60% for *S. rosetta* 1.32% for *M. brevicollis*; 1.63% for *C.owczarzaki*) (Figure 4.3).

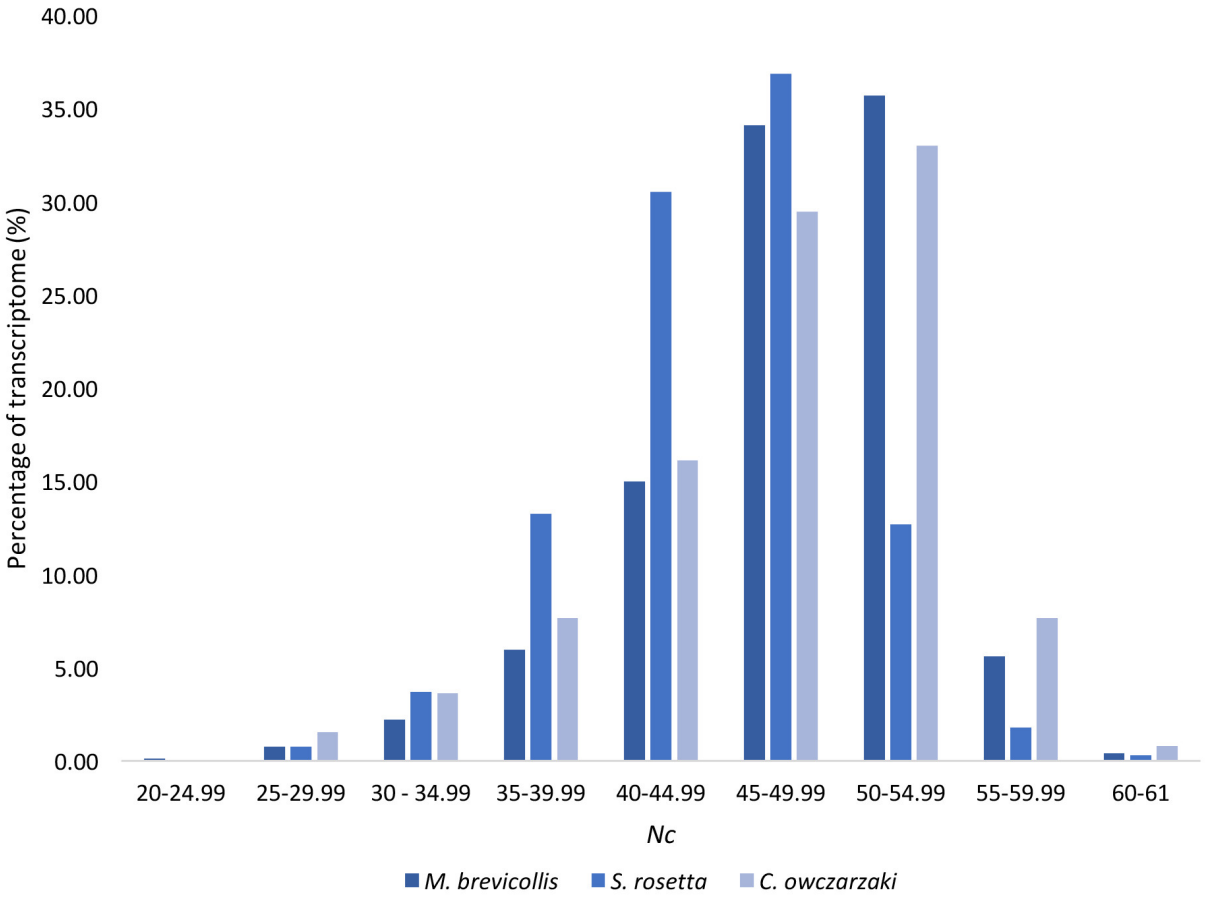


Figure 4.2: **Codon usage bias ditribution for the three transcriptomes of the holozoan protists.** *Nc* values ranging from 20-61, were categorised to blocks of five to review bias distribution percentages for each species transcriptome.

Table 4.2: **Mean codon usage statistics in the transcriptomes of the three holozoan species**. Data included the average value for effective number of codons (*Nc*), GC3s and frequency of optimal codons (F_{op}) for each species. Standard deviation was calculated for each data set.

Species	Nc	GC3s	Fop
M. brevicollis	48.05±5.62	0.638±0.060	0.572±0.080
S. rosetta	44.79±5.37	0.707±0.073	0.576±0.079
C. owczarzaki	47.60±6.45	0.653±0.075	0.494±0.096

S. rosetta and *C. owczarzaki* showed a negative correlation between GC3s and *Nc*. With this, the lower *Nc* value (the more highly biased), the greater the GC3s value. The same general trend was seen in *M. brevicollis*, with the exception of a small number of genes highlighted, which did not fit the recognised distribution (Figure 4.3). 127 genes were identified to not fit the trend as seen in Figure 4.3, with genes found to be highly biased, with a lower GC3s value between 0.35 and 0.65. The small subset of genes was decreased in comparison to the 200 genes analysed in Southworth et al. (2018). Based on reciprocal BLAST data in Southworth et al. (2018), of the 127 genes investigated, 102 genes were not identified using reciprocal blast (80.3%), and only 4 genes were identified as putatively functional. The analysis supports that the 127 genes which do not fit the expected trend for *M. brevicollis* are in fact not genuine genes, and therefore false positives.

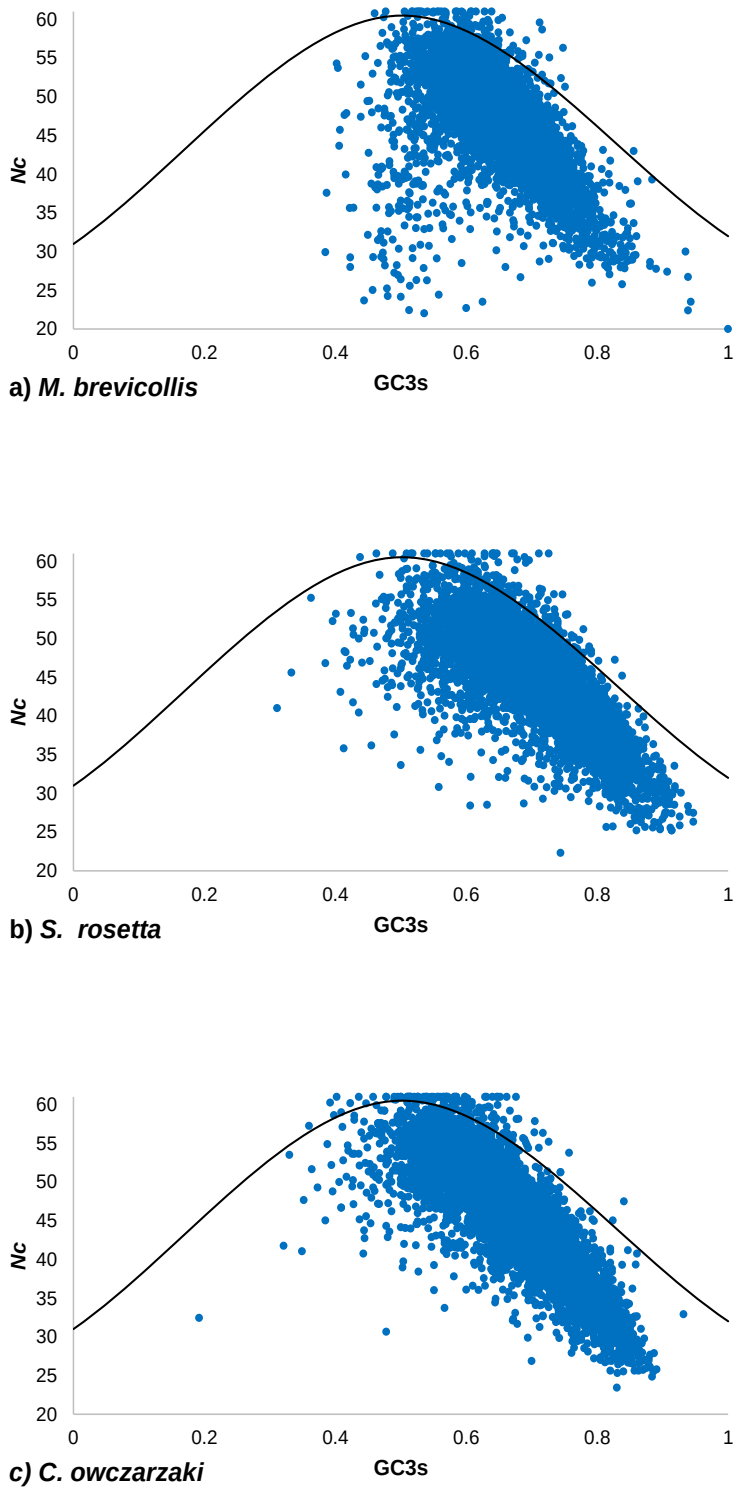


Figure 4.3: ***Nc* plots for the genes of the holozoan protists.** *Nc* values were plotted against GC3s for a) *M. brevicollis*; b) *S. rosetta* and c) *C. owczarzaki*. The modified equation, *Nc*=2+S+29/[S2+(1-S)2], from Wright (1990), with S=GC3s, was used to create the parabolic curve on each *Nc* plot (Wright, 1990).

For the *M. brevicollis* transcriptome, over 70% of genes have been annotated with KOG status, allowing for placement in gene categories and groups based on functionality (Table 4.3 and Figure 4.4). The KOG annotation work was ran to assess the function of genes in the *M. brevicollis* genome, to investigate potential trends between functionality and codon usage. As expression data was not publicly available for the choanoflagellate species, and therefore gene expression could not be determined for each bias category, the KOG analysis provided an additional avenue to explore to hypothesise the employment of optimal codons based on gene functionality. A significant difference was seen for all categories when comparing High-Mid values and High-Low values, except for genes involved in cellular process and signalling', where the data was not seen as significantly different (P=>0.05) (Table 4.3). The most significant difference in datasets was seen for the genes in the poorly characterised category; involving general function prediction only and unknown function (P=<0.0001) (Figure 4.4 and Table 4.3). Genes were enriched in the poorly characterised KOG category, for both mid and low bias genes, whereas very few high bias genes were assigned to this category (Figure 4.4). The majority of high biased genes were found to be assigned to information storage and processing, and metabolism categories. From the 25 categories, '(J) Translation, ribosomal structure and biogenesis' made up the greatest proportion of highly bias genes, with a significant difference observed when compared to mid and low bias genes (P=<0.0001).

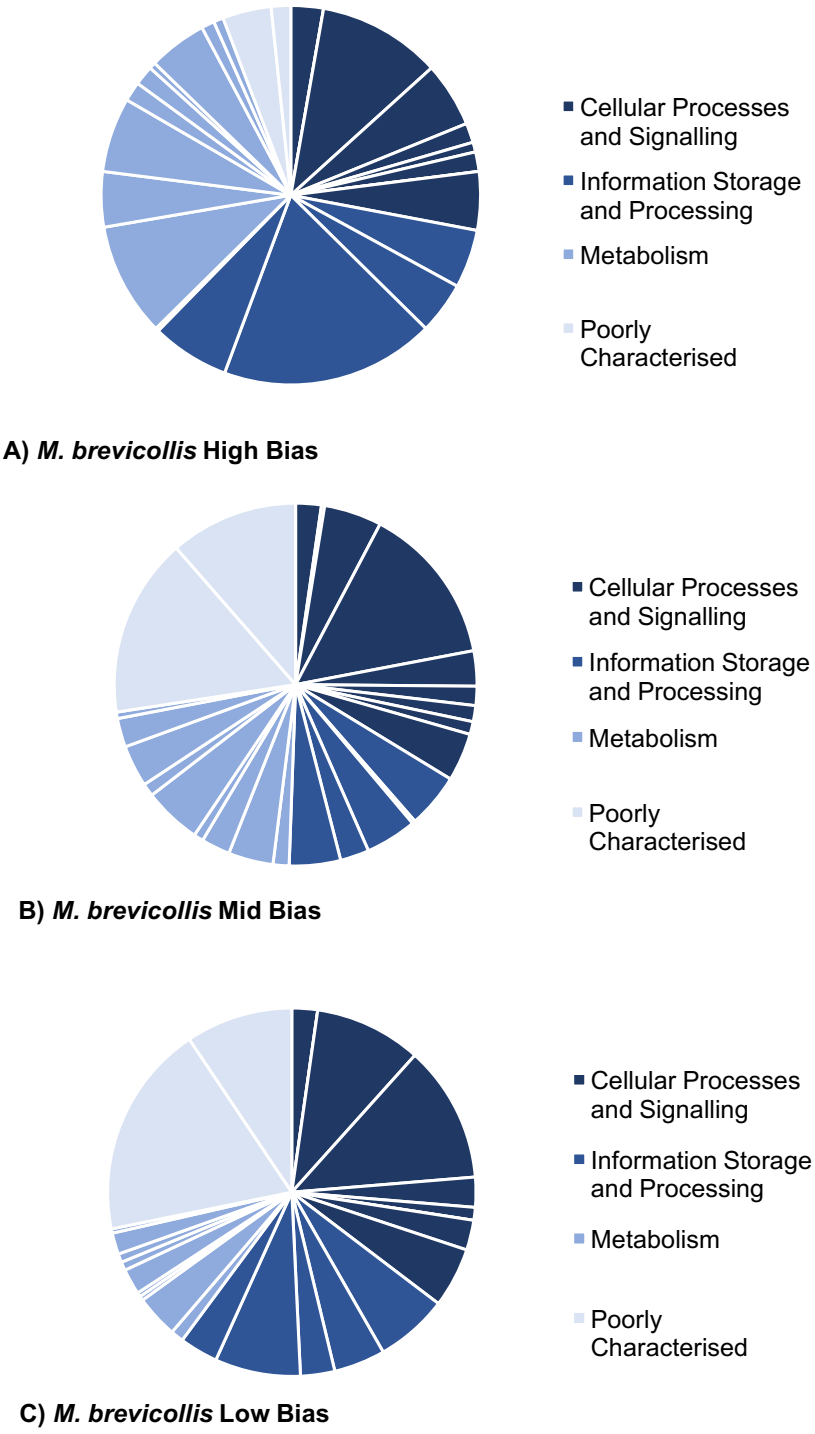


Figure 4.4: **Gene groupings and categories for each bias gene bias category in *M. brevicollis* based on KOG annotation.** Each gene category was ran through KOG annotation to investigate trends between codon usage bias, and gene category in the *M. brevicollis* genome. Category assignment is shown for A) High biased genes; (B) Mid biased genes; C) Low biased genes, which were annotated to be one of four KOG categories (Cellular Processes and Signalling; Information Storage and Processing; Metabolism; Poorly Characterised).

Table 4.3: **Gene ontology percentage distribution in the three bias categories, based on *Nc* in the genome of *M. brevicollis*.** The P value statistics were calculated using Fisher exact test (https://www.graphpad.com/quickcalcs/contingency1.cfm). The proportion of high genes were compared to the mid category, and the low category respectively. Data that was found to not be statistically significant was denoted with a '-'.

KOG Category	High Bias	Mid Bias	Low Bias	P values (High–Mid/High-Low)
Cellular Processes and Signalling	27.98	33.71	35.34	-
(M) Cell wall/membrane/envelope biogenesis	2.77	2.29	2.26	-
(N) Cell motility	0.00	0.29	0.00	-
(O) Posttranslational modification, protein turnover, chaperones	10.53	5.14	9.40	0.01/-
(T) Signal transduction mechanisms	5.54	14.29	12.03	0.0001/0.005
(U) Intracellular trafficking, secretion and vesicular transport	1.66	3.14	2.63	-
(V) Defence mechanisms	0.83	1.71	1.13	-
(W) Extracellular structures	1.66	1.43	0.00	-
(Y) Nuclear structure	0.00	1.14	2.63	-
(Z) Cytoskeleton	4.99	4.29	5.26	-
Information Storage and Processing	34.63	16.86	24.81	0.0001/0.02
(A) RNA processing and modification	4.99	4.86	6.39	-
(B) Chromatic structure and dynamics	4.43	0.29	4.51	0.0002/-
(J) Translation, ribosomal structure and biogenesis	18.28	4.57	3.01	0.0001/0.0001
(K) Transcription	6.65	2.57	7.52	-
(L) Replication, recombination and repair	0.28	4.57	3.38	0.0001/0.0005
Metabolism	31.58	22.00	11.65	0.005/0.0001
(C) Energy production and conversion	9.70	1.43	1.13	0.0001/0.0001
(D) Cell cycle control, cell division, chromosome partitioning	4.71	4.00	3.76	-
(E) Amino acid transport and metabolism	6.37	2.57	0.38	0.02/0.0001
(F) Nucleotide transport and metabolism	1.66	0.86	0.38	-
(G) Carbohydrate transport and metabolism	1.66	5.14	2.26	0.01/-
(H) Coenzyme transport and metabolism	0.55	1.14	0.75	-
(I) Lipid transport and metabolism	4.99	3.71	0.75	-/0.002
(P) Inorganic ion transport and metabolism	1.11	2.57	1.88	-
(Q) Secondary metabolites biosynthesis, transport and catabolism	0.83	0.57	0.38	-
Poorly Characterised	5.82	27.43	28.20	0.0001/0.0001
(R) General function prediction only	4.16	16.00	18.80	0.0001/0.0001
(S) Function Unknown	1.66	11.43	9.40	0.0001/0.0001

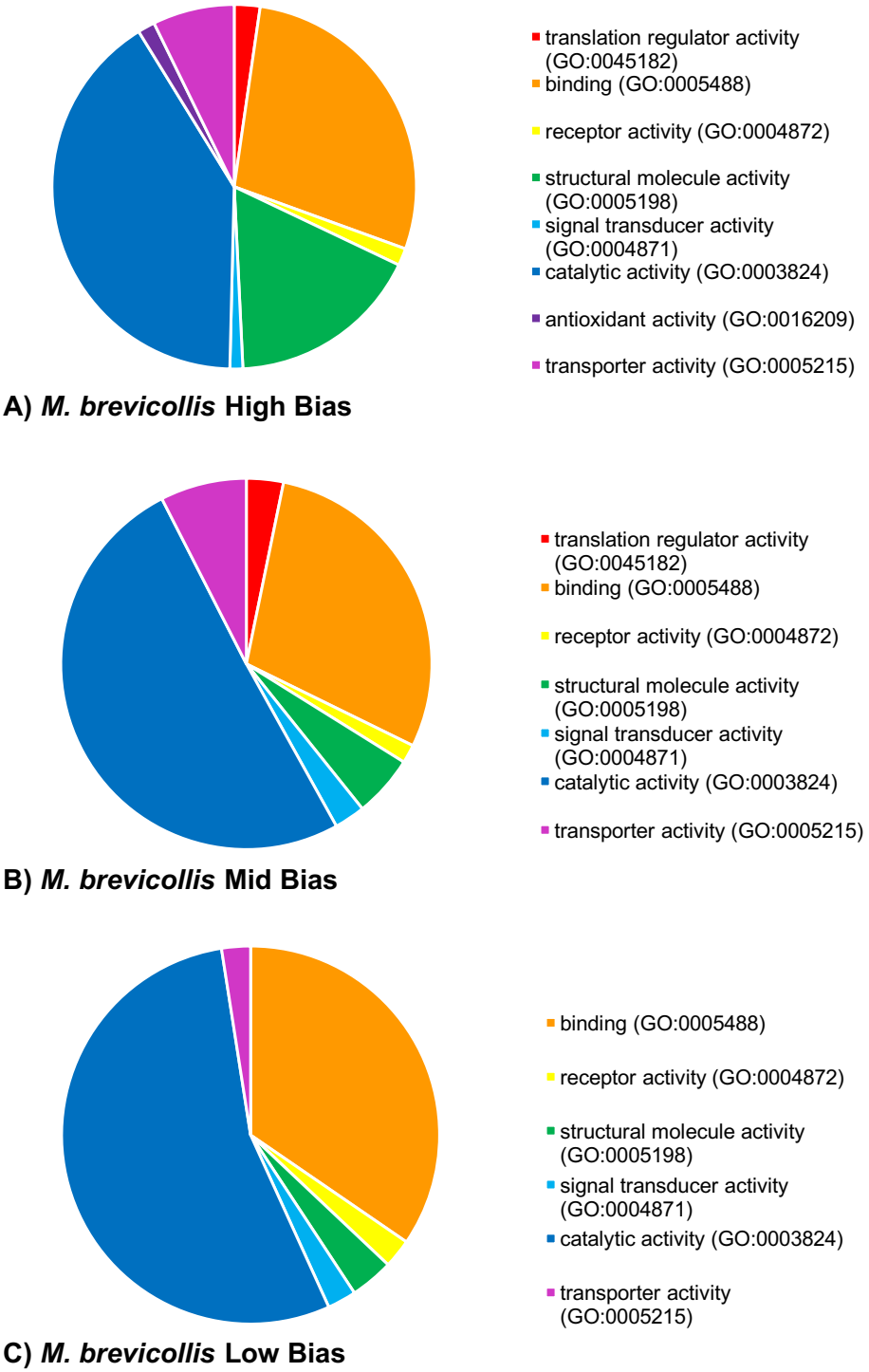


Figure 4.5: **Gene functionality for each bias gene bias category in *M. brevicollis* based on KOG annotation.** Each gene category was ran through KOG annotation to investigate trends between codon usage bias, and gene function in the *M. brevicollis* genome. Gene function assignment is shown for A) High biased genes; (B) Mid biased genes; C) Low biased genes, which were assigned to be one of six function categories (Binding; Receptor Activity; Structural Molecule Activity; Signal Transductor Activity; Catalytic Actvity and Transporter Activity.

The trends observed in gene enrichment for the KOG categories support the codon usage statistics which favoured translational efficiency in high bias genes for *M. brevicollis* (Table 4.2). With this, additional analysis on specific molecular function for the bias categories of *M. brevicollis* was reviewed. From the 361 high bias genes which were available on the KOG database, 262 were identified with molecular function (Figure 4.5). In contrast, a small proportion were annotated for molecular function in the mid and low bias categories (186/350;81/266). Genes were assigned to one of eight categories of molecular function, with catalytic activity highly enriched with genes from all three bias categories (Figure 4.5). Similarity was seen in functional distribution across the three bias categories, however, genes assigned to antioxidant activity were found to be absent in mid and low bias categories, as well as translation regulator activity in low bias genes (Figure 4.5).

In Southworth et al. (2018) for each holozoan species, the 5% highest, lowest and mid-biased genes were categorised, based on *Nc* values and used as an avenue for trends in codon usage for each of the three transcriptomes. From this, 1% categories were reviewed in this project to determine whether the patterns observed were still evident in a smaller subset of genes. For each gene category per species, codon usage statistics were ran for several variations of the bias-genes and compared to the 5% categories seen in Southworth et al. (2018). Both flanking DNA and introns were extracted per gene to determine whether evidence of mutation pressure was still evident, in the smaller subset of genes. To determine if codon usage in the holozoan species was driven by selection alone, signatures for mutation pressure were investigated here. Signature for mutation pressures include a relationship between GC3s and non-coding GC content, which would provide evidence for a mutational bias which is driving a GC-preference across the entirety of gene, rather than for any translational benefit.

4.3.2 Mutational bias hypothesis driving codon usage

The trends seen between GC3s and *Nc* were consistent with mutation pressure, however could be a signature for selection, mutation pressure or both, as drivers of codon usage bias (Southworth et al., 2018). Southworth et al. (2018) extracted the flanking DNA and introns for each gene in each bias category were measured for GC-content and GC3s value, to address the hypothesis, highly biased genes show bias towards GC across all nucleotide positions, compared to low bias genes, driven by mutational pressures. To investigate if the same findings were supported with smaller gene categories, mean values of GC3s for all genes, and GC content of flanking DNA and introns were calculated for 1% bias category for all three species.

For each holozoan species, GC3s decreased for each bias category, from high to low biased genes, for both the 5% and 1% biased categories as shown in Figure 4.6 and 4.7. This provided evidence that the original bias categories of 5% in Southworth et al. (2018) was accurate, as similar values were found when analyses were repeated with 1% bias categories. GC content of both flanking DNA, and intronic DNA were found to vary across the three species (Figure 4.7). *M. brevicollis* and *S. rosetta* were found to have the highest value of GC content for flanking and intronic DNA, but no significant difference was seen between bias categories (Table 4.4 and Figure 4.7). GC content for *C. owczarzaki* was lower than the values seen for the other two species, and the trend observed for each bias category was the opposite to the decrease seen for GC3s; as the level of bias increased, GC content was found to increase from 0.44±0.044 (High bias) to 0.47±0.043 (Low bias) (Table 4.4).

The three species stop codons were also found to show no evidence for GC-bias, with a preference shown towards codon of UAA found for all highly biased genes. The absence of GC bias in non-coding DNA supports that mutational pressure, towards guanine and cytosine, is not a driver to explain the variation seen in GC3s values Table 4.4). With this, signatures for selection were explored to determine if the prediction of selection bias could explain the data shown.

Table 4.4: **Flanking DNA and intronic GC content and GC3s value for 5% and 1% three bias categories for three holozoan species.**

Species	GC3s	Flanking GC Content (±sd)	Intron GC Content (±sd)
5% Categories			
M. brevicollis			
High Bias	0.72±0.109	0.53±0.043	0.54±0.049
Mid Bias	0.64±0.031	0.53±0.040	0.54±0.031
Low Bias	0.58±0.043	0.52±0.049	0.53±0.040
S. rosetta			
High Bias	0.82±0.054	0.54±0.030	0.54±0.036
Mid Bias	0.70±0.044	0.52±0.031	0.53±0.023
Low Bias	0.61±0.054	0.52±0.034	0.53±0.026
C. owczarzaki			
High Bias	0.80±0.051	0.44±0.042	0.47±0.055
Mid Bias	0.64±0.038	0.45±0.048	0.48±0.041
Low Bias	0.55±0.051	0.46±0.047	0.49±0.042
1% Categories			
M. brevicollis			
High Bias	0.72±0.162	0.52±0.048	0.53±0.058
Mid Bias	0.63±0.033	0.53±0.033	0.55±0.032
Low Bias	0.58±0.048	0.52±0.052	0.53±0.042
S. rosetta			
High Bias	0.86±0.058	0.53±0.0256	0.54±0.031
Mid Bias	0.70±0.043	0.52±0.033	0.53±0.024
Low Bias	0.59±0.056	0.53±0.033	0.53±0.026
C. owczarzaki			
High Bias	0.84±0.030	0.44±0.044	0.48±0.062
Mid Bias	0.64±0.043	0.45±0.049	0.47±0.037
Low Bias	0.55±0.056	0.47±0.043	0.49±0.056

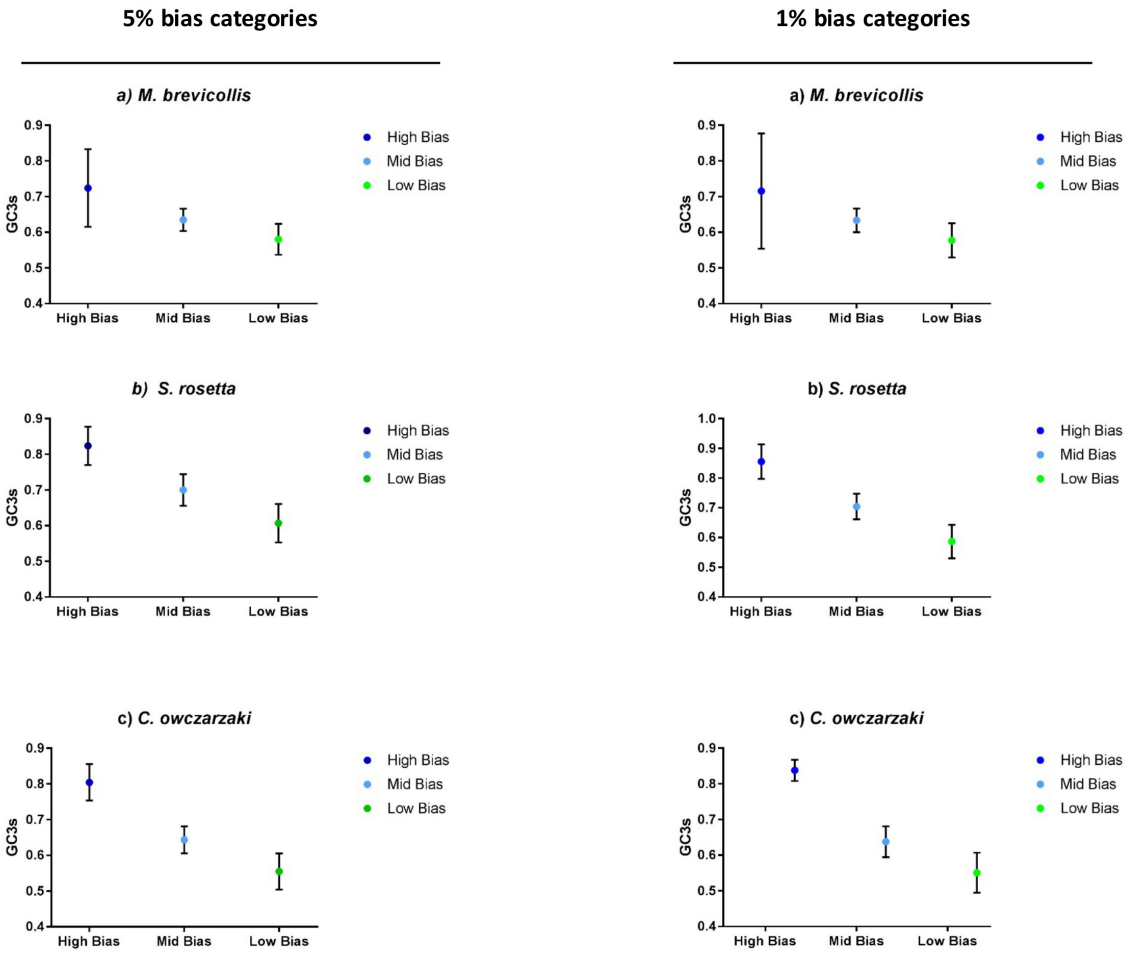


Figure 4.6: **Average GC3s value for both 5% and 1% bias categories for the three holozoan species**. Error bars were included for each bias category to show values of standard deviation per dataset.

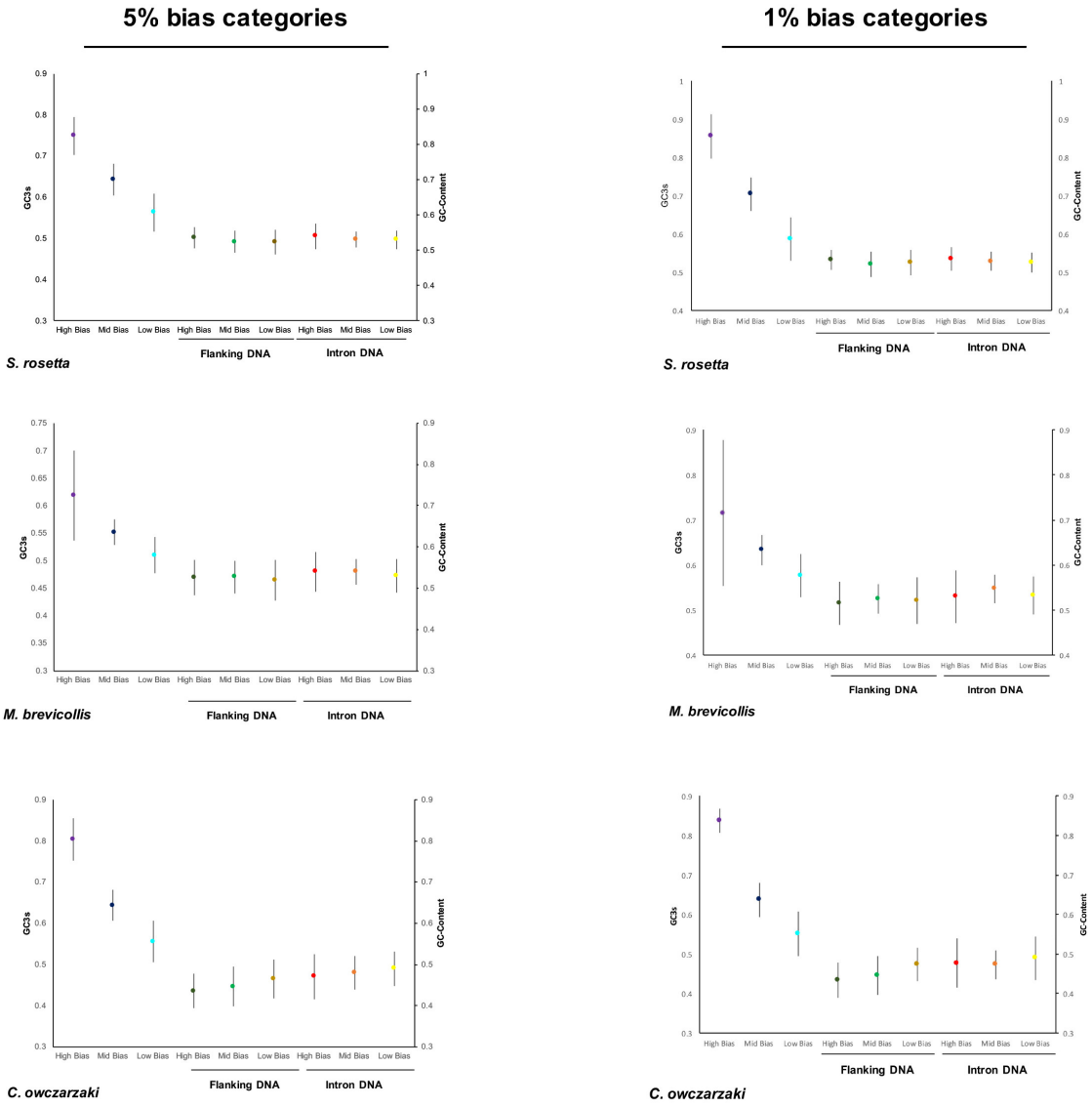


Figure 4.7: **Plots to show comparison of mean GC3s and non-coding DNA GC-content in *C. owczarzaki*, *S. rosetta* and *M. brevicollis* for both 5% and 1% bias categories].** GC3s values are plotted on the left axis (high bias is shown with a purple dot; mid bias, blue; low bias, light blue). Non coding GC content is plotted on the right y axis (Flanking DNA and Intron DNA are annotated on the x axis). Flanking DNA is annotated as follows (high bias, dark green; mid bias, light green; low bias, mustard yellow). Intron DNA is represented with the colours (high bias, red; mid bias, orange; low bias, yellow). Intronic data is taken from Southworth et al. (2018).

4.3.3 Optimal Codons and Major tRNA Genes in three holozoan species

The evidence against mutational bias led to the investigation of selection as the main driver of codon usage in the holozoan species (Southworth et al., 2018). Optimal codons were determined for each species using CodonW, and by the comparison of high and weakly expressed genes, for *S. rosetta* and *C. owczarzaki* due to transcriptome availability. The two methods employed to review optimal codons produced the same results, except for one codon in *C. owczarzaki*, and two in *S. rosetta* (Table 4.5). The complimentary results for both methods in the two species suggested that the optimal codons found for *M. brevicollis* were likely to be accurate.

The CodonW analysis found that the majority of optimal codons were found to be GC-ending, with CodonW estimates showing 60 of the 68 optimal codons having GC at the third position from across the three species (Southworth et al., 2018). It was also found that each species was found to have at least one optimal codon with uracil at the third position, with no species showing adenine-ending codons. Codon usage conservation was shown across all three species, with the majority of amino acids having only one optimal codon which was found to be identical across species (12/18 degenerate amino acids had conserved single optimal codons).Anticodons of major tRNA genes were identified for each holozoan species to show comparison with optimal codons (Table 4.5). Multicopy tRNA genes for a specific amino acid were defined as major tRNA genes, or if they were the most abundant amino acid representative. It was found that the total abundance of tRNA genes across all three species were similar, with a range from 114 - 139 genes. The choanoflagellate species showed the higher number of tRNA genes (Southworth et al., 2018). Identified tRNA genes are listed for all three species in Appendix C.

Optimal codons and major tRNA genes were found to correlate identically in all two-fold amino acids, except for lysine in *M. brevicollis* (Southworth et al., 2018). GGC was found to be the optimal codon for glycine in all three holozoan species, with the complementary anticodon found in the most abundant glycine tRNA genes. However, for the remaining degenerate amino acids, ranging from three to sixfold, very few optimal codons matched the major tRNA genes for standard Watson-Crick base pairing, with only six being complementary across the species (Southworth et al., 2018). The six complementary codons ended in uracil, however the encoded amino acids were found to show a strong preference for GC-ending optimal codons in the 5% (Southworth et al., 2018) and 1% categories of highly biased genes. The majority of the major tRNA genes coding for the 3-fold and above degenerate amino acids, were found to have adenine at the first position of the

Table 4.5: **Optimal codons assigned for the three holozoan species based on expression level and CodonW COA analysis.** Optimal codons were selected for each degenerate amino acid per species based on COA analysis performed by CodonW, and then compared to results based on expression data for *S. rosetta* and *C. owczarzaki*. An asterisk by a codon denotes that the codon would be optimal to the major tRNA gene if the adenosine at the 3rd wobble position underwent deamination modification. Optimal codons which complement the major tRNA genes are written in bold. Data based on (Southworth et al., 2018).

	S. rosetta		***M. brevicollis***		***C. owczarzaki***	
	Expression Level	**CodonW**	**Expression Level**	**CodonW**	**Expression Level**	**CodonW**
Amino Acid						
Phe	**UUC**	**UUC**	n/a	**UUC**	**UUC**	**UUC**
Leu	CUC*, **CUG**	CUC*, **CUG**	n/a	CUC*, **CUG, CUU**	CUC*	CUC*, **CUU**
Ile	AUC*	AUC*	n/a	**AUU**, AUC*	AUC*	AUC*,**AUU**
Val	GUC*	GUC*	n/a	GUC*,**GUU**	GUC*	GUC*
Ser	**UCU**, UCC*,UCG	AGC, **UCU**, UCC*,UCG	n/a	AGC, UCC*, UCG	UCC*,UCG	UCC*, UCG, **AGC**
Pro	CCC*, CCG	CCG	n/a	CCC*	CCC*,	CCC*
Thr	ACC, ACG	ACG	n/a	ACC	ACC,	ACC*
Ala	GCC*	GCC*	n/a	GCC*, **GCU**	GCC*	GCC*
Tyr	**UAC**	**UAC**	n/a	**UAC**	**UAC**	**UAC**
His	**CAC**	**CAC**	n/a	**CAC**	**CAC**	**CAC**
Gln	**CAG**	**CAG**	n/a	**CAG**	**CAG**	**CAG**
Asn	**AAC**	**AAC**	n/a	**AAC**	**AAC**	**AAC**
Lys	**AAG**	**AAG**	n/a	**AAG**	**AAG**	**AAG**
Asp	**GAC**	**GAC**	n/a	**GAC**	**GAC**	**GAC**
Glu	**GAG**	**GAG**	n/a	**GAG**	**GAG**	**GAG**
Cys	**UGC**	**UGC**	n/a	**UGC**	**UGC**	**UGC**
Arg	CGC*	CGC*	n/a	**CGU**, CGC*	**CGU**,CGC*	**CGU**, CGC*
Gly	**GGC**	**GGC**	n/a	GGU, **GGC**	**GGC**	**GGC**

anticodon. If tRNA molecules were deaminated, the adenosine would be modified to inosine, which was complementary to optimal codons with cytosine at the 3rd position. Evidence of deamination in eukaryotic taxa has been uncovered, with a greater abundance of tRNA modification in multicellular species, than unicellular organisms (Rafels-Ybern et al., 2017). tRNA modification of adenosine at the 1st position of the anticodon has been shown to deaminate the nucleotide base to inosine, which is complementary to adenine, cytosine and uracil. With this, it is supported that the codons may depend on the wobble effect hypothesis for binding to their complementary tRNA gene to take place. At the wobble position of the codon, the deamination of adenosine to inosine will enable base pairing to the optimal codons cytosine nucleotides (Southworth et al., 2018). The complement observed between optimal codons and host major tRNA genes is expected under the selection model, and therefore results uncovered were consistent with selection bias.

4.3.4 Evidence for translational accuracy in holozoan species

To explore whether selection is also being driven for translational accuracy, as well as efficiency, as previously described, the level of codon usage bias was compared per gene for domain codons, and non-domain codons, for each 5% bias category in (Southworth et al., 2018). The domains were identified using gene annotation available on NCBI, and the level of codon usage bias was determined using F_{op} via CodonW. The analysis was repeated here for the 1% bias categories to determine if similar trends were seen in more concise regions of bias (Figure 4.8).

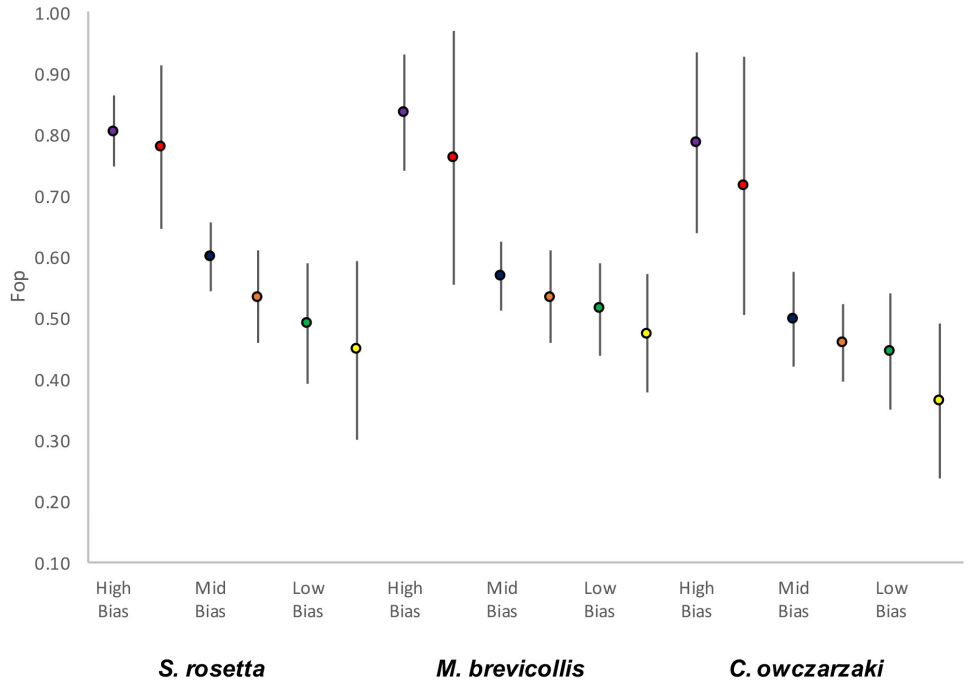


Figure 4.8: **Average Fop values for the three 1% bias categories per species based on domain encoding and non-domain encoding codons.** The 1% high bias Fop values are represented by the purple dots for domain codons, and red dots show the mean value for non domain codons. Mid-biased genes are annotated as blue dots for domain values, and orange for non-domain. Green dots represent mean value for domain codons, and yellow dot shows the mean value for non-domain codons for the low biased gene categories.

The mean Fop values for domain codons were found to be significantly higher than those calculated for non-domain codons for all three holozoan species (Figure 4.8 and Table 4.6). P values were calculated for each gene category using t test ($P<0.0001$).

Table 4.6: **Mean F_{op} values for each 5% and 1% bias categories for domain codons and non-domain codons in each of the three holozoan species.** Average F_{op} values and standard deviation is shown for each 5% and 1% bias category for domain and non-domain codons for *C. owczarzaki, S. rosetta* and *M. brevicollis* 5% were taken fronm Southworth et al. (2018).

Species	**Domain Codons F_{op} (±sd)**	**Non-Domain Codons Fop (±sd)**
5% Categories		
C. owczarzaki		
High Bias	0.740±0.102	0.648±0.163
Mid Bias	0.506±0.062	0.457±0.062
Low Bias	0.420±0.069	0.380±0.073
S. rosetta		
High Bias	0.751±0.068	0.716±0.125
Mid Bias	0.602±0.055	0.542±0.067
Low Bias	0.516±0.069	0.475±0.093
M. brevicollis		
High Bias	0.787±0.068	0.714±0.146
Mid Bias	0.576±0.047	0.537±0.059
Low Bias	0.519±0.062	0.471±0.066
1% Categories		
C. owczarzaki		
High Bias	0.786±0.147	0.715±0.212
Mid Bias	0.498±0.077	0.458±0.063
Low Bias	0.443±0.095	0.364±0.128
S. rosetta		
High Bias	0.805±0.057	0.779±0.133
Mid Bias	0.600±0.057	0.533±0.076
Low Bias	0.489±0.098	0.446±0.147
M. brevicollis		
High Bias	0.835±0.097	0.760±0.207
Mid Bias	0.568±0.056	0.534±0.075
Low Bias	0.514±0.075	0.474±0.098

4.4 Discussion and concluding remarks

4.4.1 Codon usage conservation across the holozoan protists

Codon usage has been reviewed in diverse taxa of the opisthokonts, with great emphasis on multicellular organisms (Jia and Xue, 2009; Lerat et al., 2003; Galtier et al., 2018). With this, Southworth et al. (2018) found a niche left to explore, with the genome availability of unicellular protists; *S. rosetta, C. owczarzaki* and *M. brevicollis*. Analysis of the holozoans provided an insight into whether selection and/or mutation bias drives codon usage in the unicellular organisms, with ancestral bias predicted as the closest living relatives to the metazoans. Southworth et al. (2018) found that the three protists presented with highly conserved codon usage bias, despite the distant ancestry between the holozoans (Parfrey et al., 2011). The analysis reviewed in (Southworth et al., 2018) was repeated here to show that patterns can be determined with a smaller dataset, and to assess whether decreasing the high, mid and low bias categories from 5% to 1% is still representative of the results seen previously.

As seen in Southworth et al. (2018), the holozoan species were found to have a strong bias towards codons and optimal codons, ending in GC, as well as shared optimal codon preferences, which supported conservation of codon usage bias. The bias categories were reconstructed to review if trends seen in Southworth et al. (2018) were repeated when comparing the original 5% categories to a smaller dataset of 1%. As seen for the 5% categories, there was a reduction in GC3s when comparing the gene categories, high to mid bias, and mid to low bias. Furthermore, the trend was not seen in non coding GC content across the bias categories, as reflected in Southworth et al. (2018). It can be determined that mutational pressure is not enough to drive codon usage bias, and this is also shown when analysing a smaller dataset of bias genes, which was employed for independent research during this study.

Southworth et al. (2018) noted that gene expression data for *S. rosetta* and *C. owczarzaki* provided evidence for translational efficiency in both species, with patterns being consistent with selection. As expression data is not available for the choanoflagellate *M. brevicollis*, the genes were reviewed by KOG category, and it was found that the majority of highly biased genes, which were categorised by *Nc*, were found to be putatively expressed protein translation genes in the host genome (Table 4.3). Also, it was found that in the three holozoan species, that optimal codons, identified by CodonW, and expression levels in *S. rosetta* and *C. owczarzaki* were found to be

complementary to major tRNA genes of the host species. This is another marker of translational efficiency (Ikemura, 1985), providing evidence that co-evolution of tRNA genes and highly biased genes has occurred for efficient protein translation in all three species.

As outlined in Southworth et al. (2018) in *S. rosetta* and *C. owczarzaki* it was found that genes that presented with levels of low expression possessed a lower frequency of optimal codons, which provided evidence for selection of translational efficiency. However, despite the lower abundance of optimal codons in genes of low expression, it was found that the genes were still showing evidence for translational accuracy, with functional domain regions of the genes having a greater enrichment of optimal codons, in comparison to non functional domains (Table 4.6). With this, if is supported that both translational efficiency and accuracy are major drivers of codon usage bias in filastereans and chaonoflagellate species.

Conservation of codon usage bias across the choanoflagellates species and filasterean, was an unexpected finding based upon the varied ancestry and life histories of the three holozoan species. Despite this, the greatest difference in the three protists was found between the two choanoflagellate species, rather than the choanoflagellates and the filastereans. As noted in Southworth et al. (2018),for the two choanoflagellates, it was found that codon usage bias was stronger in *S. rosetta* compared to *M. brevicollis*. If the difference in codon usage bias is accurate between the two choanoflagellates, the stronger bias in *S. rosetta* could be of a larger, effective population size. However, due to lack of population data for both choanoflagellate species, it can not be determined if strength of codon usage bias on the idealised population reviewed here was to represent the real population.

Furthermore, as reviewed by Carr et al. (2017), *M. brevicollis* is solely unicellular, whereas *S. rosetta* can form multicellular colonies, with an outcome of five varied cell types (three of which are multicellular), as well as the unicellular morphology (Dayel et al., 2011). With this, it would be plausible that the formation of ephemeral colonies in *S. rosetta* could influence codon usage bias, however limited data on rates of cell division in choanoflagellates leaves the hypothesis unresolved.

4.4.2 Conservation of codon usage in unicellular holozoans

Genomic analyses presented here were consistent with natural selection driving codon usage bias in the three holozoan protists. With this, the observed conservation was a contributor to the expansive knowledge of unicellular evolution prior to multicellularity. The findings supported that

selection was influencing both translational efficiency, and accuracy which drove the preference to GC-ending codons seen here. Furthermore, evidence for tRNA modification through deamination of adenosine to inosine was found to be present in the unicellular protists for higher degeneracy amino acids, and therefore presumably diverged in from the last common ancestor of Holozoa.

Chapter 5

Codon usage of transposable element families in three holozoan species

During the revision of this thesis, work described in this chapter was published in Southworth, J., Grace, C.A., Marron, A.O, Fatima, N. and Carr. M (2019). 'A genomic survey of transposable elements in the choanoflagellate *Salpingoeca rosetta* reveals selection on codon usage', *Mobile DNA* **10** (44), 1 – 19.

5.1 Introduction

5.1.1 Codon usage bias in TEs

As the genomic influence TEs have on a host species is well established, the relevance of studying codon usage in the mobile elements has been addressed (Shields and Sharp, 1989; Lerat et al., 2002; Jiang et al., 2006; Jia and Xue, 2009). Past studies on TE codon usage bias mainly have failed to detect the impact of natural selection; TE families have shown to predominantly show weak codon usage bias in a range of organisms, with a slight bias towards AT-ending codons (Lerat et al., 2002; Jia and Xue, 2009). An exception was found in the class I elements of the stramenopile genus, *Phytophora* (Jiang et al., 2006). The work outlined by Jiang et al. (2006) found that TE families of two *Phytophora* species, were found to shown a preference for GC-ending codons, which was reflective of the species host genes. It was also detailed that stronger codon usage bias was found in element families which were revealed with higher copy number when compared to families with low copy number (Jiang et al., 2006). As TE investigations have primarily focused on multicellular eukaryotes, the selection bias observed in the TE families of *Phytophora*, can not be determined as characteristic or unexpected for unicellular eukaryotes. The limitation outlined has

prompted the study here of TE codon usage in the three holozoan species; *S. rosetta*, *M. brevicollis* and *C. owczarzaki*. With the codon usage review detailed in Chapter 4 based on Southworth et al. (2018), the work here will illustrate whether TE codon usage reflects the patterns seen for host genes in the three species studied.

5.1.2 Codon usage bias in holozoan species

Patterns of relative synonymous codon usage for genes in *S. rosetta*, *M. brevicollis* and *C. owczarzaki* genomes were detailed in Chapter 4. Southworth et al. (2018) found that in all three holozoan species, the host genes were found to show bias toward GC-ending codons, as well as strong association between GC3s and *Nc* value; it was found that genes that were found to show stronger bias, determined by a low *Nc value*, were also likely to present with a higher GC3s value. Furthermore, codon usage seemed to be mainly driven by selection, with no evidence of mutation pressure on the host genes.

As the optimal codons in the three holozoan species were predominantly GC-ending, it is hypothesised that any selection employed for codon usage in the TE families uncovered in each species, should contrast with the observed weak bias towards AT-ending codons which have previously been found in the majority of mobile elements in eukaryotic species. The holozoans therefore provide an opportunity to determine if selection for codon usage is present in the TE families of a broader range of eukaryotes.

5.1.3 Experiment overview

Codon usage analyses were performed on the ORFs of all LTR retrotransposon and DNA transposon families identified in *S. rosetta*, *M. brevicollis* and *C. owczarzaki* detailed in Chapter 3, in order to determine whether selection and/or mutation were driving TE codon usage. The work outlined would allow comparison to be drawn between host genes and mobile elements, whether data supports the influence of natural selection driving TE codon usage, as seen in the species host genes.

5.2 Methods

5.2.1 Analysis of codon usage bias in transposable element families

TE sequences for *M. brevicollis* and *C. owczarzaki* were taken from Carr, Nelson, Leadbeater and Baldauf (2008) and Carr and Suga (2014). The method for *S. rosetta* TE annotation is outlined in Chapter 3. For each TE family in the holozoan genomes, values of *Nc*, Fop and GC3s were calculated for all coding sequences using CodonW 1.4.4 (Peden, 1999). The fop.coa file for all species was taken from Southworth et al. (2018); *S. rosetta* and *C. owczarzaki* fop.coa files were based on expression data, and *M. brevicollis* was based on CodonW analysis (Peden, 1999). In *S. rosetta*, for *Srospv6*, *Sroscv4* and *Sroscv5* the *gag* and *pol* ORFs are separated and therefore were concatenated to provide comparable values with the other LTR retrotransposon families. The same was done for *Cocv1*, *Cocv2*, *Cocv3* and all *CoL* elements in *C. owczarzaki*. Values from *S. rosetta*, *C. owczarzaki* and *M. brevicollis* host genes were taken from Southworth et al. (2018).

The GC content of non-coding DNA from each TE family was also determined in CodonW (Peden, 1999). Codons from overlapping ORFs were excluded from *Mbcv*,*Cocv1-3* and *CoL1* prior to determining codon usage statistics. In order to assess the contribution of selection on translational accuracy, the codons which encode amino acids in functional domains were separated from those that encode non-domain amino acids, to assess whether the frequency of optimal codons was greater in highly expressed genes (Southworth et al., 2018). Domain regions were identified by analysing each Pol and Transposase protein sequences in BLASTp through NCBI (Sayers et al., 2009). Fop values were then determined for the domain and non-domain codons in each family. Non coding data was also collated by running non-coding DNA for each TE family with CodonW, to determine GC content. For LTR retrotransposons, 5' LTRs and untranslated regions (UTRs) were concaternated, per family, to assess non-coding GC content. The same method was employed for DNA transposon families, except 5' ITRs were included, as well as UTRs and introns where applicable.

5.2.2 Optimal codons and major tRNA screening

Abundant codons for each TE family were generated by correspondence analysis (COA), using relative synonymous codon usage (RSCU), with employment of default parameters. The abundant codons identified were then assessed comparatively to the anticodons of the tRNA genes for each

holozoan species to determine if the TE families were found to be consistent with selection. Host gene optimal codons, and genome tRNA genes were identified in Chapter 4.

5.2.3 Determining TE family expression levels

The protocol to determine *S. rosetta* TE family expression is outlined in Chapter 3, with the employment of SMALT v. 0.2.6 (Ponstingl, 2014). Expression levels for the *C. owczarzaki* TE families were taken from (Carr and Suga, 2014).

5.3 Results

5.3.1 Transposable element families show biased synonymous codon usage in three holozoan species

For the TE codon usage analysis, a remarkably similar association between GC3s and *Nc* was observed for the TE ORFs of *S. rosetta* and *C. owczarzaki* (Appendix D). As only three LTR retrotransposon families were found in *M. brevicollis*, it could not be determined if the TEs followed the same trend (Appendix D).

In contrast to earlier studies on TE codon usage, which have reported an AT preference in a broad range of eukaryotes (Lerat et al., 2002; Jia and Xue, 2009), the TE families all exhibit an excess of GC-ending codons (Table 5.1 and 5.2). A stronger GC3s bias is observed in the LTR retrotransposons than for DNA transposons for *S.rosetta* and *C. owczarzaki* (Figure 5.1 and 5.2 and Table 5.1 and 5.2). GC-bias in synonymous 3rd positions in LTR retrotransposons is reflected in their stronger codon usage bias. A strong negative correlation was seen between GC3s and *Nc* of the TEs in *S. rosetta*, when reviewed separately to host genes, with GC3s value increased in families with stronger codon usage bias (a smaller value of *Nc*) (R^2= 0.831) (Figure 5.1). A much weaker negative relationship was observed in *C. owczarzaki* (R^2= 0.486), with no relationship seen in *M. brevicollis* (Figure 5.2 and 5.3).

For the *S. rosetta* TEs, the mean *Nc* for LTR retrotransposon families (45.40±5.59) is similar to the genome-wide mean *Nc* (44.79±5.37); however, the mean *Nc* for DNA transposon families (52.19±4.37) was found to be higher, and therefore evidence of less bias for the class II elements (Table 5.1). Similarly, in *C. owczarzaki* the mean *Nc* value for LTR and non-LTR retrotransposon families (46.34±3.87;44.49±1.76) was close to the genome-wide mean *Nc* value (47.60±6.45) (Table 5.2). The DNA transposons showed a higher mean *Nc* value, as seen in *S. rosetta*, of 52.84±2.69, and therefore showed less bias. For *M.brevicollis*, the mean *Nc* value for the genome, reflected that seen for the TE families, with a genome value of 48.05±5.62, and a similar TE value of 48.83±1.74 (Table 5.2).

Table 5.1: **Codon usage statistics for the LTR retrotransposon and DNA transposon families in the *S. rosetta* genome.** Mean values for GC3s, Nc and Fop are listed for each class of TE.

Family	GC3s	*Nc*	Fop
LTR retrotransposon Families			
Sroscv1	0.634	49.14	0.541
Sroscv2	0.627	45.28	0.56
Sroscv3	0.592	55.5	0.509
Sroscv4	0.763	41.99	0.648
Sroscv5	0.702	46.23	0.592
Srosgyp1	0.791	40.27	0.667
Srosgyp2	0.768	40.14	0.63
Srospv1	0.682	46.83	0.574
Srospv2	0.784	38.09	0.661
Srospv3	0.804	37.1	0.686
Srospv4	0.57	48.09	0.407
Srospv5	0.639	49.19	0.509
Srospv6	0.613	52.27	0.477
	Mean=0.690±0.083	**Mean=45.40±5.59**	**Mean=0.574±0.08**
Transposon Families			
SrosH	0.743	44.84	0.595
SrosM	0.609	53.73	0.428
SrosT1	0.71	49.44	0.589
SrosT2	0.647	50.8	0.525
SrosT3	0.53	58.55	0.426
SrosTig1	0.578	54.9	0.493
SrosTig2	0.583	53.05	0.465
	Mean=0.629±0.080	**Mean=52.19±4.37**	**Mean=0.503±0.070**

Table 5.2: **Codon usage statistics for the all transposable element families in the *M. brevicollis* and *C. owczarzaki* genome.** Mean values for GC3s, *Nc* and Fop are listed for each class of TE.

Family	GC3s	Nc	Fop
M. brevicollis			
Mbcv	0.659	48.45	0.588
Mbpv1	0.618	50.73	0.570
Mbpv2	0.606	47.31	0.583
	Mean=0.627±0.028	**Mean=48.83±1.74**	**Mean=0.580±0.009**
C. owczarzaki			
LTR retrotransposon Families			
Cocv1	0.698	45.33	0.442
Cocv2	0.600	51.78	0.436
Cocv3	0.698	48.59	0.446
Cocv4	0.747	43.97	0.562
Cocv5	0.620	42.03	0.572
	Mean=0.672±0.061	**Mean=46.34±3.87**	**Mean=0.491±0.070**
Non-LTR retrotransposon Families			
CoL1	0.711	46.08	0.505
CoL2	0.714	42.34	0.57
CoL3	0.694	44.03	0.534
CoL4	0.695	45.9	0.551
	Mean=0.704±0.010	**Mean=44.59±1.76**	**Mean=0.540±0.028**
DNA transposon Families			
Com1	0.635	54.59	0.428
Com2	0.602	53.79	0.46
Cop1	0.636	53.96	0.471
Cop2	0.656	51.73	0.492
Cop3	0.627	54.28	0.434
Cop4	0.582	56.33	0.401
Cop5	0.634	54.72	0.44
CoTc1	0.662	51.53	0.521
CoTc2	0.714	46.99	0.575
Cobalt1	0.597	52.86	0.456
Cobalt2	0.616	52.23	0.488
Cobalt3	0.659	47.82	0.526
CoCACTA1	0.512	53.33	0.318
CoCACTA2	0.626	55.64	0.471
	Mean=0.626±0.046	**Mean=52.84±2.69**	**Mean=0.463±0.06**

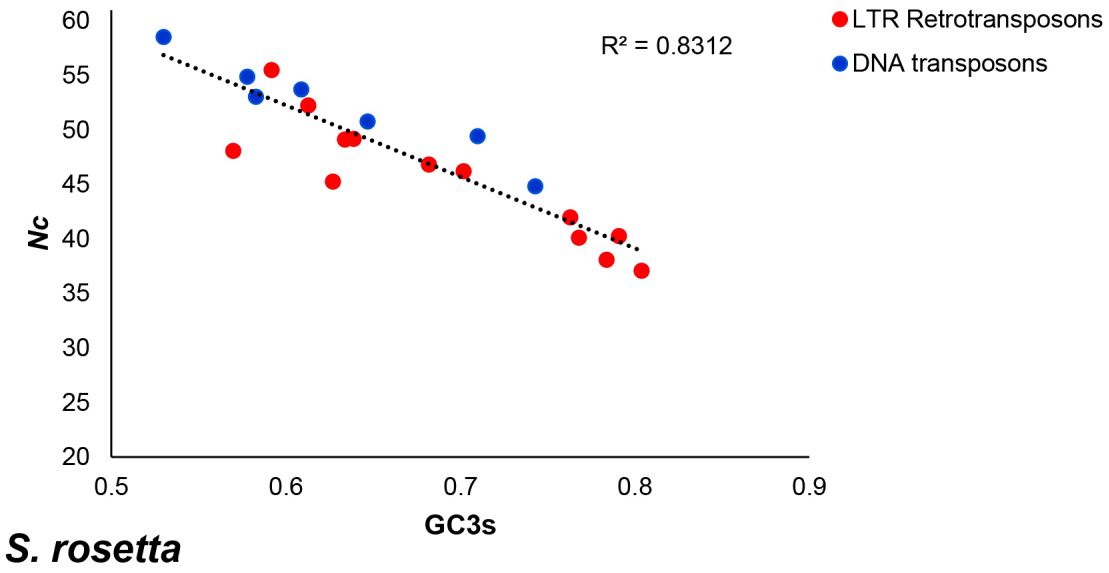


Figure 5.1: **Relationship between GC3s and *Nc* for the 20 TE families in the *S. rosetta* genome**. A linear trend line, with R^2 value was added to the graph to assess the strength of the negative relationship. LTR retrotransposons are shown in red, and DNA transposon families in blue.

The LTR retrotransposons in the choanoflagellate species showed an association between copy number and the strength of codon usage bias (Figure 5.4 and 5.5), with high copy number families tending to show stronger codon usage biases than low copy number families. *M. brevicollis* families were found to have the greatest association between *Nc* and copy number (R^2=0.994), whereas *S. rosetta* showed a weaker negative relationship (R^2=0.314) (Figure 5.4). No relationship was observed between strength of bias and copy number for both LTR and non-LTR retrotransposon families in the filasterean species (R^2=0.018) (Figure 5.4).

A similar trend was seen between Fop and copy number, with a positive correlation observed between family copy number and Fop in both *S. rosetta* and *M. brevicollis* (R^2=0.306;0.604), whereas no correlation was seen for *C. owczarzaki* (R^2=0.0035) (Figure 5.5). However, within *C. owczarzaki*, if you categorise the Class I elements further to LTR retrotransposons and non-LTR retrotransposons, the four non-LTR retrotransposon families also show a strong positive relationship between Fop and copy number (R^2 = 0.962, Appendix D). In all three holozoan species, the DNA transposons showed to have weaker/no such association between copy number and codon usage bias (Fop/*Nc*) (Appendix D).

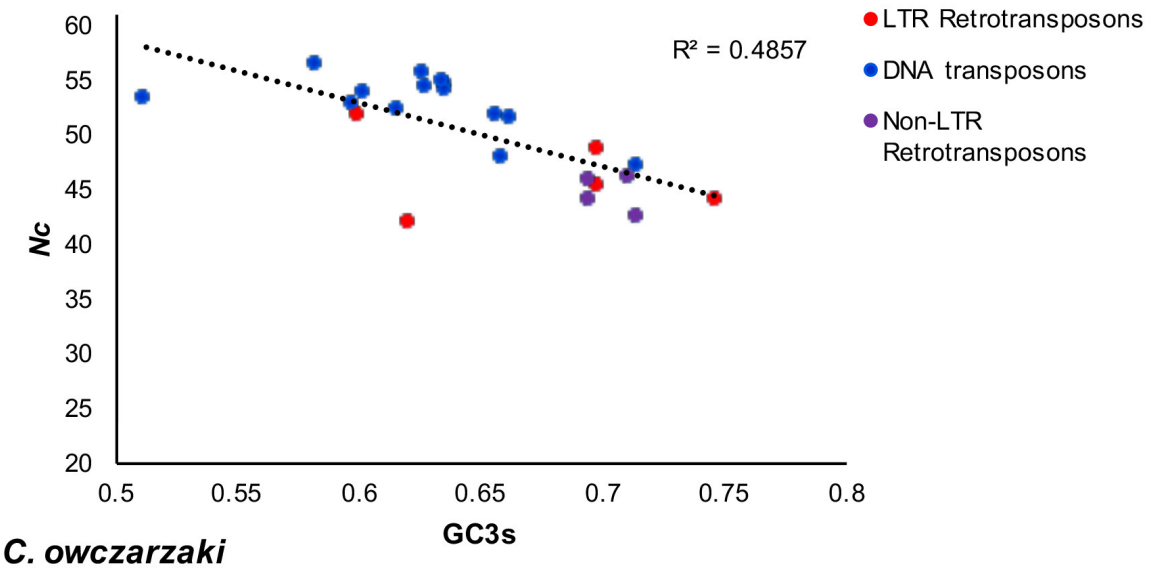


Figure 5.2: **Relationship between GC3s and *Nc* for the 23 TE families in the *C. owczarzaki* genome**. A linear trend line, with R^2 value was added to the graph to assess the strength of the negative relationship. LTR retrotransposons are shown in red, and DNA transposon families in blue.

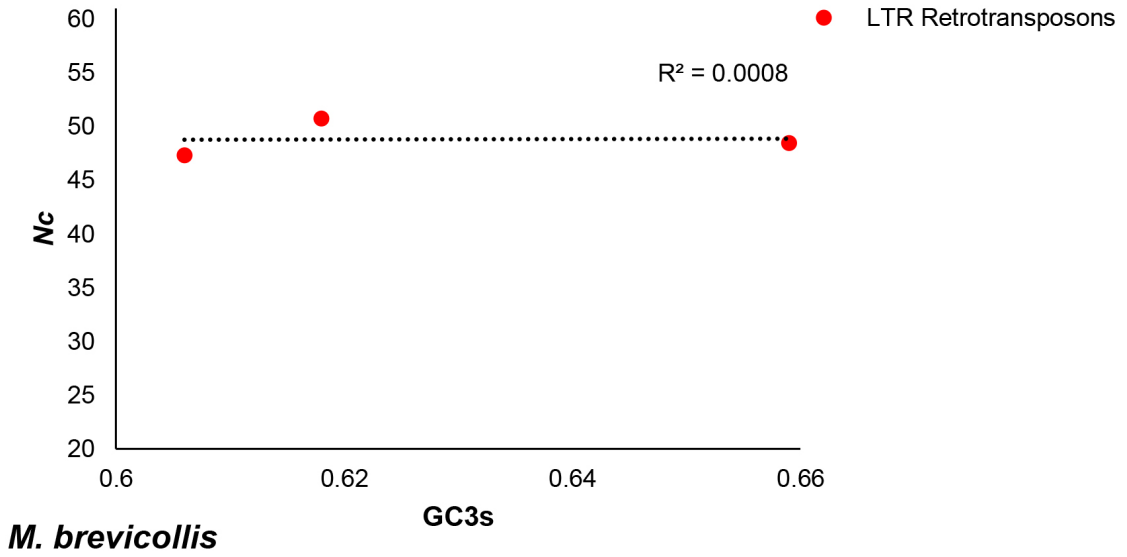


Figure 5.3: **Relationship between GC3s and *Nc* for the 3 TE families in the *M. brevicollis* genome**. A linear trend line, with R^2 value was added to the graph to assess the strength of the relationship.

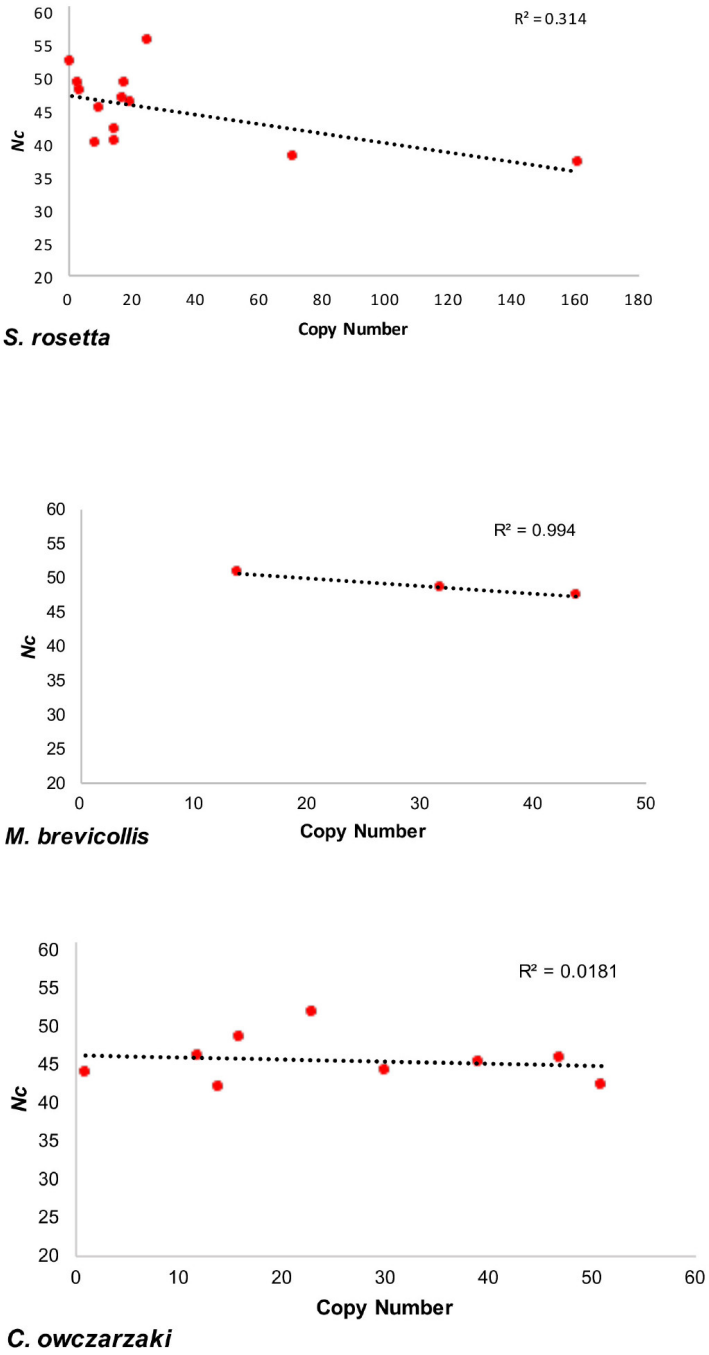


Figure 5.4: **Relationship between copy number of LTR retrotransposon families and effective number of codons (*Nc*).** Copy number of LTR retrotransposon families was plotted against *Nc* for all three holozoan species investigated.

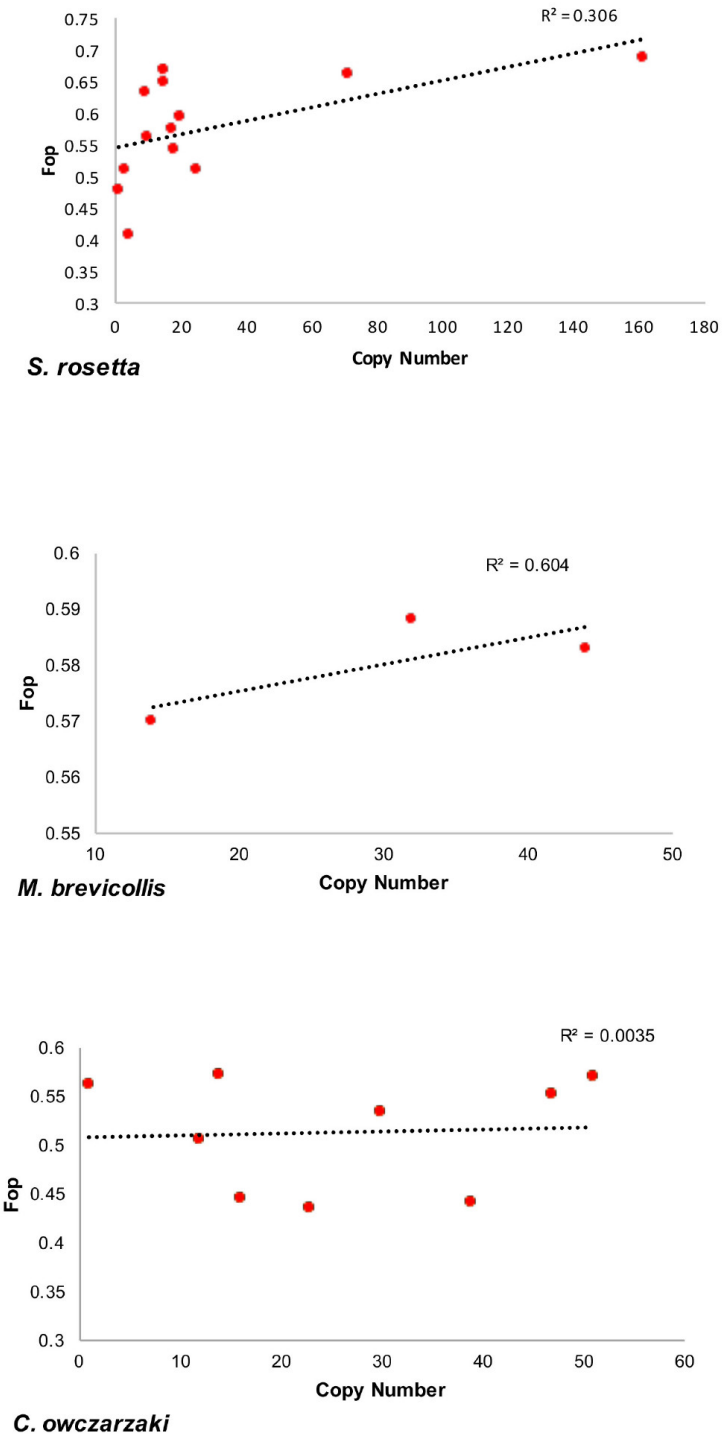


Figure 5.5: **Relationship between copy number of LTR retrotransposon families and frequency of optimal codons (Fop).** Copy number of LTR retrotransposon families was plotted against Fop for all three holozoan species investigated.

5.3.2 Evaluating the role of mutation pressure on codon usage bias

The codon usage of TE families in *S. rosetta*, *M. brevicollis* and *C. owczarzaki* is likely to be driven by one of three mechanisms; as discussed in Chapter 3. For all three species, genetic drift cannot be discarded, however it would appear that drift as the main driver of codon usage is unlikely, as the mechanism is a random process, and all TE families show an excess of GC-ending codons, which does not appear random (Table 5.1). The exception to this was *Srospv4* in *S. rosetta*, where the family was found to have a GC3s value of <0.58, and Fop value of 0.407, which was found to be lower than the other TE families in all species investigated (Table 5.1). Southworth et al. (2018) uncovered the tRNA genes for the three holozoan species, allowing the identification of amino acids which are bias towards using the codons that complement the major tRNA gene transcripts. Abundant codons were determined for each TE family and compared to the major tRNA genes of the host species identified in Chapter 4. In all three species TE families, the majority of amino acids employed preferred codons that complement the major tRNA genes (Appendix D) (Southworth et al., 2018), suggesting that codon usage in all families has been shaped by the host translational machinery in order to facilitate efficient translation.

If mutation pressure was driving codon usage bias from AT to GC would be expected to influence non-coding TE DNA, in addition to synonymous 3rd codon positions. It was found that for *S. rosetta* there was no observed relationship between the non-coding GC-content and GC3s across all of the TE families (R^2=0.033) (Figure 5.6; A). The families were categorised to observe relationships within classes and groups, as well as for all mobile elements, as the different families may be under different constraints within the host genome. The *copia-like* families do not show a relationship between GC3s and non-coding GC-content (R^2=4E-07), indicating that mutation pressure is not a major driver of codon usage bias within the six families (Figure 5.6; B). However, a positive correlation was observed between non-coding GC-content and GC3s for the chromoviral families (R^2=0.519), therefore it appears that within these families mutation pressure towards guanine and cytosine contributes towards codon usage bias as both coding and non-coding DNA exhibit similar GC-bias (Figure 5.6; C). In contrast, the transposon families were found to show no relationship between non-coding GC-content and GC3s (Figure 5.6; R^2=0.043), indicating that mutational pressure does not drive codon usage bias for these families. Firm conclusions are difficult to draw from the non-chromoviral *gypsy-like* families, *Srosgyp1* and *Srosgyp2*, as there were only two families. The families show an inverse relationship between non-coding GC-content and GC3s,

which does not appear to be consistent with mutational pressure driving codon usage (Table 5.1) (data not shown).

For *C. owczarzaki*, similar patterns were seen between non coding GC-content and GC3s. As with *S. rosetta*, no correlation is seen for all TE families between GC3s and non-coding GC content (R^2=0.081) (Figure 5.7). In contrast, the chromoviral families were found to show the strongest positive trend (R^2=0.650), with non-LTR retrotransposons showing weak associations between GC3s and non-coding GC, and DNA transposons showing no association (R^2=0.350;0.014). With this, as seen in *S. rosetta* it is suggested that the chromoviral families mutation pressure towards guanine and cytosine which contributes to codon usage bias, and that the other TE families are not subjected to the same proposed mutational bias.

Only three TE families are present in *M. brevicollis*, and therefore the trends observed are speculative. However, a positive correlation was seen between GC3s and non coding GC content, which supports that all families in the choanoflagellate are driven by mutation pressure of cytosine and guanine, as seen in the chromoviral families in *S. rosetta* and *C. owczarzaki* (R^2=0.857) (Figure 5.8).

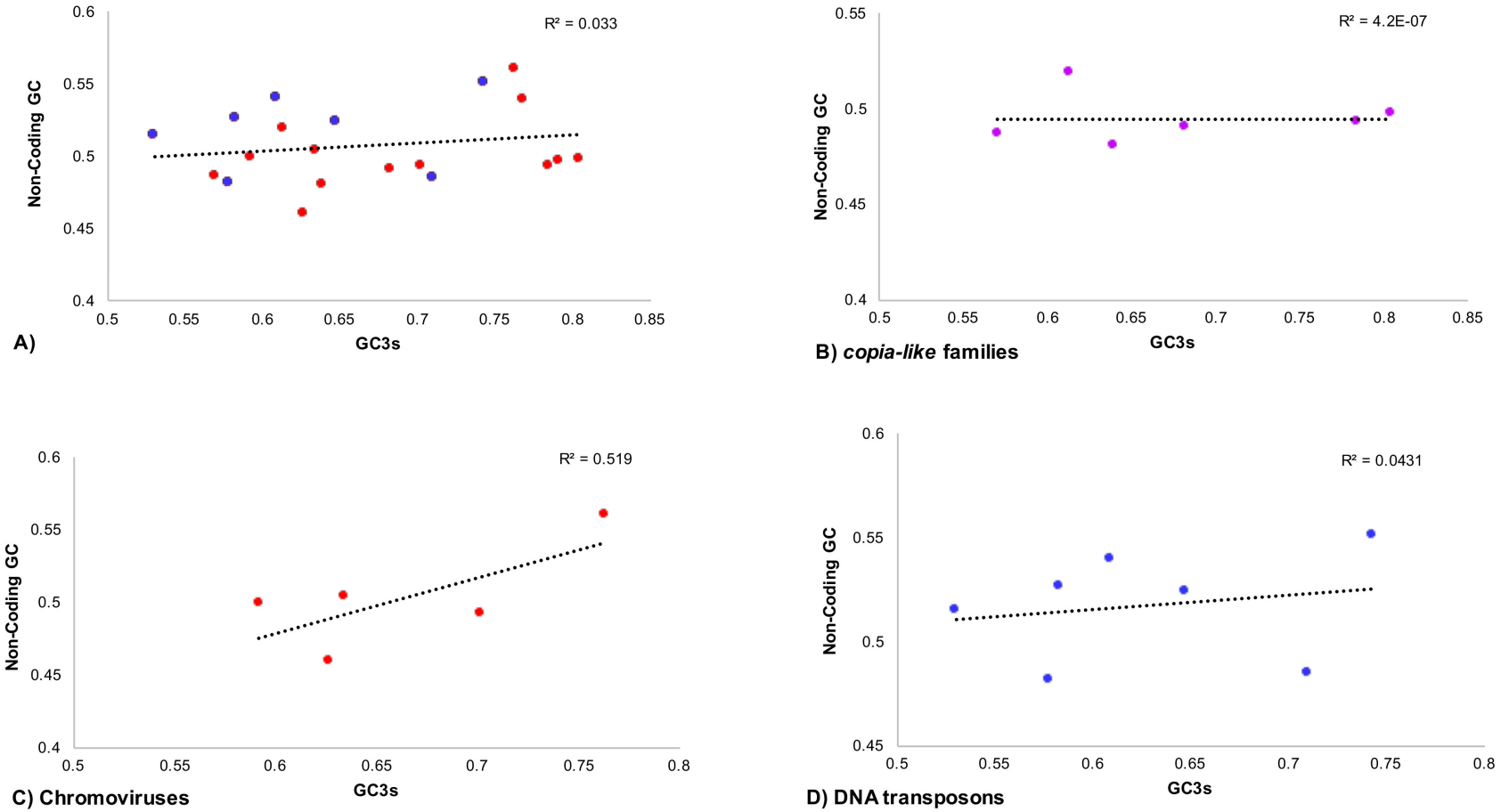


Figure 5.6: **Relationships between GC3s and GC- content of non-coding DNA for transposable element families in *S. rosetta* genome** GC3s values for all 20 transposable element families was plotted against GC content value of non-coding DNA for each family. Non-coding DNA included LTRs, ITRs, UTRs and introns, which were family specific. A) All TE families; B) *copia*-like families; C) Chromoviral families; D) DNA transposon families.

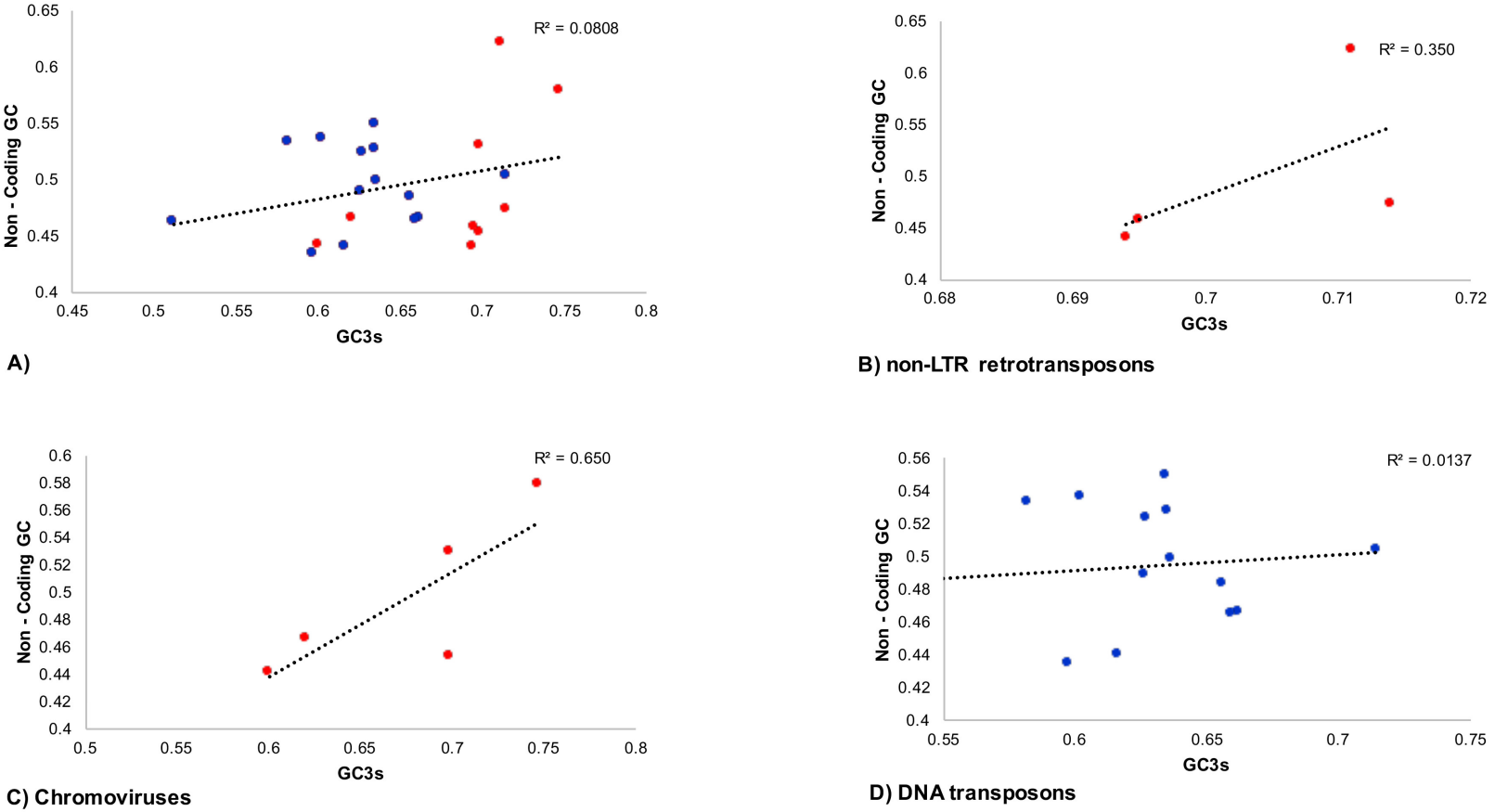


Figure 5.7: **Relationships between GC3s and GC- content of non-coding DNA for transposable element families in *C. owczarzaki* genome** GC3s values for all 23 transposable element families was plotted against GC content value of non-coding DNA for each family. Non-coding DNA included LTRs, ITRs, UTRs and introns, which were family specific. A) All TE families; B) *copia*-like families; C) Chromoviral families; D) DNA transposon families.

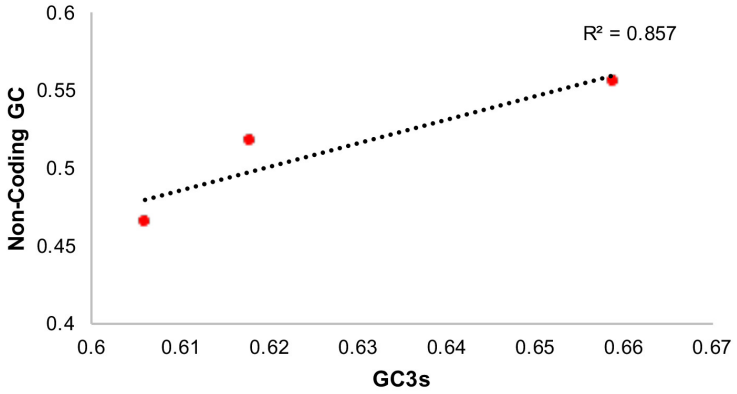


Figure 5.8: **Relationships between GC3s and GC- content of non-coding DNA for transposable element families in *M. brevicollis* genome** GC3s values for all 3 transposable element families was plotted against GC content value of non-coding DNA for each family. Non-coding DNA included LTRs and UTRs which were family specific.

5.3.3 Evaluating the role of natural selection on translational efficiency

Of the TE families within the *S. rosetta*,*C. owczarzaki* and *M. brevicollis* genomes, it was found that only the chromoviruses showed evidence for codon usage bias being influenced by mutation pressure, and LTR retrotransposons in *M. brevicollis*. However, values of Fop indicate that selection may be operating upon TE codon usage in *S. rosetta*, as both LTR retrotransposon and DNA transposon families are enriched for optimal codons (mean Fop = 0.574± 0.08;0.503± 0.070, Table 5.1). The same trend is seen in both *C. owczarzaki* and *M. brevicollis*. All class of TE in *C. owczarzaki* were found to be enriched for optimal codons when reviewing Fop values for LTR retrotransposon, non-LTR retrotransposon and rransposon families (mean Fop = 0.491±0.070; 0.540±0.028; 0.463±0.06) (Table 5.2). *M. brevicollis* showed the highest value of Fop, with a mean score of 0.580±0.009 (Table 5.2).

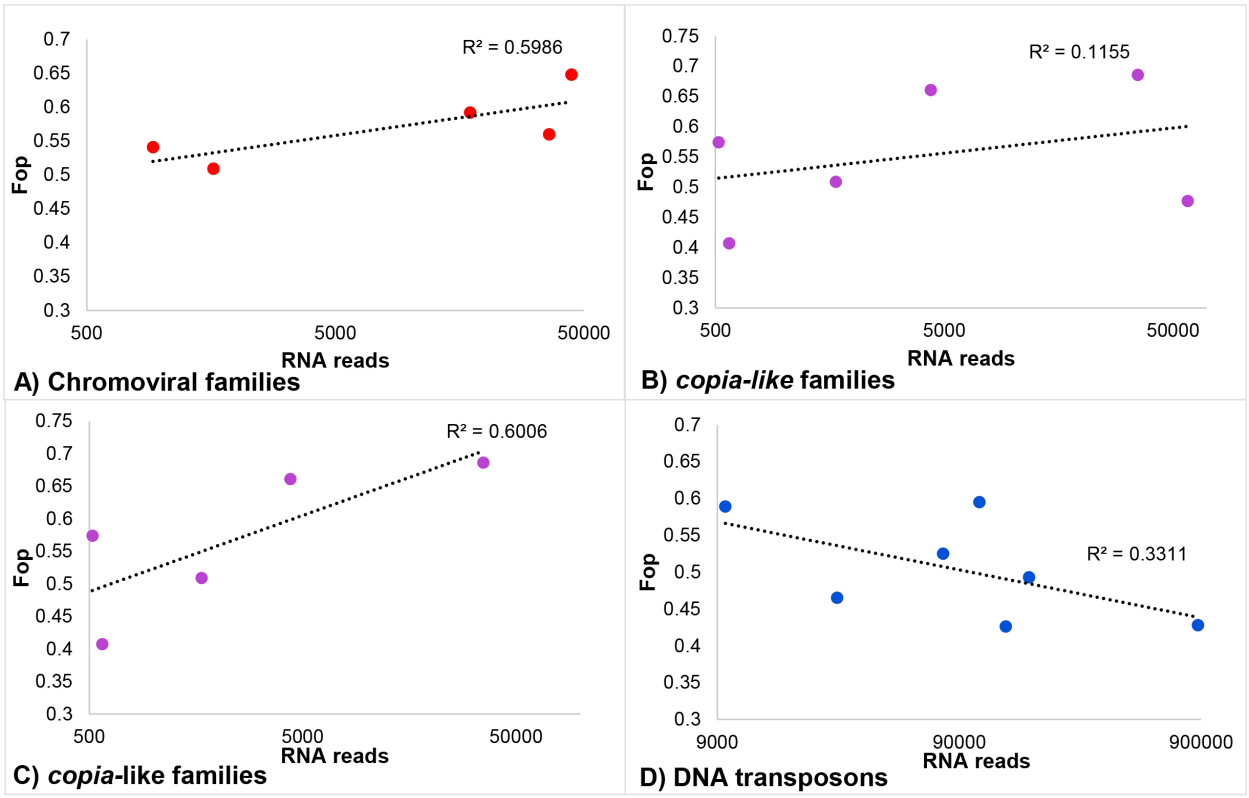


Figure 5.9: **Relationship between number of sequencing reads against Fop for the TE families in the *S. rosetta* genome** The trend line and x axis were calculated using a logarithmic scale. A) All TE families; B) *copia-like* families; C) *copia-like* families, without *Srospv6*; D) DNA transposon families.

Translational accuracy or efficiency can be driven by selection, with the accuracy observed in even weakly biased genes in the three holozoan species. Southworth et al. (2018) provided evidence for selection influencing translational efficiency on host genes, by the positive association observed between gene expression level and the strength of codon usage bias (dos Reis and Wernisch, 2009; Southworth et al., 2018). This relationship was investigated here, reviewing the TE families per species against RNAseq reads. For *S. rosetta*, plotting the number of reads against Fop for the chromoviral families revealed a positive relationship (R^2=0.599), indicating that both mutation pressure and selection for translational efficiency interplay to determine codon usage for the families (Figure 5.9). Within the *copia-like* families, only a weak positive relationship was observed between expression level and Fop (R^2=0.115) (Figure 5.9). *Srospv6* was an outlier amongst the *copia-like* families, with a high level of expression despite only being present as a single copy within the genome. If this family was excluded from the analysis the association between Fop (Figure 5.9) and the number of sequencing reads became considerably stronger (R^2=0.600), suggesting that selection for translational efficiency plays a role in their codon usage bias. The transposon families did not show a positive relationship between Fop and expression;

the negative relationship that was recovered however was very weak (R^2=0.331) (Figure 5.9). However, when the DNA transposon families were further categorised, the unclassified transposons (*SrosT1-T3*) were found to show a stronger negative correlation between Fop and expression levels (R_2=0.825).

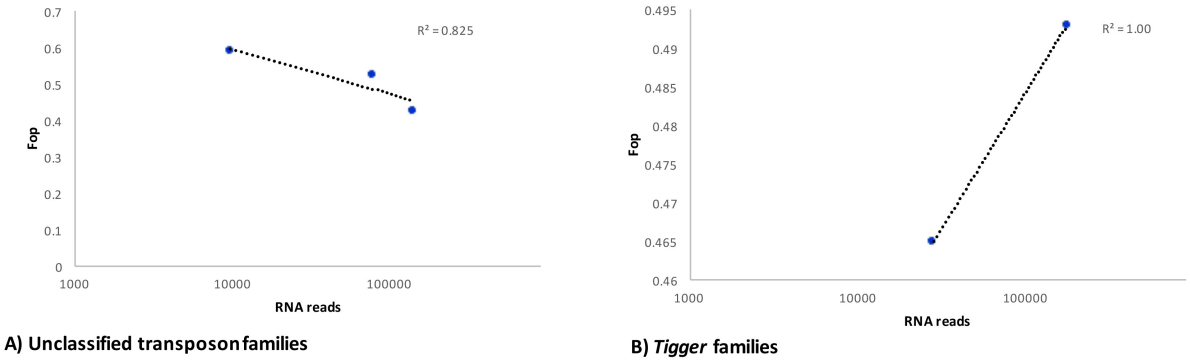


Figure 5.10: **Relationship between number of sequencing reads against Fop for the unclassified transposon and *Tigger* families in the *S. rosetta* genome.** The trend line and x axis were calculated using a logarithmic scale. A) Unclassified transposon families; B) *Tigger* families.

In contrast, for *C. owczarzaki*, no relationship was observed between Fop values and expression for all TE families (Figure 5.11). The strongest correlation was observed between Fop and RNA reads for *C. owczarzaki*, with a negative trend with an R^2 value of 0.324 for the LTR retrotransposon families, however this is still a weak association (Figure 5.11).

Southworth et al. (2018) showed that the majority of anticodons from the major tRNA genes were complementary to host gene optimal codons in *S. rosetta*, *M. brevicollis* and *C. owczarzaki*. This is consistent with selection for translational efficiency, as the most abundant tRNA molecules will be available to bind to optimal codons thereby facilitating rapid protein synthesis. The most abundant, or preferred, codons for each amino acid in each family are shown in Appendix D. No TE families were found to show a perfect match between optimal codons and the *S. rosetta* major tRNA genes. However, some of the families appear to be highly adapted to the host translation machinery with 17 out of the 18 degenerate amino acids showing a match between the preferred codon and either the major tRNA gene or a host optimal codon. These included *Sroscv4*, *Srospv2* and *Srospv3*. The remaining families showed a similar adaptation, with the weakest correlation for the families of *Srosp4*, *Srospv6*, *SrosT3* and *SrosTig1*,however the amino acids were still found to match the host species major tRNA genes 12 out of the 18 amino acids (Appendix D).

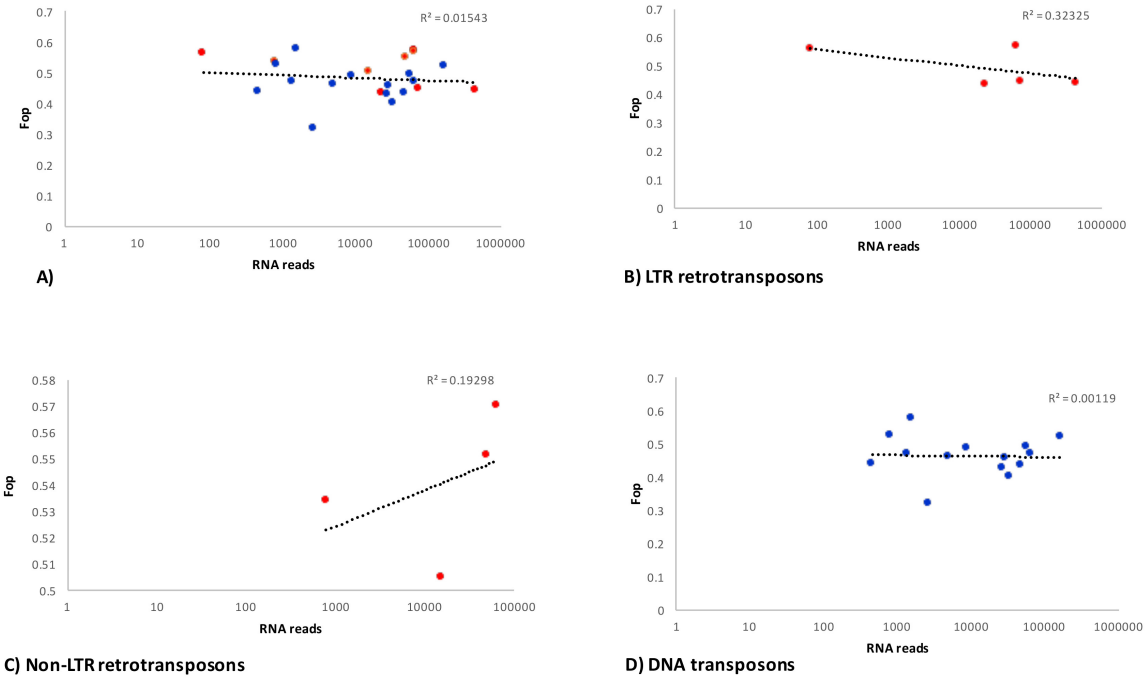


Figure 5.11: **Relationship between number of sequencing reads against Fop for the TE families in the *C. owczarzaki* genome** The trendline and x axis were calculated using a logarithmic scale. A) All TE families; B) LTR retrotransposon families; C) non-LTR retrotransposon families; D) DNA transposon families.

A similar trend was observed for *C. owczarzaki* TE families, with all families found to show complement between the majority of optimal codons for the elements, and major tRNA genes of the host. The identified families found to be highly adapted to the host translation machinery are *Cocv4*, *Cocv5*, *CoL2*, *CoL4* and *Cop2*, with 17 out of 18 degenerate amino acids optimal codons found to match the tRNA gene anticodons identified in the *C. owczarzaki* genome.

Similarly, the three families of *M. brevicollis* were found to show high complement between optimal codons and tRNA genes of the host. *Mbpv2* was found to have the highest complement, with 17 out of 18 codons matching the host tRNA genes.

5.3.4 Evaluating the role of natural selection on translational accuracy

Selection for translational efficiency is likely to affect codons across entire transposable element ORFs, however if selection is operating at the level of translational accuracy one expectation is that codons encoding functional domains will show stronger selection than other regions of genes. Southworth et al. (2018) observed evidence for selection on translational accuracy all three holozoan species. Translational accuracy was even inferred for the most weakly biased genes in the *S. rosetta* genome.

In contrast to the host genes, the *gypsy* families in the *S. rosetta* genome show very similar values of Fop in concatenated domain and non-domain codons (Figure 5.12, Appendix D), indicating that there is no clear evidence that the genes are evolving under translational accuracy. The *copia-like* families appear again to differ from the *gypsy* families, as the *pseudovirus* elements show elevated Fop in domain regions when compared to non-domain codons for all families, except *Srospv6* (Figure 5.12).The results are therefore consistent with selection for translational accuracy playing an important role in codon usage bias within the *copia-like* families.

No functional domain could be identified with *SrosT2*, however across five of the transposon families domain regions of ORFs showed elevated Fop compared to the non-domain codons (Figure 5.13). The only exception was seen for *SrosT3*, where elevated Fop was seen in non domain regions of the mobile element. The data shows that, with the probable exception of *gypsy* families, selection for translational accuracy plays a role in determining codon usage in the majority of *S. rosetta* TE families (14/19 families included in the analysis).

For *C. owczarzaki*, the LTR retrotransposons, less of a pattern emerged, within consistent Fop values per family. The chromoviral families were unresolved, with *Cocv4* and *Cocv3* having elevated Fop values for domain codons, but the remaining three families having elevated Fop for non-domain codons. The majority of non-LTR retrotransposons presented elevated Fop for domain regions, except *CoL1* (Figure 5.14). However, the same pattern observed in *S. rosetta* DNA transposon families was seen for the class II elements of *C. owczarzaki*, except *Cobalt1*, that the functional domains Fop value was greater than the non-domain codons (Figure 5.15). *CoCACTA2* was not included in analyses, as no functional domain could be identified.

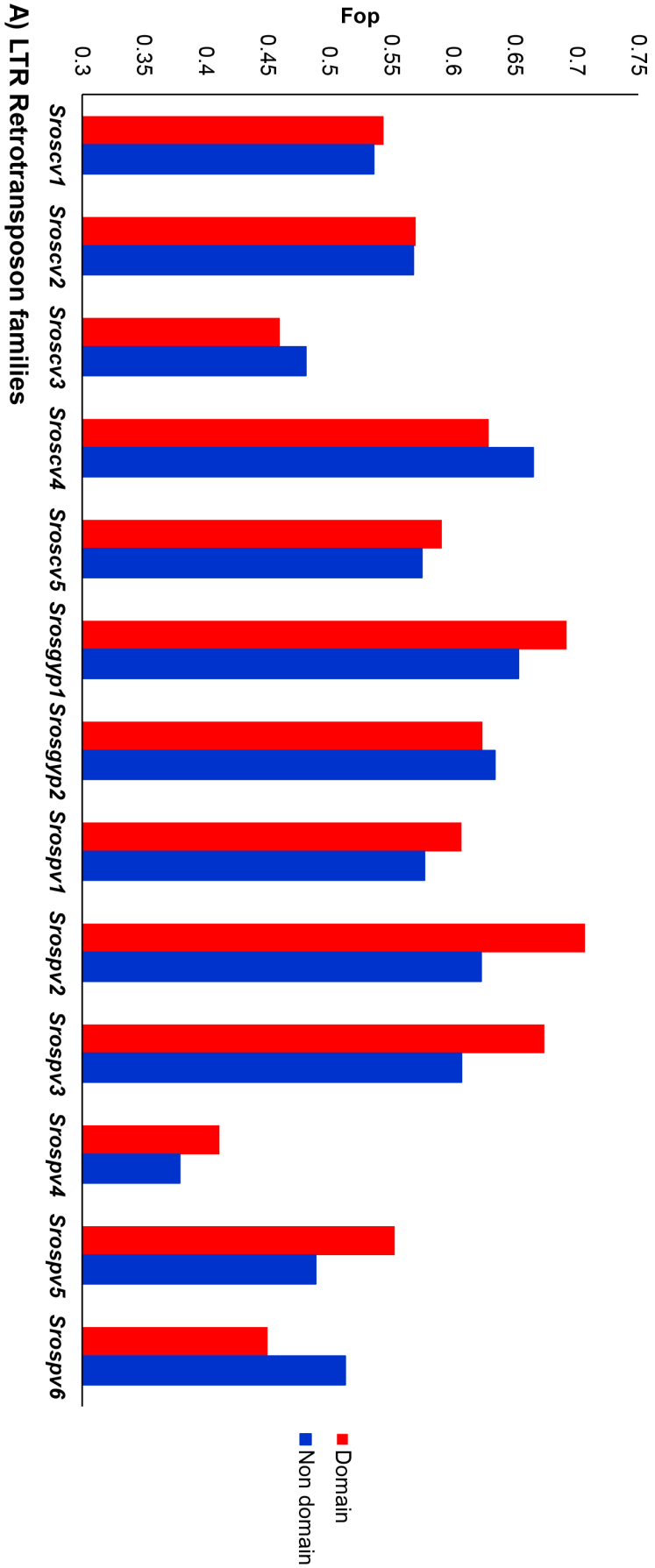


Figure 5.12: **Relationship between Fop values in domain and non-domain codons for LTR retrotransposons families in *S. rosetta***

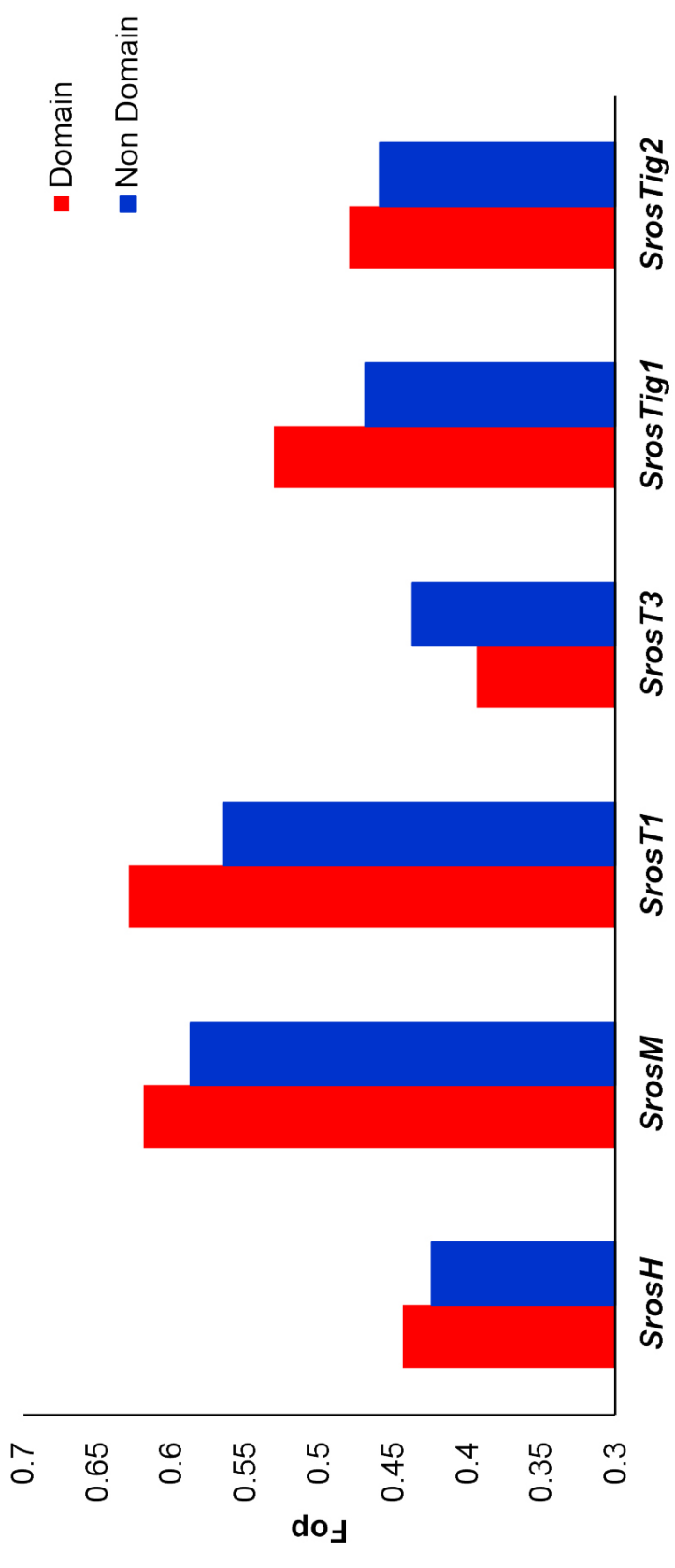


Figure 5.13: **Relationship between Fop values in domain and non-domain codons for DNA transposons families in *S. rosetta***

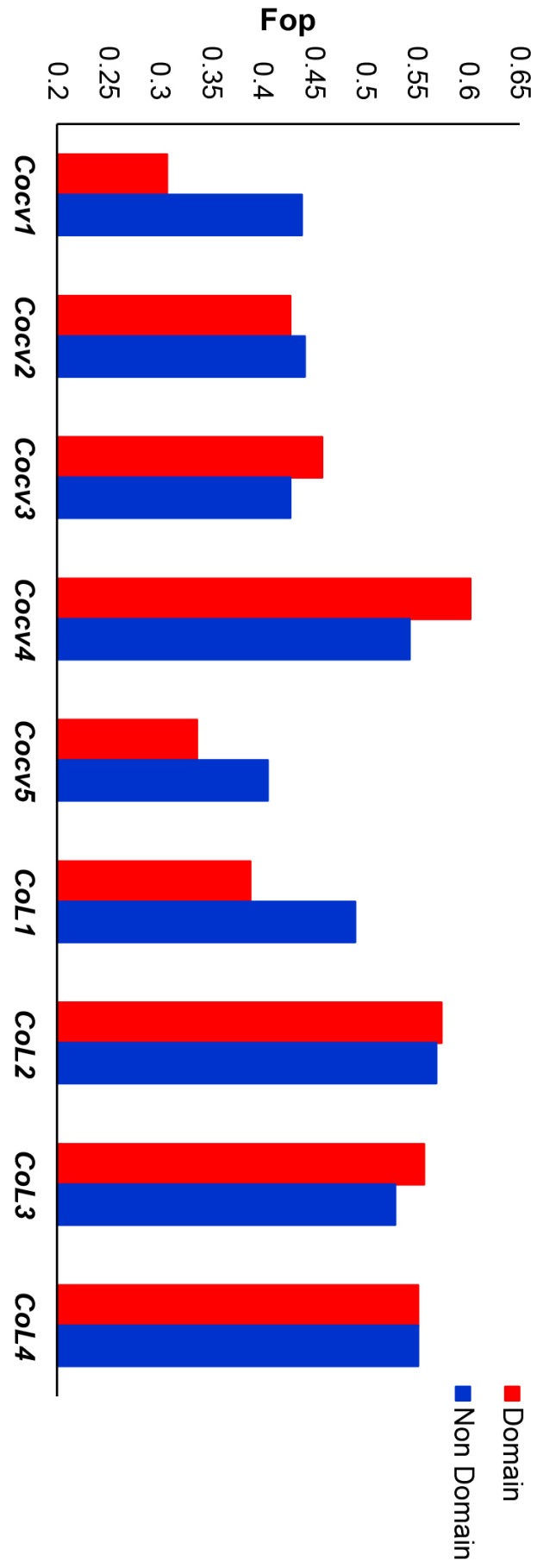


Figure 5.14: **Relationship between Fop values in domain and non-domain codons for LTR retrotransposons families in *C. owczarzaki***

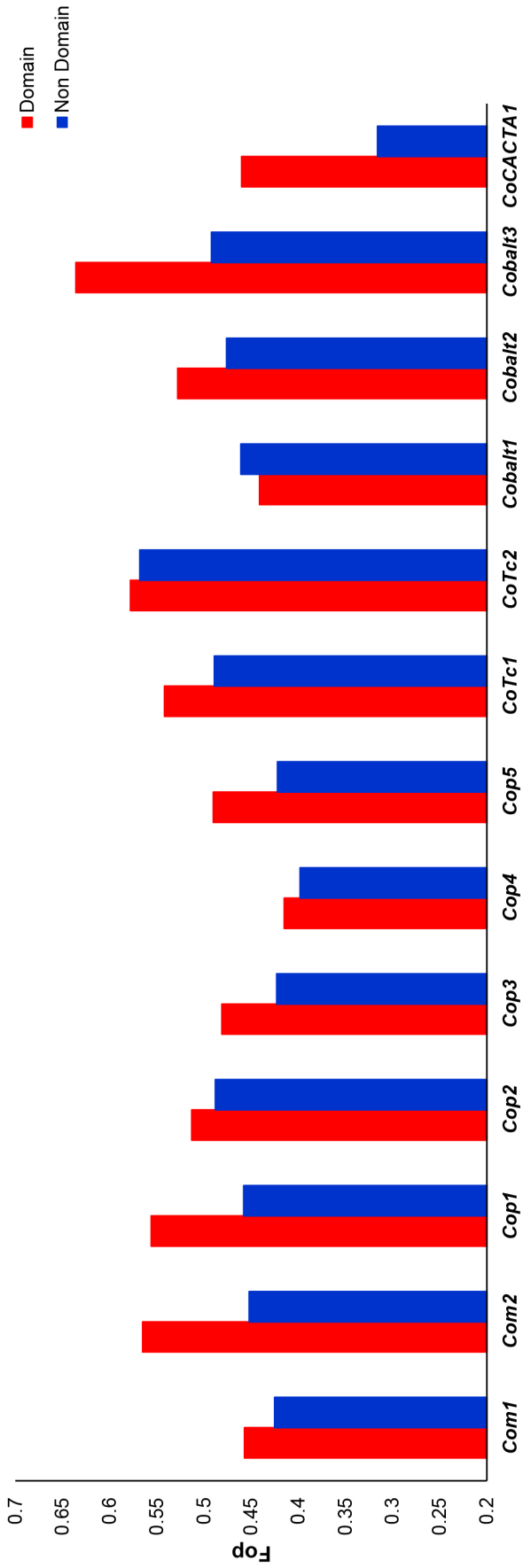


Figure 5.15: **Relationship between Fop values in domain and non-domain codons for DNA transposons families in *C. owczarzaki***

The associations were also reviewed in the three LTR retrotransposon families in *M. brevicollis* (Figure 5.16). For all families, the value of Fop was elevated in domain codons, compared to non-domain. This supports that the TE families in *M. brevicollis* are under selection pressure for translational accuracy, in line with both the majority of *S. rosetta* and *C. owczarzaki* families.

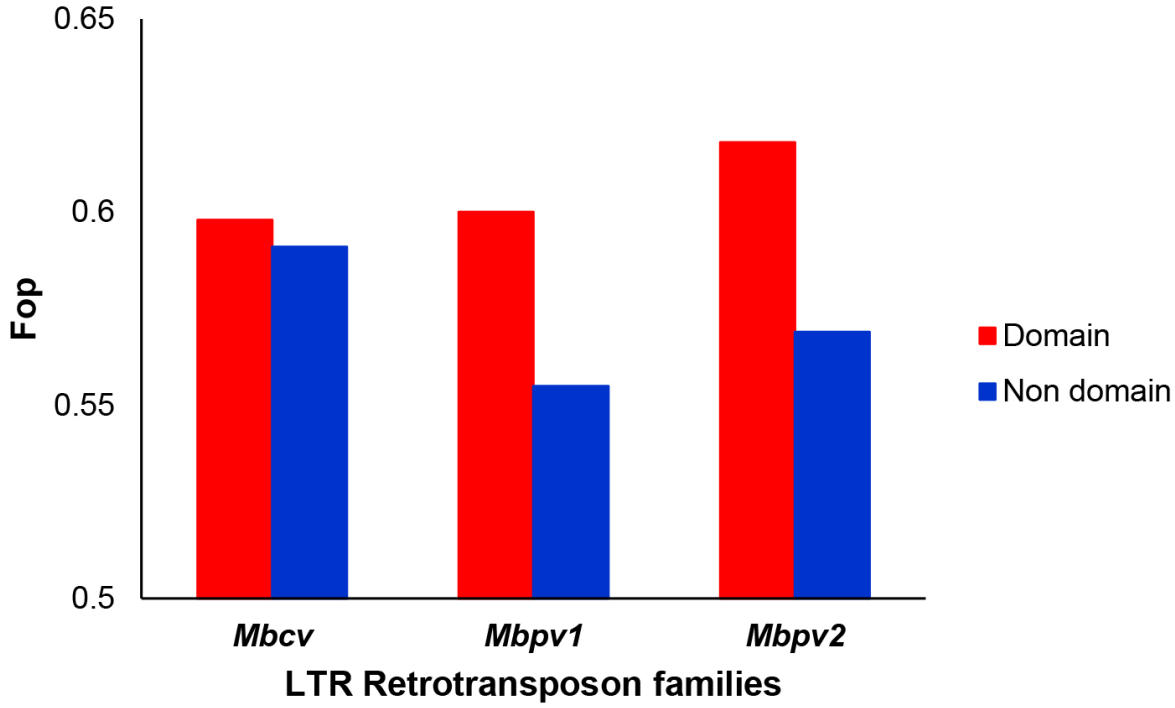


Figure 5.16: **Relationship between Fop values in domain and non-domain codons for LTR retrotransposons families in *M. brevicollis***

5.4 Discussion

5.4.1 Analysis of codon usage bias in transposable element families of three holozoan species

As outlined in Southworth et al. (2018), the whole genome availability of the three holozoan species gave insight into the codon usage of the unicellular protists, and which evolutionary forces drive bias in the species host genes. Even though the last common ancestor of the three species is predicted to be over 1 billion years ago (Parfrey et al., 2011), it was found that patterns of codon usage was highly conserved in the choanoflagellates and *C. owczarzaki* (Southworth et al., 2018). Following the analysis outlined for this project, it was found that each species' TE families were also found to show similar patterns of codon usage, with conservation observed across the three holozoan protists.

The TEs of the holozoan species were found to show an overall GC preference, which was in contrast to the majority of previous literature that uncovered an AT bias in TE families of varied eukaryotic taxa (Lerat et al., 2002; Jia and Xue, 2009). Similar patterns were drawn between the holozoan species families, and the bias patterns seen in *Phytophora* elements, as described by Jiang et al. (2006). All families were found to be GC-rich at synonymous third positions, with average GC3s values over 0.62 (Table 5.1 and 5.2). As the bias for GC at the 3rd position is also seen in the mobile elements of the three holozoan species, this would also support that codon bias is driven by natural selection, but could also be a signature for mutation bias. However, the mean non coding GC content of mobile elements for each species ranged between 0.49-0.52, which does not support that mutation pressure has a major influence codon usage bias, as a preference to GC in non coding DNA of th elements is not seen.

Furthermore, for the TE families of *S. rosetta* and *C. owczarzaki*, a strikingly similar trend was observed between GC3s and *Nc*, when compared to the host genes in Southworth et al. (2018). This could not be determined for *M. brevicollis*, as only three LTR retrotransposon families have been uncovered in this choanoflagellate species, so any trends observed are speculative. The conservation between host codon usage bias, and the bias indicated by the TE families suggested that the elements are influenced by the same selection pressure observed in the holozoan species (Southworth et al., 2018). However, the patterns seen could be driven by a variety of forces, and therefore, the signatures for both selection and mutation bias were reviewed here.

5.4.2 Selection for abundant codons is a driver of codon usage bias

The first line of evidence supported that codon usage is driven by natural selection for translational accuracy, in the TE, as well as the host genes, as seen in (Southworth et al., 2018). In order to assess the contribution of selection on translational accuracy, the codons which encode amino acids in functional domains were separated from those that encode non-domain amino acids for all TE families. Domain regions were identified by analysing each Pol and Tnpase protein sequence in BLASTp through NCBI. Fop values were then determined for the domain and non-domain codons in each family. For the three holzoan species, it was found that functional domains were found to be enriched by optimal codons in the majority of TE families, however the pattern was not as consistent as seen in Southworth et al. (2018) for the host genes. For TEs which showed a higher level of Fop in domain codons, the findings supported that selection for translation accuracy seems to be driving codon usage bias in the mobile elements of the unicellular protists, as well as the host genes. However, in chromoviral families, values of Fop were found to be similar for domain and non-domain codons. This finding supports that the chromoviruses could be evolving under different selection pressures compared to the host genes, and other TE families. Mutation bias would influence the frequency of optimal codons in elements, however little difference would be seen between domain and non-domain regions within the TE. The results uncovered therefore support that mutation bias could be driving bias in the chromoviruses rather than selection for accuracy.

5.4.3 The role of mutation bias on the TEs of the three holozoan species

Little evidence was found to support mutational pressure in the three holozoan species (Southworth et al., 2018), and a similar trend was uncovered for the majority of TE families investigated here. However, evidence to suggest mutation may be driver of codon usage bias was observed for the chromoviral families of *S. rosetta* and *C. owczarzaki*, as well as three families in *M. brevicollis*. Positive correlations between GC3s and non-coding GC content was found chromoviruses and families of *M. brevicollis*, which the weakest association seen for *S. rosetta* (R^2=0.519) (Figure 5.6), followed by *C. owczarzaki* (R^2=650), and the strongest relationship between the *M. brevicollis* TE families (R^2=0.857).

Jia and Xue (2009) outlined evidence for multiple evolutionary pressures influencing TE families in two plant species (*Oryza sativa* and *Arabidopsis thaliana*), with the combination of selection and

mutation bias driving codon usage in the mobile elements. It is proposed that the same could be said for the chromoviruses reviewed here. Overall, the chromoviral families show greater codon usage bias, in comparison to the other families in the holozoan species. Although the relationship was not strongly supported by an R^2 value, it is still evident that a positive correlation is found between GC-content of non coding DNA and overall GC3s for the mobile elements, which supports that the GC preference has shaped the codon usage of the entirety of the full length elements.

5.4.4 Concluding remarks

Of the three species studied, *S. rosetta* TEs show the strongest level of codon usage bias, followed by *C. owczarzaki* elements, with the second choanoflagellate *M. brevicollis* retrotransposons having the weakest level of bias based on *Nc* and Fop values. However, although *M. brevicollis* showed the weakest codon usage, it was the only species to show a relationship between GC3s and non-coding GC (R2= 0.857) (Figure 5.8), which supported that mutation pressure may also be driving bias in mobile elements of the choanoflagellate species . Several TE families were found to show signatures for selection as seen in the host genes of the holozoan species (Southworth et al., 2018). However, contrasting patterns were seen for the chromoviral families, which have signatures of mutation pressure, but limited evidence for selection.

Overall, it was found that LTR retrotransposon families typically have higher values of Fop and GC3s (Table 5.1 and 5.2). Evidence for recent transposition previously noted in Chapter 3 for *S. rosetta*, and Carr and Suga (2014) for the elements of *C. owczarzaki*, would suggest that retrotransposons may be more active than DNA transposons in the two holozoan protists. As the transposons uncovered in *M. brevicollis* may be pseudogenes, only values for the known LTR retrotransposons were reviewed here and therefore superfamily comparisons can not be drawn for the choanoflagellate species. The duplicative nature of retrotransposons may be an explanation as to why codon usage is found to be greater in the class I elements. It is thought that mobile elements individual copies are expected to evolve under a neutral model within the host genome, and the family as a whole may be evolving under selection. With this, it could be plausible that the utilisation of optimal codons observed in the class I elements may be as a result of a greater rate of transposition within the host species.

The TE analysis outlined in this project highlighted the influence of natural selection as the driver of codon usage in mobile elements, as well as host genes (Southworth et al., 2018), in the

three holozoan species. Despite the distant ancestry, codon usage has remained conserved in the choanoflagellates and the filasterean, indicating that GC-bias have been selected for translation accuracy in premetazoan species, and that TE are also under the same selective forces. Further research on a variety of unicellular eukaryotes would be required to determine if selection is the main driver of codon usage in other holozoan species. Previous research has provided evidence against selection for TEs within the Opisthokonts, such as the *Ty* elements of *S. cerevisiae* (Lerat et al., 2002). Therefore, hopefully the work here will inspire further non-domain analyses of codon usage of mobile elements to determine if selection bias is more prevalent in unicellular eukaryotes than first thought.

Chapter 6

Discussion

The field of evolutionary genomics is expansive, and adaptive, with advances in bioinformatic technology employed to refine methods to produce the most accurate and objective outcomes. As sequencing methods have become more accessible to researchers, it is found that very few taxonomic groups do not have multiple species representatives with genome availability. Prior to the findings reported here, the genus *Kazachstania* and choanoflagellate species *S. rosetta*, had limited information regarding genomic characteristics, including codon usage bias, and TE annotation. The work presented throughout this thesis has allowed for comparative genomics of novel yeast species, as well as three holozoan protists.

6.1 Genomic survey of novel *Kazachstania* species

The genus *Kazachstania* is a branch of the Saccharomycetaceae superfamily which was found to have limited genome availability. To address this limitation, the genomes of four *Kazachstania* species were de novo sequenced via PacBio sequencing: *K. bovina*, *K. exigua*, *K. lodderae* and *K. viticola*, selected for their diverse placement in the species phylogeny established by Kurtzman (2003). The genomes of four additional species (*K. africana*, *K. naganishii*, *K. saulgeensis* and *K. servazzii*) were publicly available at the time of this research; however little examination of the genus had previously been performed. Chapter 2 therefore details the genomic exploration of all eight species.

As outlined in Chapter 2, *K. exigua* revealed unexpected genome characteristics, with a much greater genome size of 24.8Mb, and a total number of 9964 coding genes annotated which far exceeded all other currently sequenced *Kazachstania* species. It was found that the increased genome size of *K. exigua* could not be attributed to whole genome duplication as seen in other species (Wolfe et al., 2015), or differing ploidy. The orthologous review of *K. exigua* found that

86% of genes were annotated and therefore assumed to be functional, with no duplicated areas of the genome uncovered. However, syntenic blocks were assessed between *K. exigua* and the other *Kazachstania* species, and no evidence of a duplication event was found.

Typically, synteny is documented to be strongly conserved within the Saccharomycetaceae family (Wolfe et al., 2015; Dujon and Louis, 2017). The work here found that all *Kazachstania* species, with the exception of *K. exigua*, were found to fit within the typical characteristics of Saccharomycetaceae species. As detailed in Chapter 2, genomic size ranged between 10.8Mb and 13Mb, with total number of coding genes also highly conserved, from 5300-6000 genes. However, it was found that the revealed data did not fit the general trend of Saccharomycetaceae species, with yeast members of the family with a genome size ranging from 10 - 13Mb, and total number of coding genes from 4500 - 6500 genes (Dujon and Louis, 2017). Further comparative study would need to be reviewed to construct a plausible explanation as to why the *Kazachstania* species genomic characteristics are so vastly different. As discussed in Wolfe et al. (2015), the understanding of Saccharomycetaceae genome evolution will remain unresolved as many clades do not have multiple strains per species. With this, it is known that intraspecies polymorphism of gene content is extensive (Song et al., 2015; Wolfe et al., 2015). However, the worked presented in Chapter 2 provided a detailed insight into genus specific characteristics that had previously not been explored. The additional genome sequences constructed have doubled the WGS availability for the genus, with eight species now sequenced for comparative study.

6.2 TE review in unicellular eukaryotes

Consistent with previous investigations into the TE content of Saccharomycetaceae (Kim et al., 1998; Carr et al., 2012; Bleykasten-Grosshans et al., 2013), *Kazachstania* species were found to have LTR retrotransposons, with a complete absence of DNA transposons documented for all eight species. With regards to TE content, the newly sequenced *Kazachstania* species were found to have a greater overall genomic TE content, equating to >2% in three of the four species, similar to the *S. cerevisiae* reference strain (S288c) (Carr et al., 2012). In contrast, the publicly available *Kazachstania* species, were found to have a smaller TE genome content percentage, ranging from 0.15-0.70%. Although the overall percentage of TEs is relatively low in *Kazachstania* species, copy number, as well as content, was seen to be greater in the three of the four *Kazachstania* species sequenced here. Although diversification is expected across a genus in relation to TE content and

proliferation, as seen in many host yeast species across the superfamily (Neuvéglise et al., 2002), another explanation that was considered to explain the variation is the sequencing method for the genomes.

The work detailed here, included PacBio sequencing of four novel yeast species, *K. bovina*, *K. exigua*, *K. lodderae* and *K. viticola*. In contrast, *K. africana*, *K. naganishii*, *K. saulgeensis*, which were downloaded from NCBI (Sayers et al., 2009), were all sequenced via Illumina sequencing. Sequencing methods have vastly developed since the first eukaryotic genome sequencing of *S. cerevisiae*, and with this, contention has been expressed to depict which platform provides the most accurate genome sequencing (Liu et al., 2012). Furthermore, TE annotation is typically homology based, and therefore the quality of sequencing and assembly of the query genome is paramount to the validity of the results. The two platforms detailed in Chapter 2, were analysed regarding TE annotation due to interspecies availability for the *Kazachstania* species, *K. servazzii*. The public availability of the *K. servazzii* allowed for interspecies comparison, as two strains were available for this species with different sequencing methods; *K. servazzii* (CBA6004) was sequenced using Illumina, and *K. servazzii* (SRCM102023) was sequenced with PacBio sequencing. Prior to the TE annotation in Chapter 2, it was initially considered that due to the increased read length of PacBio sequencing, that TE annotation would be more accurate with the employment of this sequencing platform. However, interspecies comparison of two *K. servazzii* strains, found a small difference in TE genome content of 0.15% between sequencing methods. With this, the findings did not support that the PacBio is necessarily more accurate for assembly, and thus TE annotation.

Chapter 3 detailed the repertoire of TEs uncovered in the choanoflagellate species, *S. rosetta*. The genome TE annotation revealed a minimum of 20 TE families, with both retrotransposons and DNA transposons families uncovered. *S. rosetta* was found to have a more diverse TE range in comparison to the first publicly available choanoflagellate species, *M. brevicollis* (Carr, Nelson, Leadbeater and Baldauf, 2008), as well as *My. fluctuans*, which was assessed using ESTs by Carr, Nelson, Leadbeater and Baldauf (2008), . However, as only two choanoflagellates species currently have whole genome availability, patterns of TE abundance can not be drawn for the class at present. However, Chapter 3 explored plausible hypotheses to explain the increase in TE abundance and genome percentage content.

As the sister group to the metazoans (King, 2005), it would be presumed that TE families in choanoflagellates would be diverse, if elements were found to be inherited vertically, in line with

the known abundance in the majority of metazoan species (Kidwell and Lisch, 1997; Arkhipova and Morrison, 2001). Alternatively, the proliferation of TEs seen in metazoans could be due to expansive transfer events to allow for proliferation within this kingdom, however this seemed less plausible. If ancestral choanoflagellates to Metazoa were found to have mobile elements in abundance, the decreased number of families documented in *M. brevicollis* (Carr, Nelson, Leadbeater and Baldauf, 2008), is due to loss within this species. Comparatively, it is also found that *S. rosetta* can present multicolonal, whereas *M. brevicollis* only has a single unicellular form (King et al., 2008). The multicellular colonies of *S. rosetta* could result facilitation of HTT within the species, reducing the chance of stochastic loss that is hypothesised to have occurred in *M. brevicollis*.

The wider repertoire of elements detailed in Chapter 3 for *S. rosetta* shows closer similarity to *C. owczarzaki*, with 23 families revealed in the filasterean species (Carr and Suga, 2014). However, Carr and Suga (2014) found that the elements uncovered in *C. owczarzaki* had closely related orthologous families in other species of Opisthokonta, which supported the presence of diverse TEs in the last common ancestor of the opisthokonts. In contrast, the work here found that *S. rosetta* elements showed a more complex inheritance pathway. In Chapter 3, the majority of elements were predicted to be acquired by vertical transfer, with the exception of *SrosT1*, which revealed evidence to support acquisition by horizontal transfer. The work supported the transfer to be facilitated between *S. rosetta* and a stramenopile donor species. Horizontal transfer events have been previously documented in choanoflagellate species (Yue et al., 2013; Tucker et al., 2015), with varied unicellular donor species, which have been acquired via a predator-prey relationship (King, 2005). Phagocytosis of unicellular prey species would enable the transfer of genetic material between organelles, allowing for transfer across species barriers. The potential transposon family of *M. brevicollis* (*MbT1*) uncovered in Chapter 3, would support that the horizontal transfer was ancient within the choanoflagellate lineage.

The work uncovered in Chapter 2 and 3 found that all *Kazachstania* species and *S. rosetta* revealed both *gypsy* and *copia* elements, with a higher abundance of *gypsy* elements found in yeast species, and *copia* elements with higher copy number revealed in the choanoflagellate. The abundance of *gypsy* elements in yeast species provided further support to the exhaustive research which has revealed the proliferation of this TE family in several other fungal species (Kim et al., 1998; Muszewska et al., 2011; Wolfe et al., 2015). In addition, the *S. rosetta* TE annotation here

has added to a limited dataset of choanoflagellate species which would continue to be expanded upon genome availability. With the presence of both classes of TE families in both choanoflagellate species, the work has provided further support that the last common ancestor of the two species possessed class I and class II elements.

Typically, TEs have been found to maintain high copy number in multicellular organisms, including metazoan and plant species (Lu et al., 2017). It is noted that, the TEs found in the holozoan species (the choanoflagellates and filasterean species *C. owczarzaki*) (Carr, Nelson, Leadbeater and Baldauf, 2008; Carr et al., 2012), are present in low copy number, with only one family found to show proliferation of over 100 copies within the *S. rosetta* genome. A similar observation was found for the *Kazachstania* species, with TE copy number across the genus ranging from 1 - 30. It is proposed that the protistan species are found to have large effective population sizes (Snoke et al., 2006), and therefore are more likely to allow elimination mechanisms of individual TE insertions within a species. The same theory could be plausible for the yeast species, with presumed large effective population size (Tsai et al., 2008), who present with active elements with low copy number. However, as no population data is available for *Kazachstania*, the theory can not be extrapolated to multiple species in different genera.

6.3 Codon usage in unicellular eukaryotes

Presented here is the first work on the comparative analysis of codon usage patterns for the genus *Kazachstania*. The findings have provided an insight to the evolution of bias in these species and high conservation was seen with *S. cerevisiae* (Sharp et al., 1986). In all degenerate amino acids, except leucine, identical optimal codons were employed for *S. cerevisiae* (Sharp et al., 1986) and *K. africana*, *K. exigua*, *K. lodderae* and *K. viticola*. Variation was observed in species *K. bovina* and *K. naganishii*, which showed contrasting patterns of codon usage. Chapter 2 outlined that *K. bovina* showed the greatest overall bias, however, optimal codons had lower complement to the major tRNA genes of the species compared to the other *Kazachstania* species. Furthermore, *K. naganishii* was found to show a preference for GC-ending codons, compared to an overall AT-preference uncovered for the other *Kazachstania* species. Predominantly, as seen in *S. cerevisiae* (Sharp et al., 1986), the results of Chapter 2 revealed signatures of selection were seen for the *Kazachstania* species. Firstly, the major tRNA complementing the CodonW optimal codons for the *Kazachstania* species points to selection being a driver of codon usage bias, as well

as the support provided by the reciprocated optimal codons defined by correspondence analysis and abundant codons of EF1a. However, although markers for selection have been uncovered, mutation bias cannot be dismissed as a driver of codon usage bias. With no evidence reviewed to eliminate mutation for the *Kazachstania* species, it can not be concluded that selection is the main driver of codon usage. However, from the trends documented, it is plausible that selection is a contributor of bias in the yeast species.

In Chapter 2 and 4, the review of codon usage bias in the unicellular eukaryotes revealed high conservation across the *Kazachstania* species, as well as the holozoan species (Southworth et al., 2018). In contrast, as outlined by Southworth et al. (2018), selection is seen to be the main driver of codon usage bias in holozoan species, with conservation observed in the three protists host genes, as well as evidence to discount a mutational model driving bias. Signatures for selection included a significant difference observed in between GC3s and non-coding GC, as well as a higher frequency of optimal codons found in domain regions of host genes (Southworth et al., 2018). The work of Southworth et al. (2018) was repeated here, as well as analysis of bias categories for 1% of the host genes, rather than 5% which was analysed previously. Chapter 4 showed that the decrease in bias category was found to have little effect on the trends and results published by Southworth et al. (2018). The finding supported that the smaller bias categories were found to be representative of the data..

Novoa et al. (2012) stated that studies have claimed that the most accurate measure of codon usage bias across diverse organisms is by the measure of GC content across species. The work of Chapter 4 explored this, and a conserved GC-preference for the holozoan species, which was in contrast to the *Kazachstania* species, who predominantly showed an AT-bias for host genes based upon correspondence analysis. In Chapter 2, a positive correlation was observed between *Nc* and GC3s which supported that genes of greater bias were found to harbour AT-ending codons. The AT-preference seen in *Kazachstania* species was an expected finding, with similar trends observed in closely related yeast species (Sharp et al., 1986; LaBella et al., 2019).

Although a striking contrast is seen between nucleotide preference, a similarity is drawn between the unicellular species for deamination. Evidence for deamination of adenosine was uncovered in all species reviewed for codon usage through the review of tRNA major genes anticodons, and the optimal codons calculated by CodonW based on correspondence analysis. Southworth et al. (2018) found that the most abundant major tRNA genes of two-fold degenerate amino acids

were found to be complementary to the optimal codons of *S. rosetta*, *C. owczarzaki* and all amino acids except lysine *M. brevicollis*. For all amino acids over two-fold degeneracy, the tRNA genes were found to have adenosine at the wobble position, with the majority of optimal codons found to reveal cytosine at the degenerate position (Southworth et al., 2018). Findings showed that complementation could be found through the deamination of the major tRNA genes. As described in Chapter 4, through tRNA modification, adenosine at the wobble position would be converted to inosine, which would allow for pairing with cytosine by standard base pairing (Southworth et al., 2018). Similarly, evidence for tRNA modification was also uncovered in all *Kazachstania* species, in high degeneracy amino acids. Ile, Leu, Seri, Thr and Val, were revealed to have adenosine at the wobble position allowing for complementary base pairing to optimal codons with cytosine at the 3rd position, as seen in the holozoan species (Southworth et al., 2018). In other eukaryotic organisms including species for metazoan, fungal and plant kingdoms, evidence to support tRNA modification is extensive, with highly expressed genes found to be enriched with codons that are complementary to tRNA molecules which have been deaminated (Rafels-Ybern et al., 2017; Novoa et al., 2012).

As detailed in Southworth et al. (2018) it is suggested that deamination of tRNA molecules evolved prior to multicellularity within Holozoa. However, deamination is proposed to have evolved early within the opisthokonts, as the event has also been uncovered in fungal species, which are outside of the holozoan grouping. The findings outlined in Chapter 2 revealed that evidence of deamination is present in the *Kazachstania* species, as seen in *S. cerevisiae* (Gerber et al., 1998). The findings of Chapter 2 that uncovered tRNA modifications in the *Kazachstania* species have provided further insight into the evolutionary pathway of deamination, and that the usage of modified tRNA molecules evolved early in the LCA of Holomycota. Furthermore, the evidence of large-scale usage of deaminated tRNA molecules could be found to be an ancestral trait present in the last common ancestor of the opisthokonts, with signatures for deamination uncovered in both holozoan and Holomycota species.The work presented here, has provided further insight into the necessity of tRNA modification in codon usage, and established an avenue for additional research to depict the origin of the differentiating feature of translational efficiency.

LaBella et al. (2019) recently described codon usage bias across the budding yeast subphylum, Saccharomycotina. In contrast to the work outlined here, which focused on a particular species, the research addressed biodiversity across an entire subphylum for arching patterns of bias (LaBella

et al., 2019). It is acknowledged that the research employed alternative methods for calculating codon usage bias, and no specific trends were outlined for focused areas within the subphylum (LaBella et al., 2019). In contrast, the work presented in Chapter 2 was more detailed, looking at specific patterns between closely related species and the model organism, *S. cerevisiae*. LaBella et al. (2019) found that the majority of genomes deviated from neutral expectation when examining patterns between GC3s and *Nc*, as seen in the work described here. Furthermore, evidence of selection for translational efficiency was found to be prevalent across the budding yeast genomes of Saccharomycotina (LaBella et al., 2019), which supported the markers of selection uncovered in the *Kazachstania* species.

6.3.1 Codon usage of mobile elements in unicellular opisthokonts

As with codon usage of the *Kazachstania species*, the work presented here is the first study of codon usage patterns for the mobile elements uncovered in species of the genus *Kazachstania*, as well as elements revealed in holozoan species *S. rosetta*, *M. brevicollis* (Carr, Nelson, Leadbeater and Baldauf, 2008) and *C. owczarzaki* (Carr and Suga, 2014). In Chapter 2 and 5, codon usage bias of mobile elements were predominantly found to mirror bias observed in the host genes for the unicellular species reviewed. The work was found to uncover an AT-preference in the TE families uncovered in *Kazachstania* species, and conservation was also seen between TEs and host genes for the three holozoan species.

Although drivers of codon usage bias in the *Kazachstania* species are not clear, patterns of bias were duplicated in the mobile elements of the yeast species. A bias towards AT-ending codons in TE ORFs has previously been observed in other eukaryotic species, including yeast, plants and metazoans (Lerat et al., 2002; Jia and Xue, 2009). With regards to evolutionary pressure, as with the host genes, the findings of Chapter 2 remained inconclusive as to whether selection or mutation is the main driver of bias in the mobile elements of the budding yeast species. However, evidence would suggest that the likeliness of genetic drift influencing bias is slim. The patterns of bias, including AT-preference and complement between abundant codons and host genes major tRNA molecules would support that the signatures for bias do not present as random. If genetic drift was the major driver of codon usage in the mobile elements, little to no correlations would be drawn between results, and the codon selection would not show patterns due to random allocation.

The work of Chapter 2 was found that bias for the TEs were similar to those of the host species genes, therefore signatures point towards either selection or mutation.

In contrast, TE AT-bias was not found for the elements uncovered in the three holozoan species, *S. rosetta*, *C. owczarzaki* and *M. brevicollis*, with the mobile elements mirroring a GC-preference observed for the species host genes, which is likely to be driven by selection. Average GC3s values for all TE families were greater than 0.62, with the mean non-coding GC content found to range from 0.49 - 0.52, which provides support against mutational pressure driving codon usage, as values would be closer to average GC3s if this were the case. Furthermore, high complement was seen between abundant codons of the TE families and major tRNA genes of the host species, for all three protists.

However, potential signatures for mutational bias were uncovered for the chromoviral families in *S. rosetta*, *C. owczarzaki* and *M. brevicollis*. A positive relationship was observed between GC3s and non-coding GC content for the chromoviruses and all families of *M. brevicollis*, with the elements found to have a stronger bias compared to the other families present in *S. rosetta*, *C. owczarzaki* and *M. brevicollis* respectively (R_2= 0.519; 0.650; 0.857). The work presented here proposed that the chromoviral families are potentially evolving under a more complex evolutionary model, and provoked interest as to why the family would be subject to different pressures than other elements in the same host genome.

To date, TE evolution in model organisms, *S. cerevisiae* and *D. melanogaster*, were found to show no signatures for selection driving codon usage (Lerat et al., 2002). The novel review of TEs codon usage, in *Kazachstania* species, and holozoan species presented here highlights the importance of further explorative research in addition to model organisms which frequently become the foundation of accepted knowledge.

6.4 Concluding remarks

The research aimed to compare genomic characteristics of unicellular opisthokonts, as well as a TE review of novel species. The work successfully addressed the outcomes proposed for the project, with trends drawn of varied origin. The atypical findings highlight the importance of comparative genomics, and should inspire further investigation of TE evolution and codon usage patterns in unicellular eukaryotes, to depict the revealed complexity of TE acquisition and codon bias.

Appendix A

Lab Protocols and Results

A.1 DNA/RNA extraction

A.1.1 Trizol RNA extraction

The following components were made for RNA extraction; DEPC Water, NaOH 10M, EDTA, NaOH EDTA and TBE Buffer. Autoclaving parameters were set at default (15 minutes at 121˚C).

DEPC H_2O (1L)

- 1 ml of DEPC, which was stored at 4°C in the fridge
- Milli-Q H_2O was added up to 500 ml
- The solution was shaken until the oily droplets are dispersed
- The solution was made up to 1 L and sterilised by autoclaving

NaOH 10 M (200 ml)

- 80 g NaOH (powder) was added to 200 ml with Milli-Q H2O
- The solution was sterilised by autoclaving

EDTA 0.5 M, pH 8.0 (1000 ml)

- 146.1 g of anhydrous EDTA was measured
- 800 ml of DEPC H2O was added to the EDTA
- NaOH was added slowly with stirring until the EDTA dissolved and reached a pH of 8.0
- The solution was made up to 1000 ml with DEPC H2O and sterilised by autoclaving

NaOH 0.5 M, 50 mM EDTA

- 10 ml 500 mM EDTA
- 4.493 ml of 10 M NaOH (as 50.7 mM NaOH in EDTA already tf 500 mM - 50.7 mM = 449.3 mM)
- 85.5 ml DEPC H2O
- 100 ml total volume and sterilised by autoclaving

10x TBE electrophoresis buffer (1000 ml)

- 50ml TBE
- 450 ml DEPC H2O
- The solution was then sterilised by autoclaving

Ethanol – 75% (50ml)

- 37.5ml Ethanol 100
- 12.5ml DEPC H2O

1.0% Agarose Gel (40ml)

- 40ml DEPC treated TBE buffer
- 0.40g Agarose (1.0%)
- 4µl Gel Stain

Trizol RNA extraction protocol

Protocol For the isolation of total RNA, the *Kazachstania* cells were cultured overnight at 25˚c in 10mL of YPD liquid medium (Vaughan-Martini et al., 2011). 1ml of the culture was added to 2ml microcentrifuge tube and cells were collected by centrifugation at 3000g for 4 minutes. Liquid medium was discarded and the pellet cells were washed with 5mL PBS DEPC H_2O and resuspended using the vortex for 5 seconds per sample (Xiao, 2006).

1ml of TRI reagent was used per 1ml of yeast sample, which lysed the cells and formed a homogenous lysate. 0.2ml of chloroform was added in the fume cupboard (0.2ml of chloroform/

1ml of TRI reagent). The sample was vigorously shaken by hand for 15 seconds and allowed to stand at room temperature for 15 minutes.

The sample was centrifuged at 12,000g at room temperature for 15 minutes which separated the mixture into three phases; a red organic phase (which contained protein), an interphase (DNA) and a colourless aqueous phase at the upper section of the same which contains RNA. The aqueous phase was transferred to a fresh Eppendorf tube containing 0.50ml of propan-2-ol (0.50ml Isopropanol/ 1ml of TRI Reagent) and mixed, before leaving to stand for a further 10 minutes at room temperature. The sample was then put through further centrifugation at 12000g at room temperature for 10 minutes, where an RNA pellet formed at the bottom of the tube.

The supernatant of the sample was removed, and the RNA pellet washed using 1ml of 75% ethanol and vortexed, and centrifuged at 7500g at room temperature for 5 minutes. Ethanol was removed from the sample and the RNA pellet was dried for 5 minutes at room temperature by air drying in a contained area to avoid contamination. 40µl of RNase free DEPC treated H_2O was added to the mixture and mixed by pipetting using a micropipette, before placing in a heat block at 55˚c for 15 minutes, mixing at 2 minute intervals by pipetting the sample.

5µl of re-suspended DNA was put to an aliquot for nanodrop analysis and gel electrophoresis. The nanodrop sample was placed on ice to freeze at -80˚c. The nanodrop programme was blanked using the resuspension liquid (RNase free DEPC treated H_2O), and RNA quality of the yeast samples were measured adding 1 µl of sample. Quality RNA should have a A260/A280 ratio of ≥1.7. RNA samples below 1.7 were discarded.

For gel electrophoresis, the 10x TBE Stock buffer was diluted to 1x TBE using DEPC H2O. For a 1% Agarose gel, 0.40g of LE Agarose was added to 40ml DEPC TBE buffer, and the solution was boiled in the microwave until dissolved (1minute on 700W, stirring intermittently). 4 µl of SYBR-SAFE (*Thermofisher*, 2017) gel stain was added. The gel casting tray for the mini RNA specific gel tank was set appropriately and hot gel added to the tray, before leaving to set for 30 minutes. Once set, the gel tank was loaded with DEPC TBE buffer, and gel was placed in gel tank. DEPC TBE buffer was added to the gel tank to maximum level and combs were removed from the gel, ensuring the wells had set sufficiently. 1 ul loading dye was added to 3 µl RNA sample before loading the samples to the wells. For a quality RNA sample, 18S and 28S rRNA bands should be visible on the gel image, with 28S intensity twice that of 18S.

Results varied with the Trizol extraction protocol. Although concentration and yield was high,

absorbance ratios were less than required submission rations for WGS. The low 260/230nm was hypothesised to be due to phenol or ethanol contamination. Results led to an alternative methodology to be employed for RNA extraction.

A.1.2 RNASwift recipes

The RNASwift protocol was employed with the optimisation of RNA yield from yeast. The following solution were utilised in the protocol (Nwokeoji et al., 2016).

LB1 Lysis Reagent

- 4% SDS (ph 7.5)
- 0.5M Sodium Chloride (100µl)

Purification Reagent

- 40ul 5M Sodium Chloride
- 250ul 1M Guanadine HCl
- 250ul Isopropanol

Wash Buffer

- 15mM Tris HCl
- 85% ethanol (ph 7.4)

RNASwift protocol

2ml of culture was pipetted in a 2ml microcentrifuge tube. The samples were centrifuged for 4500x*g* for 10 minutes to form a pellet. Liquid media was discarded. 100µl of LB1 Lysis Reagent was warmed and added to the pellet. The resuspension was heated for 4 minutes at 90 ˚c using a heat block, and mixed by pipetting. The samples were then centrifuged for 4 minutes at 12470x*g*. Post centrifugation, the supernatant was transferred to a new column. 540ul of Purification Reagent was added to each sample, and centrifuged for a further minute at 12470x*g*. The flow through was discarded, and 700µl of Wash Buffer was added to each column. The samples underwent centrifugation for a further 1 minute at 12470x*g*. The flow through was again discarded, and dry column was recentrifuged at 12470x*g* for 1 minute to ensure all Wash Buffer was removed from

the column. The samples were then eluted, with the addition of 100µl RNase free water. The columns were added to 2ml microcentifuge tubes for sample collection followed by centrifugation for 1 minute at 12470x*g*, and eluted RNA was stored at -80˚c.

A.1.3 DNA extraction using DNeasy Qiagen kit

Buffer ATL (Qiagen, 2017)

- Sodium Dodecyl Sulphate

Buffer AL (Qiagen, 2017)

- Guanidine Hydrochloride
- Maleic Acid

A.2 Yeast Husbandry

YPD Broth Constituents (Aldrich, 2017)

- Bacteriological peptone (g/L), 20
- Yeast extract (g/L), 10
- Glucose (g/L), 20

DNeasy Qiagen protocol

Cells were centrifuged at 5000x*g* for 10 minutes at room temperature to harvest the yeast cells and the supernatant was discarded. The yeast pellet was resuspended in 600 µl of sorbitol buffer, and add 200 µl of lyticase to the mixture to lyse the yeast cell walls. The mixture was then incubated at 30˚c for 30 minutes. The mixture was then pelleted by centrifugation at 300g for 10 minutes at 4˚c and resuspended in 18 µl of Buffer ATL (constiuents in Appendix A). 20 µl of proteinase K was then added to the spheroplasts, and vortexed to ensure thorough mixing. The solution was then incubated at 56˚C for 15 minutes. The mixture was mixed by pipetting at intervals of two minutes to ensure sample dispersion. When lysis was completed, the sample was vortexed for 15 seconds. 200 µl of Buffer AL (constituents in Appendix A) was added to the sample and mixed thoroughly by further vortexing. 200 µl of ethanol (100%) was then added and mixed to form a homogenous

solution. The mixture was then pipetted into the DNeasy Mini spin column, which was paced in a 2ml collection tube. Centrifugation was performed at 6000g for 1 minute. The DNeasy Mini spin column was placed in a new 2ml collection tube, and 500 µl of Buffer AW1, and centrifuged for a further 1 minute at 6000g. The mini spin column was placed in another 2ml collection tube, and 500 µl of Buffer AW2 and centrifuged for 3 minutes at 20000g. The DNeasy Mini spin column was placed in a 2ml microcentrifuge tube, and 200 µl of Buffer AE was added directly on the DNeasy membrance, which was incubated for 1 minute at room temperature, and then centrifuged at 6000g at 1 minute.

DNA extraction LiOAC-SDS method

One yeast colony from fridge stocks was added to 5ml YPD broth and incubated overnight at 25˚c. 500 µl of sample was added to a clean 2ml microcentrifuuge tube, and pelleted by centrifugation at 5000g for 5 minutes. The supernatant was removed and the cells were suspended in 100 µl of 200mM LiOAc), 1% SDS solution, and incubated at 70˚c for 5 minutes. 300 µl of ethanol (100%) was added and the sample was vortexed. The DNA and cell debris was spun by centrifugation at 15,000g for 3 minutes and supernatant removed. The pellet was then washed in ethanol (70%), and air dried for 2 minutes. The pellet was then dissolved in 100ul RNase free water, and left to incubate at room temperature for 30 minutes. A final centrifugation step was added for 30 seconds at 15,000g. The gDNA (supernatant) was removed and stored at -20˚c.

A.3 QC results for *Kazachstania* species

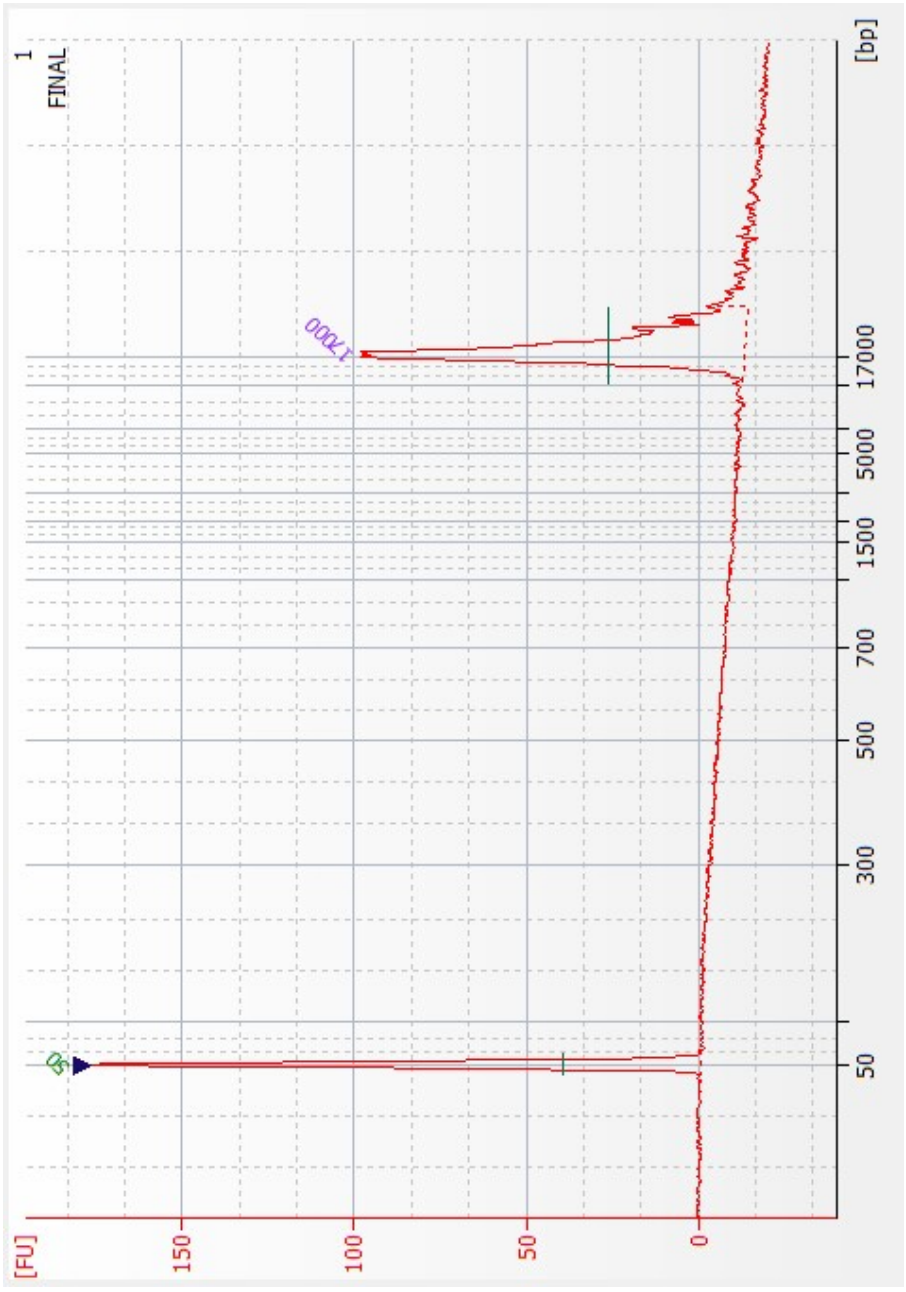


Figure A.1: **gDNA separation of QC DNA Bioanalyser 12000 chip and DNA QC-Picogen for *K. bovina*.** DNA size (bp) is plotted on the x axis, and fluorescence units (FU) on the y-axis. A lower and upper marker were ran with the DNA samples to bracket the DNA sizing analysis.

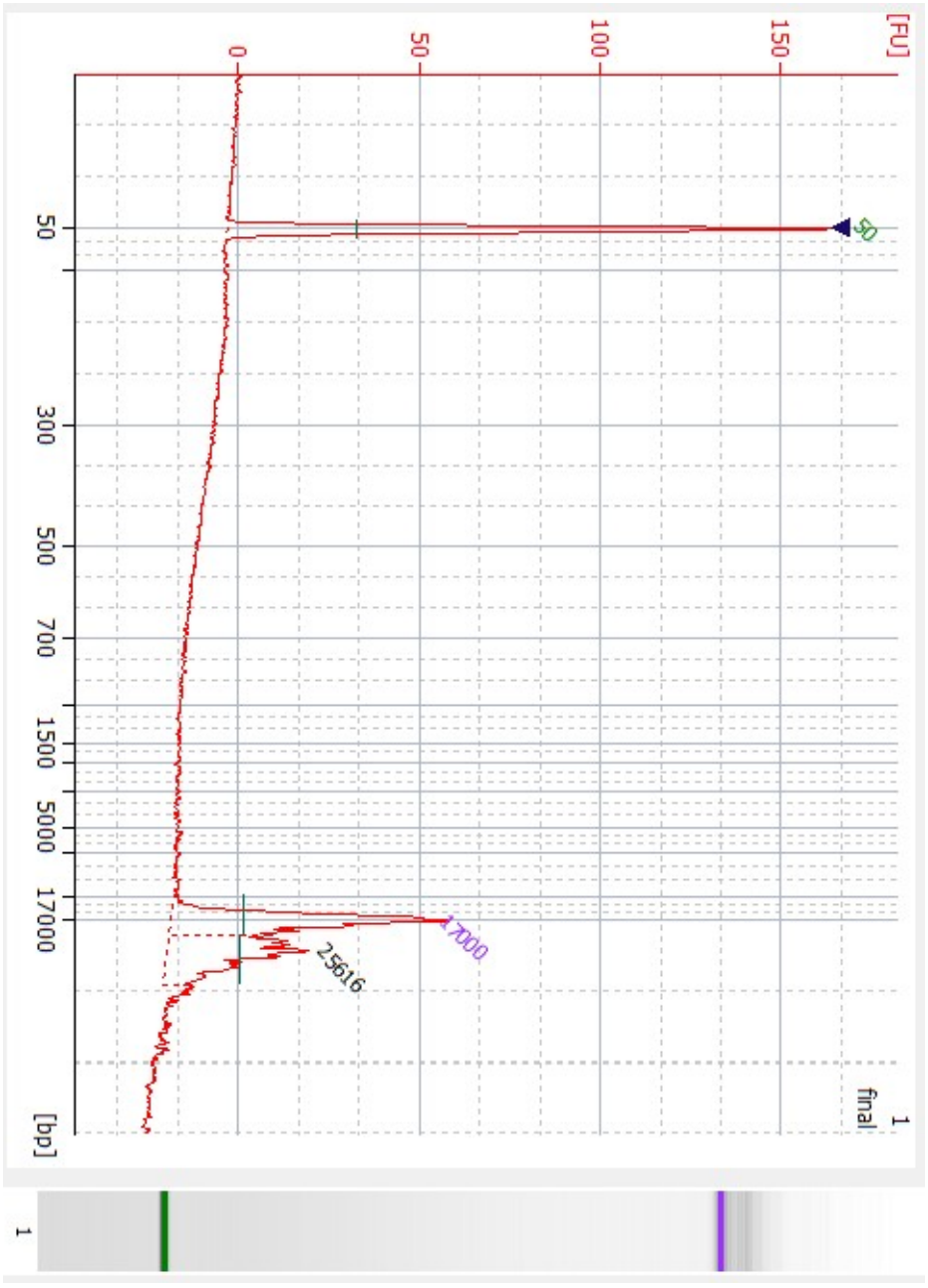


Figure A.2: **gDNA separation results of QC DNA Bioanalyser 12000 chip and DNA QC-Picogen for *K. exigua*.** Format is stated in Figure A.1.

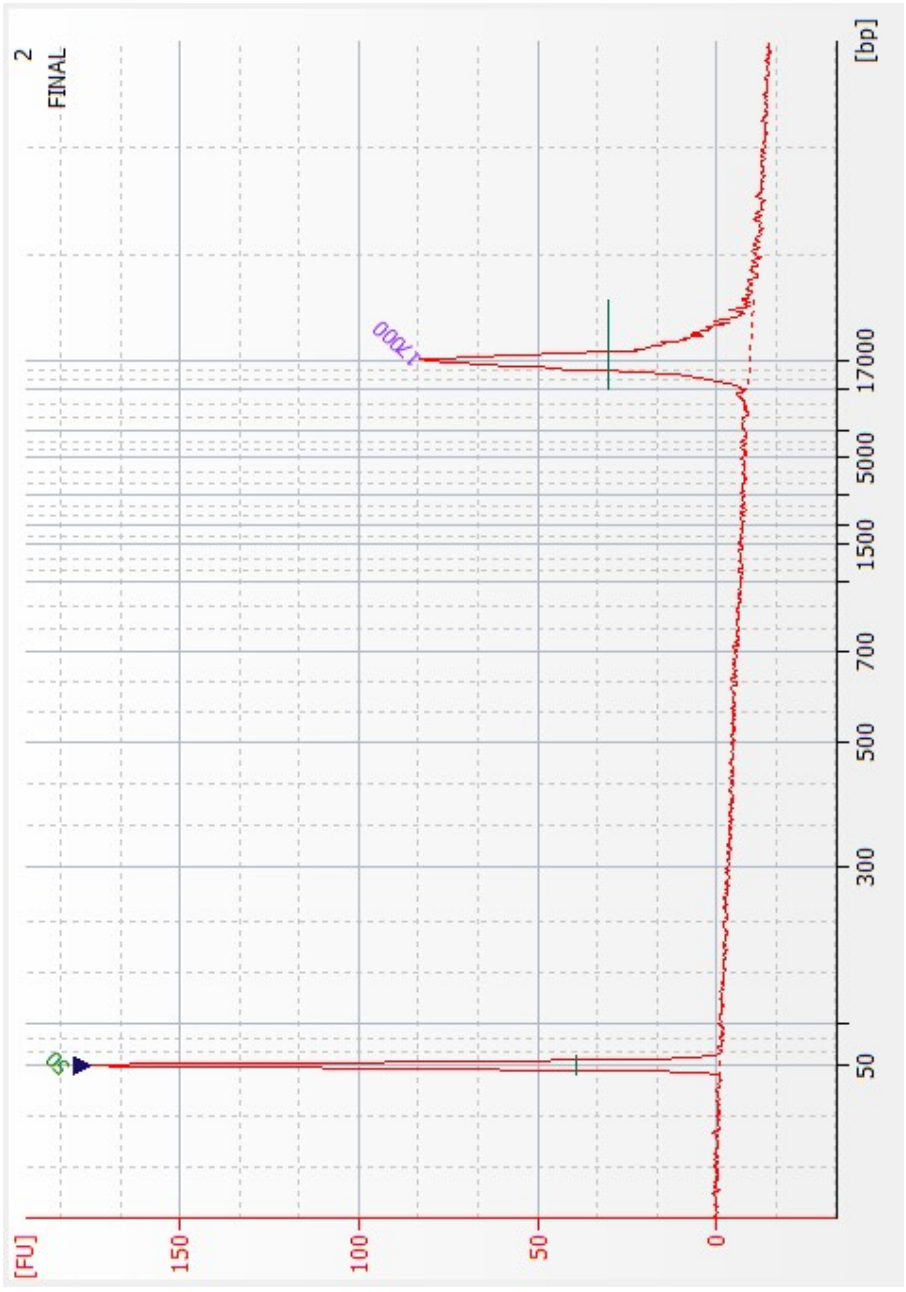


Figure A.3: **Results of QC DNA Bioanalyser 12000 chip and DNA QC-Picogen for *K. lodderae*.** Format is stated in Figure A.1.

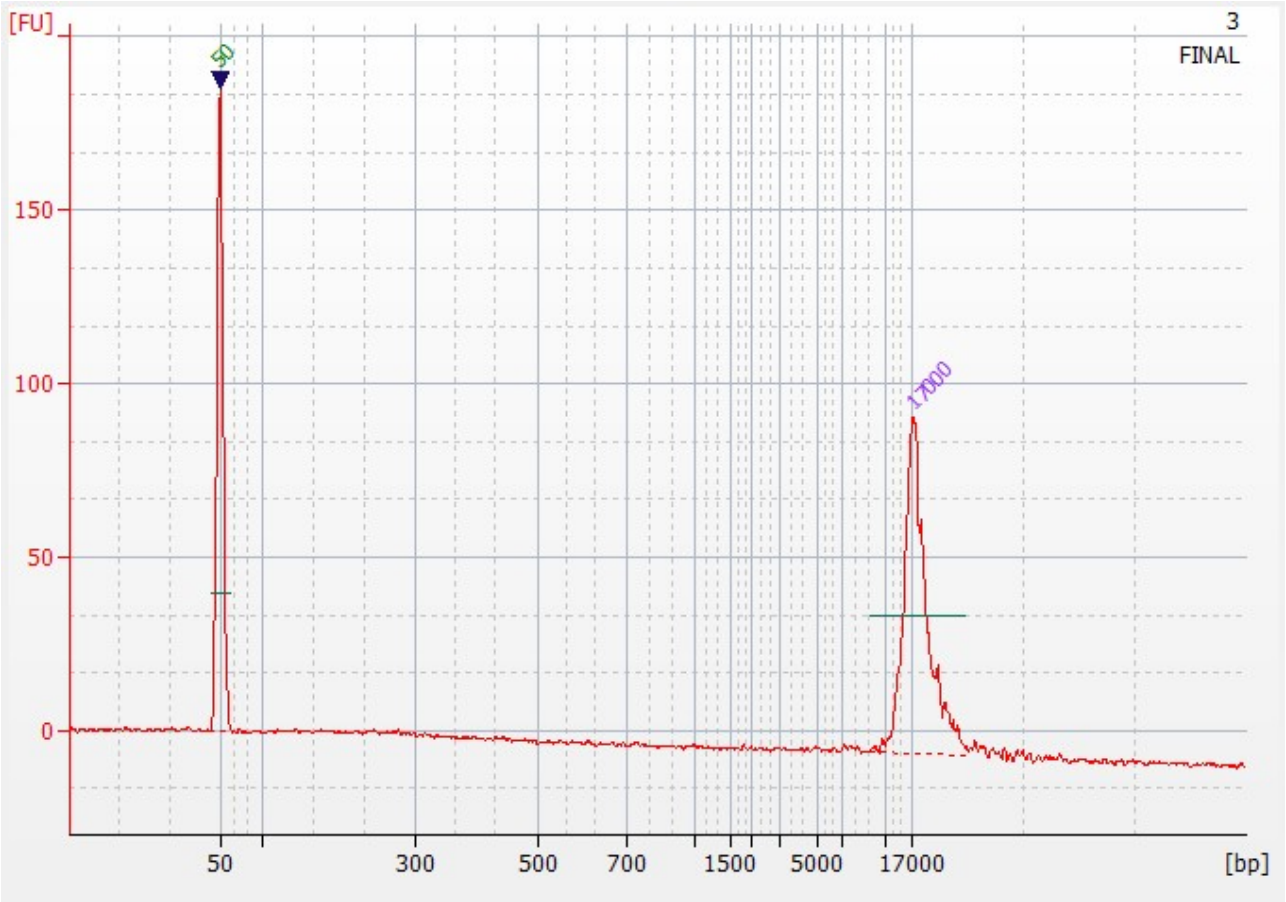


Figure A.4: **gDNA separation results of QC DNA Bioanalyser 12000 chip and DNA QC-Picogen for *K. viticola*.** Format is stated in Figure A.1.

A.4 Novel *Kazachstania* genomes

A.4.1 Genome assembly data

Table A.1: **Summary of assembly data for *Kazachstania* species**. The following assembly data are arranged in the table as follows, provided by Macrogen (2018): Contigs : The number of contigs assembled; Total Length : The total length of contigs; .N50 : 50% of all bases come from contigs longer than this value; Max Length : The length of maximum contig; Min Length : The length of minimum contig; Avg Length : The average length of contigs assembled.

Kazachstania species	Contigs	Total Length	N50	Max Length	Min Length	Avg Length
K. bovina	24	11,441,739	1,199,098	2,187,917	14,594	476,739
K. exigua	32	24,805,022	1,158,375	2,690,170	6,863	775,156
K. lodderae	17	12,438,264	1,171,635	1,514,085	5,353	731,662
K. viticola	30	11,546,415	552,107	1,132,070	12,728	384,880

Table A.2: **Results of assembly data for *K. bovina***. The following assembly data are arranged in the table as follows, provided by Macrogen (2018): Length(bp): The number of bases in each contig; GC%: the percentage of GC content for each contig; Depth: The number of reads which overlap each contig.

Contig Name	Length (bp)	GC %	Depth
contig1	2,187,917	28.40	67
contig2	1,792,149	28.30	71
contig3	1,441,763	28.40	70
contig4	1,199,098	28.10	72
contig5	1,132,946	28.20	67
contig6	1,052,801	28.10	67
contig7	564,760	28.90	62
contig8	526,411	27.90	68
contig9	339,562	28.60	68
contig10	319,215	28.70	60
contig11	247,320	28.30	60
contig12	167,536	27.10	54
contig13	63,798	28.60	35
contig14	61,597	28.0	24
contig15	53,164	27.90	25
contig16	41,325	22.60	464
contig17	38,884	31.20	12
contig18	37,917	27.10	23
contig19	35,972	27.50	36
contig20	35,804	26.90	30
contig21	34,377	29.40	15
contig22	27,174	30.80	14
contig23	25,655	28.90	21
contig24	14,594	28.90	14
Total	11,441,739	28.27	67

Table A.3: **Results of assembly data for *K. exigua***. Format is stated in A.2.

Contig Name	Length (bp)	GC %	Depth
contig1	2,690,170	32.40	34
contig2	2,100,254	31.60	38
contig3	1,569,051	33.20	32
contig4	1,467,721	31.70	39
contig5	1,423,959	33.0	34
contig6	1,304,938	32.90	32
contig7	1,213,532	31.60	46
contig8	1,158,375	31.60	35
contig9	1,049,099	33.20	32
contig10	1,039,954	33.40	34
contig11	982,772	31.70	33
contig12	878,286	31.80	32
contig13	856,810	33.10	30
contig14	827,808	31.60	33
contig15	827,116	33.70	35
contig16	690,675	31.50	34
contig17	668,569	33.30	34
contig18	621,239	32.30	36
contig19	588,913	32.30	32
contig20	576,670	32.60	65
contig21	531,489	33.10	30
contig22	482,016	31.50	28
contig23	382,152	35.60	30
contig24	358,041	33.0	33
contig25	204,536	34.20	45
contig26	137,340	34.40	138
contig27	69,103	35.0	50
contig28	63,842	34.20	64
contig29	14,393	44.90	1,562
contig30	11,153	37.50	2
contig31	8,183	44.40	3
contig32	6,863	29.80	0
Total	24,805,022	32.50	36

Table A.4: **Results of assembly data for *K. Iodderae***. Format is stated in A.2.

Contig Name	Length (bp)	GC %	Depth
contig1	1,514,085	33.9	86
contig2	1,308,446	34.2	99
contig3	1,291,226	33.9	82
contig4	1,242,119	33.5	86
contig5	1,171,635	33.7	83
contig6	899,158	33.8	84
contig7	810,586	34.1	82
contig8	807,199	33.8	84
contig9	787,865	33.7	89
contig10	672,296	34.1	88
contig11	642,064	33.9	82
contig12	511,507	33.9	87
contig13	392,444	33.8	82
contig14	328,278	33.9	86
contig15	31,631	21.0	396
contig16	22,372	34.0	22
contig17	5,353	43.4	7
Total	12,438,264	33.84	86

Table A.5: **Results of assembly data for *K. viticola***. Format is stated in A.2.

Contig Name	Length (bp)	GC %	Depth
contig1	1,132,070	32.80	92
contig2	1,083,222	32.90	86
contig3	954,296	33.20	90
contig4	869,152	32.70	83
contig5	705,520	32.80	89
contig6	696,975	32.80	89
contig7	552,107	33.10	90
contig8	521,844	32.60	89
contig9	517,668	33.20	180
contig10	493,643	32.50	87
contig11	479,321	32.50	83
contig12	464,270	32.60	87
contig13	434,284	33.10	84
contig14	338,307	33.0	94
contig15	330,165	32.80	83
contig16	328,661	32.70	85
contig17	313,821	32.40	85
contig18	299,257	33.20	83
contig19	286,982	32.60	91
contig20	201,438	32.80	83
contig21	182,978	31.70	65
contig22	66,924	16.50	272
contig23	63,584	32.30	100
contig24	56,965	31.20	74
contig25	42,244	32.90	36
contig26	34,247	32.10	87
contig27	32,330	32.60	87
contig28	25,805	32.80	31
contig29	25,607	32.30	38
contig30	12,728	31.90	22
Total	11,546,415	32.71	91

Appendix B

Chapter 2 Appendix

B.1 Comparative genomics and transposable element data for *Kazachstania* species

B.1.1 Predicted tRNA genes for the *Kazachstania* species

Table B.1: **Output of tRNAscan-SE from whole genome contigs for predicted *K. africana* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinates
HE650821.1	Ala	AGC	GCU	Yes	173388-173460
HE650821.1	Ala	AGC	GCU	Yes	457499-457571
HE650822.1	Ala	AGC	GCU	Yes	1354117-1354189
HE650823.1	Ala	AGC	GCU	Yes	411368-411440
HE650823.1	Ala	AGC	GCU	Yes	140332-140404
HE650826.1	Ala	AGC	GCU	Yes	508387-508459
HE650827.1	Ala	AGC	GCU	Yes	657382-657454
HE650827.1	Ala	AGC	GCU	Yes	447095-447167
HE650828.1	Ala	AGC	GCU	Yes	565511-565583
HE650828.1	Ala	AGC	GCU	Yes	567011-567083
HE650828.1	Ala	AGC	GCU	Yes	408975-409047
HE650830.1	Ala	AGC	GCU	Yes	122114-122186
HE650822.1	Ala	TGC	GCA	No	247102-247174
HE650824.1	Ala	TGC	GCA	No	608880-608952
HE650826.1	Ala	TGC	GCA	No	516470-516542
HE650826.1	Ala	TGC	GCA	No	368058-368130

Table B.1: **Output of tRNAscan-SE from whole genome contigs for predicted *K. africana* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinates
HE650832.1	Ala	TGC	GCA	No	202338-202410
HE650821.1	Arg	ACG	CGU	No	520771-520843
HE650822.1	Arg	ACG	CGU	No	1357529-1357601
HE650822.1	Arg	ACG	CGU	No	482397-482469
HE650828.1	Arg	ACG	CGU	No	433389-433461
HE650821.1	Arg	CCG	CGG	No	1286140-1286211
HE650824.1	Arg	CCT	AGG	No	883158-883229
HE650822.1	Arg	TCT	AGA	Yes	1374360-1374431
HE650823.1	Arg	TCT	AGA	Yes	48410-48481
HE650824.1	Arg	TCT	AGA	Yes	298968-299039
HE650825.1	Arg	TCT	AGA	Yes	326261-326332
HE650825.1	Arg	TCT	AGA	Yes	402988-403059
HE650825.1	Arg	TCT	AGA	Yes	376659-376730
HE650826.1	Arg	TCT	AGA	Yes	612486-612557
HE650826.1	Arg	TCT	AGA	Yes	249780-249851
HE650827.1	Arg	TCT	AGA	Yes	479175-479246
HE650828.1	Arg	TCT	AGA	Yes	230334-230405
HE650831.1	Arg	TCT	AGA	Yes	225305-225376
HE650821.1	Asn	GTT	AAC	Yes	1137738-1137811
HE650822.1	Asn	GTT	AAC	Yes	315405-315478
HE650822.1	Asn	GTT	AAC	Yes	1362887-1362960
HE650823.1	Asn	GTT	AAC	Yes	685783-685856
HE650823.1	Asn	GTT	AAC	Yes	453056-453129
HE650824.1	Asn	GTT	AAC	Yes	333687-333760
HE650828.1	Asn	GTT	AAC	Yes	581985-582058
HE650829.1	Asn	GTT	AAC	Yes	110458-110531
HE650829.1	Asn	GTT	AAC	Yes	495991-496064
HE650830.1	Asn	GTT	AAC	Yes	354437-354510

Table B.1: **Output of tRNAscan-SE from whole genome contigs for predicted *K. africana* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinates
HE650822.1	Asp	GTC	GAC	No	1374271-1374342
HE650822.1	Asp	GTC	GAC	No	909693-909764
HE650823.1	Asp	GTC	GAC	No	48319-48390
HE650824.1	Asp	GTC	GAC	No	298877-298948
HE650825.1	Asp	GTC	GAC	No	326352-326423
HE650825.1	Asp	GTC	GAC	No	402897-402968
HE650825.1	Asp	GTC	GAC	No	376568-376639
HE650826.1	Asp	GTC	GAC	No	612577-612648
HE650827.1	Asp	GTC	GAC	No	604876-604947
HE650827.1	Asp	GTC	GAC	No	479084-479155
HE650828.1	Asp	GTC	GAC	No	230243-230314
HE650830.1	Asp	GTC	GAC	No	165036-165107
HE650832.1	Asp	GTC	GAC	No	61831-61902
HE650821.1	Cys	GCA	UGC	No	232340-232411
HE650821.1	Cys	GCA	UGC	No	339314-339385
HE650821.1	Cys	GCA	UGC	No	1058113-1058184
HE650823.1	Cys	GCA	UGC	No	475123-475194
HE650825.1	Cys	GCA	UGC	No	618500-618571
HE650822.1	Gln	CTG	CAG	No	795757-795828
HE650821.1	Gln	TTG	CAA	Yes	193772-193843
HE650821.1	Gln	TTG	CAA	Yes	488121-488192
HE650821.1	Gln	TTG	CAA	Yes	184775-184846
HE650823.1	Gln	TTG	CAA	Yes	104607-104678
HE650827.1	Gln	TTG	CAA	Yes	295219-295290
HE650828.1	Gln	TTG	CAA	Yes	387160-387231
HE650828.1	Gln	TTG	CAA	Yes	560479-560550
HE650829.1	Gln	TTG	CAA	Yes	399483-399554
HE650832.1	Gln	TTG	CAA	Yes	142971-143042

Table B.1: **Output of tRNAscan-SE from whole genome contigs for predicted *K. africana* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinates
HE650823.1	Glu	CTC	GAG	No	1212915-1212986
HE650828.1	Glu	CTC	GAG	No	87896-87967
HE650821.1	Glu	TTC	GAA	Yes	432450-432521
HE650821.1	Glu	TTC	GAA	Yes	1305480-1305551
HE650821.1	Glu	TTC	GAA	Yes	972747-972818
HE650821.1	Glu	TTC	GAA	Yes	945376-945447
HE650821.1	Glu	TTC	GAA	Yes	869484-869555
HE650822.1	Glu	TTC	GAA	Yes	331102-331173
HE650822.1	Glu	TTC	GAA	Yes	478303-478374
HE650822.1	Glu	TTC	GAA	Yes	269330-269401
HE650823.1	Glu	TTC	GAA	Yes	238884-238955
HE650824.1	Glu	TTC	GAA	Yes	518796-518867
HE650825.1	Glu	TTC	GAA	Yes	263356-263427
HE650826.1	Glu	TTC	GAA	Yes	743619-743690
HE650826.1	Glu	TTC	GAA	Yes	354525-354596
HE650827.1	Glu	TTC	GAA	Yes	571184-571255
HE650827.1	Glu	TTC	GAA	Yes	755072-755143
HE650825.1	Gly	CCC	GGG	No	726601-726672
HE650821.1	Gly	GCC	GGC	No	564256-564326
HE650821.1	Gly	GCC	GGC	No	1383638-1383708
HE650821.1	Gly	GCC	GGC	No	390631-390701
HE650822.1	Gly	GCC	GGC	No	961437-961507
HE650822.1	Gly	GCC	GGC	No	864762-864832
HE650822.1	Gly	GCC	GGC	No	580061-580131
HE650822.1	Gly	GCC	GGC	No	366952-367022
HE650823.1	Gly	GCC	GGC	No	888412-888482
HE650823.1	Gly	GCC	GGC	No	1181502-1181572
HE650823.1	Gly	GCC	GGC	No	909796-909866

Table B.1: **Output of tRNAscan-SE from whole genome contigs for predicted *K. africana* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinates
HE650824.1	Gly	GCC	GGC	No	341914-341984
HE650825.1	Gly	GCC	GGC	No	411118-411188
HE650828.1	Gly	GCC	GGC	No	608892-608962
HE650828.1	Gly	GCC	GGC	No	543590-543660
HE650829.1	Gly	GCC	GGC	No	566773-566843
HE650831.1	Gly	GCC	GGC	No	207465-207535
HE650831.1	Gly	GCC	GGC	No	305379-305449
HE650821.1	Gly	TCC	GGA	No	1654735-1654806
HE650823.1	Gly	TCC	GGA	No	317560-317631
HE650823.1	Gly	TCC	GGA	No	414809-414880
HE650825.1	Gly	TCC	GGA	No	234360-234435
HE650821.1	His	GTG	CAC	Yes	755707-755778
HE650822.1	His	GTG	CAC	Yes	1081343-1081414
HE650822.1	His	GTG	CAC	Yes	433807-433878
HE650823.1	His	GTG	CAC	Yes	568045-568116
HE650826.1	His	GTG	CAC	Yes	801407-801478
HE650826.1	His	GTG	CAC	Yes	545762-545833
HE650822.1	Ile	AAT	AUU	Yes	94373-94446
HE650822.1	Ile	AAT	AUU	Yes	226511-226584
HE650822.1	Ile	AAT	AUU	Yes	437206-437279
HE650822.1	Ile	AAT	AUU	Yes	1382182-1382255
HE650822.1	Ile	AAT	AUU	Yes	1401362-1401435
HE650823.1	Ile	AAT	AUU	Yes	380766-380839
HE650823.1	Ile	AAT	AUU	Yes	563630-563703
HE650826.1	Ile	AAT	AUU	Yes	608774-608847
HE650828.1	Ile	AAT	AUU	Yes	279963-280036
HE650828.1	Ile	AAT	AUU	Yes	637378-637451
HE650829.1	Ile	AAT	AUU	Yes	445642-445715

Table B.1: **Output of tRNAscan-SE from whole genome contigs for predicted *K. africana* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinates
HE650830.1	Ile	AAT	AUU	Yes	235842-235915
HE650832.1	Ile	AAT	AUU	Yes	215287-215360
HE650822.1	Ile	TAT	AUA	No	1304284-1304422
HE650828.1	Ile	TAT	AUA	No	337427-337564
HE650825.1	iMe	CAT	AUG	Yes	126585-126656
HE650826.1	iMe	CAT	AUG	Yes	251822-251893
HE650827.1	iMe	CAT	AUG	Yes	441118-441189
HE650828.1	iMe	CAT	AUG	Yes	521998-522069
HE650821.1	Leu	AAG	CUU	No	1076694-1076775
HE650831.1	Leu	AAG	CUU	No	352556-352637
HE650824.1	Leu	CAA	UUG	No	806727-806844
HE650826.1	Leu	CAA	UUG	No	694414-694531
HE650821.1	Leu	TAA	UUA	Yes	191220-191302
HE650821.1	Leu	TAA	UUA	Yes	1174552-1174634
HE650821.1	Leu	TAA	UUA	Yes	631543-631625
HE650823.1	Leu	TAA	UUA	Yes	102083-102165
HE650823.1	Leu	TAA	UUA	Yes	762058-762140
HE650823.1	Leu	TAA	UUA	Yes	207257-207339
HE650824.1	Leu	TAA	UUA	Yes	148565-148647
HE650824.1	Leu	TAA	UUA	Yes	180679-180761
HE650826.1	Leu	TAA	UUA	Yes	655848-655930
HE650826.1	Leu	TAA	UUA	Yes	591730-591812
HE650827.1	Leu	TAA	UUA	Yes	706737-706819
HE650828.1	Leu	TAA	UUA	Yes	118109-118191
HE650830.1	Leu	TAA	UUA	Yes	489209-489291
HE650831.1	Leu	TAA	UUA	Yes	42887-42969
HE650827.1	Leu	TAG	CUA	No	210068-210169
HE650829.1	Leu	TAG	CUA	No	262543-262646

Table B.1: **Output of tRNAscan-SE from whole genome contigs for predicted *K. africana* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinates
HE650821.1	Lys	CTT	AAG	Yes	105766-105838
HE650821.1	Lys	CTT	AAG	Yes	686768-686840
HE650821.1	Lys	CTT	AAG	Yes	938356-938428
HE650822.1	Lys	CTT	AAG	Yes	592190-592262
HE650823.1	Lys	CTT	AAG	Yes	686659-686731
HE650823.1	Lys	CTT	AAG	Yes	1110799-1110871
HE650824.1	Lys	CTT	AAG	Yes	986907-986979
HE650824.1	Lys	CTT	AAG	Yes	245185-245257
HE650827.1	Lys	CTT	AAG	Yes	619809-619881
HE650831.1	Lys	CTT	AAG	Yes	263834-263906
HE650831.1	Lys	CTT	AAG	Yes	188193-188265
HE650831.1	Lys	CTT	AAG	Yes	168072-168144
HE650821.1	Lys	TTT	AAA	No	189828-189926
HE650823.1	Lys	TTT	AAA	No	117038-117136
HE650826.1	Lys	TTT	AAA	No	302087-302185
HE650826.1	Lys	TTT	AAA	No	762660-762758
HE650827.1	Lys	TTT	AAA	No	66111-66209
HE650828.1	Lys	TTT	AAA	No	70299-70397
HE650829.1	Lys	TTT	AAA	No	119508-119606
HE650831.1	Lys	TTT	AAA	No	266503-266601
HE650822.1	Met	CAT	AUG	Yes	1025165-1025237
HE650823.1	Met	CAT	AUG	Yes	1029386-1029458
HE650824.1	Met	CAT	AUG	Yes	969147-969219
HE650825.1	Met	CAT	AUG	Yes	816177-816249
HE650830.1	Met	CAT	AUG	Yes	201450-201522
HE650821.1	Phe	GAA	UUC	Yes	757696-757791
HE650822.1	Phe	GAA	UUC	Yes	1089357-1089452
HE650822.1	Phe	GAA	UUC	Yes	365798-365893

Table B.1: **Output of tRNAscan-SE from whole genome contigs for predicted *K. africana* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinates
HE650826.1	Phe	GAA	UUC	Yes	708904-708998
HE650829.1	Phe	GAA	UUC	Yes	189566-189661
HE650830.1	Phe	GAA	UUC	Yes	55345-55440
HE650830.1	Phe	GAA	UUC	Yes	426772-426867
HE650831.1	Phe	GAA	UUC	Yes	466109-466204
HE650831.1	Phe	GAA	UUC	Yes	427646-427741
HE650829.1	Pro	AGG	CCU	No	477348-477419
HE650821.1	Pro	TGG	CCA	Yes	559105-559214
HE650822.1	Pro	TGG	CCA	Yes	627311-627420
HE650823.1	Pro	TGG	CCA	Yes	805797-805906
HE650824.1	Pro	TGG	CCA	Yes	911279-911388
HE650824.1	Pro	TGG	CCA	Yes	961120-961230
HE650825.1	Pro	TGG	CCA	Yes	508532-508641
HE650826.1	Pro	TGG	CCA	Yes	293417-293526
HE650828.1	Pro	TGG	CCA	Yes	146496-146605
HE650829.1	Pro	TGG	CCA	Yes	462701-462810
HE650829.1	Pro	TGG	CCA	Yes	553103-553212
HE650821.1	Ser	AGA	UCU	Yes	1042296-1042377
HE650821.1	Ser	AGA	UCU	Yes	168128-168209
HE650823.1	Ser	AGA	UCU	Yes	1079778-1079859
HE650823.1	Ser	AGA	UCU	Yes	1150373-1150454
HE650823.1	Ser	AGA	UCU	Yes	152961-153042
HE650825.1	Ser	AGA	UCU	Yes	116255-116336
HE650826.1	Ser	AGA	UCU	Yes	477198-477279
HE650826.1	Ser	AGA	UCU	Yes	730038-730119
HE650826.1	Ser	AGA	UCU	Yes	626927-627008
HE650830.1	Ser	AGA	UCU	Yes	106823-106904
HE650830.1	Ser	AGA	UCU	Yes	132432-132513

Table B.1: **Output of tRNAscan-SE from whole genome contigs for predicted *K. africana* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinates
HE650826.1	Ser	CGA	UCG	No	568399-568509
HE650821.1	Ser	GCT	AGC	No	1334983-1335084
HE650821.1	Ser	GCT	AGC	No	1104357-1104458
HE650824.1	Ser	GCT	AGC	No	957832-957933
HE650826.1	Ser	GCT	AGC	No	640409-640510
HE650822.1	Ser	TGA	UCA	No	262247-262328
HE650824.1	Ser	TGA	UCA	No	531454-531535
HE650828.1	Ser	TGA	UCA	No	164962-165043
HE650832.1	Ser	TGA	UCA	No	175682-175763
HE650821.1	Thr	AGT	ACU	Yes	355276-355348
HE650821.1	Thr	AGT	ACU	Yes	345133-345205
HE650821.1	Thr	AGT	ACU	Yes	242549-242621
HE650823.1	Thr	AGT	ACU	Yes	307523-307595
HE650823.1	Thr	AGT	ACU	Yes	132796-132868
HE650824.1	Thr	AGT	ACU	Yes	276267-276339
HE650827.1	Thr	AGT	ACU	Yes	546909-546981
HE650827.1	Thr	AGT	ACU	Yes	71034-71106
HE650828.1	Thr	AGT	ACU	Yes	646867-646939
HE650830.1	Thr	AGT	ACU	Yes	261194-261266
HE650831.1	Thr	AGT	ACU	Yes	184357-184429
HE650831.1	Thr	CGT	ACG	No	527254-527325
HE650821.1	Thr	TGT	ACA	No	1103343-1103414
HE650822.1	Thr	TGT	ACA	No	932314-932385
HE650831.1	Thr	TGT	ACA	No	314620-314691
HE650821.1	Trp	CCA	UGG	Yes	906332-906432
HE650823.1	Trp	CCA	UGG	Yes	820888-820988
HE650825.1	Trp	CCA	UGG	Yes	858257-858357
HE650827.1	Trp	CCA	UGG	Yes	453206-453306

Table B.1: **Output of tRNAscan-SE from whole genome contigs for predicted *K. africana* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinates
HE650831.1	Trp	CCA	UGG	Yes	265761-265861
HE650832.1	Trp	CCA	UGG	Yes	88005-88105
HE650821.1	Tyr	GTA	UAC	Yes	596654-596745
HE650823.1	Tyr	GTA	UAC	Yes	865346-865437
HE650823.1	Tyr	GTA	UAC	Yes	900262-900353
HE650823.1	Tyr	GTA	UAC	Yes	462464-462555
HE650824.1	Tyr	GTA	UAC	Yes	885297-885388
HE650825.1	Tyr	GTA	UAC	Yes	81661-81752
HE650825.1	Tyr	GTA	UAC	Yes	567444-567535
HE650829.1	Tyr	GTA	UAC	Yes	405198-405289
HE650832.1	Tyr	GTA	UAC	Yes	368616-368708
HE650821.1	Val	AAC	GUU	Yes	992668-992741
HE650821.1	Val	AAC	GUU	Yes	400287-400360
HE650822.1	Val	AAC	GUU	Yes	528944-529017
HE650822.1	Val	AAC	GUU	Yes	444896-444969
HE650822.1	Val	AAC	GUU	Yes	402354-402427
HE650823.1	Val	AAC	GUU	Yes	532857-532930
HE650823.1	Val	AAC	GUU	Yes	1048749-1048822
HE650823.1	Val	AAC	GUU	Yes	495876-495949
HE650829.1	Val	AAC	GUU	Yes	256355-256428
HE650831.1	Val	AAC	GUU	Yes	472062-472135
HE650831.1	Val	AAC	GUU	Yes	511218-511291
HE650831.1	Val	AAC	GUU	Yes	355562-355635
HE650831.1	Val	AAC	GUU	Yes	299026-299099
HE650825.1	Val	CAC	GUG	No	255581-255653
HE650822.1	Val	TAC	GUA	No	304091-304163
HE650822.1	Val	TAC	GUA	No	939313-939385

Table B.2: **Output of tRNAscan-SE from whole genome contigs for predicted *K. bovina* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig2	Ala	TGC	GCA	No	44830-44925
contig2	Ala	TGC	GCA	No	307127-307198
contig5	Ala	TGC	GCA	No	417035-417106
contig5	Ala	TGC	GCA	No	1110564-1110637
contig6	Ala	TGC	GCA	No	1701474-1701545
contig6	Ala	TGC	GCA	No	2018422-2018493
contig6	Ala	TGC	GCA	No	2053607-2053702
contig6	Ala	TGC	GCA	No	2187753-2187826
contig7	Ala	TGC	GCA	No	1812541-1812613
contig10	Ala	TGC	GCA	No	1764295-1764366
contig12	Ala	TGC	GCA	No	1609804-1609876
contig16	Ala	TGC	GCA	No	1385611-1385684
contig2	Arg	ACG	CGT	No	914718-914810
contig2	Arg	ACG	CGT	No	806880-807001
contig4	Arg	ACG	CGT	No	794079-794152
contig4	Arg	CCG	CGG	No	548171-548243
contig6	Arg	CCT	AGG	No	283820-283892
contig6	Arg	TCT	AGA	Yes	3275-3346
contig6	Arg	TCT	AGA	Yes	136087-136178
contig8	Arg	TCT	AGA	Yes	446955-447027
contig8	Arg	TCT	AGA	Yes	448184-448255
contig8	Arg	TCT	AGA	Yes	508330-508401
contig9	Arg	TCT	AGA	Yes	520846-520919
contig16	Arg	TCT	AGA	Yes	760666-760749
contig16	Arg	TCT	AGA	Yes	794169-794242
contig19	Arg	TCT	AGA	Yes	884282-884354
contig1	Asn	GTT	AAC	No	885199-885270
contig2	Asn	GTT	AAC	No	1168169-1168242

Table B.2: **Output of tRNAscan-SE from whole genome contigs for predicted *K. bovina* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig2	Asn	GTT	AAC	No	1440227-1440309
contig3	Asn	GTT	AAC	No	1764439-1764511
contig3	Asn	GTT	AAC	No	1783830-1783901
contig6	Asn	GTT	AAC	No	1785484-1785555
contig8	Asn	GTT	AAC	No	1737266-1737357
contig10	Asn	GTT	AAC	No	1638336-1638419
contig16	Asn	GTT	AAC	No	1592484-1592556
contig21	Asn	GTT	AAC	No	1573593-1573688
contig23	Asn	GTT	AAC	No	1573496-1573569
contig1	Asp	GTC	GAC	No	1475008-1475091
contig2	Asp	GTC	GAC	No	1378584-1378657
contig2	Asp	GTC	GAC	No	1378501-1378573
contig3	Asp	GTC	GAC	No	1236272-1236345
contig3	Asp	GTC	GAC	No	1082356-1082427
contig4	Asp	GTC	GAC	No	1017170-1017241
contig4	Asp	GTC	GAC	No	822485-822556
contig5	Asp	GTC	GAC	No	752260-752331
contig6	Asp	GTC	GAC	No	728199-728280
contig11	Asp	GTC	GAC	No	536670-536741
contig11	Asp	GTC	GAC	No	395145-395216
contig16	Asp	GTC	GAC	No	297806-297879
contig2	Cys	GCA	TGC	No	100049-100120
contig5	Cys	GCA	TGC	No	7156-77229
contig5	Cys	GCA	TGC	No	7420-77501
contig1	Gln	CTG	CAG	No	49173-149246
contig2	Gln	TTG	CAA	Yes	8568-78638
contig3	Gln	TTG	CAA	Yes	25116-125189
contig3	Gln	TTG	CAA	Yes	69062-269133

Table B.2: **Output of tRNAscan-SE from whole genome contigs for predicted *K. bovina* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig6	Gln	TTG	CAA	Yes	24978-325049
contig8	Gln	TTG	CAA	Yes	338943-1339013
contig1	Glu	CTC	GAG	No	1093739-1093810
contig2	Glu	TTC	GAA	Yes	439601-1439682
contig2	Glu	TTC	GAA	Yes	1424906-1424987
contig2	Glu	TTC	GAA	Yes	1323639-1323710
contig4	Glu	TTC	GAA	Yes	1188128-1188209
contig4	Glu	TTC	GAA	Yes	897848-897919
contig4	Glu	TTC	GAA	Yes	723578-723649
contig4	Glu	TTC	GAA	Yes	186743-186866
contig6	Glu	TTC	GAA	Yes	121387-121460
contig8	Glu	TTC	GAA	Yes	104680-104753
contig10	Glu	TTC	GAA	Yes	42761-42831
contig1	Gly	GCC	GGC	No	39186-39269
contig3	Gly	GCC	GGC	No	37821-37892
contig3	Gly	GCC	GGC	No	594824-594895
contig3	Gly	GCC	GGC	No	806099-806170
contig3	Gly	GCC	GGC	No	806183-806280
contig3	Gly	GCC	GGC	No	856276-856348
contig4	Gly	GCC	GGC	No	873885-873956
contig4	Gly	GCC	GGC	No	915344-915416
contig6	Gly	GCC	GGC	No	915430-915500
contig8	Gly	GCC	GGC	No	915509-915580
contig8	Gly	GCC	GGC	No	969584-969655
contig11	Gly	TCC	GGA	No	42676-42747
contig11	Gly	TCC	GGA	No	37905-37975
contig16	Gly	TCC	GGA	No	969667-969763
contig2	His	GTG	CAC	No	1009235-1009307

Table B.2: **Output of tRNAscan-SE from whole genome contigs for predicted *K. bovina* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig6	His	GTG	CAC	No	1040472-1040555
contig8	His	GTG	CAC	No	1040569-1040640
contig9	His	GTG	CAC	No	1161791-1161874
contig1	Ile	AAT	ATT	Yes	1067299-1067370
contig1	Ile	AAT	ATT	Yes	858527-858597
contig2	Ile	AAT	ATT	Yes	858447-858518
contig2	Ile	AAT	ATT	Yes	857299-857370
contig3	Ile	AAT	ATT	Yes	194229-194300
contig3	Ile	AAT	ATT	Yes	293866-293937
contig5	Ile	AAT	ATT	Yes	374339-374420
contig5	Ile	AAT	ATT	Yes	551414-551485
contig6	Ile	AAT	ATT	Yes	595959-596030
contig6	Ile	AAT	ATT	Yes	733448-733539
contig10	Ile	AAT	ATT	Yes	803681-803753
contig12	Ile	TAT	ATA	No	1002613-1002685
contig16	Ile	TAT	ATA	No	593977-594068
contig2	iMet	CAT	ATG	Yes	1011937-1012018
contig3	iMet	CAT	ATG	Yes	1004973-1005054
contig1	iMet	CAT	ATG	Yes	784121-784202
contig2	Leu	AAG	CTT	No	144444-144525
contig2	Leu	AAG	CTT	No	987950-988031
contig2	Leu	CAA	TTG	No	416950-417021
contig2	Leu	CAG	CTG	No	372009-372082
contig2	Leu	CAG	CTG	No	690676-690748
contig3	Leu	TAA	TTA	Yes	449344-449416
contig3	Leu	TAA	TTA	Yes	371926-371998
contig4	Leu	TAA	TTA	Yes	184138-184209
contig4	Leu	TAA	TTA	Yes	372531-372602

Table B.2: **Output of tRNAscan-SE from whole genome contigs for predicted *K. bovina* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig5	Leu	TAA	TTA	Yes	547967-548039
contig8	Leu	TAA	TTA	Yes	911197-911278
contig8	Leu	TAA	TTA	Yes	913231-913322
contig8	Leu	TAA	TTA	Yes	991119-991192
contig11	Leu	TAA	TTA	Yes	991203-991275
contig16	Leu	TAA	TTA	Yes	1007994-1008075
contig19	Leu	TAA	TTA	Yes	932454-932525
contig19	Leu	TAG	CTA	No	325987-326059
contig1	Lys	CTT	AAG	No	911671-911742
contig2	Lys	CTT	AAG	No	909380-909452
contig2	Lys	CTT	AAG	No	409971-410042
contig4	Lys	CTT	AAG	No	398788-398861
contig4	Lys	CTT	AAG	No	381236-381308
contig4	Lys	CTT	AAG	No	249258-249331
contig5	Lys	CTT	AAG	No	26874-26946
contig6	Lys	TTT	AAA	No	932187-932268
contig8	Lys	TTT	AAA	No	911835-911907
contig8	Lys	TTT	AAA	No	911751-911821
contig8	Lys	TTT	AAA	No	768923-768996
contig10	Lys	TTT	AAA	No	686780-686853
contig11	Lys	TTT	AAA	No	686697-686769
contig14	Lys	TTT	AAA	No	458935-459006
contig16	Lys	TTT	AAA	No	11738-11810
contig18	Lys	TTT	AAA	No	35369-35441
contig20	Lys	TTT	AAA	No	520361-520432
contig2	Met	CAT	ATG	Yes	63751-63822
contig2	Met	CAT	ATG	Yes	310772-310845
contig6	Met	CAT	ATG	Yes	310854-310925

Table B.2: **Output of tRNAscan-SE from whole genome contigs for predicted *K. bovina* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig14	Met	CAT	ATG	Yes	312468-312538
contig16	Met	CAT	ATG	Yes	366392-366483
contig16	Met	CAT	ATG	Yes	406795-406867
contig16	Met	CAT	ATG	Yes	425705-425776
contig24	Met	CAT	ATG	Yes	495338-495433
contig2	Phe	GAA	TTC	Yes	495226-495297
contig4	Phe	GAA	TTC	Yes	478084-478157
contig4	Phe	GAA	TTC	Yes	406345-406428
contig4	Phe	GAA	TTC	Yes	368529-368612
contig6	Phe	GAA	TTC	Yes	365315-365385
contig9	Phe	GAA	TTC	Yes	324274-324370
contig9	Phe	GAA	TTC	Yes	314299-314370
contig16	Phe	GAA	TTC	Yes	314190-314286
contig2	Pro	TGG	CCA	Yes	311352-311423
contig2	Pro	TGG	CCA	Yes	223472-223553
contig3	Pro	TGG	CCA	Yes	193027-193108
contig5	Pro	TGG	CCA	Yes	62839-62911
contig7	Pro	TGG	CCA	Yes	21031-21102
contig8	Pro	TGG	CCA	Yes	119675-119746
contig8	Pro	TGG	CCA	Yes	220679-220750
contig10	Pro	TGG	CCA	Yes	263182-263254
contig3	Ser	AGA	TCT	Yes	57588-57684
contig3	Ser	AGA	TCT	Yes	243073-243146
contig3	Ser	AGA	TCT	Yes	102582-102653
contig5	Ser	AGA	TCT	Yes	49133-49204
contig5	Ser	AGA	TCT	Yes	24980-25051
contig5	Ser	AGA	TCT	Yes	30054-30124
contig5	Ser	AGA	TCT	Yes	30136-30207

Table B.2: **Output of tRNAscan-SE from whole genome contigs for predicted *K. bovina* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig6	Ser	CGA	TCG	No	102695-102790
contig6	Ser	GCT	AGC	No	295998-296071
contig6	Ser	GCT	AGC	No	28739-28822
contig6	Ser	TGA	TCA	Yes	289962-290035
contig8	Ser	TGA	TCA	Yes	21961-22032
contig10	Ser	TGA	TCA	Yes	242990-243062
contig12	Ser	TGA	TCA	Yes	145641-145722
contig12	Ser	TGA	TCA	Yes	24896-24966
contig16	Ser	TGA	TCA	Yes	156104-156200
contig16	Sup	TCA	TGA	Yes	245101-245172
contig1	Thr	AGT	ACT	Yes	129645-129726
contig1	Thr	AGT	ACT	Yes	150190-150271
contig2	Thr	AGT	ACT	Yes	146402-146475
contig2	Thr	AGT	ACT	Yes	146319-146391
contig2	Thr	AGT	ACT	Yes	34228-34301
contig4	Thr	AGT	ACT	Yes	9985-10057
contig4	Thr	AGT	ACT	Yes	10061-10133
contig6	Thr	CGT	ACG	No	8744-8816
contig6	Thr	TGT	ACA	No	23327-23398
contig6	Thr	TGT	ACA	No	28170-28261
contig11	Thr	TGT	ACA	No	10160-10231
contig16	Thr	TGT	ACA	No	15599-15672
contig3	Trp	CCA	TGG	Yes	16102-16180
contig8	Trp	CCA	TGG	Yes	20092-20164
contig10	Trp	CCA	TGG	Yes	20168-20252
contig13	Trp	CCA	TGG	Yes	22642-22713
contig1	Tyr	GTA	TAC	No	22732-22805
contig1	Tyr	GTA	TAC	No	22900-22970

Table B.2: **Output of tRNAscan-SE from whole genome contigs for predicted *K. bovina* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig2	Tyr	GTA	TAC	No	24848-24931
contig8	Tyr	GTA	TAC	No	25035-25107
contig10	Tyr	GTA	TAC	No	25291-25364
contig1	Val	AAC	GTT	Yes	27752-27824
contig2	Val	AAC	GTT	Yes	29560-29630
contig2	Val	AAC	GTT	Yes	19321-19412
contig2	Val	AAC	GTT	Yes	19700-19783
contig3	Val	AAC	GTT	Yes	27934-28017
contig3	Val	AAC	GTT	Yes	29456-29552
contig4	Val	AAC	GTT	Yes	11132-11206
contig6	Val	AAC	GTT	Yes	9209-9282
contig8	Val	CAC	GTG	No	19260-19332
contig9	Val	TAC	GTA	No	27833-27903
contig16	Val	TAC	GTA	No	12868-12941

Table B.3: **Output of tRNAscan-SE from whole genome contigs for predicted *K. exigua* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig1	Ala	AGC	GCT	Yes	462191-462264
contig2	Ala	AGC	GCT	Yes	484478-484549
contig3	Ala	AGC	GCT	Yes	601251-601333
contig3	Ala	AGC	GCT	Yes	608615-608697
contig4	Ala	AGC	GCT	Yes	706942-707013
contig5	Ala	AGC	GCT	Yes	717715-717786
contig5	Ala	AGC	GCT	Yes	744823-744894
contig5	Ala	AGC	GCT	Yes	940231-940304
contig6	Ala	AGC	GCT	Yes	986645-986716
contig6	Ala	AGC	GCT	Yes	1415062-1415162
contig7	Ala	AGC	GCT	Yes	1917396-1917468

Table B.3: **Output of tRNAscan-SE from whole genome contigs for predicted *K. exigua* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig8	Ala	AGC	GCT	Yes	1955303-1955384
contig12	Ala	AGC	GCT	Yes	2124459-2124531
contig14	Ala	AGC	GCT	Yes	2152237-2152309
contig15	Ala	AGC	GCT	Yes	2246455-2246536
contig15	Ala	AGC	GCT	Yes	2425974-2426047
contig16	Ala	TGC	GCA	No	138329-138400
contig16	Ala	TGC	GCA	No	519553-519634
contig17	Ala	TGC	GCA	No	734055-734126
contig18	Ala	TGC	GCA	No	986555-986626
contig21	Ala	TGC	GCA	No	1474669-1474760
contig22	Ala	TGC	GCA	No	1741259-1741332
contig22	Ala	TGC	GCA	No	1908032-1908102
contig1	Arg	ACG	CGT	Yes	2282518-2282590
contig1	Arg	ACG	CGT	Yes	2114335-2114408
contig3	Arg	ACG	CGT	Yes	2071057-2071128
contig3	Arg	ACG	CGT	Yes	1380875-1380947
contig4	Arg	ACG	CGT	Yes	727436-727523
contig4	Arg	ACG	CGT	Yes	357663-357756
contig5	Arg	ACG	CGT	Yes	342474-342547
contig5	Arg	ACG	CGT	Yes	378196-378276
contig5	Arg	CCG	CGG	No	352544-352615
contig7	Arg	CCG	CGG	No	198475-198561
contig10	Arg	CCT	AGG	No	2058430-2058500
contig11	Arg	CCT	AGG	No	483410-483492
contig11	Arg	TCT	AGA	Yes	2471755-2471873
contig11	Arg	TCT	AGA	Yes	2395946-2396017
contig12	Arg	TCT	AGA	Yes	2290586-2290659
contig14	Arg	TCT	AGA	Yes	2255142-2255228

Table B.3: **Output of tRNAscan-SE from whole genome contigs for predicted *K. exigua* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig14	Arg	TCT	AGA	Yes	2201920-2201991
contig15	Arg	TCT	AGA	Yes	1340518-1340590
contig17	Arg	TCT	AGA	Yes	885549-885636
contig17	Arg	TCT	AGA	Yes	754546-754633
contig18	Arg	TCT	AGA	Yes	612078-612160
contig19	Arg	TCT	AGA	Yes	64219-64327
contig20	Arg	TCT	AGA	Yes	146711-146783
contig20	Arg	TCT	AGA	Yes	211995-212076
contig22	Arg	TCT	AGA	Yes	278682-278753
contig23	Arg	TCT	AGA	Yes	441709-441800
contig24	Arg	TCT	AGA	Yes	675549-675630
contig1	Asn	GTT	AAC	Yes	897867-897953
contig1	Asn	GTT	AAC	Yes	961183-961317
contig1	Asn	GTT	AAC	Yes	1089760-1089831
contig2	Asn	GTT	AAC	Yes	1222464-1222534
contig2	Asn	GTT	AAC	Yes	1250112-1250184
contig2	Asn	GTT	AAC	Yes	1608288-1608359
contig3	Asn	GTT	AAC	Yes	1695808-1695879
contig4	Asn	GTT	AAC	Yes	1931338-1931481
contig4	Asn	GTT	AAC	Yes	1515872-1515953
contig8	Asn	GTT	AAC	Yes	1505393-1505474
contig15	Asn	GTT	AAC	Yes	1237217-1237287
contig16	Asn	GTT	AAC	Yes	1206192-1206265
contig19	Asn	GTT	AAC	Yes	386341-386413
contig21	Asn	GTT	AAC	Yes	324523-324612
contig25	Asn	GTT	AAC	Yes	144634-144707
contig1	Asp	GTC	GAC	Yes	132643-132716
contig3	Asp	GTC	GAC	Yes	201055-201128

Table B.3: **Output of tRNAscan-SE from whole genome contigs for predicted *K. exigua* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig4	Asp	GTC	GAC	Yes	443412-443499
contig7	Asp	GTC	GAC	Yes	590460-590532
contig11	Asp	GTC	GAC	Yes	681213-681285
contig13	Asp	GTC	GAC	Yes	729405-729476
contig15	Asp	GTC	GAC	Yes	841670-841743
contig15	Asp	GTC	GAC	Yes	849776-849848
contig15	Asp	GTC	GAC	Yes	879692-879778
contig18	Asp	GTC	GAC	Yes	937389-937460
contig19	Asp	GTC	GAC	Yes	1032007-1032080
contig19	Asp	GTC	GAC	Yes	1078730-1078801
contig19	Asp	GTC	GAC	Yes	1094066-1094136
contig20	Asp	GTC	GAC	Yes	1524954-1525025
contig20	Asp	GTC	GAC	Yes	1402223-1402296
contig20	Asp	GTC	GAC	Yes	1232118-1232190
contig21	Asp	GTC	GAC	Yes	1189073-1189154
contig2	Cys	GCA	TGC	No	1020870-1020942
contig5	Cys	GCA	TGC	No	991232-991304
contig7	Cys	GCA	TGC	No	885698-885779
contig10	Cys	GCA	TGC	No	692633-692704
contig14	Cys	GCA	TGC	No	605274-605347
contig21	Cys	GCA	TGC	No	403983-404054
contig1	Gln	CTG	CAG	No	231710-231783
contig3	Gln	TTG	CAA	Yes	335512-335585
contig4	Gln	TTG	CAA	Yes	255026-255097
contig6	Gln	TTG	CAA	Yes	208127-208200
contig7	Gln	TTG	CAA	Yes	170811-170884
contig7	Gln	TTG	CAA	Yes	132075-132146
contig10	Gln	TTG	CAA	Yes	131985-132056

Table B.3: **Output of tRNAscan-SE from whole genome contigs for predicted *K. exigua* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig11	Gln	TTG	CAA	Yes	113595-113667
contig12	Gln	TTG	CAA	Yes	99458-99530
contig1	Glu	CTC	GAG	No	9783-9856
contig1	Glu	CTC	GAG	No	1277037-1277108
contig3	Glu	CTC	GAG	No	1050747-1050819
contig3	Glu	CTC	GAG	No	521230-521316
contig3	Glu	TTC	GAA	Yes	285028-285110
contig5	Glu	TTC	GAA	Yes	288965-289047
contig7	Glu	TTC	GAA	Yes	380211-380282
contig8	Glu	TTC	GAA	Yes	387169-387240
contig11	Glu	TTC	GAA	Yes	570123-570196
contig11	Glu	TTC	GAA	Yes	622855-622926
contig13	Glu	TTC	GAA	Yes	622945-623016
contig13	Glu	TTC	GAA	Yes	950306-950378
contig15	Glu	TTC	GAA	Yes	1039560-1039632
contig15	Glu	TTC	GAA	Yes	1073951-1074020
contig16	Glu	TTC	GAA	Yes	1356293-1356364
contig18	Glu	TTC	GAA	Yes	1096643-1096714
contig19	Glu	TTC	GAA	Yes	965922-965995
contig19	Glu	TTC	GAA	Yes	395309-395395
contig21	Glu	TTC	GAA	Yes	362059-362132
contig22	Glu	TTC	GAA	Yes	225364-225445
contig23	Glu	TTC	GAA	Yes	193063-193160
contig23	Glu	TTC	GAA	Yes	99947-100037
contig24	Glu	TTC	GAA	Yes	285613-285702
contig24	Glu	TTC	GAA	Yes	327343-327414
contig1	Gly	CCC	GGG	No	605318-605388
contig1	Gly	CCC	GGG	No	247029-247114

Table B.3: **Output of tRNAscan-SE from whole genome contigs for predicted *K. exigua* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig1	Gly	GCC	GGC	No	468203-468275
contig2	Gly	GCC	GGC	No	789293-789365
contig2	Gly	GCC	GGC	No	808391-808463
contig3	Gly	GCC	GGC	No	873277-873385
contig3	Gly	GCC	GGC	No	1080647-1080717
contig4	Gly	GCC	GGC	No	1259278-1259386
contig5	Gly	GCC	GGC	No	1330322-1330393
contig5	Gly	GCC	GGC	No	1382274-1382346
contig5	Gly	GCC	GGC	No	1293276-1293348
contig6	Gly	GCC	GGC	No	1267389-1267460
contig6	Gly	GCC	GGC	No	961184-961276
contig8	Gly	GCC	GGC	No	290820-290891
contig10	Gly	GCC	GGC	No	183891-183988
contig12	Gly	GCC	GGC	No	125251-125322
contig12	Gly	GCC	GGC	No	10616-10688
contig12	Gly	GCC	GGC	No	365298-365395
contig14	Gly	GCC	GGC	No	677867-677939
contig15	Gly	GCC	GGC	No	763139-763211
contig15	Gly	GCC	GGC	No	1029354-1029427
contig19	Gly	TCC	GGA	No	684105-684177
contig20	Gly	TCC	GGA	No	963802-963872
contig22	Gly	TCC	GGA	No	1197598-1197670
contig22	Gly	TCC	GGA	No	208071-208142
contig1	His	GTG	CAC	Yes	1230343-1230413
contig1	His	GTG	CAC	Yes	1266954-1267025
contig1	His	GTG	CAC	Yes	1267595-1267739
contig2	His	GTG	CAC	Yes	1239740-1239810
contig4	His	GTG	CAC	Yes	1062554-1062626

Table B.3: **Output of tRNAscan-SE from whole genome contigs for predicted *K. exigua* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig5	His	GTG	CAC	Yes	993903-993985
contig9	His	GTG	CAC	Yes	856082-856155
contig13	His	GTG	CAC	Yes	552641-552714
contig16	His	GTG	CAC	Yes	278863-278944
contig18	His	GTG	CAC	Yes	165848-165919
contig22	His	GTG	CAC	Yes	331037-331108
contig1	Ile	AAT	ATT	Yes	382594-382666
contig1	Ile	AAT	ATT	Yes	458995-459066
contig3	Ile	AAT	ATT	Yes	633871-633944
contig3	Ile	AAT	ATT	Yes	702638-702709
contig3	Ile	AAT	ATT	Yes	806044-806130
contig3	Ile	AAT	ATT	Yes	808294-808366
contig3	Ile	AAT	ATT	Yes	842253-842324
contig4	Ile	AAT	ATT	Yes	1163590-1163662
contig4	Ile	AAT	ATT	Yes	1148371-1148444
contig6	Ile	AAT	ATT	Yes	1080134-1080225
contig7	Ile	AAT	ATT	Yes	987992-988092
contig7	Ile	AAT	ATT	Yes	889702-889775
contig8	Ile	AAT	ATT	Yes	708722-708794
contig11	Ile	AAT	ATT	Yes	654732-654803
contig11	Ile	AAT	ATT	Yes	103032-103113
contig13	Ile	AAT	ATT	Yes	433729-433862
contig17	Ile	AAT	ATT	Yes	721251-721332
contig19	Ile	TAT	ATA	No	785194-785267
contig19	Ile	TAT	ATA	No	228282-228355
contig21	Ile	TAT	ATA	No	513361-513447
contig21	Ile	TAT	ATA	No	945698-945771
contig2	iMet	CAT	ATG	Yes	1074527-1074599

Table B.3: **Output of tRNAscan-SE from whole genome contigs for predicted *K. exigua* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig3	iMet	CAT	ATG	Yes	985466-985536
contig4	iMet	CAT	ATG	Yes	897884-897956
contig5	iMet	CAT	ATG	Yes	714097-714170
contig7	iMet	CAT	ATG	Yes	700346-700427
contig10	iMet	CAT	ATG	Yes	690036-690117
contig14	iMet	CAT	ATG	Yes	612098-612189
contig1	Leu	CAA	TTG	No	247760-247878
contig1	Leu	CAA	TTG	No	518872-518953
contig1	Leu	CAA	TTG	No	554242-554323
contig1	Leu	GAG	CTC	No	395880-395951
contig1	Leu	GAG	CTC	No	586747-586818
contig1	Leu	TAA	TTA	Yes	323847-323918
contig4	Leu	TAA	TTA	Yes	263588-263681
contig4	Leu	TAA	TTA	Yes	183134-183234
contig6	Leu	TAA	TTA	Yes	300150-300237
contig7	Leu	TAA	TTA	Yes	808556-808628
contig8	Leu	TAA	TTA	Yes	890108-890197
contig8	Leu	TAA	TTA	Yes	1018972-1019058
contig8	Leu	TAA	TTA	Yes	233468-233610
contig8	Leu	TAA	TTA	Yes	103811-103882
contig12	Leu	TAA	TTA	Yes	230205-230277
contig12	Leu	TAA	TTA	Yes	306572-306644
contig12	Leu	TAA	TTA	Yes	676743-676814
contig12	Leu	TAA	TTA	Yes	502781-502852
contig12	Leu	TAG	CTA	Yes	333592-333664
contig13	Leu	TAG	CTA	Yes	1003836-1003917
contig13	Leu	TAG	CTA	Yes	936641-936712
contig18	Leu	TAG	CTA	Yes	816479-816560

Table B.3: **Output of tRNAscan-SE from whole genome contigs for predicted *K. exigua* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig19	Leu	TAG	CTA	Yes	756121-756212
contig21	Leu	TAG	CTA	Yes	77993-78064
contig21	Leu	TAG	CTA	Yes	143694-143775
contig1	Lys	CTT	AAG	Yes	110962-111034
contig1	Lys	CTT	AAG	Yes	93401-93498
contig2	Lys	CTT	AAG	Yes	76271-76357
contig2	Lys	CTT	AAG	Yes	226354-226426
contig3	Lys	CTT	AAG	Yes	295718-295790
contig3	Lys	CTT	AAG	Yes	384921-384992
contig4	Lys	CTT	AAG	Yes	699121-699194
contig4	Lys	CTT	AAG	Yes	786694-786767
contig5	Lys	CTT	AAG	Yes	860023-860094
contig5	Lys	CTT	AAG	Yes	860113-860184
contig6	Lys	CTT	AAG	Yes	891888-891960
contig6	Lys	CTT	AAG	Yes	598697-598783
contig6	Lys	CTT	AAG	Yes	102932-103004
contig7	Lys	CTT	AAG	Yes	88448-88545
contig10	Lys	CTT	AAG	Yes	61286-61372
contig10	Lys	CTT	AAG	Yes	96976-97046
contig11	Lys	CTT	AAG	Yes	273157-273229
contig11	Lys	TTT	AAA	No	76368-76439
contig12	Lys	TTT	AAA	No	636206-636277
contig14	Lys	TTT	AAA	No	762453-762524
contig14	Lys	TTT	AAA	No	808844-808917
contig18	Lys	TTT	AAA	No	829092-829165
contig20	Lys	TTT	AAA	No	878573-878645
contig23	Lys	TTT	AAA	No	815751-815824
contig24	Lys	TTT	AAA	No	134944-135025

Table B.3: **Output of tRNAscan-SE from whole genome contigs for predicted *K. exigua* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig1	Met	CAT	ATG	Yes	345597-345679
contig1	Met	CAT	ATG	Yes	349979-350061
contig3	Met	CAT	ATG	Yes	495172-495243
contig5	Met	CAT	ATG	Yes	524270-524351
contig10	Met	CAT	ATG	Yes	602677-602748
contig11	Met	CAT	ATG	Yes	742563-742663
contig12	Met	CAT	ATG	Yes	808924-809016
contig14	Met	CAT	ATG	Yes	705413-705485
contig20	Met	CAT	ATG	Yes	663912-663984
contig1	Phe	GAA	TTC	Yes	592246-592328
contig1	Phe	GAA	TTC	Yes	584731-584813
contig2	Phe	GAA	TTC	Yes	488784-488856
contig5	Phe	GAA	TTC	Yes	306121-306194
contig7	Phe	GAA	TTC	Yes	106385-106456
contig8	Phe	GAA	TTC	Yes	71652-71723
contig8	Phe	GAA	TTC	Yes	462063-462195
contig9	Phe	GAA	TTC	Yes	538176-538262
contig12	Phe	GAA	TTC	Yes	777038-777109
contig13	Phe	GAA	TTC	Yes	837766-837852
contig16	Phe	GAA	TTC	Yes	822691-822777
contig17	Phe	GAA	TTC	Yes	817624-817705
contig17	Phe	GAA	TTC	Yes	692671-692742
contig22	Phe	GAA	TTC	Yes	559570-559641
contig23	Phe	GAA	TTC	Yes	552546-552617
contig24	Phe	GAA	TTC	Yes	352312-352384
contig1	Pro	GGG	CCC	No	338710-338781
contig1	Pro	GGG	CCC	No	282197-282289
contig1	Pro	GGG	CCC	No	703322-703393

Table B.3: **Output of tRNAscan-SE from whole genome contigs for predicted *K. exigua* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig1	Pro	TGG	CCA	Yes	264465-264583
contig1	Pro	TGG	CCA	Yes	195077-195177
contig2	Pro	TGG	CCA	Yes	219457-219527
contig2	Pro	TGG	CCA	Yes	530917-531006
contig3	Pro	TGG	CCA	Yes	578225-578311
contig3	Pro	TGG	CCA	Yes	621384-621455
contig4	Pro	TGG	CCA	Yes	645270-645367
contig4	Pro	TGG	CCA	Yes	728452-728541
contig5	Pro	TGG	CCA	Yes	536581-536652
contig7	Pro	TGG	CCA	Yes	497996-498067
contig9	Pro	TGG	CCA	Yes	360039-360111
contig9	Pro	TGG	CCA	Yes	264255-264344
contig10	Pro	TGG	CCA	Yes	152071-152143
contig11	Pro	TGG	CCA	Yes	45973-46045
contig11	Pro	TGG	CCA	Yes	29628-29700
contig13	Pro	TGG	CCA	Yes	54444-54516
contig14	Pro	TGG	CCA	Yes	144944-145014
contig18	Pro	TGG	CCA	Yes	236428-236500
contig21	Pro	TGG	CCA	Yes	337439-337510
contig1	Ser	AGA	TCT	Yes	482329-482401
contig1	Ser	AGA	TCT	Yes	491614-491685
contig1	Ser	AGA	TCT	Yes	500867-500938
contig2	Ser	AGA	TCT	Yes	535256-535337
contig2	Ser	AGA	TCT	Yes	438398-438468
contig2	Ser	AGA	TCT	Yes	109755-109826
contig2	Ser	AGA	TCT	Yes	416605-416678
contig2	Ser	AGA	TCT	Yes	440544-440617
contig3	Ser	AGA	TCT	Yes	323112-323205

Table B.3: **Output of tRNAscan-SE from whole genome contigs for predicted *K. exigua* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig3	Ser	AGA	TCT	Yes	318790-318861
contig4	Ser	AGA	TCT	Yes	62059-62131
contig6	Ser	AGA	TCT	Yes	85017-85090
contig7	Ser	AGA	TCT	Yes	127731-127822
contig8	Ser	AGA	TCT	Yes	461743-461814
contig9	Ser	AGA	TCT	Yes	572438-572519
contig9	Ser	AGA	TCT	Yes	523248-523320
contig9	Ser	AGA	TCT	Yes	446854-446963
contig10	Ser	AGA	TCT	Yes	69609-69690
contig10	Ser	CGA	TCG	No	116049-116121
contig11	Ser	GCT	AGC	No	380968-381038
contig12	Ser	GCT	AGC	No	601447-601519
contig13	Ser	GCT	AGC	No	523661-523733
contig15	Ser	GCT	AGC	No	220660-220760
contig17	Ser	GCT	AGC	No	324856-324948
contig17	Ser	GCT	AGC	No	513219-513291
contig18	Ser	TGA	TCA	No	456705-456776
contig19	Ser	TGA	TCA	No	524182-524253
contig20	Ser	TGA	TCA	No	189361-189434
contig26	Ser	TGA	TCA	No	309372-309445
contig1	Thr	AGT	ACT	Yes	193120-193228
contig2	Thr	AGT	ACT	Yes	521404-521490
contig2	Thr	AGT	ACT	Yes	564437-564509
contig3	Thr	AGT	ACT	Yes	502900-502981
contig3	Thr	AGT	ACT	Yes	374376-374447
contig3	Thr	AGT	ACT	Yes	235948-236019
contig4	Thr	AGT	ACT	Yes	178976-179047
contig5	Thr	AGT	ACT	Yes	125937-126009

Table B.3: **Output of tRNAscan-SE from whole genome contigs for predicted *K. exigua* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig6	Thr	AGT	ACT	Yes	242231-242312
contig7	Thr	AGT	ACT	Yes	458281-458354
contig9	Thr	AGT	ACT	Yes	537259-537332
contig9	Thr	AGT	ACT	Yes	547623-547705
contig10	Thr	AGT	ACT	Yes	523685-523818
contig11	Thr	AGT	ACT	Yes	311471-311542
contig11	Thr	AGT	ACT	Yes	282092-282163
contig12	Thr	AGT	ACT	Yes	253829-253900
contig13	Thr	AGT	ACT	Yes	242895-242965
contig14	Thr	AGT	ACT	Yes	312787-312858
contig17	Thr	CGT	ACG	No	229812-229883
contig18	Thr	CGT	ACG	No	404147-404217
contig18	Thr	TGT	ACA	No	461688-461759
contig21	Thr	TGT	ACA	No	330485-330556
contig23	Thr	TGT	ACA	No	446969-447040
contig24	Thr	TGT	ACA	No	274548-274619
contig2	Trp	CCA	TGG	Yes	319813-319885
contig5	Trp	CCA	TGG	Yes	404101-404172
contig5	Trp	CCA	TGG	Yes	434847-434928
contig17	Trp	CCA	TGG	Yes	346663-346734
contig18	Trp	CCA	TGG	Yes	346573-346644
contig22	Trp	CCA	TGG	Yes	274267-274339
contig22	Trp	CCA	TGG	Yes	20628-20699
contig23	Trp	CCA	TGG	Yes	41728-41801
contig24	Trp	CCA	TGG	Yes	116319-116390
contig25	Trp	CCA	TGG	Yes	217201-217273
contig1	Tyr	GTA	TAC	Yes	253908-253979
contig2	Tyr	GTA	TAC	Yes	400619-400692

Table B.3: **Output of tRNAscan-SE from whole genome contigs for predicted *K. exigua* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig2	Tyr	GTA	TAC	Yes	483469-483551
contig4	Tyr	GTA	TAC	Yes	474481-474554
contig5	Tyr	GTA	TAC	Yes	464213-464345
contig5	Tyr	GTA	TAC	Yes	393004-393085
contig8	Tyr	GTA	TAC	Yes	302949-303022
contig9	Tyr	GTA	TAC	Yes	205664-205750
contig9	Tyr	GTA	TAC	Yes	200001-200074
contig13	Tyr	GTA	TAC	Yes	122412-122484
contig13	Tyr	GTA	TAC	Yes	62580-62651
contig14	Tyr	GTA	TAC	Yes	5310-5416
contig14	Tyr	GTA	TAC	Yes	93287-93357
contig14	Tyr	GTA	TAC	Yes	198387-198457
contig1	Val	AAC	GTT	Yes	353813-353921
contig1	Val	AAC	GTT	Yes	419924-419995
contig3	Val	AAC	GTT	Yes	471749-471821
contig3	Val	AAC	GTT	Yes	384543-384615
contig3	Val	AAC	GTT	Yes	360351-360422
contig6	Val	AAC	GTT	Yes	91033-91125
contig6	Val	AAC	GTT	Yes	85237-85307
contig7	Val	AAC	GTT	Yes	186278-186350
contig7	Val	AAC	GTT	Yes	241195-241266
contig7	Val	AAC	GTT	Yes	206105-206212
contig8	Val	AAC	GTT	Yes	95553-95645
contig8	Val	AAC	GTT	Yes	97743-97814
contig8	Val	AAC	GTT	Yes	174473-174545
contig11	Val	AAC	GTT	Yes	167861-167933
contig11	Val	AAC	GTT	Yes	107984-108077
contig11	Val	AAC	GTT	Yes	71466-71574

Table B.3: **Output of tRNAscan-SE from whole genome contigs for predicted *K. exigua* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig12	Val	AAC	GTT	Yes	5114-5187
contig13	Val	AAC	GTT	Yes	49925-50037
contig15	Val	CAC	GTG	No	184182-184254
contig16	Val	CAC	GTG	No	253591-253662
contig16	Val	TAC	GTA	No	313060-313132
contig21	Val	TAC	GTA	No	253455-253526
contig21	Val	TAC	GTA	No	230389-230460
contig21	Val	TAC	GTA	No	197959-198066

Table B.4: **Output of tRNAscan-SE from whole genome contigs for predicted *K. lodderae* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig1	Ala	AGC	GCT	Yes	302700-302771
contig2	Ala	AGC	GCT	Yes	331840-331912
contig3	Ala	AGC	GCT	Yes	353224-353296
contig3	Ala	AGC	GCT	Yes	433470-433552
contig3	Ala	AGC	GCT	Yes	438716-438797
contig3	Ala	AGC	GCT	Yes	481283-481353
contig4	Ala	AGC	GCT	Yes	657693-657763
contig4	Ala	AGC	GCT	Yes	713351-713423
contig5	Ala	AGC	GCT	Yes	729838-729909
contig5	Ala	AGC	GCT	Yes	1136900-1136973
contig6	Ala	AGC	GCT	Yes	1201071-1201144
contig7	Ala	AGC	GCT	Yes	1287514-1287586
contig7	Ala	AGC	GCT	Yes	910574-910647
contig8	Ala	TGC	GCA	No	155014-155085
contig9	Ala	TGC	GCA	No	797562-797633
contig9	Ala	TGC	GCA	No	797653-797724
contig14	Ala	TGC	GCA	No	864390-864462

Table B.4: **Output of tRNAscan-SE from whole genome contigs for predicted *K. lodderae* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig14	Ala	TGC	GCA	No	1176212-1176285
contig15	Ala	TGC	GCA	No	866824-866932
contig1	Arg	ACG	CGT	Yes	827889-827960
contig1	Arg	ACG	CGT	Yes	151276-151347
contig1	Arg	ACG	CGT	Yes	275078-275159
contig2	Arg	ACG	CGT	Yes	691048-691121
contig2	Arg	ACG	CGT	Yes	962801-962872
contig2	Arg	CCG	CGG	No	851712-851784
contig3	Arg	CCT	AGG	No	861775-861846
contig3	Arg	TCT	AGA	Yes	491851-491922
contig4	Arg	TCT	AGA	Yes	447193-447265
contig5	Arg	TCT	AGA	Yes	391644-391764
contig6	Arg	TCT	AGA	Yes	216059-216132
contig7	Arg	TCT	AGA	Yes	103371-103471
contig8	Arg	TCT	AGA	Yes	389773-389844
contig10	Arg	TCT	AGA	Yes	571754-571824
contig11	Arg	TCT	AGA	Yes	615542-615637
contig11	Arg	TCT	AGA	Yes	631918-631989
contig13	Arg	TCT	AGA	Yes	723822-723893
contig13	Arg	TCT	AGA	Yes	914162-914247
contig15	Arg	TCT	AGA	Yes	1038944-1039017
contig2	Asn	GTT	AAC	Yes	1048398-1048533
contig2	Asn	GTT	AAC	Yes	1193323-1193394
contig4	Asn	GTT	AAC	Yes	1217871-1217944
contig4	Asn	GTT	AAC	Yes	1111685-1111758
contig4	Asn	GTT	AAC	Yes	987382-987453
contig6	Asn	GTT	AAC	Yes	809513-809586
contig7	Asn	GTT	AAC	Yes	742726-742799

Table B.4: **Output of tRNAscan-SE from whole genome contigs for predicted *K. lodderae* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig9	Asn	GTT	AAC	Yes	699922-699995
contig9	Asn	GTT	AAC	Yes	681590-681661
contig11	Asn	GTT	AAC	Yes	654263-654334
contig14	Asn	GTT	AAC	Yes	553125-553195
contig15	Asn	GTT	AAC	Yes	357219-357314
contig1	Asp	GTC	GAC	Yes	354363-354434
contig2	Asp	GTC	GAC	Yes	343296-343369
contig3	Asp	GTC	GAC	Yes	319469-319540
contig3	Asp	GTC	GAC	Yes	273183-273278
contig6	Asp	GTC	GAC	Yes	43679-43752
contig7	Asp	GTC	GAC	Yes	58342-58414
contig7	Asp	GTC	GAC	Yes	113583-113654
contig8	Asp	GTC	GAC	Yes	287004-287099
contig9	Asp	GTC	GAC	Yes	324468-324563
contig10	Asp	GTC	GAC	Yes	504805-504902
contig10	Asp	GTC	GAC	Yes	505359-505440
contig11	Asp	GTC	GAC	Yes	561077-561150
contig11	Asp	GTC	GAC	Yes	639161-639232
contig13	Asp	GTC	GAC	Yes	660873-660970
contig13	Asp	GTC	GAC	Yes	662983-663055
contig3	Cys	GCA	TGC	No	672541-672613
contig3	Cys	GCA	TGC	No	743011-743084
contig3	Cys	GCA	TGC	No	782083-782155
contig5	Cys	GCA	TGC	No	852758-852847
contig7	Cys	GCA	TGC	No	858354-858427
contig15	Cys	GCA	TGC	No	912933-913005
contig1	Gln	TTG	CAA	Yes	973605-973675
contig1	Gln	TTG	CAA	Yes	997604-997675

Table B.4: **Output of tRNAscan-SE from whole genome contigs for predicted *K. lodderae* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig1	Gln	TTG	CAA	Yes	1067047-1067118
contig2	Gln	TTG	CAA	Yes	1248044-1248115
contig3	Gln	TTG	CAA	Yes	1215038-1215124
contig5	Gln	TTG	CAA	Yes	1164019-1164090
contig5	Gln	CTG	CAG	Yes	945487-945558
contig8	Gln	TTG	CAA	Yes	938667-938739
contig8	Gln	TTG	CAA	Yes	921323-921404
contig9	Gln	TTG	CAA	Yes	918963-919034
contig1	Glu	CTC	GAG	No	806695-806766
contig2	Glu	CTC	GAG	No	203186-203258
contig2	Glu	TTC	GAA	Yes	845311-845383
contig2	Glu	TTC	GAA	Yes	743867-743975
contig2	Glu	TTC	GAA	Yes	721105-721176
contig2	Glu	TTC	GAA	Yes	659631-659702
contig3	Glu	TTC	GAA	Yes	618843-618924
contig3	Glu	TTC	GAA	Yes	608068-608159
contig4	Glu	TTC	GAA	Yes	545987-546058
contig6	Glu	TTC	GAA	Yes	497902-497973
contig8	Glu	TTC	GAA	Yes	492687-492773
contig8	Glu	TTC	GAA	Yes	479480-479553
contig10	Glu	TTC	GAA	Yes	428679-428750
contig11	Glu	TTC	GAA	Yes	416264-416335
contig11	Glu	TTC	GAA	Yes	161402-161474
contig12	Glu	TTC	GAA	Yes	95243-95316
contig12	Glu	TTC	GAA	Yes	104065-104147
contig15	Glu	TTC	GAA	Yes	156042-156139
contig1	Gly	CCC	GGG	No	630150-630222
contig1	Gly	GCC	GGC	No	401020-401101

Table B.4: **Output of tRNAscan-SE from whole genome contigs for predicted *K. lodderae* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig2	Gly	GCC	GGC	No	406874-406955
contig2	Gly	GCC	GGC	No	590250-590339
contig3	Gly	GCC	GGC	No	601481-601581
contig3	Gly	GCC	GGC	No	651261-651333
contig3	Gly	GCC	GGC	No	805412-805484
contig4	Gly	GCC	GGC	No	896133-896214
contig4	Gly	GCC	GGC	No	928311-928383
contig5	Gly	GCC	GGC	No	972560-972632
contig6	Gly	GCC	GGC	No	993788-993896
contig6	Gly	GCC	GGC	No	1037514-1037584
contig6	Gly	GCC	GGC	No	1188168-1188250
contig7	Gly	GCC	GGC	No	1117614-1117687
contig7	Gly	GCC	GGC	No	1107441-1107513
contig8	Gly	GCC	GGC	No	1029834-1029905
contig8	Gly	GCC	GGC	No	983738-983809
contig9	Gly	GCC	GGC	No	937358-937430
contig10	Gly	GCC	GGC	No	796434-796505
contig11	Gly	GCC	GGC	No	580145-580217
contig11	Gly	TCC	GGA	No	801291-801364
contig13	Gly	TCC	GGA	No	819643-819725
contig2	His	GTG	CAC	Yes	559131-559203
contig2	His	GTG	CAC	Yes	428593-428688
contig3	His	GTG	CAC	Yes	313082-313154
contig3	His	GTG	CAC	Yes	57483-57555
contig6	His	GTG	CAC	Yes	129472-129544
contig6	His	GTG	CAC	Yes	144846-144918
contig9	His	GTG	CAC	Yes	164509-164579
contig15	His	GTG	CAC	Yes	464048-464134

Table B.4: **Output of tRNAscan-SE from whole genome contigs for predicted *K. lodderae* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig1	Ile	AAT	ATT	Yes	639402-639483
contig1	Ile	AAT	ATT	Yes	679329-679401
contig1	Ile	AAT	ATT	Yes	741075-741157
contig2	Ile	AAT	ATT	Yes	748499-748570
contig2	Ile	AAT	ATT	Yes	943990-944061
contig2	Ile	AAT	ATT	Yes	969870-969951
contig2	Ile	AAT	ATT	Yes	1000195-1000266
contig3	Ile	AAT	ATT	Yes	1067555-1067646
contig5	Ile	AAT	ATT	Yes	1129034-1129130
contig6	Ile	AAT	ATT	Yes	990236-990322
contig7	Ile	AAT	ATT	Yes	884162-884243
contig9	Ile	AAT	ATT	Yes	860238-860310
contig9	Ile	AAT	ATT	Yes	712691-712764
contig9	Ile	AAT	ATT	Yes	425845-425916
contig11	Ile	GAT	ATC	Yes	351050-351121
contig14	Ile	TAT	ATA	No	867556-867647
contig15	Ile	TAT	ATA	No	1056898-1056994
contig3	iMet	CAT	ATG	Yes	189403-189474
contig5	iMet	CAT	ATG	Yes	69289-69362
contig7	iMet	CAT	ATG	Yes	32521-32592
contig11	iMet	CAT	ATG	Yes	80759-80829
contig11	iMet	CAT	ATG	Yes	119519-119592
contig1	Leu	AAG	CTT	No	334182-334252
contig1	Leu	AAG	CTT	No	620952-621024
contig1	Leu	CAA	TTG	No	240841-240938
contig3	Leu	CAA	TTG	No	677660-677751
contig4	Leu	CAA	TTG	No	590823-590944
contig4	Leu	CAA	TTG	No	538511-538611

Table B.4: **Output of tRNAscan-SE from whole genome contigs for predicted *K. Iodderae* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig4	Leu	TAA	TTA	Yes	198657-198728
contig5	Leu	TAA	TTA	Yes	367203-367275
contig5	Leu	TAA	TTA	Yes	471651-471722
contig6	Leu	TAA	TTA	Yes	503387-503458
contig6	Leu	TAA	TTA	Yes	601895-601990
contig6	Leu	TAA	TTA	Yes	632454-632525
contig8	Leu	TAA	TTA	Yes	808737-808819
contig9	Leu	TAA	TTA	Yes	802501-802592
contig11	Leu	TAA	TTA	Yes	680226-680307
contig11	Leu	TAA	TTA	Yes	654250-654331
contig11	Leu	TAA	TTA	Yes	522422-522503
contig12	Leu	TAA	TTA	Yes	510788-510861
contig12	Leu	TAA	TTA	Yes	500823-500923
contig12	Leu	TAA	TTA	Yes	496360-496433
contig13	Leu	TAG	CTA	No	324018-324114
contig14	Leu	TAG	CTA	No	746816-746907
contig15	Leu	TAG	CTA	No	872145-872271
contig1	Lys	CTT	AAG	Yes	344069-344139
contig1	Lys	CTT	AAG	Yes	312915-312988
contig3	Lys	CTT	AAG	Yes	146961-147032
contig3	Lys	CTT	AAG	Yes	179704-179776
contig3	Lys	CTT	AAG	Yes	712703-712773
contig4	Lys	CTT	AAG	Yes	715603-715675
contig4	Lys	CTT	AAG	Yes	537054-537189
contig4	Lys	CTT	AAG	Yes	405582-405653
contig5	Lys	CTT	AAG	Yes	89618-89700
contig5	Lys	CTT	AAG	Yes	92656-92727
contig5	Lys	CTT	AAG	Yes	169998-170070

Table B.4: **Output of tRNAscan-SE from whole genome contigs for predicted *K. lodderae* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig6	Lys	CTT	AAG	Yes	319459-319529
contig6	Lys	TTT	AAA	No	109083-109156
contig8	Lys	TTT	AAA	No	146870-146941
contig8	Lys	TTT	AAA	No	158648-158719
contig8	Lys	TTT	AAA	No	743823-743904
contig10	Lys	TTT	AAA	No	772981-773051
contig10	Lys	TTT	AAA	No	633826-633917
contig11	Lys	TTT	AAA	No	324222-324293
contig11	Lys	TTT	AAA	No	62176-62262
contig12	Lys	TTT	AAA	No	240395-240490
contig15	Lys	TTT	AAA	No	329228-329314
contig3	Met	CAT	ATG	Yes	361409-361495
contig4	Met	CAT	ATG	Yes	500980-501052
contig4	Met	CAT	ATG	Yes	633205-633277
contig10	Met	CAT	ATG	Yes	681691-681762
contig12	Met	CAT	ATG	Yes	638529-638625
contig15	Met	CAT	ATG	Yes	479813-479884
contig2	Phe	GAA	TTC	Yes	452342-452414
contig2	Phe	GAA	TTC	Yes	370215-370286
contig2	Phe	GAA	TTC	Yes	341462-341532
contig3	Phe	GAA	TTC	Yes	305657-305764
contig3	Phe	GAA	TTC	Yes	251631-251702
contig4	Phe	GAA	TTC	Yes	150875-150956
contig6	Phe	GAA	TTC	Yes	119737-119809
contig6	Phe	GAA	TTC	Yes	104569-104665
contig8	Phe	GAA	TTC	Yes	108292-108363
contig9	Phe	GAA	TTC	Yes	115778-115849
contig9	Phe	GAA	TTC	Yes	188210-188280

Table B.4: **Output of tRNAscan-SE from whole genome contigs for predicted *K. lodderae* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig13	Phe	GAA	TTC	Yes	217040-217111
contig15	Phe	GAA	TTC	Yes	420020-420093
contig2	Pro	GGG	CCC	No	232056-232143
contig3	Pro	TGG	CCA	Yes	540293-540374
contig3	Pro	TGG	CCA	Yes	672814-672887
contig4	Pro	TGG	CCA	Yes	680293-680393
contig5	Pro	TGG	CCA	Yes	692973-693046
contig5	Pro	TGG	CCA	Yes	589947-590042
contig6	Pro	TGG	CCA	Yes	534151-534224
contig6	Pro	TGG	CCA	Yes	493773-493868
contig6	Pro	TGG	CCA	Yes	365812-365885
contig7	Pro	TGG	CCA	Yes	287209-287281
contig9	Pro	TGG	CCA	Yes	282484-282555
contig9	Pro	TGG	CCA	Yes	229112-229184
contig9	Pro	TGG	CCA	Yes	209593-209693
contig11	Pro	TGG	CCA	Yes	200498-200585
contig1	Ser	AGA	TCT	Yes	132677-132768
contig2	Ser	AGA	TCT	Yes	120777-120868
contig3	Ser	AGA	TCT	Yes	101486-101559
contig3	Ser	AGA	TCT	Yes	98005-98076
contig4	Ser	AGA	TCT	Yes	147923-148030
contig4	Ser	AGA	TCT	Yes	280789-280860
contig4	Ser	AGA	TCT	Yes	313885-313957
contig4	Ser	AGA	TCT	Yes	628435-628508
contig5	Ser	AGA	TCT	Yes	238401-238471
contig5	Ser	AGA	TCT	Yes	172854-172926
contig6	Ser	AGA	TCT	Yes	121548-121619
contig6	Ser	AGA	TCT	Yes	76596-76667

Table B.4: **Output of tRNAscan-SE from whole genome contigs for predicted *K. lodderae* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig6	Ser	AGA	TCT	Yes	84584-84674
contig6	Ser	CGA	TCG	No	393091-393163
contig6	Ser	GCT	AGC	No	153347-153418
contig7	Ser	GCT	AGC	No	175907-176014
contig8	Ser	GCT	AGC	No	53676-53762
contig9	Ser	GCT	AGC	No	92081-92152
contig9	Ser	TGA	TCA	No	69317-69389
contig11	Ser	TGA	TCA	No	332418-332490
contig11	Ser	TGA	TCA	No	166592-166664
contig15	Ser	TGA	TCA	No	123860-123931
contig15	Sup	TCA	TGA	Yes	128002-128075
contig1	Thr	AGT	ACT	Yes	189192-189299
contig3	Thr	AGT	ACT	Yes	249561-249634
contig3	Thr	AGT	ACT	Yes	470813-470884
contig3	Thr	AGT	ACT	Yes	591768-591864
contig4	Thr	AGT	ACT	Yes	568166-568237
contig4	Thr	AGT	ACT	Yes	531232-531302
contig4	Thr	AGT	ACT	Yes	473591-473691
contig4	Thr	AGT	ACT	Yes	461597-461668
contig5	Thr	AGT	ACT	Yes	431134-431205
contig5	Thr	AGT	ACT	Yes	411284-411355
contig5	Thr	AGT	ACT	Yes	135522-135604
contig8	Thr	AGT	ACT	Yes	128490-128563
contig8	Thr	AGT	ACT	Yes	22912-22983
contig10	Thr	CGT	ACG	No	267188-267270
contig10	Thr	CGT	ACG	No	57514-57585
contig10	Thr	CGT	ACG	No	74146-74219
contig11	Thr	TGT	ACA	No	400676-400757

Table B.4: **Output of tRNAscan-SE from whole genome contigs for predicted *K. Iodderae* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig13	Thr	TGT	ACA	No	438103-438185
contig15	Thr	TGT	ACA	No	65729-65799
contig16	Thr	TGT	ACA	No	47719-47839
contig1	Trp	CCA	TGG	Yes	104927-104998
contig3	Trp	CCA	TGG	Yes	135014-135086
contig4	Trp	CCA	TGG	Yes	196546-196664
contig8	Trp	CCA	TGG	Yes	376817-376888
contig10	Trp	CCA	TGG	Yes	403042-403124
contig10	Trp	CCA	TGG	Yes	259761-259833
contig11	Trp	CCA	TGG	Yes	251267-251340
contig3	Tyr	GTA	TAC	Yes	279395-279466
contig3	Tyr	GTA	TAC	Yes	279486-279557
contig5	Tyr	GTA	TAC	Yes	295100-295172
contig5	Tyr	GTA	TAC	Yes	310979-311051
contig7	Tyr	GTA	TAC	Yes	379710-379792
contig8	Tyr	GTA	TAC	Yes	372798-372869
contig8	Tyr	GTA	TAC	Yes	246992-247062
contig9	Tyr	GTA	TAC	Yes	120894-120991
contig9	Tyr	GTA	TAC	Yes	19436-19509
contig11	Tyr	GTA	TAC	Yes	107486-107558
contig1	Val	AAC	GTT	Yes	203671-203744
contig1	Val	AAC	GTT	Yes	304970-305052
contig1	Val	AAC	GTT	Yes	273073-273145
contig2	Val	AAC	GTT	Yes	30051-30131
contig2	Val	AAC	GTT	Yes	29519-29590
contig2	Val	AAC	GTT	Yes	29442-29514
contig3	Val	AAC	GTT	Yes	29078-29148
contig3	Val	AAC	GTT	Yes	28993-29064

Table B.4: **Output of tRNAscan-SE from whole genome contigs for predicted *K. lodderae* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig3	Val	AAC	GTT	Yes	11572-11643
contig3	Val	AAC	GTT	Yes	11409-11481
contig5	Val	AAC	GTT	Yes	11196-11281
contig6	Val	AAC	GTT	Yes	8176-8246
contig6	Val	AAC	GTT	Yes	3148-3220
contig9	Val	AAC	GTT	Yes	815-886
contig10	Val	AAC	GTT	Yes	634-706
contig10	Val	AAC	GTT	Yes	528-600
contig12	Val	CAC	GTG	No	162863-162936
contig13	Val	TAC	GTA	No	11493-11564
contig14	Val	TAC	GTA	No	4570-4642
contig15	Val	TAC	GTA	No	17055-17126

Table B.5: **Output of tRNAscan-SE from whole genome contigs for predicted *K. naganishii* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinates
HE978314.1	Ala	AGC	GCU	Yes	512464-512536
HE978315.1	Ala	AGC	GCU	Yes	999907-999979
HE978315.1	Ala	AGC	GCU	Yes	978421-978493
HE978316.1	Ala	AGC	GCU	Yes	1109697-1109769
HE978318.1	Ala	AGC	GCU	Yes	614315-614387
HE978318.1	Ala	AGC	GCU	Yes	554541-554613
HE978318.1	Ala	AGC	GCU	Yes	501434-501506
HE978314.1	Ala	CGC	GCG	No	973343-973415
HE978314.1	Ala	TGC	GCA	No	899090-899162
HE978316.1	Arg	ACG	CGU	Yes	299840-299912
HE978320.1	Arg	ACG	CGU	Yes	536993-537065
HE978326.1	Arg	ACG	CGU	Yes	212644-212716
HE978315.1	Arg	CCG	CGG	No	1116479-1116550
HE978315.1	Arg	CCT	AGG	No	44586-44657
HE978319.1	Arg	TCT	AGA	Yes	523948-524019
HE978319.1	Arg	TCT	AGA	Yes	619394-619465
HE978321.1	Arg	TCT	AGA	Yes	167625-167696
HE978323.1	Arg	TCT	AGA	Yes	437945-438016
HE978324.1	Arg	TCT	AGA	Yes	424218-424291
HE978326.1	Arg	TCT	AGA	Yes	229329-229400
HE978314.1	Asn	GTT	AAC	No	1250231-1250304
HE978314.1	Asn	GTT	AAC	No	1076795-1076868
HE978315.1	Asn	GTT	AAC	No	1286065-1286138
HE978317.1	Asn	GTT	AAC	No	482575-482648
HE978318.1	Asn	GTT	AAC	No	761849-761922
HE978322.1	Asn	GTT	AAC	No	210661-210734
HE978318.1	Asp	GTC	GAC	No	182582-182653
HE978318.1	Asp	GTC	GAC	No	173107-173178

Table B.5: **Output of tRNAscan-SE from whole genome contigs for predicted *K. naganishii* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinates
HE978319.1	Asp	GTC	GAC	No	524038-524109
HE978319.1	Asp	GTC	GAC	No	619304-619375
HE978321.1	Asp	GTC	GAC	No	167534-167605
HE978323.1	Asp	GTC	GAC	No	438035-438106
HE978324.1	Asp	GTC	GAC	No	424128-424199
HE978326.1	Asp	GTC	GAC	No	229420-229491
HE978314.1	Cys	GCA	UGC	No	505812-505883
HE978314.1	Cys	GCA	UGC	No	173539-173610
HE978315.1	Cys	GCA	UGC	No	175314-175385
HE978315.1	Cys	GCA	UGC	No	1023915-1023986
HE978323.1	Gln	CTG	CUA	No	52049-52120
HE978324.1	Gln	CTG	CUA	No	428294-428365
HE978321.1	Gln	TTG	CAA	Yes	569115-569186
HE978321.1	Gln	TTG	CAA	Yes	216364-216435
HE978322.1	Gln	TTG	CAA	Yes	310513-310584
HE978323.1	Gln	TTG	CAA	Yes	156290-156361
HE978316.1	Glu	CTC	GAG	No	1221535-1221606
HE978318.1	Glu	CTC	GAG	No	719435-719506
HE978325.1	Glu	CTC	GAG	No	316008-316079
HE978315.1	Glu	TTC	GAA	Yes	1097812-1097883
HE978317.1	Glu	TTC	GAA	Yes	463009-463080
HE978318.1	Glu	TTC	GAA	Yes	809038-809109
HE978319.1	Glu	TTC	GAA	Yes	606635-606703
HE978325.1	Glu	TTC	GAA	Yes	177352-177423
HE978326.1	Glu	TTC	GAA	Yes	77569-77640
HE978315.1	Gly	CCC	GGG	No	767791-767861
HE978316.1	Gly	CCC	GGG	No	1190952-1191022
HE978315.1	Gly	GCC	GGC	No	120655-120725

Table B.5: **Output of tRNAscan-SE from whole genome contigs for predicted *K. naganishii* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinates
HE978315.1	Gly	GCC	GGC	No	50833-50903
HE978318.1	Gly	GCC	GGC	No	151544-151614
HE978318.1	Gly	GCC	GGC	No	684883-684953
HE978318.1	Gly	GCC	GGC	No	694190-694260
HE978320.1	Gly	GCC	GGC	No	146218-146288
HE978321.1	Gly	GCC	GGC	No	262884-262954
HE978323.1	Gly	GCC	GGC	No	176859-176929
HE978326.1	Gly	GCC	GGC	No	175878-175948
HE978315.1	Gly	TCC	GGA	No	125825-125896
HE978317.1	His	GTG	CAC	No	346470-346541
HE978318.1	His	GTG	CAC	No	315575-315646
HE978320.1	His	GTG	CAC	No	320005-320076
HE978326.1	His	GTG	CAC	No	435950-436021
HE978314.1	Ile	AAT	AUU	Yes	207496-207569
HE978314.1	Ile	AAT	AUU	Yes	804218-804291
HE978315.1	Ile	AAT	AUU	Yes	574853-574926
HE978316.1	Ile	AAT	AUU	Yes	808491-808564
HE978321.1	Ile	AAT	AUU	Yes	365894-365967
HE978322.1	Ile	AAT	AUU	Yes	329681-329754
HE978322.1	Ile	AAT	AUU	Yes	252029-252102
HE978314.1	Ile	TAT	AUA	No	635837-635969
HE978314.1	Ile	TAT	AUA	No	266406-266539
HE978320.1	iMe	CAT	AUG	Yes	114288-114359
HE978323.1	iMe	CAT	AUG	Yes	249848-249919
HE978322.1	Leu	AAG	CUU	Yes	572351-572456
HE978315.1	Leu	CAA	UUG	Yes	1009708-1009829
HE978317.1	Leu	CAA	UUG	Yes	691306-691427
HE978317.1	Leu	CAA	UUG	Yes	662538-662659

Table B.5: **Output of tRNAscan-SE from whole genome contigs for predicted *K. naganishii* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinates
HE978326.1	Leu	CAA	UUG	Yes	492794-492916
HE978326.1	Leu	CAA	UUG	Yes	351897-352019
HE978322.1	Leu	TAA	UUA	Yes	510840-510922
HE978323.1	Leu	TAA	UUA	Yes	152191-152273
HE978323.1	Leu	TAA	UUA	Yes	453870-453952
HE978317.1	Leu	TAG	CUA	Yes	221266-221363
HE978319.1	Leu	TAG	CUA	Yes	502656-502753
HE978323.1	Leu	TAG	CUA	Yes	511856-511953
HE978315.1	Lys	CTT	AAG	No	679664-679736
HE978315.1	Lys	CTT	AAG	No	810810-810882
HE978317.1	Lys	CTT	AAG	No	855120-855192
HE978317.1	Lys	CTT	AAG	No	402517-402589
HE978318.1	Lys	CTT	AAG	No	677458-677530
HE978319.1	Lys	CTT	AAG	No	484454-484526
HE978325.1	Lys	CTT	AAG	No	219623-219695
HE978316.1	Lys	TTT	AAA	Yes	420603-420697
HE978316.1	Lys	TTT	AAA	Yes	1240811-1240906
HE978317.1	Lys	TTT	AAA	Yes	779134-779229
HE978323.1	Lys	TTT	AAA	Yes	166748-166843
HE978319.1	Met	CAT	AUG	Yes	485924-485996
HE978322.1	Met	CAT	AUG	Yes	353305-353377
HE978315.1	Phe	GAA	UUC	No	375700-375802
HE978315.1	Phe	GAA	UUC	No	905834-905936
HE978316.1	Phe	GAA	UUC	No	593837-593939
HE978317.1	Phe	GAA	UUC	No	426247-426349
HE978324.1	Phe	GAA	UUC	No	200831-200930
HE978326.1	Phe	GAA	UUC	No	432649-432748
HE978322.1	Pro	AGG	CCU	No	213791-213862

Table B.5: **Output of tRNAscan-SE from whole genome contigs for predicted *K. naganishii* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinates
HE978314.1	Pro	TGG	CCA	Yes	1151869-1151979
HE978315.1	Pro	TGG	CCA	Yes	964924-965032
HE978315.1	Pro	TGG	CCA	Yes	760940-761048
HE978320.1	Pro	TGG	CCA	Yes	100097-100205
HE978321.1	Pro	TGG	CCA	Yes	516352-516460
HE978325.1	Pro	TGG	CCA	Yes	103080-103188
HE978315.1	Ser	AGA	UCU	Yes	989848-989929
HE978315.1	Ser	AGA	UCU	Yes	1126510-1126591
HE978315.1	Ser	AGA	UCU	Yes	962168-962249
HE978315.1	Ser	AGA	UCU	Yes	712967-713048
HE978318.1	Ser	AGA	UCU	Yes	538648-538729
HE978324.1	Ser	AGA	UCU	Yes	285924-286005
HE978324.1	Ser	AGA	UCU	Yes	317200-317281
HE978322.1	Ser	CGA	UCG	No	567245-567340
HE978315.1	Ser	GCT	AGC	No	1061284-1061384
HE978317.1	Ser	GCT	AGC	No	598698-598798
HE978322.1	Ser	GCT	AGC	No	322679-322779
HE978325.1	Ser	TGA	UCA	Yes	274294-274375
HE978321.1	Thr	AGT	ACU	Yes	442687-442759
HE978321.1	Thr	AGT	ACU	Yes	376810-376882
HE978323.1	Thr	AGT	ACU	Yes	224807-224879
HE978323.1	Thr	AGT	ACU	Yes	183282-183354
HE978324.1	Thr	AGT	ACU	Yes	108403-108475
HE978325.1	Thr	AGT	ACU	Yes	191261-191333
HE978326.1	Thr	AGT	ACU	Yes	227987-228059
HE978315.1	Thr	CGT	ACG	No	25320-25391
HE978314.1	Thr	TGT	ACA	No	1191654-1191725
HE978326.1	Thr	TGT	ACA	No	299848-299919

Table B.5: **Output of tRNAscan-SE from whole genome contigs for predicted *K. naganishii* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinates
HE978315.1	Trp	CCA	UGG	Yes	1348961-1349064
HE978319.1	Trp	CCA	UGG	Yes	458100-458202
HE978320.1	Trp	CCA	UGG	Yes	216319-216420
HE978321.1	Trp	CCA	UGG	Yes	68425-68528
HE978314.1	Tyr	GTA	UAC	No	563484-563580
HE978315.1	Tyr	GTA	UAC	No	39684-39780
HE978315.1	Tyr	GTA	UAC	No	171648-171744
HE978320.1	Tyr	GTA	UAC	No	52498-52595
HE978322.1	Tyr	GTA	UAC	No	295849-295945
HE978314.1	Val	AAC	GUU	Yes	1107360-1107433
HE978316.1	Val	AAC	GUU	Yes	1116009-1116082
HE978318.1	Val	AAC	GUU	Yes	709175-709248
HE978320.1	Val	AAC	GUU	Yes	250394-250467
HE978320.1	Val	AAC	GUU	Yes	346326-346399
HE978322.1	Val	AAC	GUU	Yes	154758-154831
HE978325.1	Val	AAC	GUU	Yes	171333-171406
HE978315.1	Val	CAC	GUG	No	910143-910215
HE978326.1	Val	CAC	GUG	No	273056-273128
HE978321.1	Val	TAC	GUA	Yes	243100-243172

Table B.6: **Output of tRNAscan-SE from whole genome contigs for predicted *K. viticola* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig1.	Ala	AGC	GCT	Yes	90021-90094
contig1.	Ala	AGC	GCT	Yes	135416-135487
contig2.	Ala	AGC	GCT	Yes	146308-146379
contig2.	Ala	AGC	GCT	Yes	146399-146470
contig2.	Ala	AGC	GCT	Yes	459191-459272
contig3.	Ala	AGC	GCT	Yes	653382-653464

Table B.6: **Output of tRNAscan-SE from whole genome contigs for predicted *K. viticola* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig3.	Ala	AGC	GCT	Yes	670612-670712
contig4.	Ala	AGC	GCT	Yes	789409-789490
contig6.	Ala	AGC	GCT	Yes	824954-825035
contig8.	Ala	AGC	GCT	Yes	901213-901285
contig8.	Ala	AGC	GCT	Yes	1028482-1028554
contig10	Ala	AGC	GCT	Yes	1047476-1047547
contig10	Ala	AGC	GCT	Yes	1077440-1077512
contig12	Ala	TGC	GCA	No	135507-135578
contig13	Ala	TGC	GCA	No	402761-402832
contig21	Ala	TGC	GCA	No	561626-561699
contig21	Ala	TGC	GCA	No	832213-832285
contig22	Ala	TGC	GCA	No	1088033-1088106
contig1.	Arg	ACG	CGT	Yes	602218-602289
contig1.	Arg	ACG	CGT	Yes	377181-377252
contig1.	Arg	ACG	CGT	Yes	303798-303869
contig1.	Arg	ACG	CGT	Yes	67671-67742
contig2.	Arg	CCG	CGG	No	24902-24973
contig2.	Arg	CCT	AGG	No	197920-197992
contig3.	Arg	TCT	AGA	Yes	919774-919845
contig3.	Arg	TCT	AGA	Yes	919683-919754
contig3.	Arg	TCT	AGA	Yes	808194-808266
contig7.	Arg	TCT	AGA	Yes	792371-792463
contig8.	Arg	TCT	AGA	Yes	633250-633322
contig9.	Arg	TCT	AGA	Yes	602127-602198
contig11	Arg	TCT	AGA	Yes	574979-575071
contig18	Arg	TCT	AGA	Yes	67762-67833
contig20	Arg	TCT	AGA	Yes	178082-178180
contig22	Arg	TCT	AGA	Yes	209742-209814

Table B.6: **Output of tRNAscan-SE from whole genome contigs for predicted *K. viticola* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig1.	Asn	GTT	AAC	Yes	215541-215612
contig3.	Asn	GTT	AAC	Yes	681952-682041
contig6.	Asn	GTT	AAC	Yes	773291-773363
contig10	Asn	GTT	AAC	Yes	686765-686836
contig12	Asn	GTT	AAC	Yes	582222-582295
contig13	Asn	GTT	AAC	Yes	561405-561477
contig16	Asn	GTT	AAC	Yes	488830-488902
contig17	Asn	GTT	AAC	Yes	477295-477367
contig22	Asn	GTT	AAC	Yes	165098-165223
contig1.	Asp	GTC	GAC	Yes	99804-99904
contig1.	Asp	GTC	GAC	Yes	74733-74806
contig1.	Asp	GTC	GAC	Yes	48915-48988
contig1.	Asp	GTC	GAC	Yes	70204-70305
contig2.	Asp	GTC	GAC	Yes	169670-169768
contig2.	Asp	GTC	GAC	Yes	268128-268199
contig3.	Asp	GTC	GAC	Yes	387960-388060
contig4.	Asp	GTC	GAC	Yes	435878-435976
contig5.	Asp	GTC	GAC	Yes	476908-477006
contig9.	Asp	GTC	GAC	Yes	555693-555775
contig10	Asp	GTC	GAC	Yes	625975-626047
contig18	Asp	GTC	GAC	Yes	639202-639283
contig29	Asp	GTC	GAC	Yes	756753-756824
contig1.	Cys	GCA	TGC	No	756844-756915
contig3.	Cys	GCA	TGC	No	865596-865667
contig8.	Cys	GCA	TGC	No	871350-871423
contig19	Cys	GCA	TGC	No	890549-890647
contig2.	Gln	TTG	CAA	Yes	901092-901164
contig4.	Gln	TTG	CAA	Yes	914062-914144

Table B.6: **Output of tRNAscan-SE from whole genome contigs for predicted *K. viticola* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig5.	Gln	TTG	CAA	Yes	857685-857756
contig10	Gln	TTG	CAA	Yes	842952-843025
contig11	Gln	CTG	CAG	No	716551-716623
contig11	Gln	TTG	CAA	Yes	611756-611829
contig12	Gln	TTG	CAA	Yes	359458-359530
contig15	Gln	TTG	CAA	Yes	169-241
contig1.	Glu	CTC	GAG	No	45329-45426
contig1.	Glu	CTC	GAG	No	409678-409749
contig4.	Glu	TTC	GAA	Yes	108241-108313
contig6.	Glu	TTC	GAA	Yes	113466-113537
contig7.	Glu	TTC	GAA	Yes	171194-171275
contig7.	Glu	TTC	GAA	Yes	194901-194973
contig10	Glu	TTC	GAA	Yes	216875-216974
contig10	Glu	TTC	GAA	Yes	389056-389128
contig11	Glu	TTC	GAA	Yes	436604-436674
contig13	Glu	TTC	GAA	Yes	485638-485719
contig14	Glu	TTC	GAA	Yes	559041-559112
contig17	Glu	TTC	GAA	Yes	791742-791843
contig17	Glu	TTC	GAA	Yes	854282-854364
contig22	Glu	TTC	GAA	Yes	542533-542606
contig3.	Gly	CCC	GGG	No	460478-460550
contig4.	Gly	GCC	GGC	No	307091-307164
contig4.	Gly	GCC	GGC	No	229196-229267
contig5.	Gly	GCC	GGC	No	199886-199956
contig8.	Gly	GCC	GGC	No	21090-21182
contig9.	Gly	GCC	GGC	No	339647-339728
contig11	Gly	GCC	GGC	No	480072-480144
contig11	Gly	GCC	GGC	No	578517-578589

Table B.6: **Output of tRNAscan-SE from whole genome contigs for predicted *K. viticola* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig11	Gly	GCC	GGC	No	620299-620370
contig12	Gly	GCC	GGC	No	553948-554041
contig12	Gly	GCC	GGC	No	493322-493403
contig14	Gly	GCC	GGC	No	338345-338415
contig17	Gly	GCC	GGC	No	284306-284377
contig18	Gly	GCC	GGC	No	242421-242493
contig20	Gly	GCC	GGC	No	429748-429820
contig21	Gly	GCC	GGC	No	607557-607655
contig21	Gly	TCC	GGA	No	2066-2137
contig22	Gly	TCC	GGA	No	405327-405400
contig28	Gly	TCC	GGA	No	564637-564708
contig2.	His	GTG	CAC	Yes	635297-635386
contig10	His	GTG	CAC	Yes	618193-618297
contig12	His	GTG	CAC	Yes	513127-513200
contig16	His	GTG	CAC	Yes	12879-12961
contig17	His	GTG	CAC	Yes	232317-232388
contig1.	Ile	AAT	ATT	Yes	238507-238579
contig2.	Ile	AAT	ATT	Yes	297891-297962
contig3.	Ile	AAT	ATT	Yes	533094-533167
contig3.	Ile	AAT	ATT	Yes	512466-512539
contig5.	Ile	AAT	ATT	Yes	507357-507429
contig10	Ile	AAT	ATT	Yes	226059-226140
contig12	Ile	AAT	ATT	Yes	330521-330602
contig12	Ile	AAT	ATT	Yes	398150-398222
contig14	Ile	AAT	ATT	Yes	462391-462463
contig15	Ile	AAT	ATT	Yes	486369-486440
contig17	Ile	TAT	ATA	No	479108-479179
contig19	Ile	TAT	ATA	No	520816-520886

Table B.6: **Output of tRNAscan-SE from whole genome contigs for predicted *K. viticola* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig1.	iMet	CAT	ATG	Yes	484406-484478
contig4.	iMet	CAT	ATG	Yes	241199-241280
contig8.	iMet	CAT	ATG	Yes	234445-234516
contig10	iMet	CAT	ATG	Yes	222321-222402
contig24	iMet	CAT	ATG	Yes	259867-259939
contig1.	Leu	AAG	CTT	No	193206-193277
contig2.	Leu	AAG	CTT	No	: 382645-382716
contig3.	Leu	CAA	TTG	No	479783-479854
contig3.	Leu	CAA	TTG	No	455499-455570
contig3.	Leu	TAA	TTA	Yes	428461-428533
contig4.	Leu	TAA	TTA	Yes	289878-289951
contig4.	Leu	TAA	TTA	Yes	258317-258389
contig7.	Leu	TAA	TTA	Yes	81726-81798
contig8.	Leu	TAA	TTA	Yes	111551-111622
contig9.	Leu	TAA	TTA	Yes	241983-242055
contig11	Leu	TAA	TTA	Yes	304413-304485
contig11	Leu	TAA	TTA	Yes	336035-336131
contig16	Leu	TAA	TTA	Yes	446172-446270
contig16	Leu	TAA	TTA	Yes	450365-450470
contig16	Leu	TAG	CTA	No	316932-317014
contig20	Leu	TAG	CTA	No	62828-62899
contig1.	Lys	CTT	AAG	Yes	: 352962-353034
contig2.	Lys	CTT	AAG	Yes	: 307491-307564
contig4.	Lys	CTT	AAG	Yes	: 279838-279909
contig4.	Lys	CTT	AAG	Yes	: 234734-234805
contig4.	Lys	CTT	AAG	Yes	72505-72576
contig4.	Lys	CTT	AAG	Yes	242881-242985
contig4.	Lys	CTT	AAG	Yes	249696-249778

Table B.6: **Output of tRNAscan-SE from whole genome contigs for predicted *K. viticola* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig5.	Lys	CTT	AAG	Yes	257876-257948
contig7.	Lys	CTT	AAG	Yes	335558-335630
contig10	Lys	CTT	AAG	Yes	383183-383254
contig10	Lys	CTT	AAG	Yes	: 280671-280741
contig12	Lys	CTT	AAG	Yes	: 255649-255720
contig15	Lys	TTT	AAA	No	: 243566-243639
contig17	Lys	TTT	AAA	No	: 89932-90003
contig17	Lys	TTT	AAA	No	331188-331259
contig22	Lys	TTT	AAA	No	368656-368726
contig24	Lys	TTT	AAA	No	380974-381045
contig29	Lys	TTT	AAA	No	: 374273-374355
contig1.	Met	CAT	ATG	Yes	: 240387-240491
contig2.	Met	CAT	ATG	Yes	: 208322-208394
contig3.	Met	CAT	ATG	Yes	105065-105138
contig22	Met	CAT	ATG	Yes	129184-129254
contig22	Met	CAT	ATG	Yes	212907-212980
contig2.	Phe	GAA	TTC	Yes	267148-267221
contig3.	Phe	GAA	TTC	Yes	370332-370404
contig3.	Phe	GAA	TTC	Yes	400403-400492
contig3.	Phe	GAA	TTC	Yes	381973-382044
contig3.	Phe	GAA	TTC	Yes	353348-353448
contig6.	Phe	GAA	TTC	Yes	339791-339880
contig10	Phe	GAA	TTC	Yes	: 299806-299879
contig17	Phe	GAA	TTC	Yes	: 245866-245939
contig22	Phe	GAA	TTC	Yes	: 211553-211624
contig1.	Pro	AGG	CCT	No	44341-44422
contig1.	Pro	TGG	CCA	Yes	: 118959-119029
contig4.	Pro	TGG	CCA	Yes	: 85090-85162

Table B.6: **Output of tRNAscan-SE from whole genome contigs for predicted *K. viticola* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig5.	Pro	TGG	CCA	Yes	237878-237949
contig6.	Pro	TGG	CCA	Yes	281688-281761
contig16	Pro	TGG	CCA	Yes	366919-367008
contig16	Pro	TGG	CCA	Yes	104657-104729
contig19	Pro	TGG	CCA	Yes	171305-171378
contig20	Pro	TGG	CCA	Yes	276836-276906
contig21	Pro	TGG	CCA	Yes	186845-186926
contig21	Pro	TGG	CCA	Yes	129541-129612
contig24	Pro	TGG	CCA	Yes	70256-70329
contig1.	Ser	AGA	TCT	Yes	8928-9001
contig1.	Ser	AGA	TCT	Yes	185349-185438
contig1.	Ser	AGA	TCT	Yes	244164-244261
contig1.	Ser	AGA	TCT	Yes	258393-258464
contig2.	Ser	AGA	TCT	Yes	249489-249562
contig3.	Ser	AGA	TCT	Yes	229615-229686
contig3.	Ser	AGA	TCT	Yes	212595-212695
contig4.	Ser	AGA	TCT	Yes	120715-120793
contig4.	Ser	AGA	TCT	Yes	152150-152221
contig5.	Ser	AGA	TCT	Yes	304109-304180
contig5.	Ser	CGA	TCG	No	277330-277403
contig8.	Ser	GCT	AGC	No	118631-118766
contig8.	Ser	GCT	AGC	No	56387-56458
contig9.	Ser	GCT	AGC	No	153979-154061
contig12	Ser	GCT	AGC	No	134145-134218
contig13	Ser	TGA	TCA	No	225050-225142
contig14	Ser	TGA	TCA	No	306403-306485
contig16	Ser	TGA	TCA	No	9302-9428
contig18	Ser	TGA	TCA	No	233876-233949

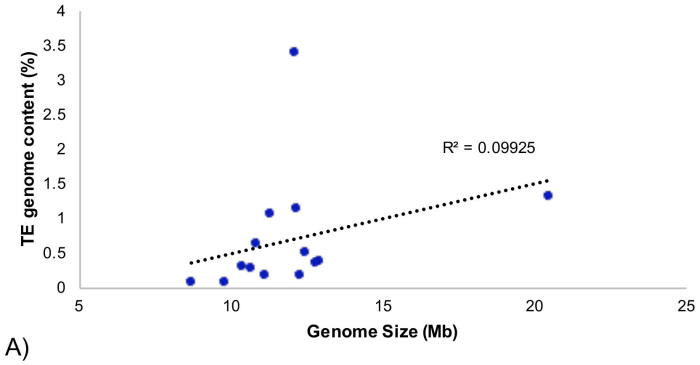
Table B.6: **Output of tRNAscan-SE from whole genome contigs for predicted *K. viticola* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig22	Ser	TGA	TCA	No	259894-259965
contig22	Sup	TCA	TGA	Yes	231887-231957
contig1.	Thr	AGT	ACT	Yes	210876-210948
contig4.	Thr	AGT	ACT	Yes	: 181093-181191
contig4.	Thr	AGT	ACT	Yes	211682-211763
contig5.	Thr	AGT	ACT	Yes	272482-272586
contig6.	Thr	AGT	ACT	Yes	297048-297119
contig8.	Thr	AGT	ACT	Yes	297139-297210
contig9.	Thr	AGT	ACT	Yes	234393-234463
contig9.	Thr	AGT	ACT	Yes	127777-127848
contig10	Thr	AGT	ACT	Yes	50694-50786
contig11	Thr	AGT	ACT	Yes	49051-49122
contig11	Thr	AGT	ACT	Yes	102863-102944
contig11	Thr	CGT	ACG	No	165335-165427
contig20	Thr	TGT	ACA	No	6255-6352
contig20	Thr	TGT	ACA	No	232993-233128
contig22	Thr	TGT	ACA	No	83525-83614
contig30	Thr	TGT	ACA	No	127106-127177
contig6.	Trp	CCA	TGG	Yes	66911-66981
contig10	Trp	CCA	TGG	Yes	53260-53349
contig11	Trp	CCA	TGG	Yes	13213-13284
contig11	Trp	CCA	TGG	Yes	86802-86894
contig18	Trp	CCA	TGG	Yes	113158-113230
contig23	Trp	CCA	TGG	Yes	123997-124067
contig30	Trp	CCA	TGG	Yes	57982-58074
contig2.	Tyr	GTA	TAC	Yes	31933-32005
contig12	Tyr	GTA	TAC	Yes	21920-21990
contig12	Tyr	GTA	TAC	Yes	15652-15723

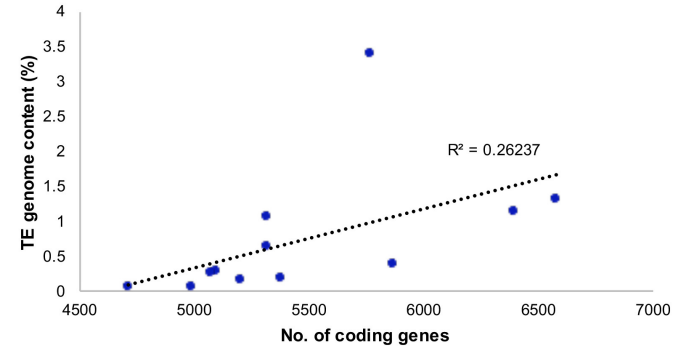
Table B.6: **Output of tRNAscan-SE from whole genome contigs for predicted *K. viticola* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal codon	Co-ordinates
contig13	Tyr	GTA	TAC	Yes	23023-23094
contig15	Tyr	GTA	TAC	Yes	23465-23537
contig20	Tyr	GTA	TAC	Yes	23806-23876
contig20	Tyr	GTA	TAC	Yes	25203-25274
contig1.	Val	AAC	GTT	Yes	25936-26008
contig1.	Val	AAC	GTT	Yes	35953-36024
contig2.	Val	AAC	GTT	Yes	38892-38978
contig2.	Val	AAC	GTT	Yes	40430-40502
contig4.	Val	AAC	GTT	Yes	41133-41203
contig4.	Val	AAC	GTT	Yes	41312-41385
contig7.	Val	AAC	GTT	Yes	11738-11809
contig7.	Val	AAC	GTT	Yes	26985-27056
contig9.	Val	AAC	GTT	Yes	28881-28953
contig10	Val	AAC	GTT	Yes	23944-24017
contig12	Val	AAC	GTT	Yes	21954-22024
contig12	Val	AAC	GTT	Yes	6626-6697
contig14	Val	AAC	GTT	Yes	6018-6122
contig14	Val	CAC	GTG	No	35523-35593
contig17	Val	TAC	GTA	No	48135-48239
contig17	Val	TAC	GTA	No	8448-8540
contig22	Val	TAC	GTA	No	1467-1539
contig28	Val	TAC	GTA	No	8515-8587

B.1.2 Transposable Element Data for *Kazachstania* species



A)



B)

Figure B.1: **Relationship between TE genome content in Saccharomyceteceae species and additional genome characteristics.** A) Genome size is plotted on the x-axis against TE genome content (%) on the y-axis. B) Number of coding genes is plotted on the x-axis against TE genome content (%) on the y-axis.

B.1.3 Codon usage data for *Kazachstania* species

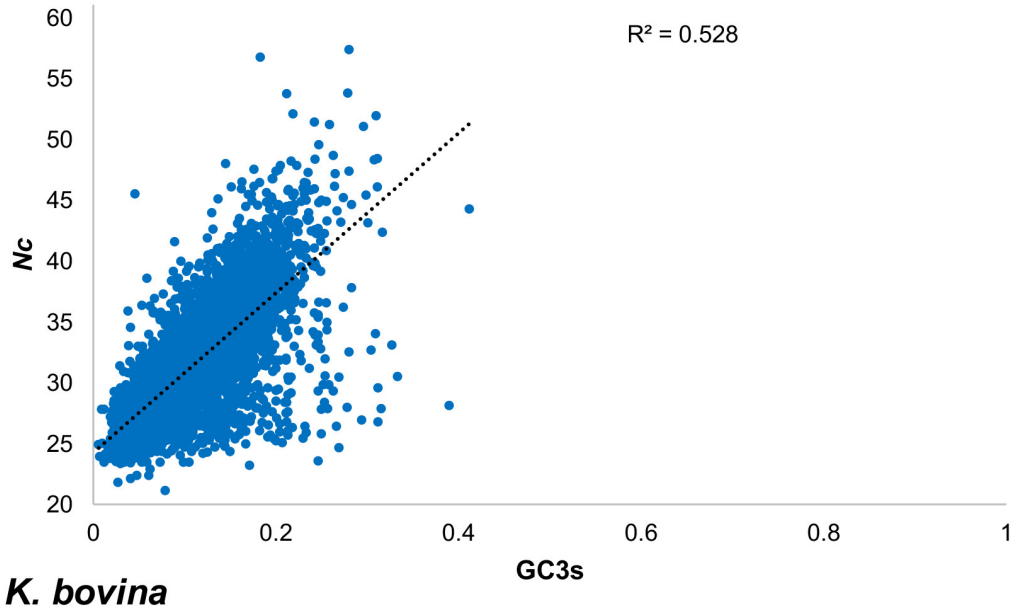


Figure B.2: **Nc plot against GC3s for the genes of *K. bovina*.**

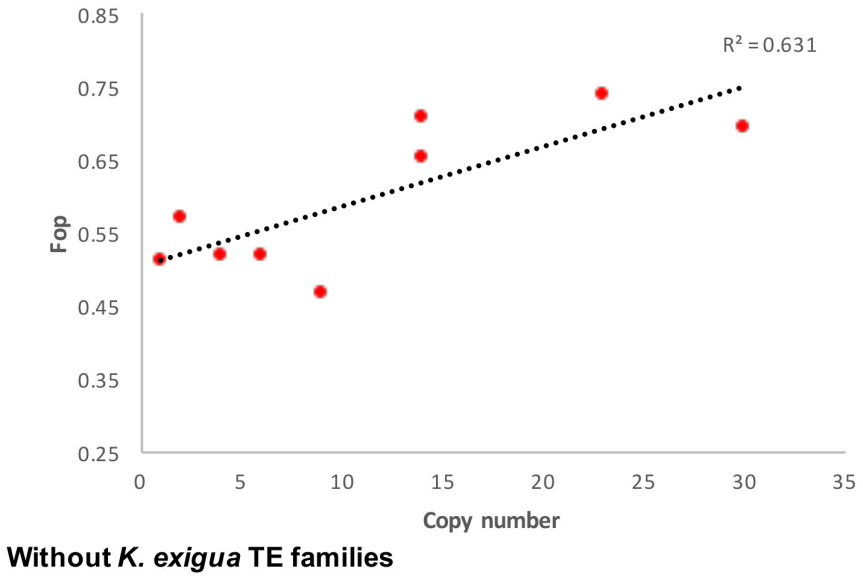


Figure B.3: **Fop x copy number without TE families from *K. exigua***

B.1.4 Syntenic data for *Kazachstania* species

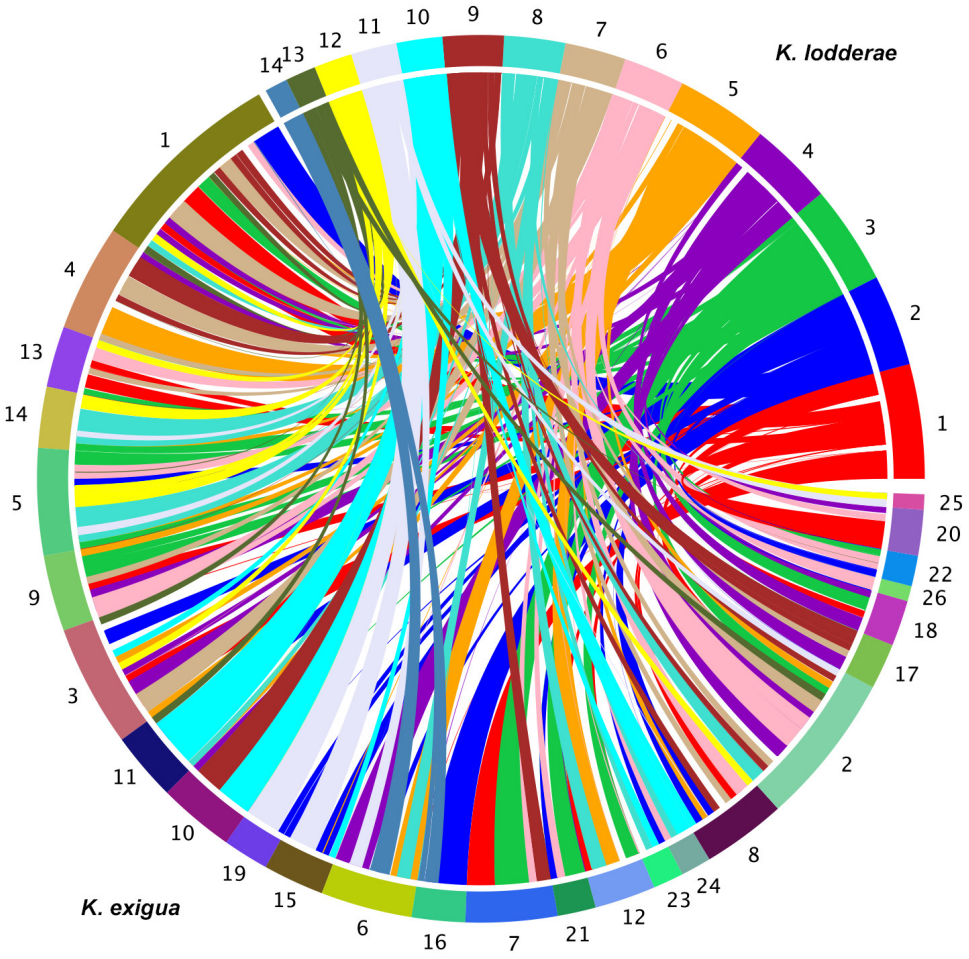


Figure B.4: **SyMap synteny analyses between *K. exigua* and *K. lodderae*.** Sytenic mapping between the contigs of *K. exigua* to the contigs of *K. lodderae*. Syntenic blocks were viewed using a circular view, with scaling based on genome size.

Chromodomain annotation

During phylogenetic analyses, the *gypsy*-like elements uncovered in additional yeast species were reviewed for the presence of a chromodomain, to determine if *gypsy* elements were likely to be chromoviruses. A predicted chromodomain was observed in *S. cerevisiae* Pol3 (AAA98435.1) downstream of the enzymatic domain Integrase within the Pol polyprotein. No putative chromodomains

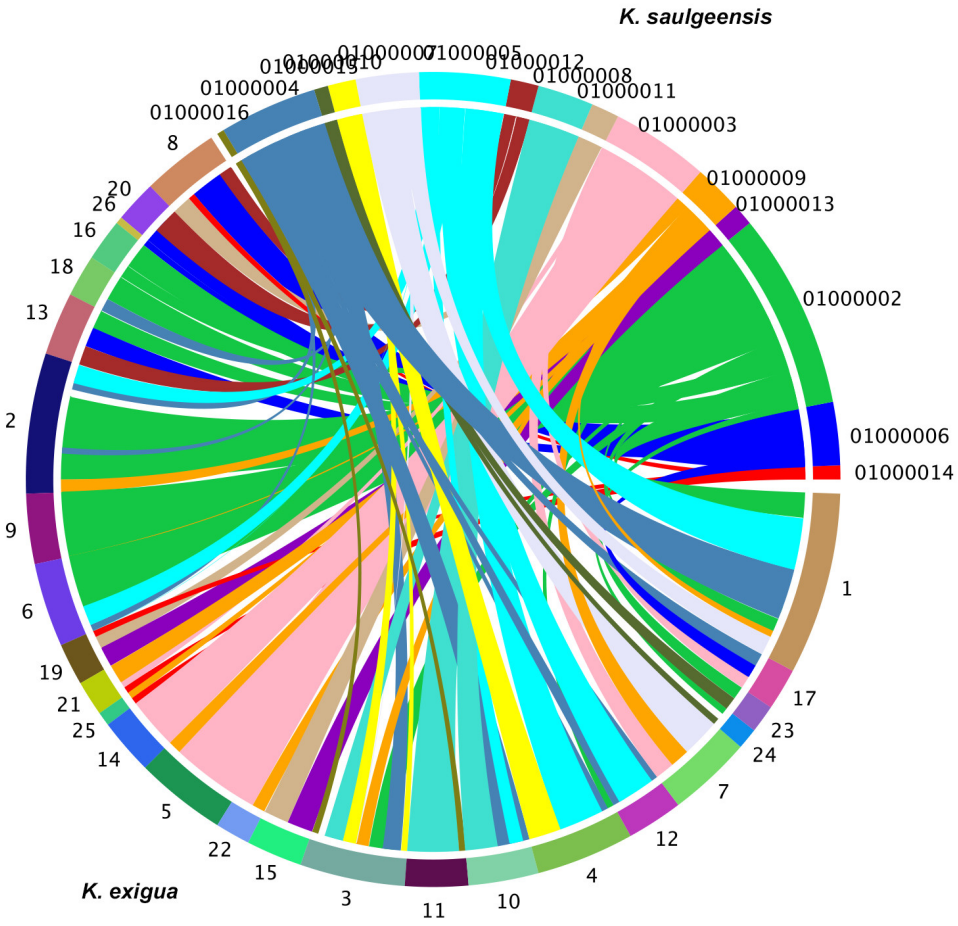


Figure B.5: **SyMap synteny analyses between *K. exigua* and *K. saulgeensis*.** Format is stated in B.4.

were uncovered in *Kazachstania species*. Details of the *S. cerevisiae* chromodomain are detailed in Appendix B. When ran against a template chromodomain crystallised structure on SWISS-MODEL (Biasini et al., 2014), the predicted sequence shared high similarity and reflected a similar secondary structure when viewed using PyMOL (Schrodinger, 2017). Similar sequences were found in several yeast species in the superfamily, and high conservation was seen throughout (Figure B.13).

The predicted chromodomain from Pol3 (AAA98435.1) was uploaded to PSI-PRED (Buchan et al., 2013; Jones, 1999) to predict the sequence secondary structure. Results showed a prediction of three beta pleated sheets, followed by an alpha helix (Figure B.14). This secondary structure has previously been documented for chromodomains (Kordis, 2005).

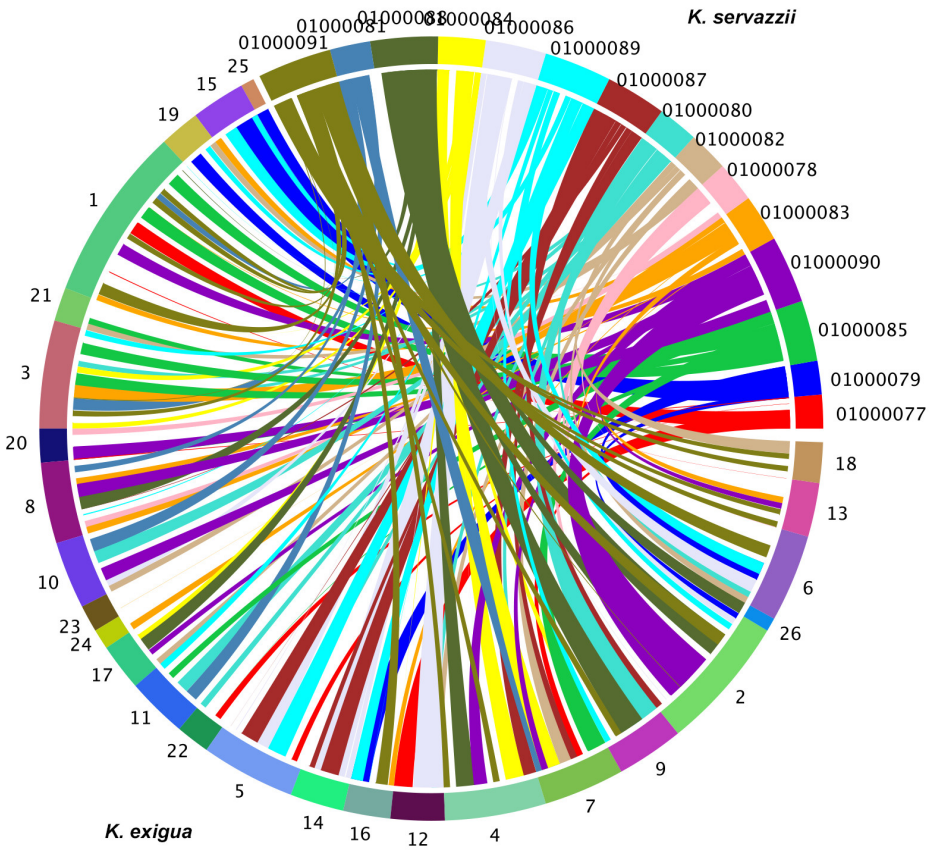


Figure B.6: **SyMap synteny analyses between *K. exigua* and *K. servazzii*.** Format is stated in B.4.

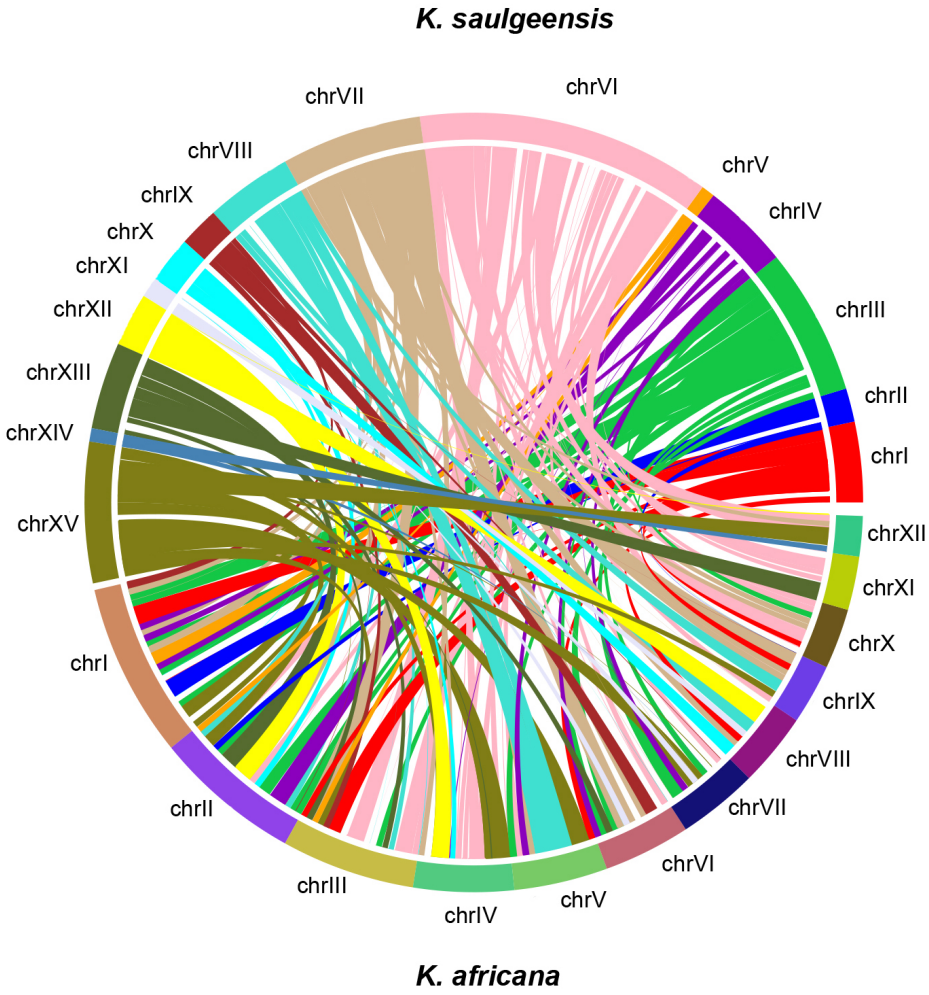


Figure B.7: **SyMap synteny analyses between *K. africana* and *K. saulgeensis*.** Format is stated in B.4.

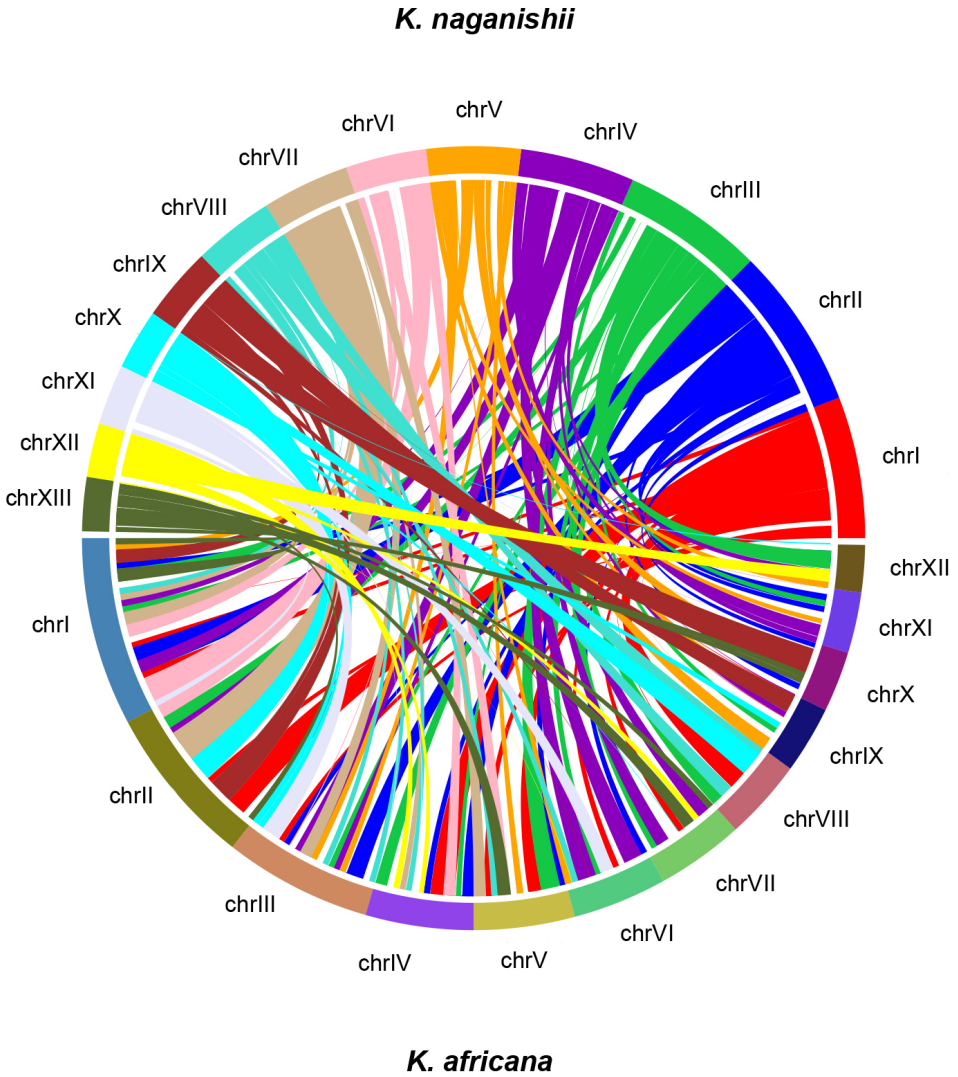


Figure B.8: **SyMap synteny analyses between *K. naganishii* and *K. africana*.** Format is stated in B.4.

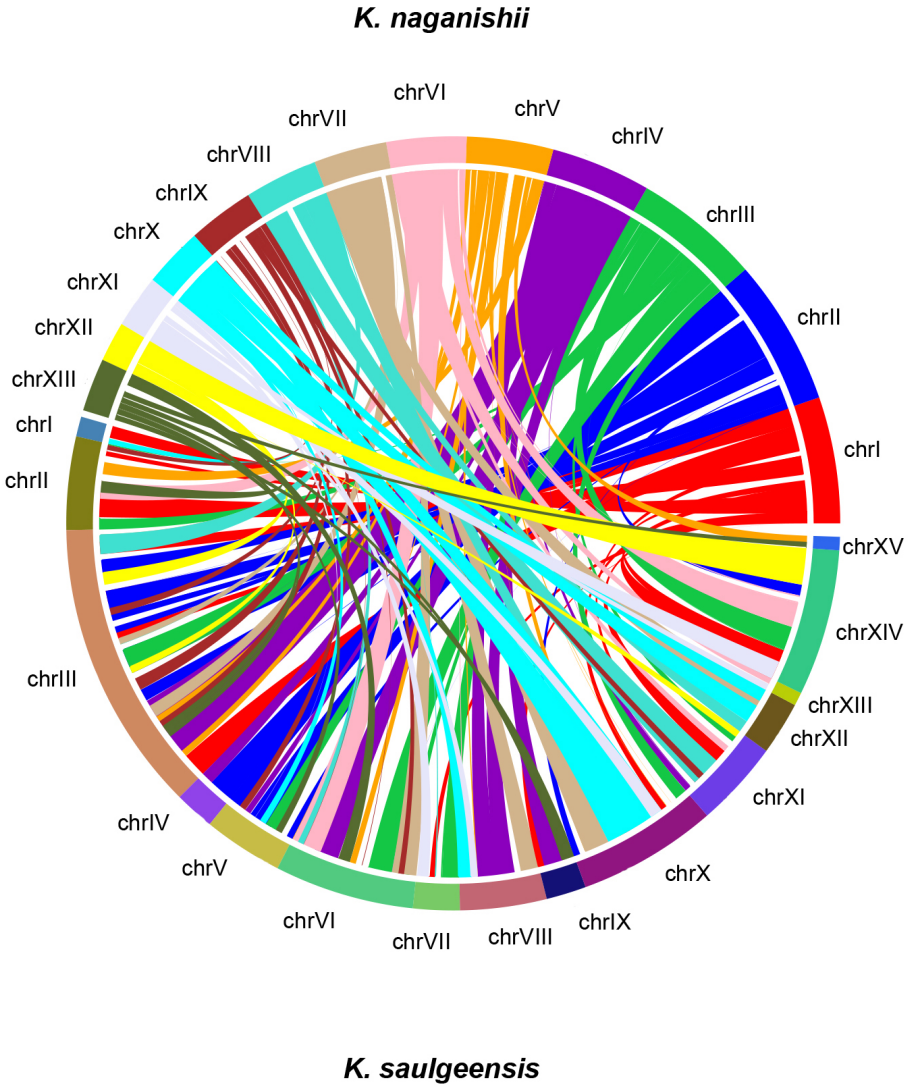


Figure B.9: **SyMap synteny analyses between *K. naganishii* and *K. saulgeensis*.** Format is stated in B.4.

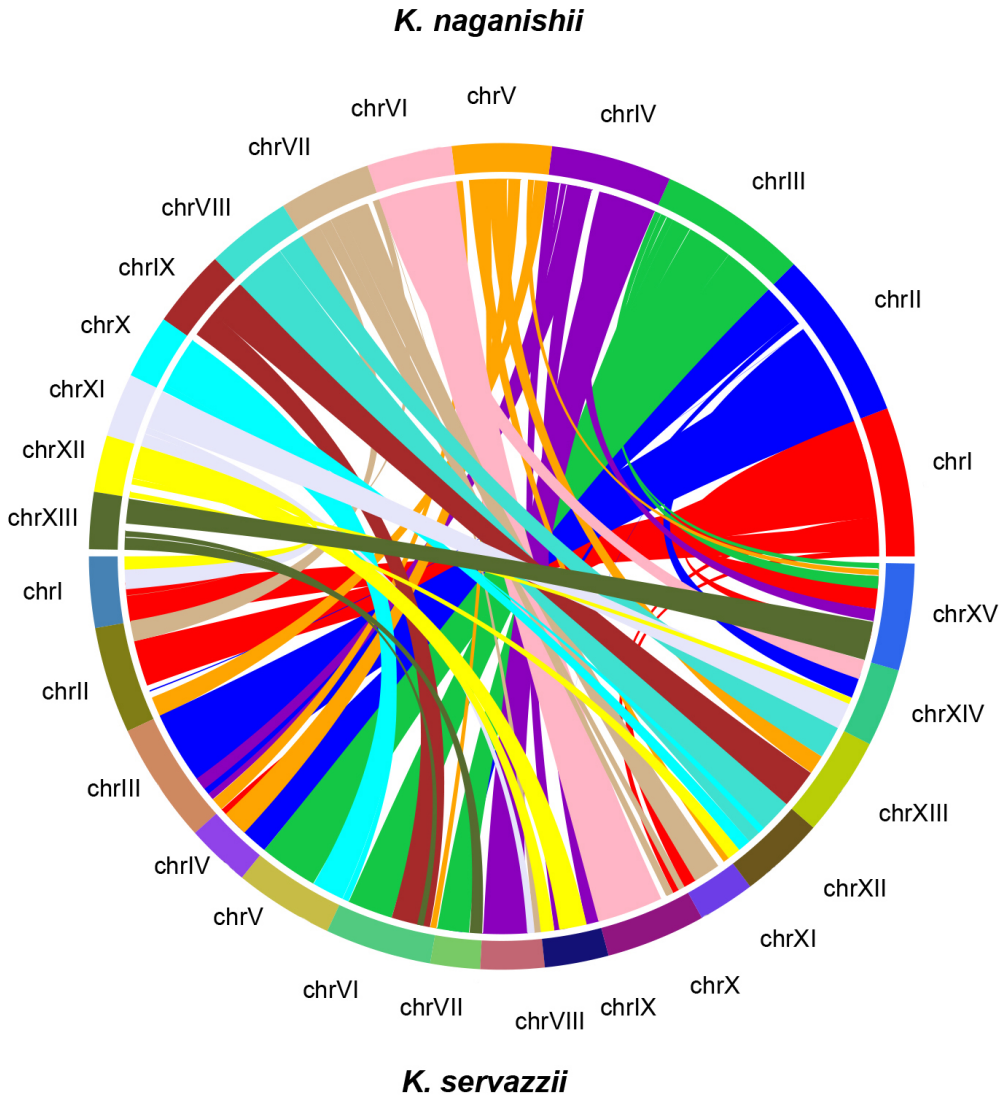


Figure B.10: **SyMap synteny analyses between *K. naganishii* and *K. servazzii*.** Format is stated in B.4.

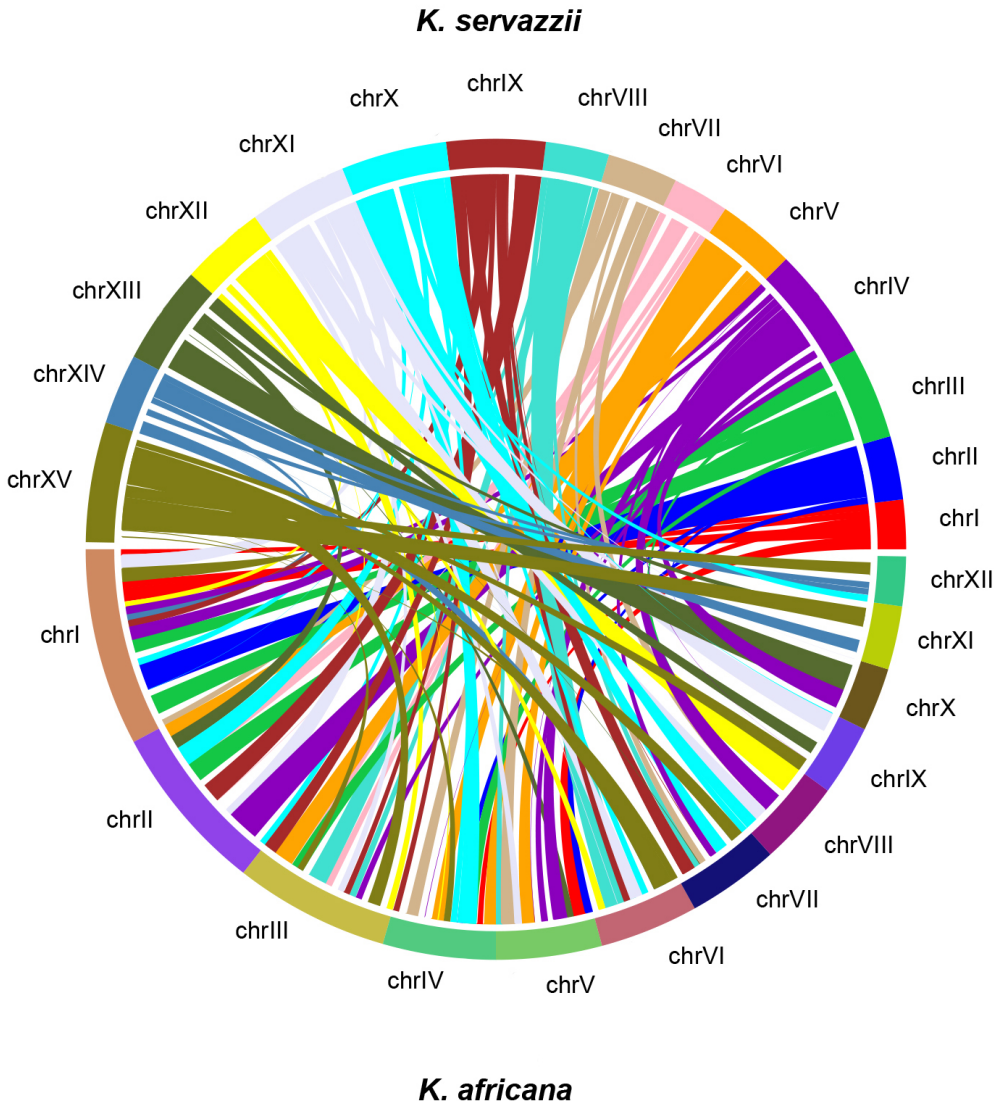


Figure B.11: **SyMap synteny analyses between *K. servazzii* and *K. africana*.** Format is stated in B.4.

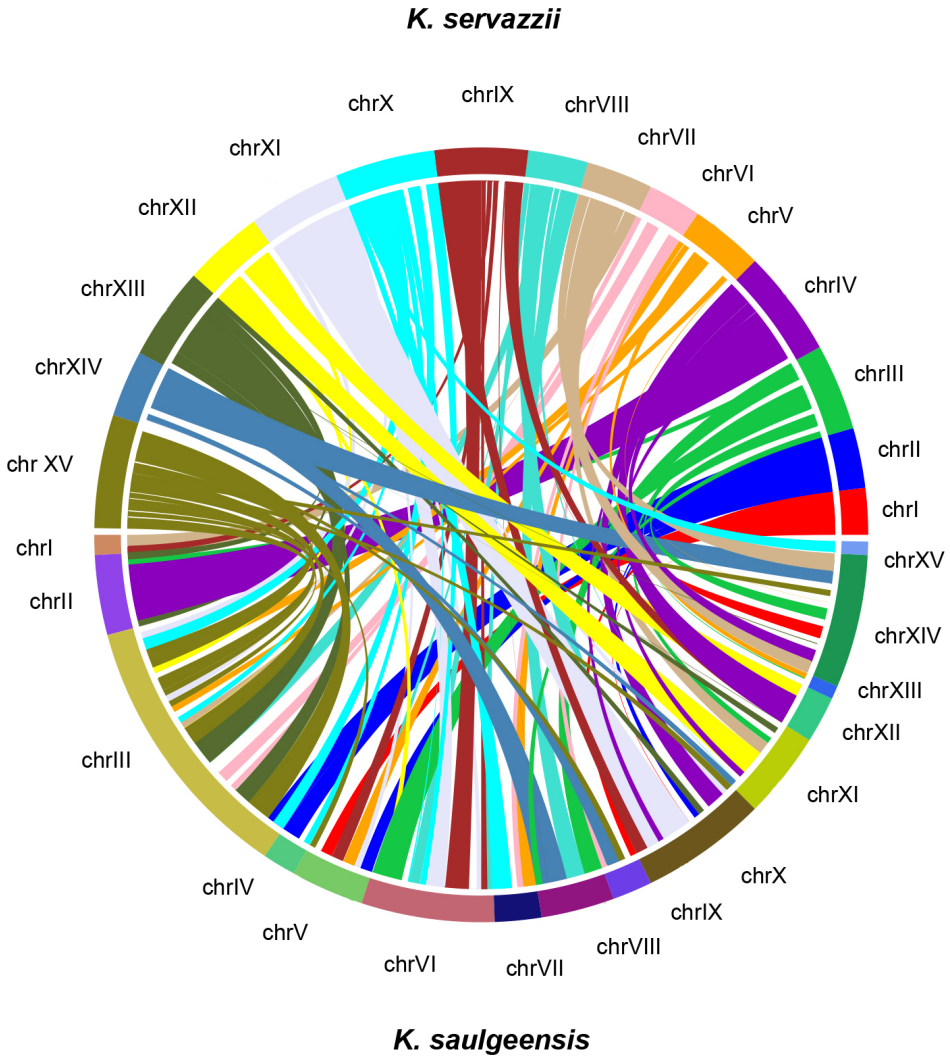


Figure B.12: **SyMap synteny analyses between *K. servazzii* and *K. saulgeensis*.** Format is stated in B.4.

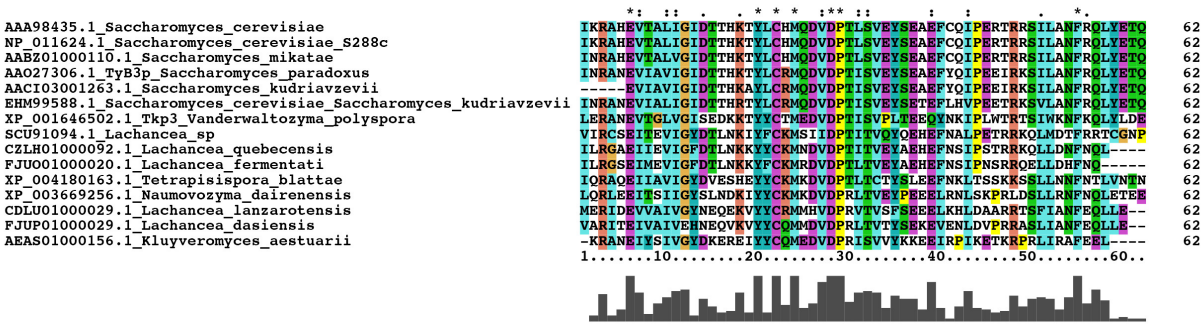


Figure B.13: **A graphic representation of the predicted chromodomain observed in chromoviral gypsy elements from the superfamily Saccharomycetaceae**.The alignment was produced using ClustalX v2.1 (Thompson et al., 2002). Sequences with high similarity to the query (AAA98435.1) were obtained using BLASTp on the NCBI server (Altschul et al., 1990; Sayers et al., 2009).

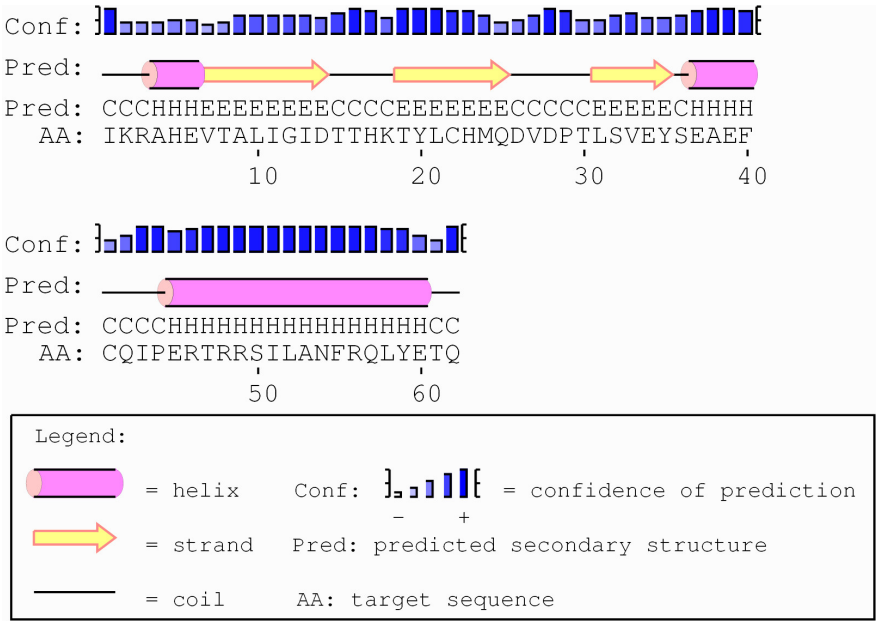


Figure B.14: **Predicted secondary structure of the predicted chromodomain in *S. cerevisiae* produced by PSIPRED (Jones, 1999).** Beta pleated sheets are represented yellow arrows; Alpha helices are represented by pink cylinders. The amino acid input sequence is displayed below the secondary structure.

Appendix C

Chapter 4 Appendix

C.1 Codon usage bias of three holozoan species

C.1.1 Predicted tRNA genes for the three holozoan species

Table C.1: **Output of tRNAscan-SE from whole genome contigs for predicted *S. rosetta* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal	Coordinates
NW_004754930.1	Ala	AGC	GCU	No	452568-452640
NW_004754912.1	Ala	AGC	GCU	No	107024-107096
NW_004754912.1	Ala	AGC	GCU	No	101134-101206
NW_004754894.1	Ala	AGC	GCU	No	134818-134890
NW_004754894.1	Ala	AGC	GCU	No	44385-44457
NW_004754923.1	Ala	CGC	GCG	No	1515588-1515659
NW_004754923.1	Ala	CGC	GCG	No	1311089-1311160
NW_004754923.1	Ala	TGC	GCA	No	1393765-1393836
NW_004754923.1	Ala	TGC	GCA	No	1511445-1511516
NW_004754918.1	Arg	ACG	CGU	No	412305-412378
NW_004754918.1	Arg	ACG	CGU	No	871624-871697
NW_004754918.1	Arg	ACG	CGU	No	883994-884067
NW_004754908.1	Arg	ACG	CGU	No	836639-836712
NW_004754908.1	Arg	ACG	CGU	No	435615-435688
NW_004754914.1	Arg	CCG	CGG	No	185725-185798
NW_004754896.1	Arg	CCT	AGG	No	119489-119571
NW_004754931.1	Arg	TCG	CGA	No	422660-422732

Table C.1: **Output of tRNAscan-SE from whole genome contigs for predicted *S. rosetta* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal	Coordinates
NW_004754927.1	Asn	GTT	AAC	Yes	1357268-1357341
NW_004754905.1	Asn	GTT	AAC	Yes	663864-663937
NW_004754899.1	Asn	GTT	AAC	Yes	651655-651728
NW_004754899.1	Asn	GTT	AAC	Yes	610981-611054
NW_004754922.1	Asp	GTC	GAC	Yes	63247-63318
NW_004754922.1	Asp	GTC	GAC	Yes	66135-66206
NW_004754922.1	Asp	GTC	GAC	Yes	130805-130876
NW_004754922.1	Asp	GTC	GAC	Yes	127942-128013
NW_004754909.1	Asp	GTC	GAC	Yes	966669-966740
NW_004754909.1	Asp	GTC	GAC	Yes	969565-969636
NW_004754924.1	Cys	GCA	UGC	Yes	60048-60119
NW_004754924.1	Cys	GCA	UGC	Yes	1607879-1607951
NW_004754930.1	Gln	CTG	CAG	Yes	1016433-1016504
NW_004754901.1	Gln	CTG	CAG	Yes	447297-447368
NW_004754901.1	Gln	CTG	CAG	Yes	446886-446957
NW_004457740.1	Gln	TTG	CAA	No	2056615-2056686
NW_004754929.1	Glu	CTC	GAG	Yes	485494-485565
NW_004754929.1	Glu	CTC	GAG	Yes	479268-479339
NW_004754927.1	Glu	CTC	GAG	Yes	299883-299954
NW_004754927.1	Glu	CTC	GAG	Yes	258209-258280
NW_004754911.1	Glu	TTC	GAA	No	273210-273283
NW_004754911.1	Glu	TTC	GAA	No	189506-189579
NW_004754906.1	Gly	CCC	GGG	No	448375-448446
NW_004754905.1	Gly	GCC	GGC	Yes	644976-645082
NW_004754905.1	Gly	GCC	GGC	Yes	664207-664313
NW_004754905.1	Gly	GCC	GGC	Yes	664007-664114
NW_004754899.1	Gly	GCC	GGC	Yes	610606-610712
NW_004754899.1	Gly	GCC	GGC	Yes	610806-610913

Table C.1: **Output of tRNAscan-SE from whole genome contigs for predicted *S. rosetta* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal	Coordinates
NW_004754899.1	Gly	GCC	GGC	Yes	639114-639220
NW_004754899.1	Gly	GCC	GGC	Yes	633177-633283
NW_004754924.1	Gly	TCC	GGA	No	775759-775850
NW_004754932.1	His	GTG	CAC	Yes	1836498-1836568
NW_004754919.1	His	GTG	CAC	Yes	650808-650878
NW_004754919.1	His	GTG	CAC	Yes	644193-644263
NW_004457740.1	Ile	AAT	AUU	No	2024814-2024887
NW_004457740.1	Ile	AAT	AUU	No	2037296-2037369
NW_004457740.1	Ile	AAT	AUU	No	2038726-2038799
NW_004754922.1	Ile	AAT	AUU	No	1096532-1096605
NW_004457740.1	Ile	TAT	AUA	No	874669-874774
NW_004754920.1	Leu	AAG	CUU	No	463901-463983
NW_004754920.1	Leu	AAG	CUU	No	469645-469727
NW_004754907.1	Leu	AAG	CUU	No	136181-136263
NW_004754907.1	Leu	AAG	CUU	No	119294-119376
NW_004754920.1	Leu	CAA	UUG	No	1416730-1416813
NW_004457740.1	Leu	CAG	CUG	Yes	2170541-2170623
NW_004457740.1	Leu	CAG	CUG	Yes	2170727-2170809
NW_004754930.1	Leu	CAG	CUG	Yes	1014228-1014311
NW_004754924.1	Leu	TAA	UUA	No	132321-132392
NW_004754894.1	Leu	TAG	CUA	No	73422-73504
NW_004754928.1	Lys	CTT	AAG	Yes	1755315-1755389
NW_004754928.1	Lys	CTT	AAG	Yes	1962031-1962105
NW_004754928.1	Lys	CTT	AAG	Yes	1962775-1962849
NW_004754928.1	Lys	CTT	AAG	Yes	1959956-1960030
NW_004754928.1	Lys	CTT	AAG	Yes	1748013-1748087
NW_004754928.1	Lys	CTT	AAG	Yes	1747255-1747329
NW_004754929.1	Lys	TTT	AAA	No	894357-894460

Table C.1: **Output of tRNAscan-SE from whole genome contigs for predicted *S. rosetta* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal	Coordinates
NW_004754926.1	Met	CAT	AUG	Yes	117006-117077
NW_004754926.1	Met	CAT	AUG	n/a	116790-116861
NW_004754918.1	Met	CAT	AUG	n/a	1192537-1192644
NW_004754917.1	Met	CAT	AUG	n/a	528284-528357
NW_004754900.1	Met	CAT	AUG	n/a	502952-503023
NW_004754900.1	Met	CAT	AUG	n/a	503309-503380
NW_004754900.1	Met	CAT	AUG	n/a	505892-505963
NW_004754900.1	Met	CAT	AUG	n/a	576610-576681
NW_004754900.1	Met	CAT	AUG	n/a	579259-579330
NW_004754882.1	Met	CAT	AUG	n/a	6245-6352
NW_004754882.1	Met	CAT	AUG	n/a	961-1068
NW_004754926.1	Phe	GAA	UUC	Yes	556718-556792
NW_004754926.1	Phe	GAA	UUC	Yes	540816-540890
NW_004754923.1	Phe	GAA	UUC	Yes	592507-592599
NW_004754925.1	Pro	AGG	CCU	No	1136201-1136272
NW_004754918.1	Pro	AGG	CCU	No	549797-549868
NW_004754918.1	Pro	AGG	CCU	No	550871-550942
NW_004754921.1	Pro	CGG	CCG	Yes	1428456-1428527
NW_004754907.1	Pro	TGG	CCA	No	427073-427144
NW_004754907.1	Pro	TGG	CCA	No	419713-419784
NW_004754898.1	SeC	TCA	UGA	No	611578-611664
NW_004754931.1	Ser	AGA	UCU	Yes	1419292-1419373
NW_004754931.1	Ser	AGA	UCU	Yes	1776086-1776167
NW_004754931.1	Ser	AGA	UCU	Yes	1425940-1426021
NW_004754907.1	Ser	CGA	UCG	Yes	145985-146069
NW_004754923.1	Ser	GCT	AGC	Yes	1445637-1445710
NW_004754898.1	Ser	GCT	AGC	Yes	503122-503198
NW_004754930.1	Ser	TGA	UCA	No	1984169-1984272

Table C.1: **Output of tRNAscan-SE from whole genome contigs for predicted *S. rosetta* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal	Coordinates
NW_004754932.1	Thr	TGT	ACA	No	2394324-2394396
NW_004754930.1	Thr	TGT	ACA	No	618517-618612
NW_004754930.1	Thr	TGT	ACA	No	619530-619625
NW_004754925.1	Thr	CGT	ACG	Yes	1263450-1263522
NW_004754925.1	Thr	CGT	ACG	Yes	1274595-1274667
NW_004754910.1	Trp	CCA	UGG	n/a	196530-196602
NW_004754910.1	Trp	CCA	UGG	n/a	201603-201675
NW_004754930.1	Tyr	GTA	UAC	Yes	2201971-2202061
NW_004754930.1	Tyr	GTA	UAC	Yes	2196947-2197034
NW_004754929.1	Tyr	GTA	UAC	Yes	305886-305975
NW_004754918.1	Tyr	GTA	UAC	Yes	34361-34448
NW_004754898.1	Tyr	GTA	UAC	Yes	160364-160451
NW_004754917.1	Val	AAC	GUU	No	1029907-1029980
NW_004754917.1	Val	AAC	GUU	No	1318199-1318272
NW_004754917.1	Val	AAC	GUU	No	1318390-1318463
NW_004754917.1	Val	AAC	GUU	No	891759-891832
NW_004754917.1	Val	CAC	GUG	No	1204577-1204650
NW_004754917.1	Val	CAC	GUG	No	1333815-1333888
NW_004754917.1	Val	CAC	GUG	No	1212563-1212636
NW_004754927.1	Val	TAC	GUA	No	263192-263264

Table C.2: **Output of tRNAscan-SE from whole genome contigs for predicted *M. brevicollis* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinates
NW_001865068.1	Ala	AGC	GCU	Yes	52832-52922
NW_001865068.1	Ala	AGC	GCU	Yes	53028-53117
NW_001865058.1	Ala	AGC	GCU	Yes	634940-635029
NW_001865053.1	Ala	AGC	GCU	Yes	886446-886535
NW_001865053.1	Ala	AGC	GCU	Yes	886655-886744

Table C.2: **Output of tRNAscan-SE from whole genome contigs for predicted *M. brevicollis* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinate
NW_001865053.1	Ala	AGC	GCU	Yes	894519-8946
NW_001865045.1	Ala	CGC	GCG	No	411478-4115
NW_001865045.1	Ala	CGC	GCG	No	411667-4117
NW_001865045.1	Ala	TGC	GCA	No	411877-4119
NW_001865045.1	Ala	TGC	GCA	No	412059-4121
NW_001865066.1	Arg	ACG	CGU	Yes	371166-3712
NW_001865040.1	Arg	ACG	CGU	Yes	1150818-1150
NW_001865040.1	Arg	ACG	CGU	Yes	1150634-1150
NW_001865040.1	Arg	ACG	CGU	Yes	1150434-1150
NW_001865040.1	Arg	ACG	CGU	Yes	1150111-1150
NW_001865040.1	Arg	ACG	CGU	Yes	1149876-1149
NW_001865084.1	Arg	CCT	AGG	No	105773-1058
NW_001865050.1	Arg	TCG	CGA	No	947505-9475
NW_001865041.1	Arg	TCG	CGA	No	315927-3159
NC_004309.1.	Arg	TCG	CGA	No	68221-6829
NW_001865049.1	Arg	TCT	AGA	No	353629-3537
NC_004309.1.	Arg	TCT	AGA	No	19351-1942
NW_001865086.1	Asn	GTT	AAC	Yes	125396-1254
NW_001865086.1	Asn	GTT	AAC	Yes	126199-1262
NW_001865086.1	Asn	GTT	AAC	Yes	131693-1317
NC_004309.1.	Asn	GTT	AAC	Yes	73074-7314
NW_001865075.1	Asp	GTC	GAC	Yes	259272-2593
NW_001865075.1	Asp	GTC	GAC	Yes	259442-2595
NW_001865075.1	Asp	GTC	GAC	Yes	259633-2597
NW_001865075.1	Asp	GTC	GAC	Yes	259813-2598
NW_001865069.1	Asp	GTC	GAC	Yes	369064-3691
NW_001865069.1	Asp	GTC	GAC	Yes	352845-3529
NC_004309.1.	Asp	GTC	GAC	Yes	27266-2733

Table C.2: **Output of tRNAscan-SE from whole genome contigs for predicted *M. brevicollis* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinates
NW_001865078.1	Cys	GCA	UGC	Yes	44495-44567
NW_001865043.1	Cys	GCA	UGC	Yes	935967-936035
NW_001865043.1	Cys	GCA	UGC	Yes	935718-935797
NW_001865070.1	Gln	CTG	CAG	Yes	354273-354344
NW_001865070.1	Gln	CTG	CAG	Yes	144029-144100
NW_001865070.1	Gln	CTG	CAG	Yes	143463-143534
NW_001865054.1	Gln	TTG	CAA	No	251303-251374
NW_001865054.1	Gln	TTG	CAA	No	251503-251574
NC_004309.1.	Gln	TTG	CAA	No	26938-27008
NW_001865066.1	Glu	CTC	GAG	Yes	310848-310919
NW_001865054.1	Glu	CTC	GAG	Yes	714063-714134
NW_001865050.1	Glu	CTC	GAG	Yes	414502-414573
NW_001865050.1	Glu	CTC	GAG	Yes	414673-414744
NW_001865050.1	Glu	CTC	GAG	Yes	414847-414918
NW_001865050.1	Glu	CTC	GAG	Yes	415537-415608
NW_001865052.1	Glu	TTC	GAA	No	945940-946011
NW_001865052.1	Glu	TTC	GAA	No	946133-946204
NC_004309.1.	Glu	TTC	GAA	No	29141-29213
NW_001865079.1	Gly	CCC	GGG	No	240281-240351
NW_001865089.1	Gly	GCC	GGC	Yes	30959-31029
NW_001865089.1	Gly	GCC	GGC	Yes	30629-30699
NW_001865079.1	Gly	GCC	GGC	Yes	241017-241087
NW_001865079.1	Gly	GCC	GGC	Yes	240796-240866
NW_001865079.1	Gly	GCC	GGC	Yes	240476-240546
NW_001865079.1	Gly	GCC	GGC	Yes	240103-240173
NW_001865044.1	Gly	GCC	GGC	Yes	431060-431130
NW_001865062.1	Gly	TCC	GGA	No	334076-334184
NC_004309.1.	Gly	TCC	GGA	No	76497-76568

Table C.2: **Output of tRNAscan-SE from whole genome contigs for predicted *M. brevicollis* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinate
NW_001865041.1	His	GTG	CAC	Yes	1805888-1805
NW_001865041.1	His	GTG	CAC	Yes	1806150-1806
NW_001865041.1	His	GTG	CAC	Yes	1806545-1806
NC_004309.1.	His	GTG	CAC	Yes	15743-1581
NW_001865051.1	Ile	AAT	AUU	Yes	1036763-1036
NW_001865040.1	Ile	AAT	AUU	Yes	2156956-2157
NW_001865040.1	Ile	AAT	AUU	Yes	2156594-2156
NC_004309.1.	Ile	GAT	AUC	Yes	25193-2526
NW_001865058.1	Ile	TAT	AUA	No	348320-3484
NW_001865053.1	iMe	CAT	AUG	n\a	423588-4236
NW_001865053.1	iMe	CAT	AUG	n\a	423951-4240
NW_001865055.1	Leu	AAG	CUU	Yes	627184-6272
NW_001865055.1	Leu	AAG	CUU	Yes	626847-6269
NW_001865055.1	Leu	AAG	CUU	Yes	626440-6265
NW_001865055.1	Leu	AAG	CUU	Yes	625638-6257
NW_001865043.1	Leu	CAA	UUG	No	964597-9646
NW_001865043.1	Leu	CAA	UUG	No	964395-9644
NW_001865043.1	Leu	CAG	CUG	Yes	986392-9864
NW_001865043.1	Leu	CAG	CUG	Yes	986572-9866
NW_001865043.1	Leu	CAG	CUG	Yes	986878-9869
NW_001865040.1	Leu	TAA	UUA	No	792410-7924
NC_004309.1.	Leu	TAA	UUA	No	29637-2971
NW_001865044.1	Leu	TAG	CUA	No	726957-7270
NC_004309.1.	Leu	TAG	CUA	No	7006-7089
NW_001865073.1	Lys	CTT	AAG	Yes	54631-5479
NW_001865063.1	Lys	CTT	AAG	Yes	250951-2510
NW_001865063.1	Lys	CTT	AAG	Yes	251372-2514
NW_001865063.1	Lys	CTT	AAG	Yes	252098-2521

Table C.2: **Output of tRNAscan-SE from whole genome contigs for predicted *M. brevicollis* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinates
NW_001865063.1	Lys	CTT	AAG	Yes	252651-252741
NW_001865042.1	Lys	CTT	AAG	Yes	857442-857521
NW_001865063.1	Lys	TTT	AAA	No	253157-253229
NC_004309.1.	Lys	TTT	AAA	No	34553-34624
NW_001865067.1	Met	CAT	AUG	n/a	542238-542310
NW_001865049.1	Met	CAT	AUG	n/a	537483-537555
NW_001865049.1	Met	CAT	AUG	n/a	537154-537226
NC_004309.1.	Met	CAT	AUG	n/a	29286-29358
NC_004309.1.	Met	CAT	AUG	n/a	33428-33501
NW_001865060.1	Phe	GAA	UUC	Yes	536169-536241
NW_001865060.1	Phe	GAA	UUC	Yes	552611-552683
NW_001865060.1	Phe	GAA	UUC	Yes	553023-553095
NW_001865060.1	Phe	GAA	UUC	Yes	553464-553536
NC_004309.1.	Phe	GAA	UUC	Yes	30013-30085
NW_001865042.1	Pro	AGG	CCU	No	695526-695597
NW_001865042.1	Pro	AGG	CCU	No	695352-695423
NW_001865042.1	Pro	AGG	CCU	No	695159-695230
NW_001865092.1	Pro	CGG	CCG	No	46063-46181
NW_001865080.1	Pro	TGG	CCA	No	229064-229135
NC_004309.1.	Pro	TGG	CCA	No	30298-30368
NW_001865040.1	SeC	TCA	UGA	No	481253-481339
NW_001865061.1	Ser	AGA	UCU	No	340568-340649
NW_001865051.1	Ser	AGA	UCU	No	220132-220213
NW_001865051.1	Ser	AGA	UCU	No	219118-219199
NW_001865051.1	Ser	CGA	UCG	Yes	234840-234921
NW_001865051.1	Ser	CGA	UCG	Yes	217486-217567
NW_001865047.1	Ser	GCT	AGC	Yes	212569-212642
NW_001865047.1	Ser	GCT	AGC	Yes	212996-213069

Table C.2: **Output of tRNAscan-SE from whole genome contigs for predicted *M. brevicollis* tRNA genes**

Contig	tRNA gene	tRNA anticodon	Codon	CodonW Optimal Codon	Coordinate
NW_001865051.1	Ser	TGA	UCA	No	231490-2315
NC_004309.1.	Ser	TGA	UCA	No	71794-7187
NC_004309.1.	Sup	TCA	UGA	No	25593-2566
NW_001865041.1	Thr	CGT	ACG	No	111542-1116
NW_001865039.1	Thr	CGT	ACG	No	2422150-2422
NW_001865049.1	Thr	TGT	ACA	No	536597-5366
NC_004309.1.	Thr	TGT	ACA	No	15143-1521
NW_001865092.1	Trp	CCA	UGG	n/a	49232-4930
NW_001865092.1	Trp	CCA	UGG	n/a	49484-4955
NW_001865058.1	Tyr	GTA	UAC	Yes	683103-6831
NW_001865058.1	Tyr	GTA	UAC	Yes	683494-6835
NW_001865049.1	Tyr	GTA	UAC	Yes	220465-2205
NC_004309.1.	Tyr	GTA	UAC	Yes	54783-5486
NW_001865062.1	Val	AAC	GUU	Yes	589695-5897
NW_001865062.1	Val	AAC	GUU	Yes	600688-6007
NW_001865062.1	Val	AAC	GUU	Yes	658228-6583
NW_001865039.1	Val	AAC	GUU	Yes	2094501-2094
NW_001865039.1	Val	CAC	GUG	No	2095019-2095
NW_001865039.1	Val	CAC	GUG	No	2094696-2094
NW_001865059.1	Val	TAC	GUA	No	406615-4066
NC_004309.1.	Val	TAC	GUA	No	31108-3117

Table C.3: **Output of tRNAscan-SE from whole genome contigs for predicted *C. owczarzaki* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal Codon	Co-ordinate
NW_011887292.1	Ala	AGC	GCU	No	315865-3159
NW_011887292.1	Ala	AGC	GCU	No	320860-3209
NW_011887292.1	Ala	AGC	GCU	No	3306684-3306

Table C.3: **Output of tRNAscan-SE from whole genome contigs for predicted *C. owczarzaki* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal Codon	Co-ordinates
NW_011887300.1	Ala	AGC	GCU	No	1184309-118438
NW_011887304.1	Ala	AGC	GCU	No	453635-453707
NW_011887305.1	Ala	AGC	GCU	No	539692-539764
NW_011887306.1	Ala	CGC	GCG	No	60262-60334
NW_011887307.1	Ala	CGC	GCG	No	375836-375908
NW_011887293.1	Ala	TGC	GCA	No	2370531-237060
NW_011887293.1	Ala	TGC	GCA	No	2370691-237076
NW_011887294.1	Arg	ACG	CGU	Yes	1183664-118373
NW_011887298.1	Arg	ACG	CGU	Yes	947727-947799
NW_011887298.1	Arg	ACG	CGU	Yes	949658-949730
NW_011887306.1	Arg	ACG	CGU	Yes	272494-272566
NW_011887292.1	Arg	CCT	AGG	No	1313812-131388
NW_011887301.1	Arg	TCG	CGA	No	935143-935213
NW_011887301.1	Arg	TCT	AGA	No	262133-262220
NW_011887295.1	Asn	GTT	AAC	Yes	691943-692015
NW_011887296.1	Asn	GTT	AAC	Yes	230228-230300
NW_011887296.1	Asn	GTT	AAC	Yes	230405-230477
NW_011887297.1	Asp	GTC	GAC	Yes	4350-4421
NW_011887298.1	Asp	GTC	GAC	Yes	563490-563561
NW_011887304.1	Asp	GTC	GAC	Yes	190311-190382
NW_011887308.1	Asp	GTC	GAC	Yes	392162-392233
NW_011887308.1	Asp	GTC	GAC	Yes	391870-391941
NW_011887308.1	Asp	GTC	GAC	Yes	376650-376721
NW_011887319.1	Asp	GTC	GAC	Yes	94-164
NW_011887295.1	Cys	GCA	UGC	Yes	1814481-181455
NW_011887299.1	Cys	GCA	UGC	Yes	659561-659632
NW_011887292.1	Gln	CTG	CAG	Yes	316153-316231
NW_011887292.1	Gln	CTG	CAG	Yes	316447-316525

Table C.3: **Output of tRNAscan-SE from whole genome contigs for predicted *C. owczarzaki* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal Codon	Co-ordinat
NW_011887307.1	Gln	CTG	CAG	Yes	240141-2402
NW_011887293.1	Gln	TTG	CAG	Yes	641170-6412
NW_011887292.1	Glu	CTC	GAG	Yes	366443-3665
NW_011887292.1	Glu	CTC	GAG	Yes	379458-3795
NW_011887292.1	Glu	CTC	GAG	Yes	366151-3662
NW_011887292.1	Glu	CTC	GAG	Yes	365879-3659
NW_011887301.1	Glu	CTC	GAG	Yes	612921-6129
NW_011887292.1	Glu	TTC	GAA	No	366698-3667
NW_011887292.1	Glu	TTC	GAA	No	366858-3669
NW_011887294.1	Gly	GCC	GGC	Yes	452651-4527
NW_011887298.1	Gly	GCC	GGC	Yes	1057510-1057
NW_011887299.1	Gly	GCC	GGC	Yes	469461-4695
NW_011887301.1	Gly	GCC	GGC	Yes	571383-5714
NW_011887306.1	Gly	GCC	GGC	Yes	728620-7286
NW_011887306.1	Gly	GCC	GGC	Yes	728308-7283
NW_011887296.1	Gly	TCC	GGA	No	1178305-1178
NW_011887306.1	Gly	TCC	GGA	No	369813-3698
NW_011887319.1	Gly	TCC	GGA	No	22-92
NW_011887294.1	His	GTG	CAC	Yes	838921-8389
NW_011887294.1	His	GTG	CAC	Yes	839087-8391
NW_011887294.1	His	GTG	CAC	Yes	970333-9704
NW_011887319.1	His	GTG	CAC	Yes	167-237
NW_011887292.1	Ile	AAT	AUU	Yes	2636595-2636
NW_011887295.1	Ile	AAT	AUU	Yes	433704-4337
NW_011887300.1	Ile	AAT	AUU	Yes	40300-4037
NW_011887300.1	Ile	AAT	AUU	Yes	40127-4020
NW_011887297.1	iMe	CAT	AUG	n/a	926241-9263
NW_011887292.1	Leu	AAG	CUU	No	2788540-2788

Table C.3: **Output of tRNAscan-SE from whole genome contigs for predicted *C. owczarzaki* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal Codon	Co-ordinates
NW_011887308.1	Leu	AAG	CUU	No	47432-47514
NW_011887310.1	Leu	AAG	CUU	No	26429-26511
NW_011887329.1	Leu	AAG	CUU	No	5125-5207
NW_011887294.1	Leu	CAA	UUG	No	550590-550671
NW_011887297.1	Leu	CAG	CUG	Yes	52912-52992
NW_011887304.1	Leu	CAG	CUG	Yes	636207-636287
NW_011887297.1	Leu	TAA	UUA	No	1442132-144221
NW_011887293.1	Leu	TAG	CUA	No	24800-24880
NW_011887296.1	Lys	CTT	AAG	Yes	46082-46154
NW_011887296.1	Lys	CTT	AAG	Yes	46310-46382
NW_011887296.1	Lys	CTT	AAG	Yes	48675-48747
NW_011887303.1	Lys	CTT	AAG	Yes	1028091-102816
NW_011887303.1	Lys	CTT	AAG	Yes	745699-745771
NW_011887296.1	Lys	TTT	AAA	No	65275-65347
NW_011887292.1	Met	CAT	AUG	n/a	870551-870623
NW_011887292.1	Met	CAT	AUG	n/a	865912-865984
NW_011887302.1	Met	CAT	AUG	n/a	466577-466648
NW_011887302.1	Met	CAT	AUG	n/a	466388-466459
NW_011887300.1	Phe	GAA	UUC	Yes	816570-816642
NW_011887300.1	Phe	GAA	UUC	Yes	817003-817075
NW_011887300.1	Phe	GAA	UUC	Yes	816833-816905
NW_011887293.1	Pro	AGG	CCU	No	1160615-116068
NW_011887293.1	Pro	AGG	CCU	No	1278319-127839
NW_011887293.1	Pro	AGG	CCU	No	1278147-127821
NW_011887296.1	Pro	CGG	CCG	No	950983-951054
NW_011887296.1	Pro	CGG	CCG	No	951129-951200
NW_011887295.1	Pro	TGG	CCA	No	249340-249415
NW_011887302.1	SeC	TCA	UGA	No	616113-616200

Table C.3: **Output of tRNAscan-SE from whole genome contigs for predicted *C. owczarzaki* tRNA genes**

Contig	tRNA gene	tRNA anti codon	Codon	CodonW Optimal Codon	Co-ordinat
NW_011887294.1	Ser	AGA	UCU	No	1028980-1029
NW_011887294.1	Ser	AGA	UCU	No	1029301-1029
NW_011887296.1	Ser	AGA	UCU	No	1478121-1478
NW_011887294.1	Ser	CGA	UCG	Yes	1031241-1031
NW_011887294.1	Ser	CGA	UCG	Yes	1029608-1029
NW_011887294.1	Ser	GCT	AGC	Yes	1468662-1468
NW_011887294.1	Ser	GCT	AGC	Yes	667552-6676
NW_011887295.1	Ser	TGA	UCA	No	71393-7147
NW_011887293.1	Thr	AGT	ACU	No	1543853-1543
NW_011887295.1	Thr	AGT	ACU	No	482476-4825
NW_011887296.1	Thr	AGT	ACU	No	311848-3119
NW_011887307.1	Thr	AGT	ACU	No	640395-6404
NW_011887296.1	Thr	CGT	ACG	No	312020-3120
NW_011887296.1	Thr	CGT	ACG	No	241446-2415
NW_011887306.1	Thr	TGT	ACA	No	245427-2455
NW_011887293.1	Trp	CCA	UGG	n/a	1702351-1702
NW_011887295.1	Trp	CCA	UGG	n/a	489010-4890
NW_011887301.1	Tyr	GTA	UAC	Yes	201635-2017
NW_011887306.1	Tyr	GTA	UAC	Yes	54050-5413
NW_011887306.1	Und	NNN	NNN	n/a	369741-3698
NW_011887293.1	Val	AAC	GUU	No	559192-5592
NW_011887297.1	Val	AAC	GUU	No	277318-2773
NW_011887297.1	Val	AAC	GUU	No	277407-2774
NW_011887301.1	Val	AAC	GUU	No	979312-9793
NW_011887296.1	Val	CAC	GUG	No	120083-1201
NW_011887292.1	Val	TAC	GUA	No	1103043-1103

C.1.2 Normal distribution graphs for bias categories

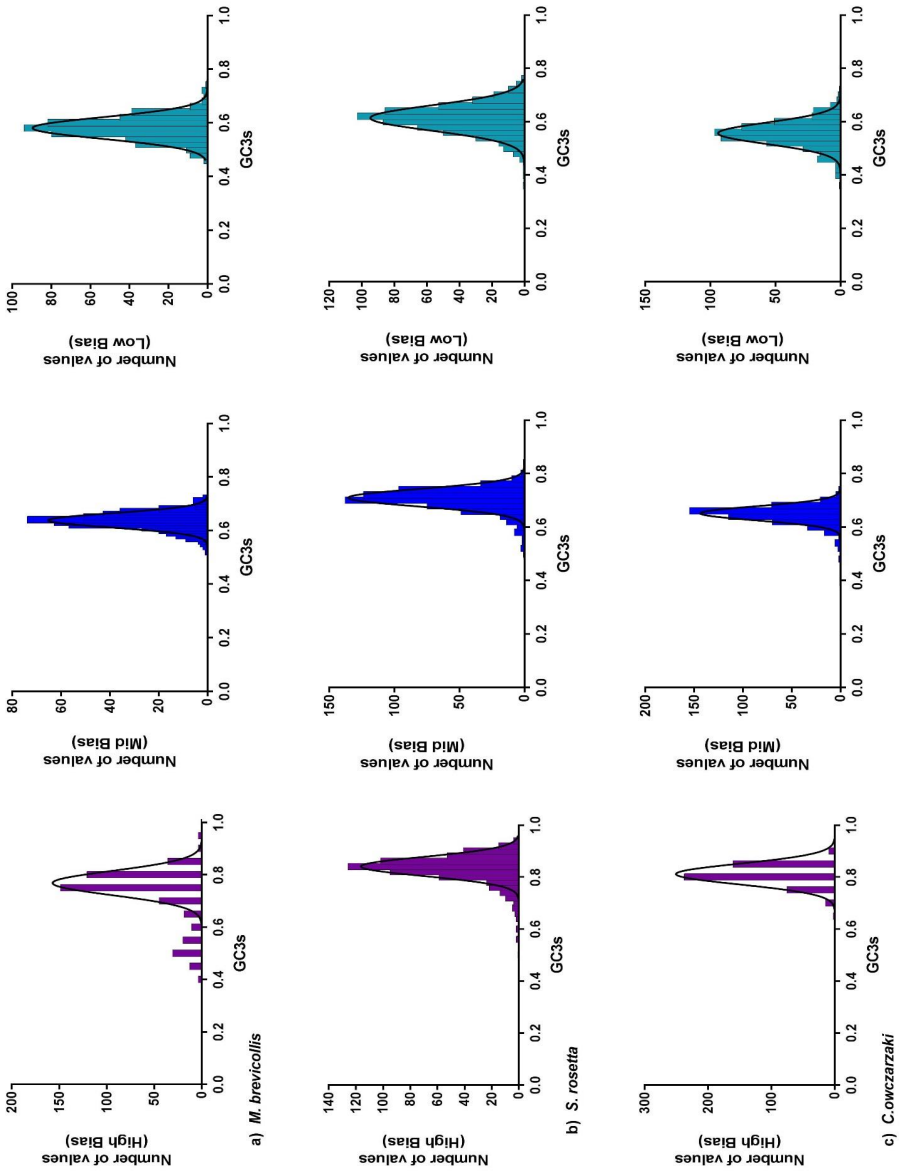


Figure C.1: **Normal distribution of GC content at the synonymous third position for the three 5% bias categories**

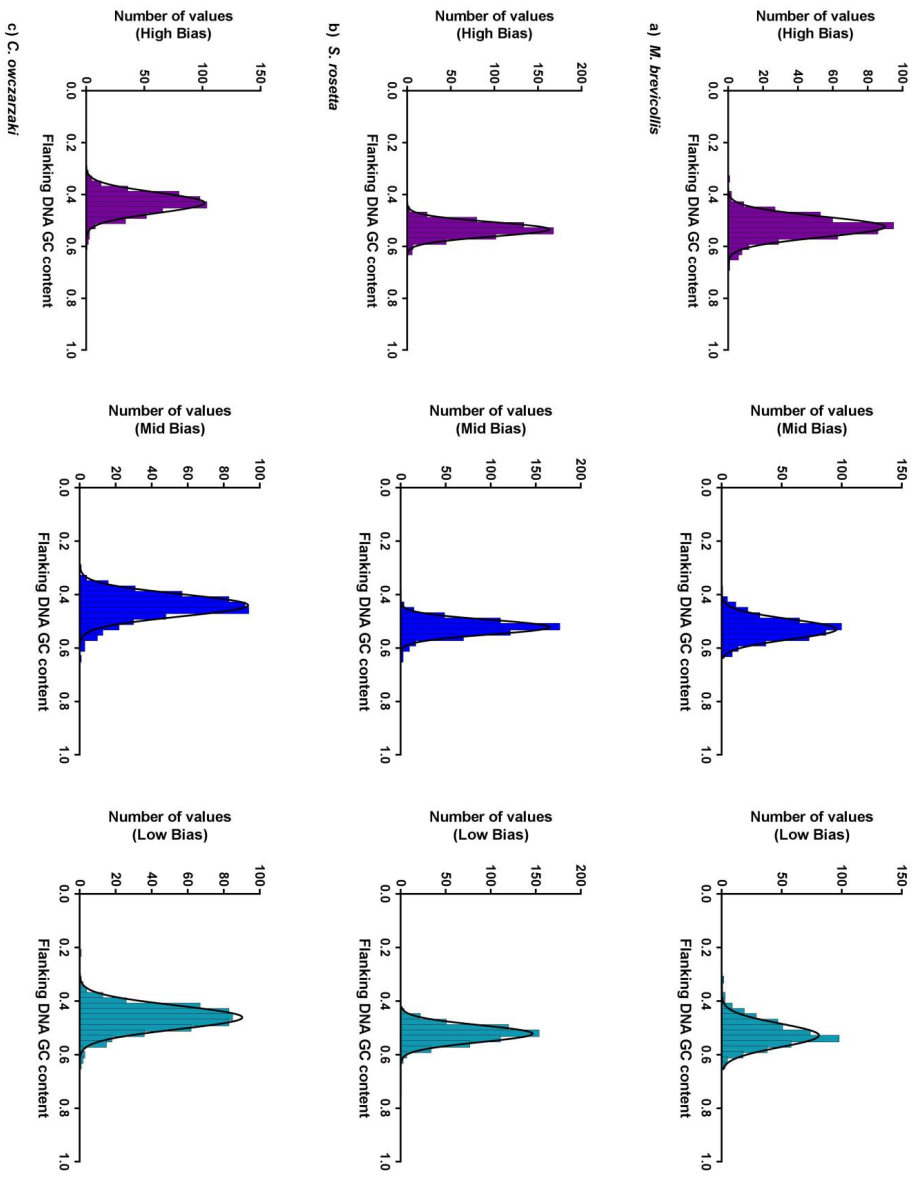


Figure C.2: **Normal distribution of non-coding flanking DNA for the three 5% bias categories**

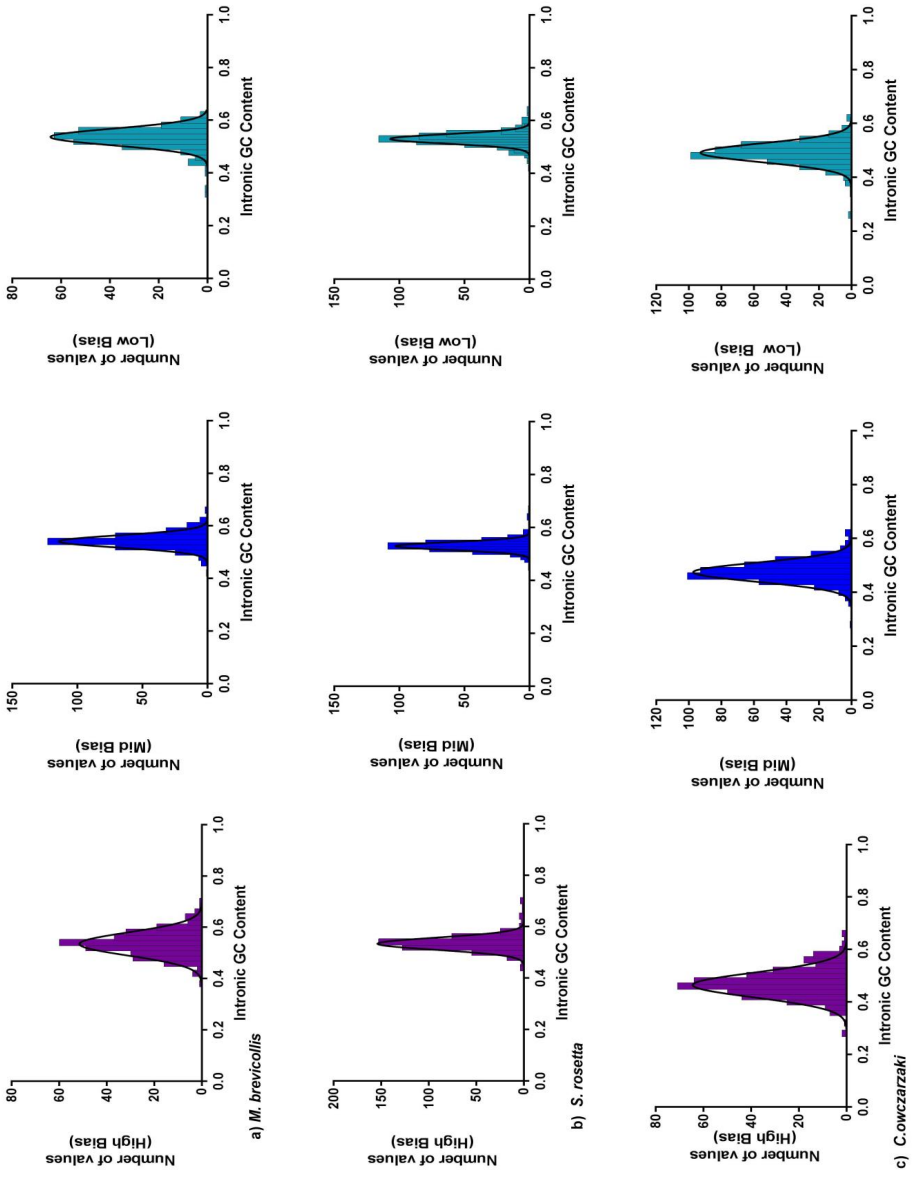


Figure C.3: **Normal distribution of non-coding intronic GC for the three 5% bias categories**

Appendix D

Chapter 5 Appendix

D.1 Codon usage statistics data for the transpoable elements of three holozoan species

D.1.1 Abundant codons for the TE families

Table D.1: **Preferred codons for each amino acid in the *S. rosetta* TE families.** Bold font denotes a favoured codon which complements the product of the major tRNA gene for the amino acid outlined in Chapter 4

Amino acid	*Srospv1*	*Srospv2*	*Srospv3*	*Srospv4*	*Srospv5*	*Srospv6*	*Srosgyp1*	*Srosgyp2*	*Sroscv1*	*Sroscv2*	*Sroscv3*	*Sroscv4*	*Sroscv5*	*SrosT1*	*SrosT2*	*SrosT3*	*SrosTig1*	*SrosTig2*	*SrosM*	*SrosH*
Phe	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**
Leu	**CUG**	**CUG**	**CUG**	**CUG**	**CUG**	**CUG**	**CUG**	**CUG**	**CUG**	**CUC**	**CUC**	**CUC**	**CUG**	**CUG**	**CUC**	**CUG**	**CUG**	**CUG**	**CUG**	**CUG**
Ile	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	AUU, **AUC**	**AUC**	**AUC**	**AUC**
Val	GUG	**GUC**	**GUC**	GUG	GUG	GUG	GUG	GUG	GUG	**GUC**	GUU	GUG	GUG	GUG	GUG	GUG	GUG	GUG	GUG	GUG
Ser	**AGC**	**AGC**	**AGC**	**AGC**	**AGC**	**AGC**	**UCC**	**AGC**	**UCC**	UCA	**UCC**	**UCC**	**UCG**	**UCC**, UCA, **AGC**	**UCG**	**UCC**	**UCC**	UCA	UCA	**UCG**
Pro	CCA	CCA	CCA	CCA	CCA	CCA	**CCC**	**CCG**	CCA	CCA	CCA	**CCC**	CCA, **CCG**	**CCC**	CCA	**CCU**	CCA	CCA	**CCG**	**CCC**
Thr	**ACG**	**ACG**	**ACG**	ACA	ACA	ACA	ACC	ACC	**ACG**	ACC	ACA, **ACG**	ACC	ACA	**ACG**	ACC	ACC, ACA	ACC	ACA	ACC	ACC
Ala	GCA	**GCC**	**GCC**	GCA	GCA	GCA	**GCC**	**GCC**	**GCC**	GCA	**GCC**	**GCC**	**GCC**	**GCA**	**GCC**	**GCU**	GCA	GCA	GCA	GCG
Tyr	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**
His	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**
Gln	**CAG**	**CAG**	**CAG**	**CAG**	**CAG**	**CAA**	**CAG**	**CAG**	**CAG**	**CAG**	**CAG**	**CAG**	**CAG**	**CAG**	**CAG**	**CAA**	**CAG**	**CAG**	**CAG**	**CAG**
Asn	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**
Lys	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**
Asp	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**
Glu	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAA, GAG**	**GAG**
Cys	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGU, UGC**	**UGC**	**UGC**	**UGC**	**UGC**
Arg	**CGC**	**CGC**	**CGC**	AGA	**CGC**	AGG	**CGC**	**CGC**	**CGC**	**CGC**	**CGC**	**CGC**	**CGC**	**CGC**	**CGC**	**CGU**	**CGC**	**CGC**	**CGG**	**CGC**
Gly	**GGC**	**GGC**	**GGC**	GGA	**GGC**, GGA	GGA	**GGC**	**GGC**	**GGC**	GGA	**GGC**	**GGC**	GGA	**GGC**	**GGC**	GGG	**GGC**	GGA	**GGC**	**GGC**

Table D.2: **Preferred codons for each amino acid in the *C. owczarzaki* TE families.** Bold font denotes a favoured codon which complements the product of the major tRNA gene for the amino acid outlined in Chapter 4

Amino acid	*Cocv1*	*Cocv2*	*Cocv3*	*Cocv4*	*Cocv5*	*CoL1*	*CoL2*	*CoL3*	*CoL4*	*Com1*	*Com2*	*Cop1*	*Cop2*	*Cop3*	*Cop4*	*Cop5*	*CoTc1*	*CoTc2*	*Cobalt1*	*Cobalt2*	*Cobalt3*	*CoCACTA1*	*CoCACTA2*
Phe	**UUC**	UUU	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	UUU, **UUC**	**UUC**	**UUC**	UUU	UUU	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**
Leu	**CUG**	**CUG**	**CUG**	**CUG**	**CUC**	**CUG**	**CUC**	**CUC**	**CUC**	**CUG**	**CUC**	**CUG**	**CUC**	**CUG**	**CUC**	**CUC**	**CUC**	**CUC**	**CUC**	**CUG**	**CUC, CUG**	**CUC**	**CUG**
Ile	**AUU, AUC**	**AUU**	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUU, AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUC**	**AUU**	**AUC**
Val	GUG	GUG	GUG	**GUC**	**GUC**	GUG	**GUC**	**GUC**	**GUC**	**GUC**, GUG	GUU	**GUC**	**GUC**	**GUC, GUG**	**GUC**	GUG	**GUC**	**GUC**	GUG	**GUC**	**GUC**	**GUC**	**GUC**
Ser	**UCG**	**UCG**	**UCG**	**UCG**	**UCC**	**AGC**	**UCC**	**UCC**	**UCC**	**AGC**	**AGC**	**UCG, AGC**	**UCG**	**UCG**	**UCG**	**UCG**	**AGC**	**UCG**	UCU, **UCC**	**UCC**	**AGC**	**AGC**	**UCC, UCA**
Pro	CCG	CCU	CCG	CCU	CCG	CCG	**CCC**	**CCC**	**CCC**	CCG	CCA	**CCC**	CCG	CCG	**CCC**, CCG	CCG	**CCC**	CCG	**CCC**	CCG	CCU, CCG	CCA	CCU, CCA
Thr	ACG	ACG	ACG	**ACC**	**ACC**	**ACC**	**ACC**	**ACC**	**ACC**	ACG	**ACC**	**ACC**	**ACC**	ACA	ACU	ACG	**ACC**	**ACC**	ACG	ACA	ACG	ACG	**ACC**
Ala	GCG	GCG	GCG	**GCC**	**GCC**	**GCC**	**GCC**	**GCC**	**GCC**	GCG	GCU	**GCC**	**GCC**	GCG	**GCC**	GCG	**GCC**	**GCC**	GCU	GCA	**GCC**	GCA	**GCC**
Tyr	**UAC**	UAU, **UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**
His	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	CAU	**CAC**	**CAC**	CAU	**CAC**
Gln	**CAG**	**CAG**	**CAG**	**CAG**	**CAG**	**CAG**	CAA	CAA	CAA	**CAG**	CAA	**CAG**	**CAG**	**CAG**	CAA	**CAG**	**CAG**	**CAG**	CAA	**CAG**	CAA	CAA	CAA
Asn	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	AAU	**AAC**	**AAC**	**AAC**	AAU	AAU	**AAC**	**AAC**	**AAC**
Lys	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**	**AAG**
Asp	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAU**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	**GAU**	**GAC**
Glu	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	GAA	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	**GAG**	GAA	GAA, **GAG**
Cys	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	**UGC**	UGU	UGU, **UGC**	**UGC**	**UGC**	**UGC**	**UGC**
Arg	CGG	**CGC**	**CGU**, CGG	**CGC**	**CGC**	**CGC**	**CGC**	**CGC**	**CGC**	**CGC**	**CGC**, CGA	**CGC**	**CGC**	**CGC**	**CGC**	**CGC**	**CGC**	**CGC**	**CGC**	**CGC**	**CGC**	**CGA**	**CGU**
Gly	**GGC**	**GGC**	**GGC**	**GGC**	**GGC**	**GGC**	**GGC**	**GGC**	**GGC**	**GGC**	**GGC**	**GGC**	**GGC**	**GGC**	**GGC**	**GGC**	GGU, **GGC**	**GGC**	**GGC**	**GGC**	**GGC**	**GGA**	**GGC**

Table D.3: **Preferred codons for each amino acid in the *M. brevicollis* TE families.** Bold font denotes a favoured codon which complements the product of the major tRNA gene for the amino acid outlined in Chapter 4

Amino acid	*Mbcv1*	*Mbpv1*	*Mbpv2*
Phe	UUU	UUU	UUU
Leu	**CUC**	**CUG**	**CUC**
Ile	**AUC**	**AUC**	**AUC**
Val	GUG	GUG	**GUC**
Ser	**UCG, AGC**	**UCG, AGC**	**UCU**
Pro	CCU, CCA	CCG	**CCC**
Thr	**ACC**	**ACC**	**ACC**
Ala	**GCC**	**GCC**	**GCC**
Tyr	**UAC**	**UAC**	**UAC**
His	**CAC**	**CAU**	**CAC**
Gln	**CAG**	**CAG**	**CAG**
Asn	**AAC**	**AAC**	**AAC**
Lys	**AAG**	**AAG**	**AAG**
Asp	**GAC**	**GAC**, GAU	**GAC**
Glu	**GAG**	**GAG**	**GAG**
Cys	**UGC**	**UGC**	**UGC**
Arg	**CGC**	**CGC**	**CGC**
Gly	**GGC**	**GGC**	**GGC**

D.1.2 Codon usage statistics

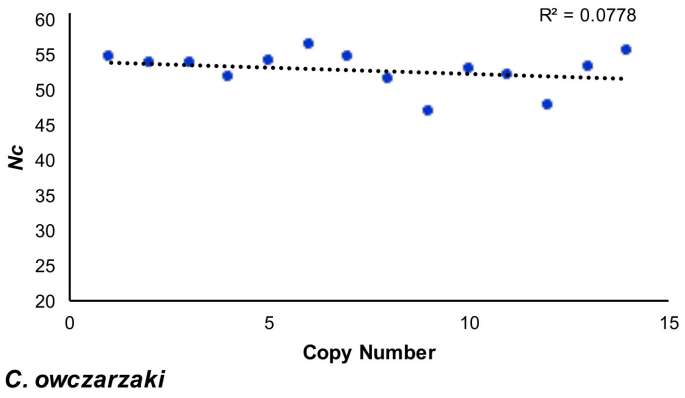


Figure D.1: **Relationship between copy number of TE families and *Nc* in *C. owczarzaki*** Copy number of each TE family was plotted against *Nc* for the elements of *C. owczarzaki*

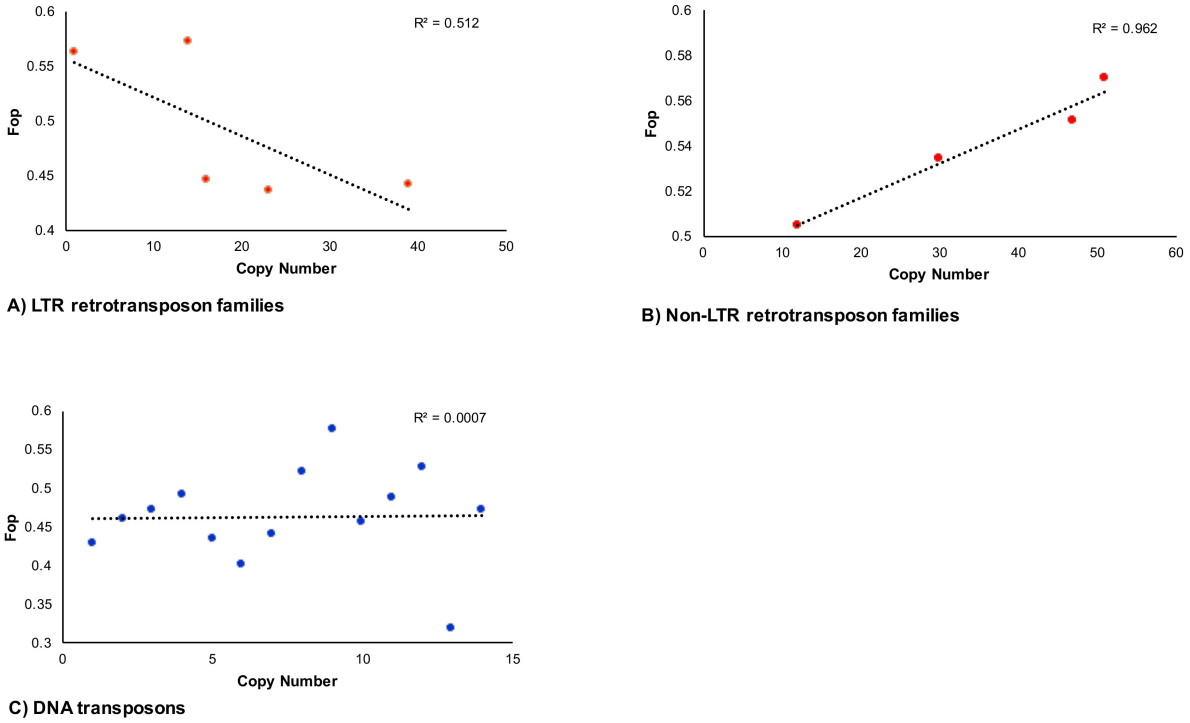


Figure D.2: **Relationship between copy number of TE families and frequency of optimal codons (Fop) in *C. owczarzaki*** Copy number of each TE family was plotted against Fop for the elements of *C. owczarzaki*

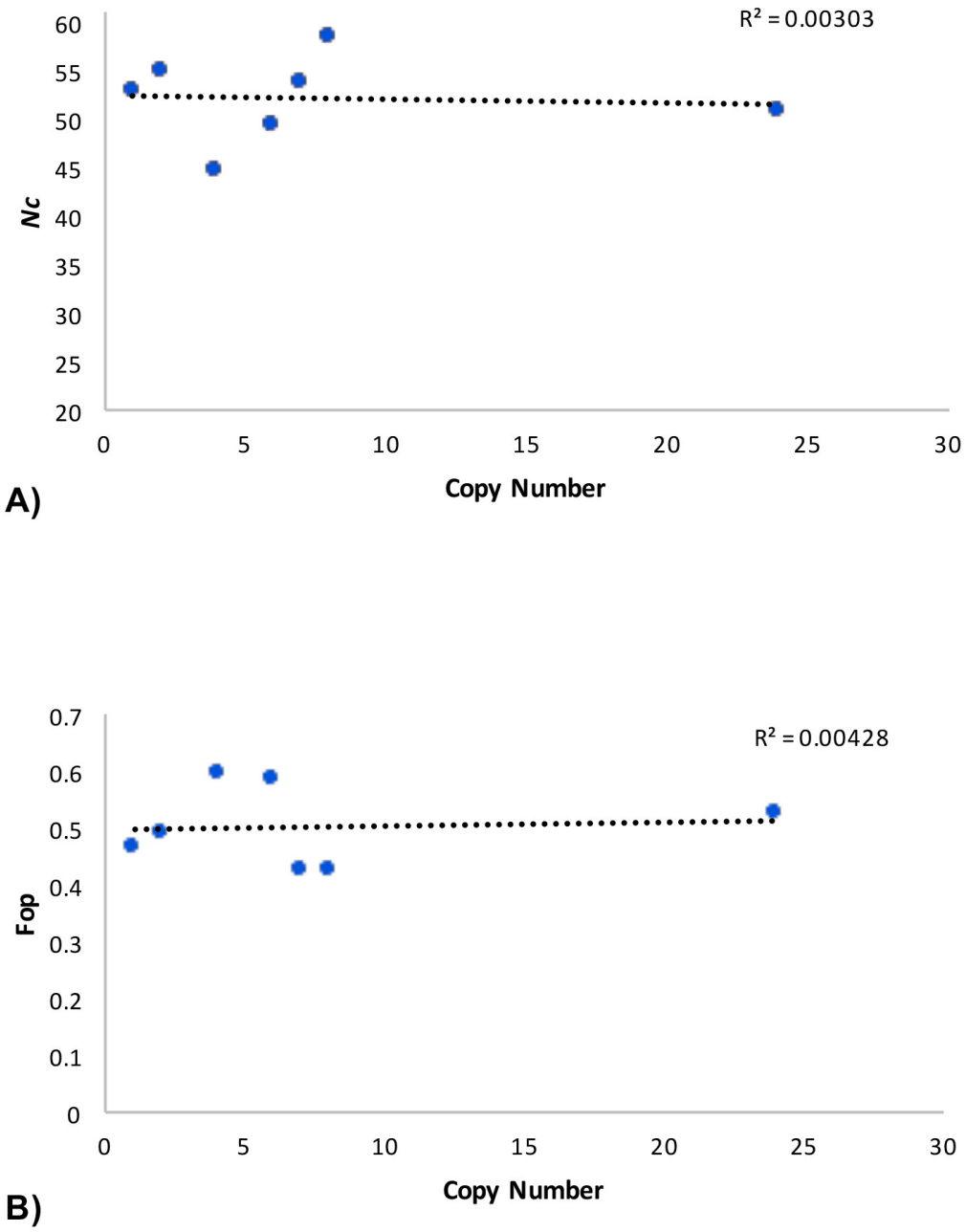
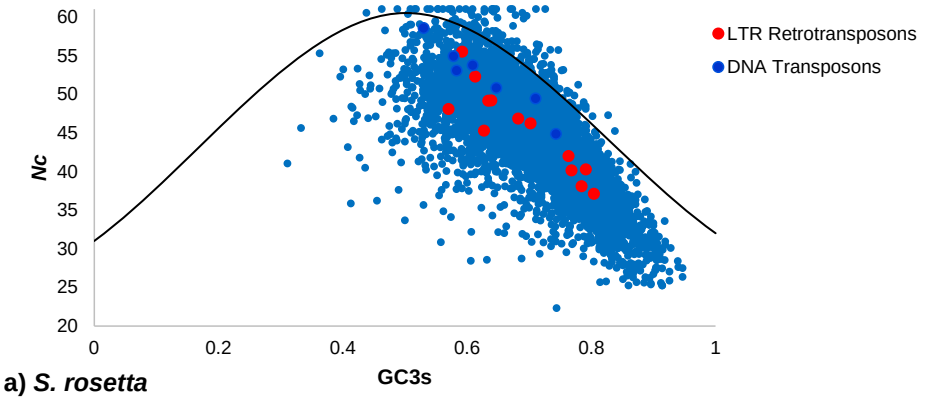
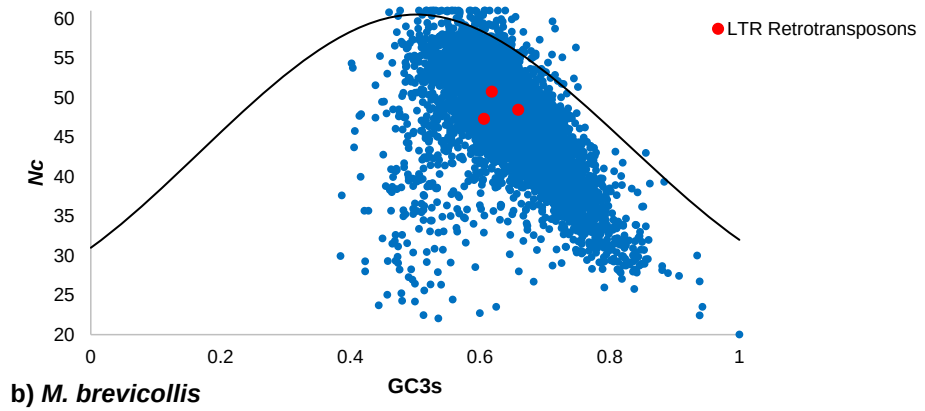


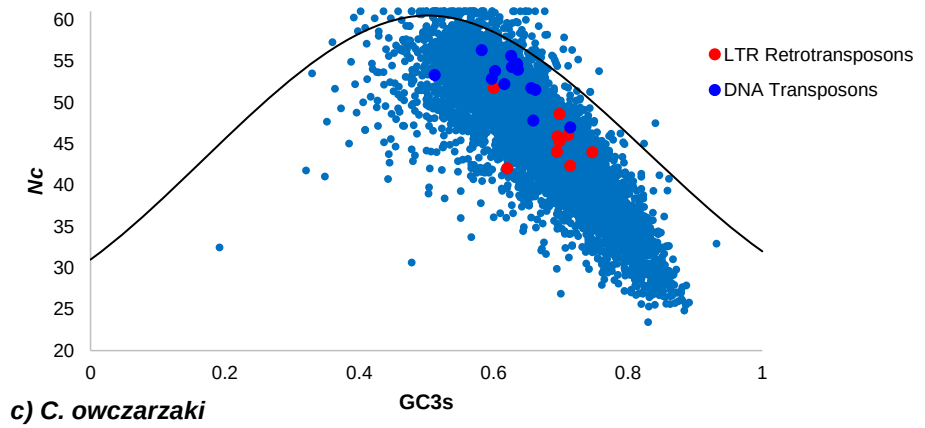
Figure D.3: **Relationship between copy number of DNA transposon families and *Nc* in *S. rosetta* and *C. owczarzaki*** Copy number of each TE family was plotted against *Nc* for the elements of *S. rosetta* (A) and elements of *C. owczarzaki* (B).



a) *S. rosetta*



b) *M. brevicollis*



c) C. owczarzaki

Figure D.4: ***Nc* plot against GC3s for the genes of *S. rosetta*, *M. brevicollis* and *C. owczarzaki*, including TE families.** *Nc* values were plotted against GC3s for the three holozoan species and TE families, which are highlighted in red and blue. The modified equation, *Nc*=2+S+29/[S2+(1-S)2], from Wright (1990), with S=GC3s, was used to create the parabolic curve on each *Nc* plot (Southworth et al., 2018).

Table D.4: **Non-coding GC values for TE families identified in the *S. rosetta* genome.**

Family	**Non-coding GC**
S. rosetta	
LTR retrotransposons	
Sroscv1	0.504
Sroscv2	0.46
Sroscv3	0.499
Sroscv4	0.56
Sroscv5	0.493
Srosgyp1	0.497
Srosgyp2	0.539
Srospv1	0.491
Srospv2	0.493
Srospv3	0.498
Srospv4	0.487
Srospv5	0.481
Srospv6	0.519
	Mean=0.502±0.026
DNA transposons	
SrosT1	0.485
SrosT2	0.524
SrosT3	0.515
SrosTig1	0.482
SrosTig2	0.527
SrosM	0.54
SrosH	0.551
	Mean=0.518±0.026

Table D.5: **Non-coding GC values for TE families identified in the *M. brevicollis* genome.**

Family	**Non-coding GC**
M. brevicollis	
LTR retrotransposons	
Mbcv	0.556
Mbpv1	0.517
Mbpv2	0.465
	Mean=0.513±0.046

Table D.6: **Non-coding GC values for TE families identified in the *C. owczarzaki* genome.**

Family	**Non-coding GC**
C. owczarzaki	
LTR retrotransposons	
Cocv1	0.454
Cocv2	0.442
Cocv3	0.53
Cocv4	0.579
Cocv5	0.466
	Mean=0.494±0.058
Non-LTR retrotransposons	
CoL1	0.622
CoL2	0.473
CoL3	0.441
CoL4	0.458
	Mean=0.499±0.083
DNA transposons	
Com1	0.528
Com2	0.537
Cop1	0.499
Cop2	0.484
Cop3	0.524
Cop4	0.533
Cop5	0.55
CoTc1	0.466
CoTc2	0.504
Cobalt1	0.435
Cobalt2	0.44
Cobalt3	0.465
CoCACTA1	0.462
CoCACTA2	0.489
	Mean=0.494±0.037

Appendix E

Bioinformatics parameters

E.1 SMALT

SMALT version 0.7.6 was employed for *S.rosetta* and *C. owczarzaki* transcriptome reads to review host gene and TE expression (Ponstingl, 2014). Default parameters were used unless specified. A text file of *S. rosetta* and *C. owczarzaki* TE families in FASTA format was employed as the SMALT index with the transcriptome reads ran to map onto the specified index. Parameters were accessed from the SMALT User Manual (Ponstingl, 2014). For the index, the word length of hashed words was set to 20 (-k 20), and sampling step size was set to 1, so all words were hashed, with the path for the FASTA file inserted, as seen in literature (Ponstingl, 2014) . The smalt map was set with SAM file format (-f sam). Individual output files were concatenated post SMALT analysis. Command: Smalt index -k 20 -s 1 Srosgenesk20s1 Desktop/Srosetta_families.fasta Smalt map -f sam -Srosgenesk20s1 Desktop/Srosetta_transcriptome.cat

E.2 Phylogeny construction

RAxML HPC2 8.2.6. on XSEDE (Stamatakis, 2014) was employed for protein phylogenies via server based, Cipres Science Gateway (Miller et al., 2010). All parameters unspecified remained on the default setting automatically employed by the tool. The ML phylogenies were employed with a PROCAT model, and 100 bootstrap interactions. The ML amino acid substitution model used for each family was determined from the output of the mixed model analysis from MrBayes. Bayesian inference protein phylogenies were produced using Mr Bayes 3.2.6 on XSEDE (Ronquist and Huelsenbeck, 2003) via Cipres Science Gateway. A mixed amino acid model was used, and employed 5000000 generations with a burnin value of 10000. 0.25 burnin fraction was employed for the samples. Nucleotide phylogenies were ran with the same parameters, except the nucleic

acid database was employed. For LTR phylogenies, maximum likelihood analyses were performed using raxmlGUI 1.5 beta (Silvestro and Michalak, 2011). The phylogenies were ran using “ML +thorough bootstrap’ with 100 runs and bootstrapped using 1000 replicates. The GTRCAT model was employed.

E.3 Basic Local Alignment Search Tool (BLAST)

For protein phylogenies, similarity searches were ran with the input of TE family query sequences against the BLAST protein and translated nucleotide databases (blastp and tBLASTn) (Altschul et al., 1990). For tBLASTn analysis the following taxa were employed for whole genome sequence analysis. Metazoa (Deuterostomia, Platyhelminthes, Protosomia, Cnidaria, Ctenophora, Placozoa, Porifera); Fungi (Blastocladiomycota, Chytridiomycota, Cryptomycota, Ascomycota, Basidimycota Entomophthoromycota, Glomeromycota, Microsporida, Neocallimastigomycota); Protist (Alveolata; Amoebozoa, Apusozoa, Rhodophyta, Stramenopiles); Plant (Chlorophyta, Mesostigmata) (Sayers et al., 2009).

E.4 RepeatMasker

RepeatMasker version open-4.0.6 (*RepeatMasker*, 1996) parameters remained at default unless specified, run with rmblastn version 2.2.27. Library employment was defined as Repbase library, downloaded from the Genetic Information Research Institute (*GIRI*, 2016), or a custom library, created specifically for Saccharomyceteceae and chonanoflagellate analysis. Parameters were selected from literature (Tempel, 2012). With the employment of Repbase, the following parameters were set for transposable element detection. Species (-fungi); No bacterial element insertion (-no_is); No masking of simple repeats (-nolow); alignment file between query and hit created in output (-a), in query orientation (-inv). When a custom library was employed for analysis, species specification was removed from the command, and path for custom library location specified (-lib). Additional Saccharomyceteceae species were ran via RepeatMasker to review TE content (*S. cerevisiae* S288c; *Candida glabarata* CDS138; *Z. rouxii* CBS732; *K. thermotolerans* CBS6340; *L. kluyveri* CBS3082; *K. lactis* CBS2359; E. gossypii ATCC10895; *D. hansenii* CBS767 and *Y. lipolytica* CBS7504) (Genolevures et al., 2009).

Appendix F

Phylogenies

F.1 Protein and nucleotide phylogenies

F.1.1 Protein phylogenies

F.1.2 Nucleotide phylogenies

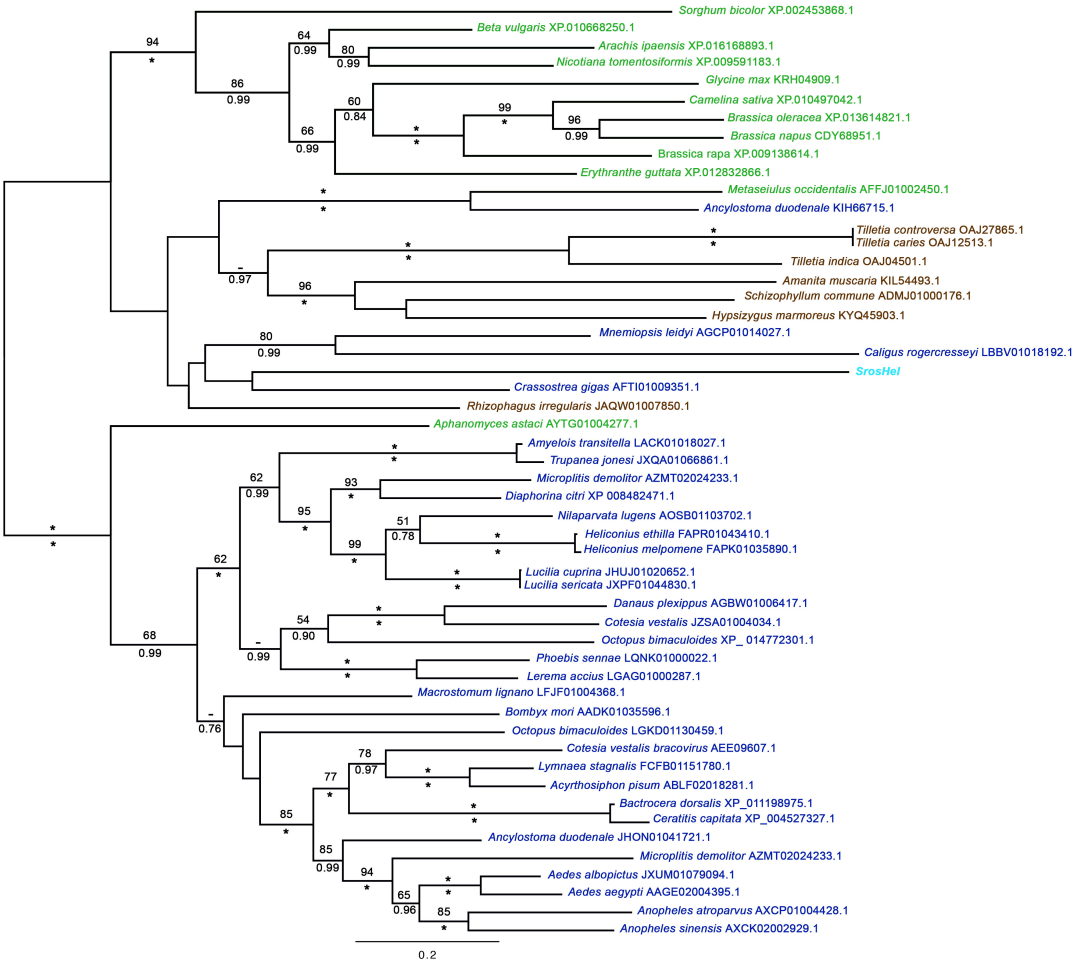


Figure F.1: **Maximum likelihood phylogeny of Helitron amino acid sequences across eukaryotic supergroups**. The phylogeny was constructed by an alignment of 466 amino acid constructs with the employment of raxmlGUI using the PROTCAT model and estimated amino acid frequencies with WAG substitution matrix. ML and biPP values are labelled above and below corresponding branches. Maximum support is annotated by '*' (100ML/1.0 biPP) and low support (<50 ML/ <0.70 biPP) are annotated '-'. The scale bar signifies the number of amino acid substitution per amino acid site. Metazoan proteins are in dark blue, choanoflagellate proteins are in light blue, fungal sequences are written in brown and plants are in green.

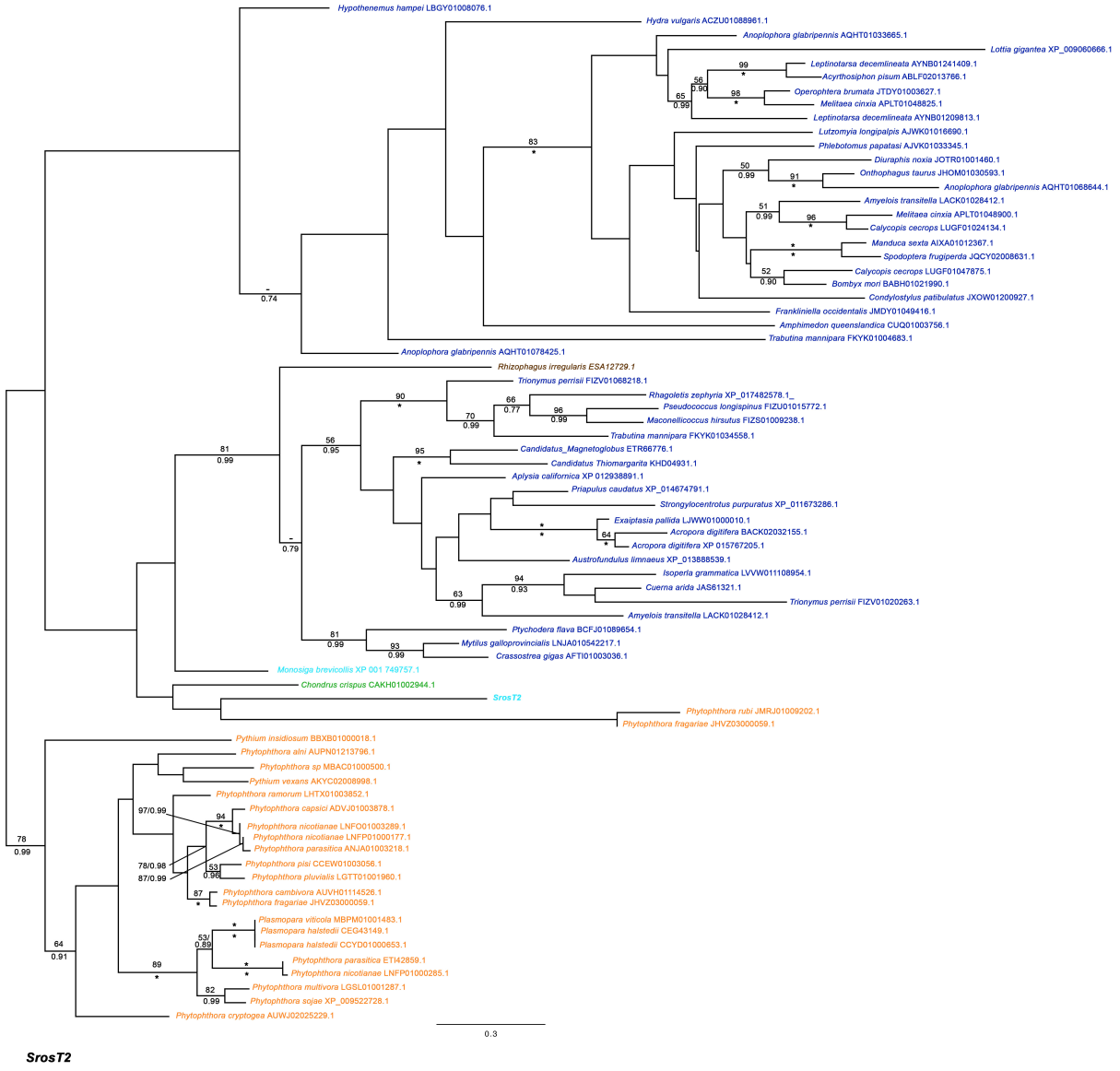
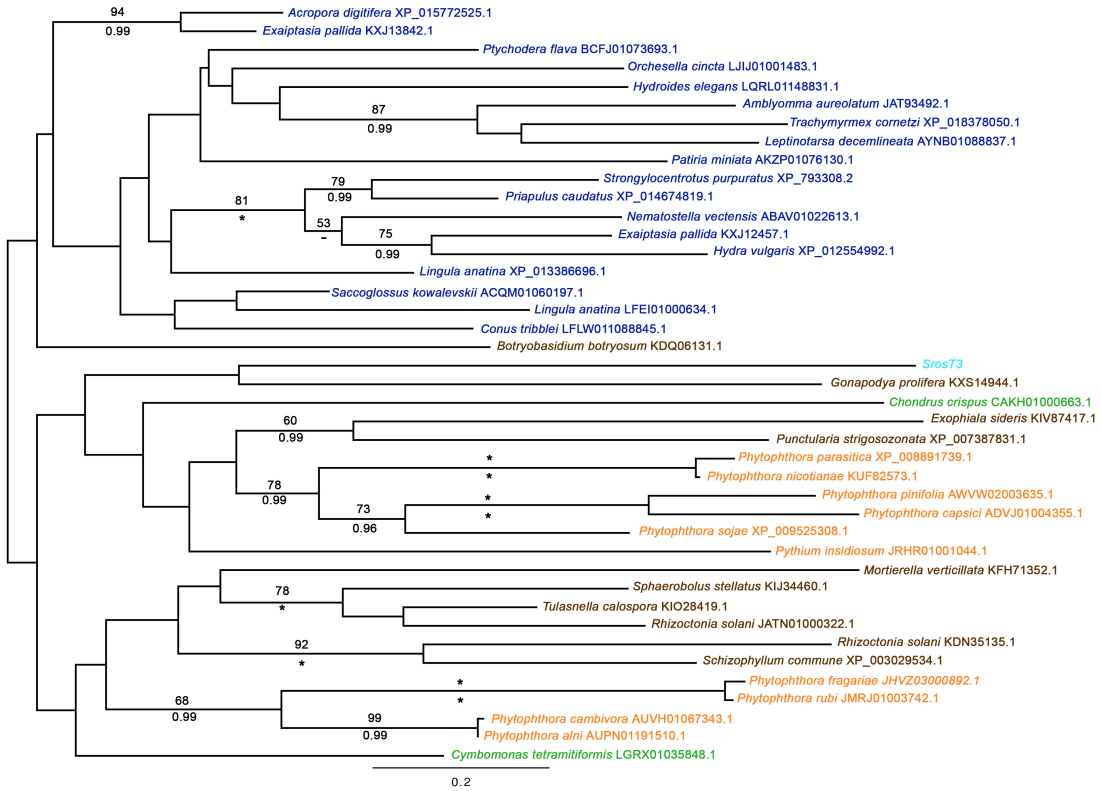


Figure F.2: **Maximum likelihood phylogeny of *Transposon-2* amino acid sequences across eukaryotic supergroups**. The phylogeny was constructed by an alignment of 127 amino acid constructs with the employment of raxmlGUI using the PROTCAT model and estimated amino acid frequencies with BLOSUM substitution matrix. Format is stated in F.1



SrosT3

Figure F.3: **Maximum likelihood phylogeny of *Transposon-3* amino acid sequences across eukaryotic supergroups**. The phylogeny was constructed by an alignment of 143 amino acid constructs with the employment of raxmlGUI using the PROTCAT model and estimated amino acid frequencies with WAG substitution matrix. Format is stated in F.1

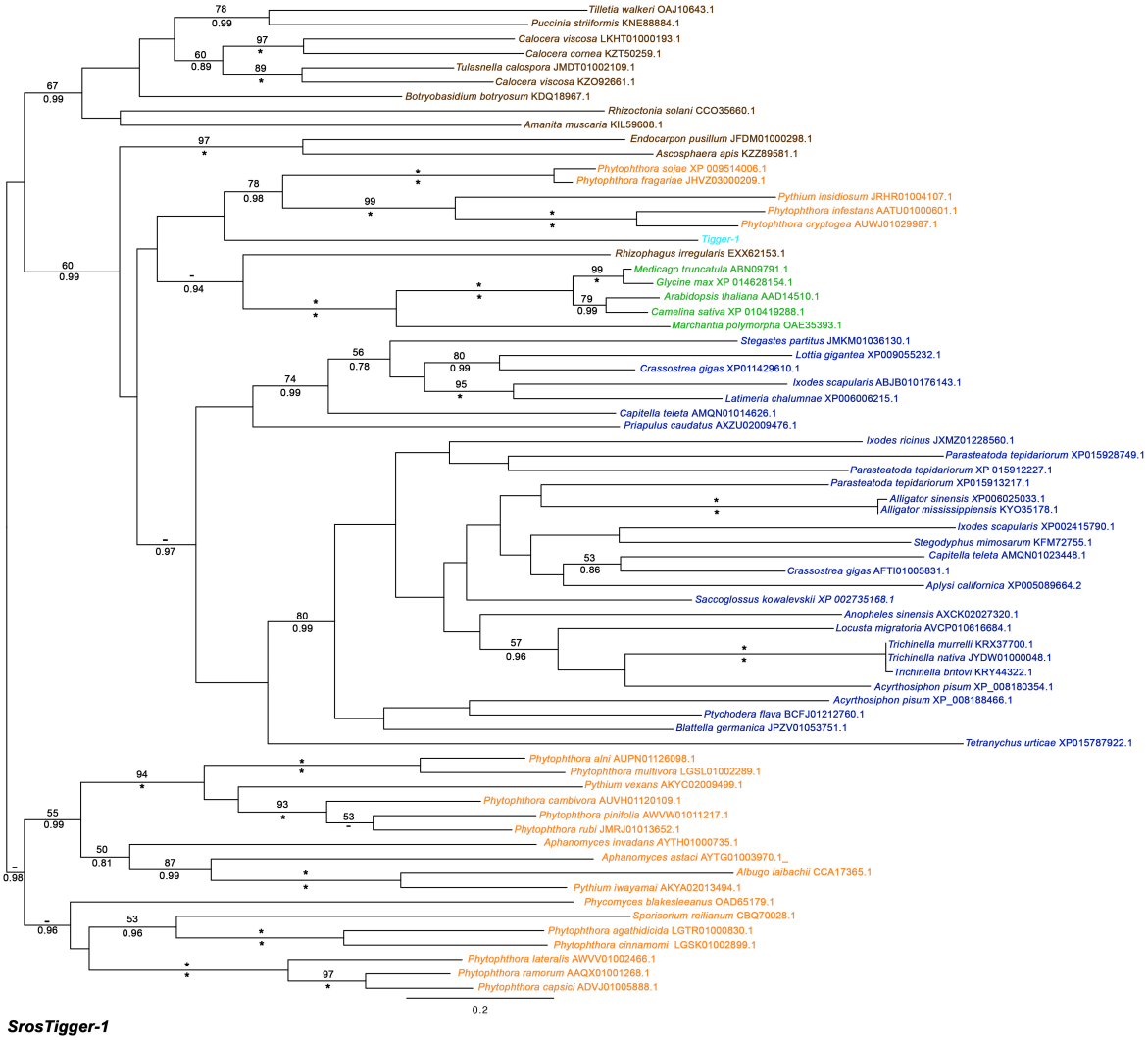


Figure F.4: **Maximum likelihood phylogeny of *Tigger-1* amino acid sequences across eukaryotic supergroups**. The phylogeny was constructed by an alignment of 288 amino acid constructs with the employment of raxmlGUI using the PROTCAT model and estimated amino acid frequencies with WAG substitution matrix. Format is stated in F.1

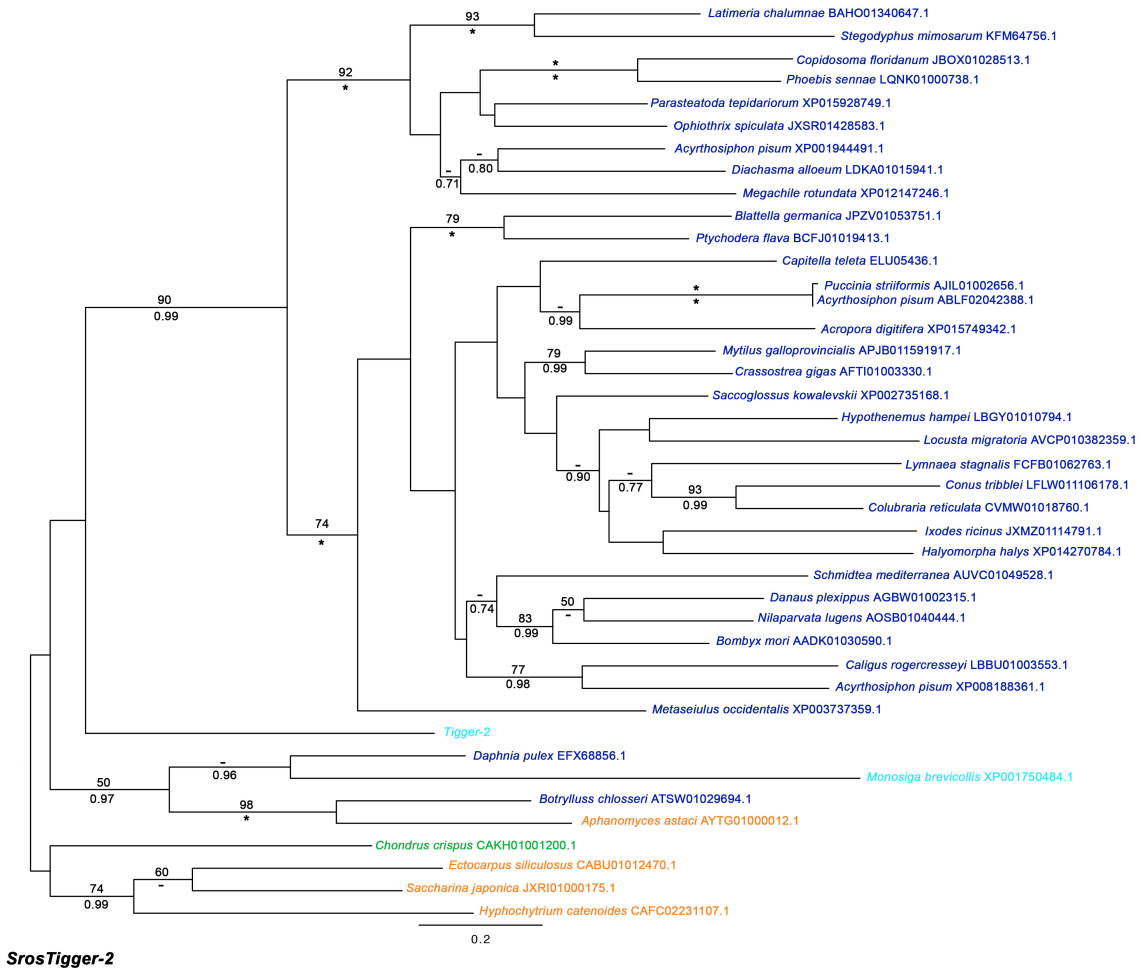


Figure F.5: **Maximum likelihood phylogeny of *Tigger-2* amino acid sequences across eukaryotic supergroups**. The phylogeny was constructed by an alignment of 217 amino acid constructs with the employment of raxmlGUI using the PROTCAT model and estimated amino acid frequencies with WAG substitution matrix. Format is stated in F.1

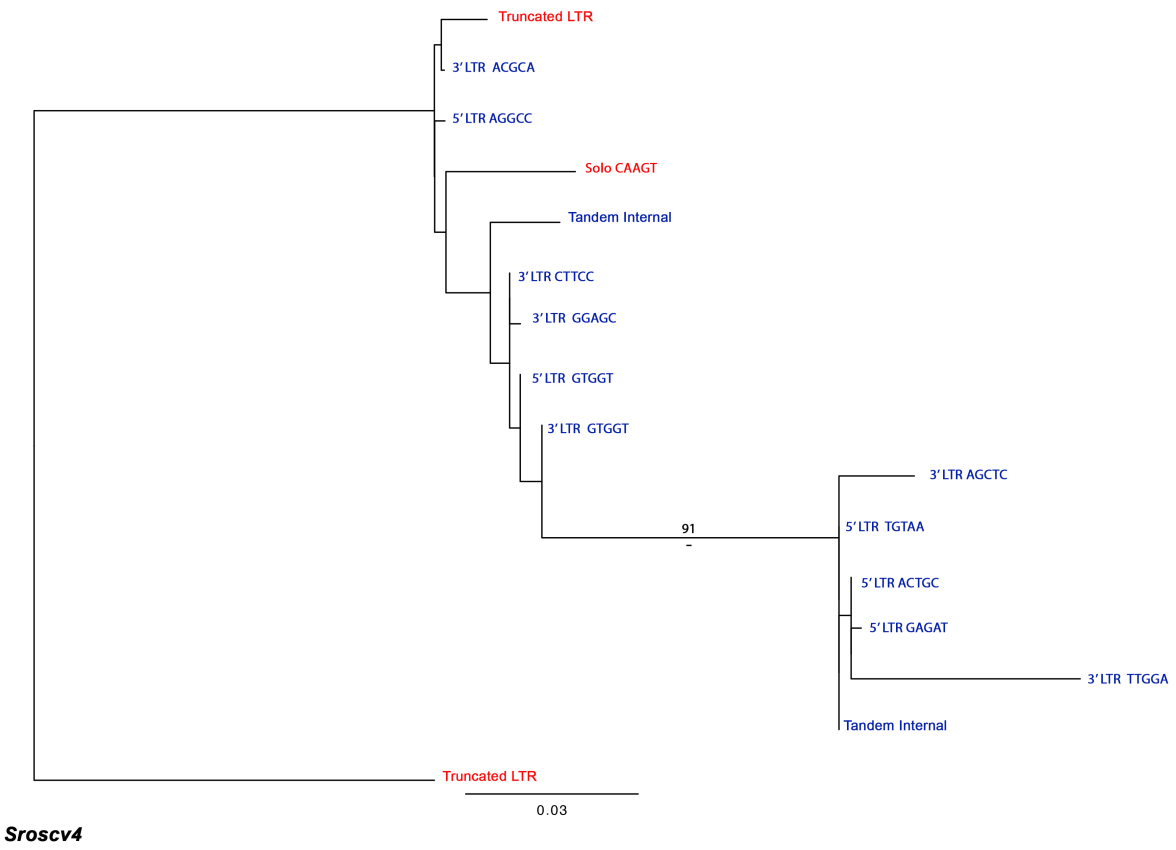
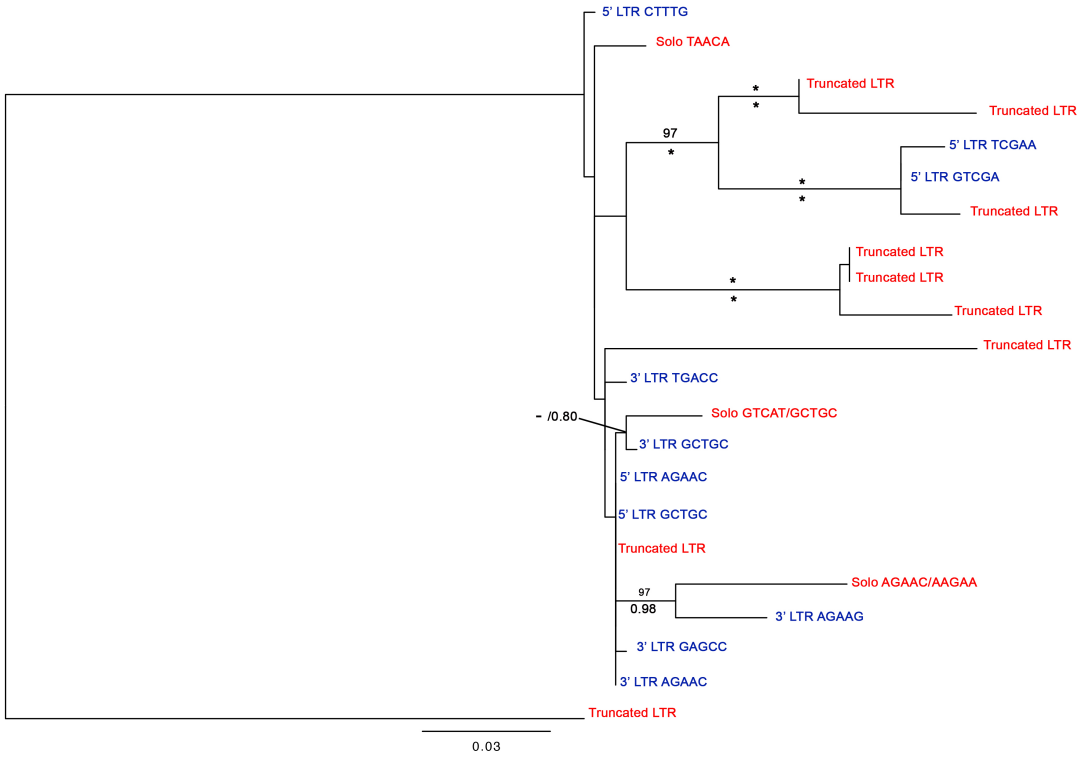
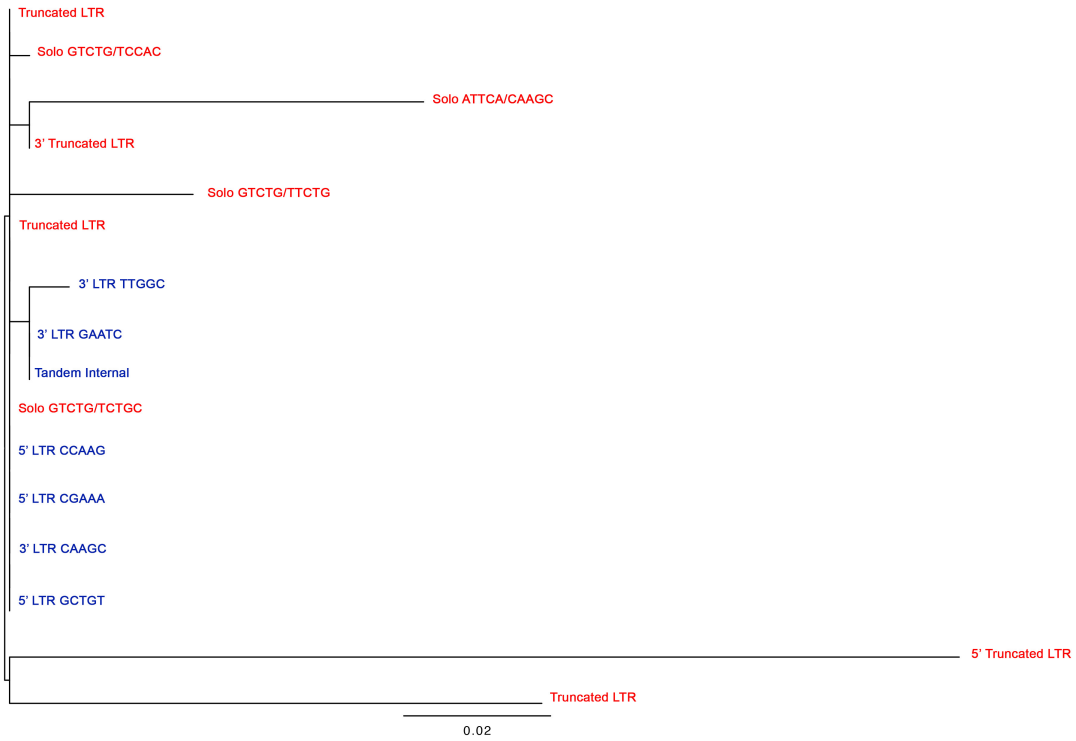


Figure F.6: **Maximum Likelihood phylogeny of individual element copies of *Sroscv4*.** The phylogeny was constructed by an alignment of 639 nucleotide constructs with the employment of raxmlGUI using the GTRCAT model. ML and biPP values are labelled above and below corresponding branches. Maximum support is annotated by '*' (100ML/1.0 biPP) and low support (<50 ML/ <0.70 biPP) are annotated '-'. The scale bar signifies the number of nucleotide substitutions per amino acid site. 5' and 3' LTR sequences are written in blue, with individual TSDs annotated on terminal branches respectively.



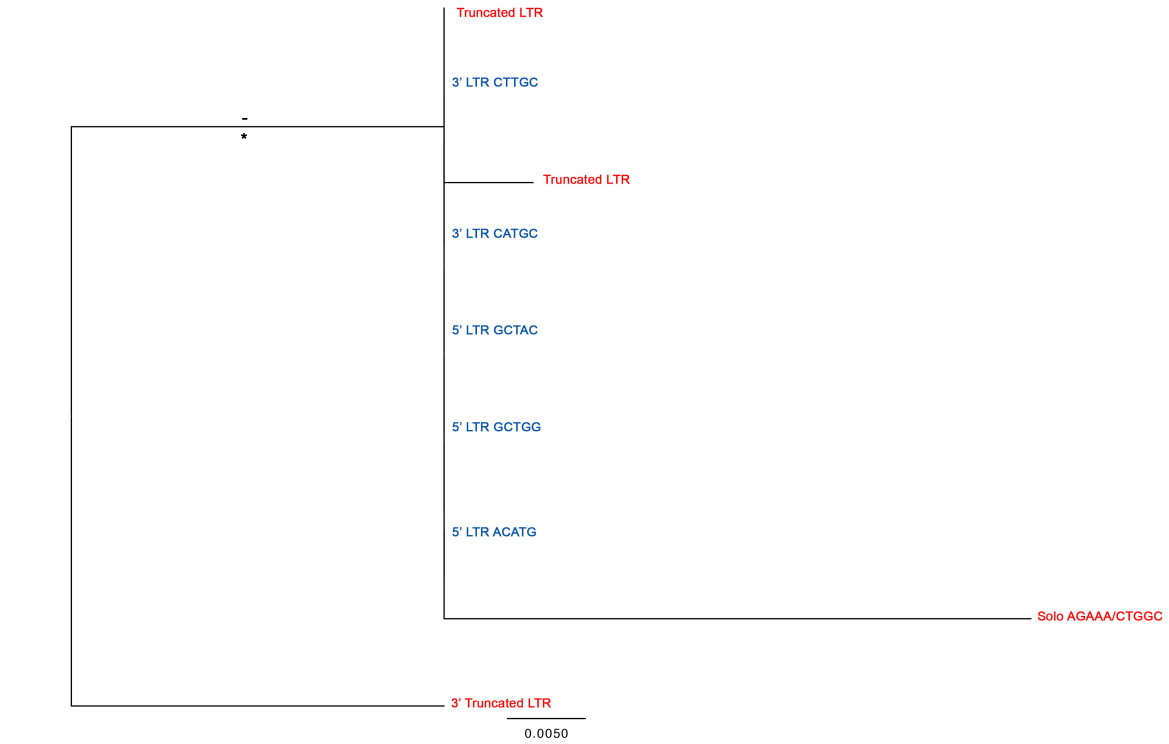
Sroscv5

Figure F.7: **Maximum Likelihood phylogeny of individual element copies of *Sroscv5*.** The phylogeny was constructed by an alignment of 440 nucleotide constructs with the employment of raxmlGUI using the GTRCAT model. Formatting is stated in F.6



Srosgyp1

Figure F.8: **Maximum Likelihood phylogeny of individual element copies of *Srosgyp1*.** The phylogeny was constructed by an alignment of 382 nucleotide constructs with the employment of raxmlGUI using the GTRCAT model. Formatting is stated in F.6



Srosgyp2

Figure F.9: **Maximum Likelihood phylogeny of individual element copies of *Srosgyp2*.** The phylogeny was constructed by an alignment of 391 nucleotide constructs with the employment of raxmlGUI using the GTRCAT model. Formatting is stated in F.6

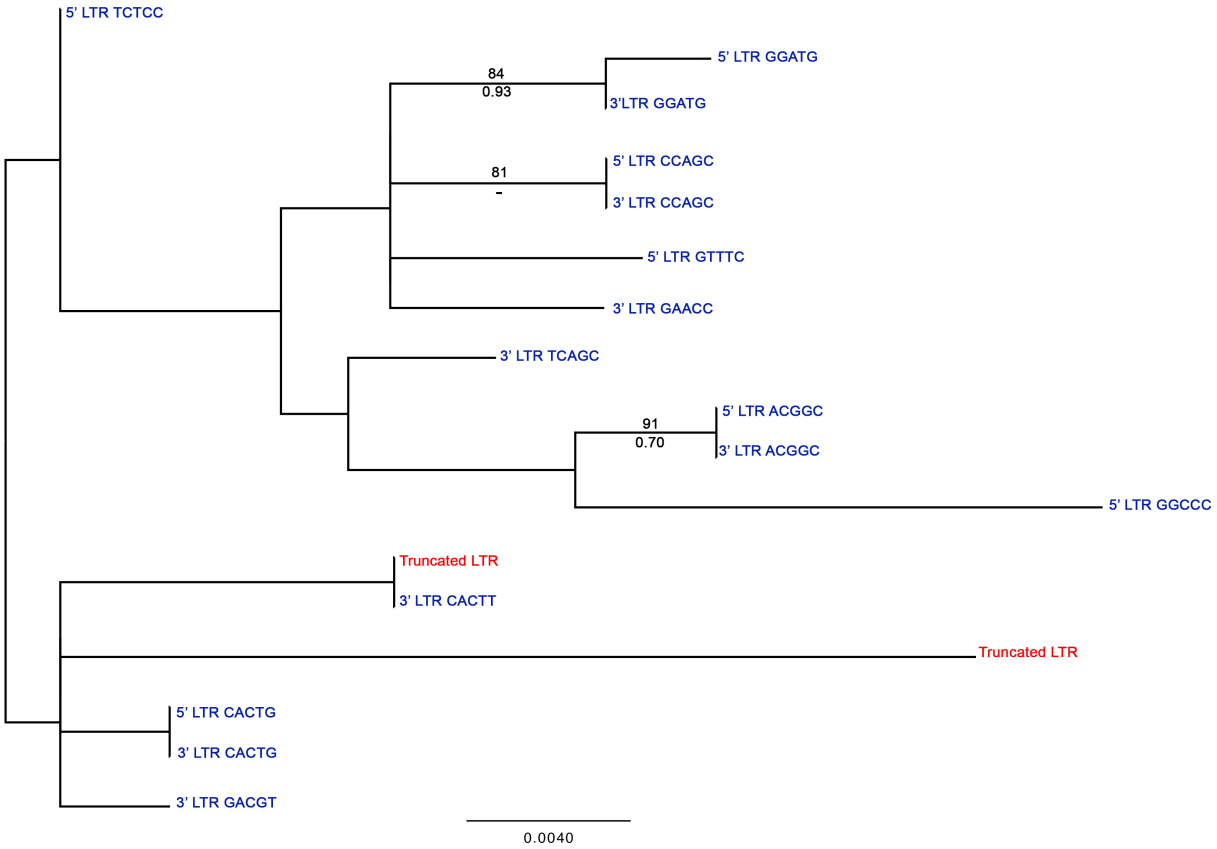
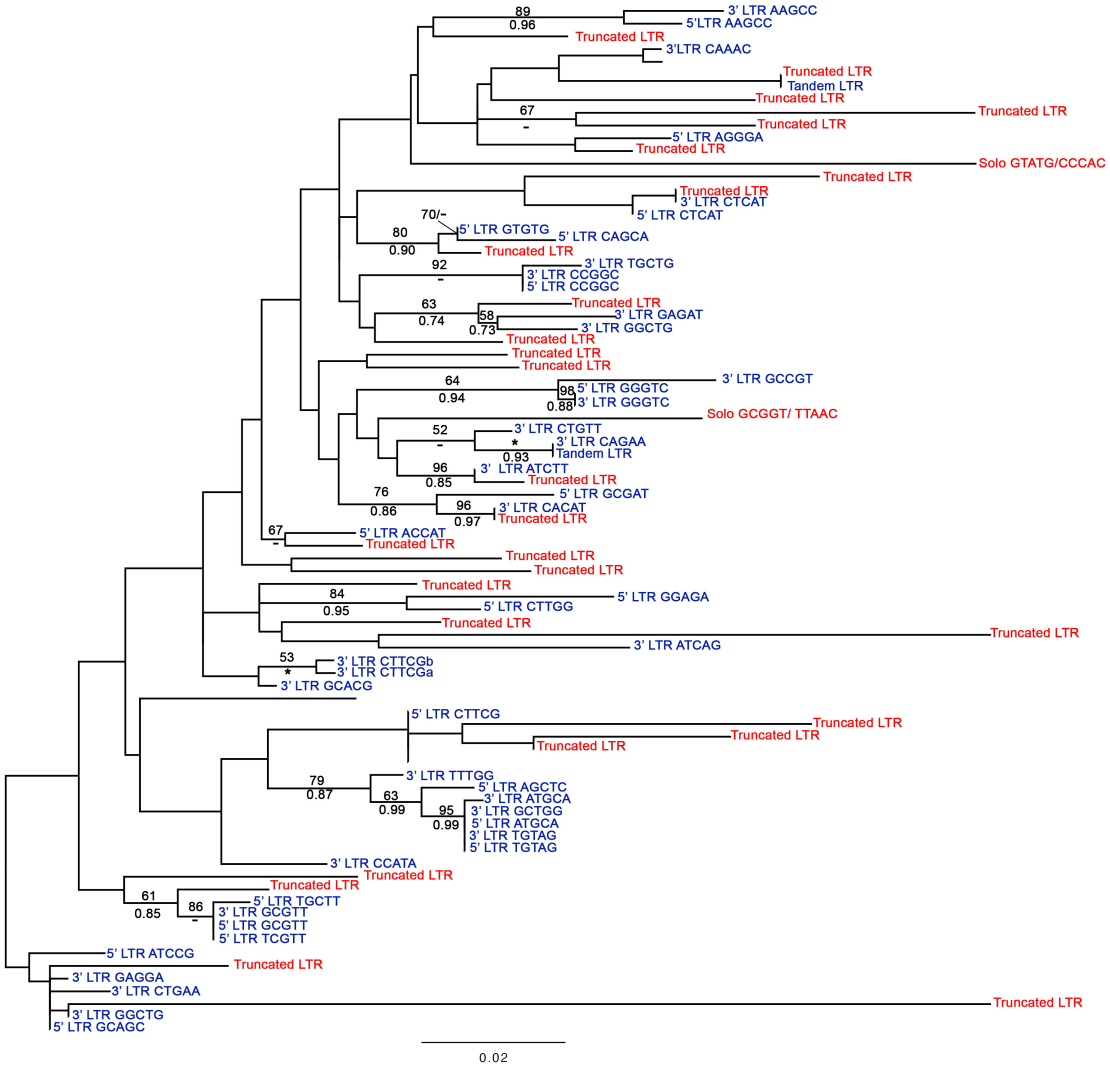
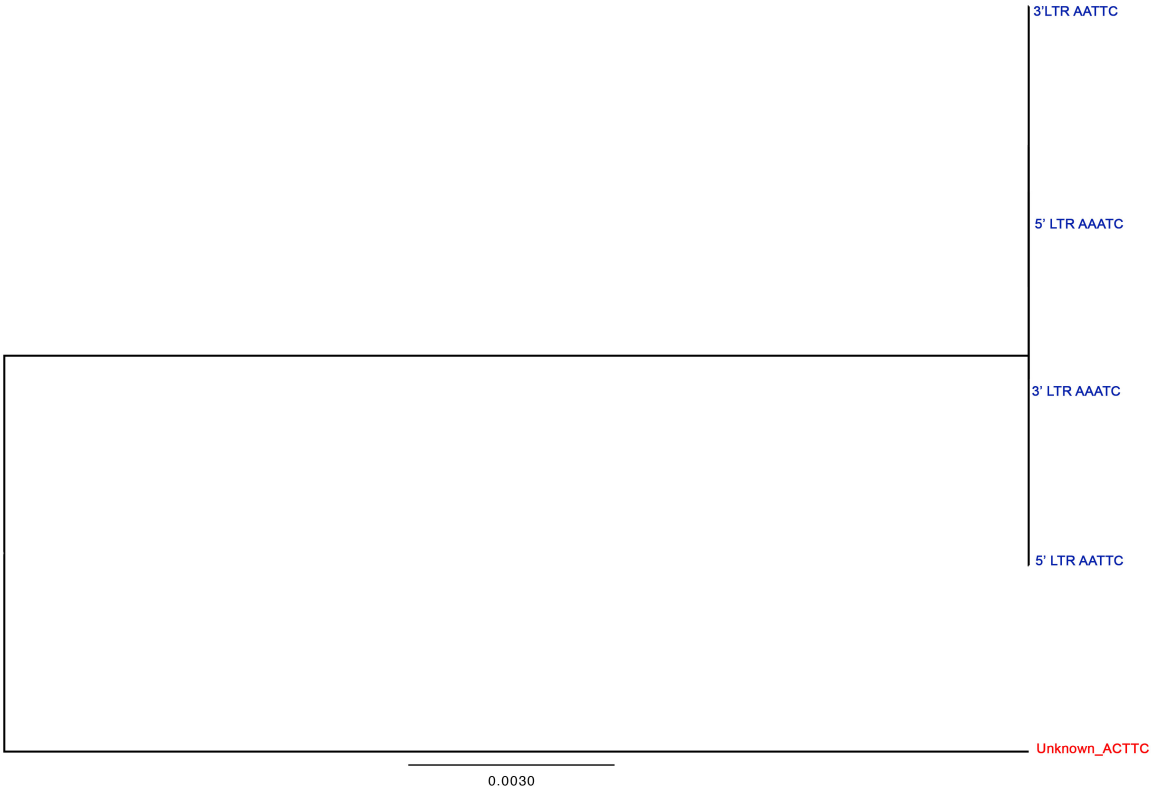


Figure F.10: **Maximum Likelihood phylogeny of individual element copies of *Srospv1*.** The phylogeny was constructed by an alignment of 391 nucleotide constructs with the employment of raxmlGUI using the GTRCAT model. Formatting is stated in F.6



Srospv2

Figure F.11: **Maximum Likelihood phylogeny of individual element copies of *Srospv2*.** The phylogeny was constructed by an alignment of 610 nucleotide constructs with the employment of raxmlGUI using the GTRCAT model. Formatting is stated in F.6



Srospv5

Figure F.12: **Maximum Likelihood phylogeny of individual element copies of *Srospv5*.** The phylogeny was constructed by an alignment of 364 nucleotide constructs with the employment of raxmlGUI using the GTRCAT model. Formatting is stated in F.6

References

Adl, S. M., Bass, D., Lane, C. E., Lukeš, J., Schoch, C. L., Smirnov, A., Agatha, S., Berney, C., Brown, M. W., Burki, F., Cárdenas, P., Čepička, I., Chistyakova, L., del Campo, J., Dunthorn, M., Edvardsen, B., Eglit, Y., Guillou, L., Hampl, V., Heiss, A. A., Hoppenrath, M., James, T. Y., Karpov, S., Kim, E., Kolisko, M., Kudryavtsev, A., Lahr, D. J. G., Lara, E., Le Gall, L., Lynn, D. H., Mann, D. G., Massana i Molera, R., Mitchell, E. A. D., Morrow, C., Park, J. S., Pawlowski, J. W., Powell, M. J., Richter, D. J., Rueckert, S., Shadwick, L., Shimano, S., Spiegel, F. W., Torruella i Cortes, G., Youssef, N., Zlatogursky, V. and Zhang, Q. (2018), 'Revisions to the classification, nomenclature, and diversity of eukaryotes', *Journal of Eukaryotic Microbiology* **66**(1), 4–119.

Agrawal, N., Dasaradhi, P. V. N., Mohmmed, A., Malhotra, P., Bhatnagar, R. K. and Mukherjee, S. K. (2003), 'Rna interference: Biology, mechanism, and applications', *Microbiology and Molecular Biology Reviews* **67**(4), 657–685.

Akashi, H. (1994), 'Synonymous codon usage in drosophila melanogaster: Natural selection and translational accuracy', *Genetics Society of America* **136**, 927–935.

Aldrich, S. (2017).
URL: *https://www.sigmaaldrich.com/united-kingdom.html*

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990), 'Basic local alignment search tool', *J Mol Biol* **215**(3), 403–410.

Amberg, D., Burke, D. and Strathern, J. (2005), *Methods in Yeast Genetics: A Cold Spring Harbor Laboratory Course Manual*, 2005 edn, Cold Spring Harbor Laboratory Press, USA.

Arkhipova, I. R. and Morrison, H. G. (2001), 'Three retrotransposon families in the genome of *Giardia lamblia*: two telomeric, one dead', *Proc Natl Acad Sci U S A* **98**(25), 14497–14502.

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., Grosdidier, A., Hernandez, C., Ioannidis, V., Kuznetsov, D., Liechti, R., Moretti, S., Mostaguir, K., Redaschi, N., Rossier, G., Xenarios, I. and Stockinger,

H. (2012), ‘Expasy: Sib bioinformatics resource portal’, *Nucleic Acids Res* **40**(Web Server issue), W597–603.

Baldauf, S. L. and Palmer, J. (1993), ‘Animals and fungi are each other’s closest relatives: Congruent evidence from multiple proteins’, *Proc. NatL Acad. Sci. USA* **90**(1), 1158–11562.

Bartolome, C., Bello, X. and Maside, X. (2009), ‘Widespread evidence for horizontal transfer of transposable elements across *Drosophila* genomes’, *Genome Biol* **10**(2), 1–11.

Bassing, C. H., Swat, W. and Alt, F. W. (2002), ‘The mechanism and regulation of chromosonal v(d)j recombination’, *Cell* **109**(1), 45 – 55.

Beauregard, A., Curcio, M. J. and Belfort, M. (2008), ‘The take and give between retrotransposable elements and their hosts’, *Annu Rev Genet* **42**, 587–617.

Belyayev, A. (2014), ‘Bursts of transposable elements as an evolutionary driving force’, *J Evol Biol* **27**(12), 2573–2584.

Benachenhou, F., Sperber, G. O., Bongcam-Rudloff, E., Andersson, G., Boeke, J. D. and Blomberg, J. (2013), ‘Conserved structure and inferred evolutionary history of long terminal repeats (LTRs)’, *Mob DNA* **4**(1), 1–15.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000), ‘The protein data bank’, *Nucleic Acids Res* **28**(1), 235 – 242.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Gallo Cassarino, T., Bertoni, M., Bordoli, L. and Schwede, T. (2014), ‘Swiss-model: modelling protein tertiary and quaternary structure using evolutionary information’, *Nucleic Acids Res* **42**(Web Server issue), W252–8.

Biemont, C. (2010), ‘A brief history of the status of transposable elements: from junk dna to major players in evolution’, *Genetics* **186**(4), 1085–1093.

Bleykasten-Grosshans, C., Friedrich, A. and Schacherer, J. (2013), ‘Genome-wide analysis of intraspecific transposon diversity in yeast’, *BMC Genomics* **14**(399), 1 – 13.

Bleykasten-Grosshans, C., Jung, P. P., Fritsch, E. S., Potier, S., de Montigny, J. and Souciet, J. L. (2011), ‘The *Ty1* LTR-retrotransposon population in *Saccharomyces cerevisiae* genome: dynamics and sequence variations during mobility’, *FEMS Yeast Res* **11**(4), 334–344.

Bock, R. (2010), 'The give-and-take of dna: horizontal gene transfer in plants', *Trends Plant Sci* **15**(1), 11–22.

Boeke, J. D. and Corces, V. G. (1989), 'Transcription and reverse transcription of retrotransposons', *Annu Rev Microbiol* **43**, 403–34.

Boeke, J. D. and Devine, S. E. (1998), 'Yeast retrotransposons: Finding a nice quiet neighborhood', *Cell* **93**, 1087 – 1089.

Botstein, D., Chervitz, S. A. and Cherry, J. M. (1997), 'Yeast as a model organism', *Science* **277**(5330), 1259 –1260.

Botstein, D. and Fink, G. R. (2011), 'Yeast: an experimental organism for 21st century biology', *Genetics* **189**(3), 695–704.

Brown, M. W., Spiegel, F. W. and Silberman, J. D. (2009), 'Phylogeny of the "forgotten" cellular slime mold, fonticula alba, reveals a key evolutionary branch within opisthokonta', *Molecular Biology and Evolution* **26**(12), 2699–2709.

Buchan, D. W., Minneci, F., Nugent, T. C., Bryson, K. and Jones, D. T. (2013), 'Scalable web services for the psipred protein analysis workbench', *Nucleic Acids Res* **41**(Web Server issue), W349–W357.

Bulmer, M. (1991), 'The selection-mutation-drift theory of synonymous codon usage', *Genetics Society of America* **129**(1), 897–907.

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A. and Yandell, M. (2007), 'Maker: An easy-to-use annotation pipeline designed for emerging model organism genomes', *Genome Research* **18**(1), 188–196.

Carr, M., Bensasson, D. and Bergman, C. M. (2012), 'Evolutionary genomics of transposable elements in *Saccharomyces cerevisiae*', *PLoS One* **7**(11), e50978.

Carr, M., Leadbeater, B. S. and Baldauf, S. L. (2010), 'Conserved meiotic genes point to sex in the choanoflagellates', *J Eukaryot Microbiol* **57**(1), 56–62.

Carr, M., Leadbeater, B. S., Nelson, M. and Baldauf, S. L. (2008), 'Molecular phylogeny of choanoflagellates, the sister group to metazoa', *PNAS* **105**(43), 16641 – 16646.

Carr, M., Nelson, M., Leadbeater, B. S. and Baldauf, S. L. (2008), 'Three families of LTR retrotransposons are present in the genome of the choanoflagellate *Monosiga brevicollis*', *Protist* **159**(4), 579–590.

Carr, M., Richter, D. J., Fozouni, P., Smith, T. J., Jeuck, A., Leadbeater, B. S. and Nitsche, F. (2017), 'A six-gene phylogeny provides new insights into choanoflagellate evolution', *Mol Phylogenet Evol* **107**, 166–178.

Carr, M. and Suga, H. (2014), 'The holozoan *Capsaspora owczarzaki* possesses a diverse complement of active transposable element families', *Genome Biol Evol* **6**(4), 949–963.

Casaregola, S., Neuve¨glise, C., Lepingle, A., Bon, E., Feynerol, C., Artiguenave, F., Wincker, P. and Gaillardin, C. (2000), 'Genomic exploration of the hemiascomycetous yeasts: 17. yarrowia lipolytica', *FEBS Letters* **487**, 95 –100.

Cavalier-Smith, T. (1987), *The origin of fungi and pseudofungi*, Cambridge University Press, Cambridge, pp. 339–353.

Cavalier-Smith, T. (2017), 'Origin of animal multicellularity: precursors, causes, consequences-the choanoflagellate/sponge transition, neurogenesis and the cambrian explosion', *Philos Trans R Soc Lond B Biol Sci* **372**(1713), 1–16.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27994119*

Cavalier-Smith, T. and Chao, E. E. (1995), 'The opalozoan apusomonas is related to the common ancestor of animals, fungi, and choanoflagellates', *Proceedings of the Royal Society of London. Series B: Biological Sciences* **261**(1360), 1–6.

Chapman, B. A., Bowers, J. E., Schulze, S. R. and Paterson, A. H. (2004), 'A comparative phylogenetic approach for dating whole genome duplication events', *Bioinformatics* **20**(2), 180–185.

Charlesworth, B., Langley, C. H. and Sniegowski, P. D. (1997), 'Transposable element distributions in drosophila', *Genetics* **147**(1), 1993–1995.

Cheng, E., Vaisica, J. A., Ou, J., Baryshnikova, A., Lu, Y., Roth, F. P. and Brown, G. W. (2012), 'Genome rearrangements caused by depletion of essential dna replication proteins in saccharomyces cerevisiae', *Genetics* **192**(1), 147–160.

Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W. and Korlach, J. (2013), 'Nonhybrid, finished microbial genome assemblies from long-read smrt sequencing data', *Nat Methods* **10**(6), 563–569.

Chin, C. S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G. R., Delledonne, M., Luo, C., Ecker, J. R., Cantu, D., Rank, D. R. and Schatz, M. C. (2016), 'Phased diploid genome assembly with single-molecule real-time sequencing', *Nat Methods* **13**(12), 1050–1054.

Clare, J. J., Belcourt, M. and Farabaugh, P. J. (1988), 'Efficient translational frameshifting occurs within a conserved sequence of the overlap between the two genes of a yeast ty1 transposon', *Proc Natl Acad Sci U S A* **85**(1), 6816 – 6820.

Clark, B. (1970), 'Darwinian evolution of proteins', *Science* **168**(3934), 1009–1011.

Cooney, A. L., Singh, B. K. and Sinn, P. L. (2015), 'Hybrid nonviral/viral vector systems for improved piggybac dna transposon in vivo delivery', *Mol Ther* **23**(4), 667–674.

Cordaux, R. and Batzer, M. A. (2009), 'The impact of retrotransposons on human genome evolution', *Nature Reviews Genetics* **10**(10), 691–703.

Crooks, R. E., Hon, G., Chandonia, J. and Brenner, S. E. (2004), 'Weblogo: A sequence logo generator', *Genome Res* **14**(1), 1188 – 1190.

Curcio M., J. and Belfort, M. (2007), 'The beginning of the end: Links between ancient retroelements and modern telomerases', *PNAS* **14**(22), 9107–9108.

Curcio, M. J., Lutz, S. and P., L. (2015), 'The ty1 ltr-retrotransposon of budding yeast, saccharomyces cerevisiae', *Microbiol Spectr* **3**(2), 1 – 35.

Dayel, M. J., Alegado, R. A., Fairclough, S. R., Levin, T. C., Nichols, S. A., McDonald, K. and King, N. (2011), 'Cell differentiation and morphogenesis in the colony-forming choanoflagellate salpingoeca rosetta', *Dev Biol* **357**(1), 73–82.

Dayel, M. J. and King, N. (2014), 'Prey capture and phagocytosis in the choanoflagellate salpingoeca rosetta', *PLoS One* **9**(5), 1–6.

deHaro, D., Kines, K. J., Sokolowski, M., Dauchy, R. T., Streva, V. A., Hill, S. M., Hanifin, J. P., Brainard, G. C., Blask, D. E. and Belancio, V. P. (2014), 'Regulation of l1 expression and retrotransposition by melatonin and its receptor: implications for cancer risk associated with light exposure at night', *Nucleic Acids Res* **42**(12), 7694–707.

Del Campo, J., Mallo, D., Massana, R., de Vargas, C., Richards, T. A. and Ruiz-Trillo, I. (2015), 'Diversity and distribution of unicellular opisthokonts along the european coast analysed using high-throughput sequencing', *Environ Microbiol* **17**(9), 3195–207.

Dhillon, B., G. N. H. R. and Goodwin, S. (2014), 'The landscape of transposable elements in the finished genome of the fungal wheat pathogen mycosphaerella graminicola', *BMC Genomics* **15**(1132), 1 – 17.

Dickinson, J. R. (2005), 'Are yeasts free-living unicellular eukaryotes?', *Lett Appl Microbiol* **41**(6), 445–447.

Doolittle, W. F. and Sapienza, C. (1980), 'Selfish genes, the phenotype paradigm and genome evolution', *Nature* **284**(5757), 601–603.

dos Reis, M. and Wernisch, L. (2009), 'Estimating translational selection in eukaryotic genomes', *Mol Biol Evol* **26**(2), 451–61.

Drinnenberg, I. A., Weinberg, D. E., Xie, K. T., Mower, J. P., Wolfe, K. H., Fink, G. R. and Bartel, D. P. (2009), 'RNAi in budding yeast', *Science* **326**(5952), 544–550.

Dujon, B. A. and Louis, E. J. (2017), 'Genome diversity and evolution in the budding yeasts (saccharomycotina)', *Genetics* **206**(2), 717–750.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuvéglise, C., Talla, E., Goffard, N., Frangeul, L., Aigle, M., Anthouard, V., Babour, A., Barbe, V., Barnay, S., Blanchin, S., Beckerich, J., Beyne, E., Bleykasten, C., Boisramé, A., Boyer, J., Cattolico, L., Confanioleri, F., De Daruvar, A., Despons, L., Fabre, E., Fairhead, C., Ferry-Dumazet, H., Groppi, A., Hantraye, F., Hennequin, C., Jauniaux, N., Joyet, P., Kachouri, R., Kerrest, A., Koszul, R., Lemaire, M., Lesur, I., Ma, L., Muller, H., Nicaud, J., Nikolski, M., Oztas, S., Ozier-Kalogeropoulos, O., Pellenz, S., Potier, S., Richard, G., Straub, M., Suleau, A., Swennen, D., Tekaia, F., Wésolowski-Louvel, M., Westhof, E., Wirth, B., Zeniou-Meyer, M., Zivanovic, I., Bolotin-Fukuhara, M., Thierry, A., Bouchier, C., Caudron, B., Scarpelli, C.,

Gaillardin, C., Weissenbach, J., Wincker, P. and Souciet, J. (2004), 'Genome evolution in yeasts', *Nature* **1**(430), 35–44.

Ehrenberg, M. and Kurland, C. G. (1984), 'Costs of accuracy determined by a maximal growth rate constraint', *Q Rev Biophys* **17**(1), 45–82.

Eickbush, T. H. and Jamburuthugoda, V. K. (2008), 'The diversity of retrotransposons and the properties of their reverse transcriptases', *Virus Res* **134**(1-2), 221–234.

Eide, D. and Anderson, P. (1988), 'Insertion and excision of caenorhabditis elegans transposable element tc1', *Molecular and Cellular Biology* **8**(2), 737–746.

El Baidouri, M., Carpentier, M. C., Cooke, R., Gao, D., Lasserre, E., Llauro, C., Mirouze, M., Picault, N., Jackson, S. A. and Panaud, O. (2014), 'Widespread and frequent horizontal transfers of transposable elements in plants', *Genome Res* **24**(5), 831–838.

Elliott, T. A. and Gregory, T. R. (2015), 'Do larger genomes contain more diverse transposable elements?', *BMC Evol Biol* **15**(69), 1–10.

Fairclough, S. R., Chen, Z., Kramer, E., Zeng, Q., Young, S., Robertson, H. M., Begovic, E., Richter, D. J., Russ, C., Westbrook, M. J., Manning, G., Lang, B. F., Haas, B., Nusbaum, C. and King, N. (2013), 'Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate salpingoeca rosetta', *Genome Biol* **14**(2), 1–15.

Fairclough, S. R., Dayel, M. J. and King, N. (2010), 'Multicellular development in a choanoflagellate', *Curr Biol* **20**(20), R875–6.

Fedoroff, N. V. (2012), 'Transposable elements, epigenetics, and genome evolution', *Science* **338**, 758 – 768.

Feschotte, C. and Pritham, E. J. (2007), 'Dna transposons and the evolution of eukaryotic genomes', *Annu Rev Genet* **41**, 331–368.

Finnegan, D. J. (1989), 'Eukaryotic transposable elements and genome evolution', *Trends in genetics* **5**(4), 103 – 107.

Fisher (2017).

URL: *https://www.fishersci.co.uk/gb/en/home.html*

Fitzpatrick, D. A. (2012), 'Horizontal gene transfer in fungi', *FEMS Microbiol Lett* **329**(1), 1–8.

Fortune, P. M., Roulin, A. and Panaud, O. (2008), 'Horizontal transfer of transposable elements in plants', *Commun Integr Biol* **1**(1), 74–77.

Freund, R. and Meselson, M. (1984), 'Long terminal repeat nucleotide sequence and specific insertion of the gypsy transposon', *Proc Natl Acad Sci U S A* **81**, 4462–4464.

Frumkin, I., Lajoie, M. J., Gregg, C. J., Hornung, G., Church, G. M. and Pilpel, Y. (2018), 'Codon usage of highly expressed genes affects proteome-wide translation efficiency', *Proceedings of the National Academy of Sciences* **115**(21), E4940–E4949.

Gaillardin, C., Duchateau-Nguyen, G., Tekaia, F., Llorente, B., Casaregola, S., Toffano-Nioche, C., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., Bon, E., Brottier, P., de Montigny, J., Dujon, B., Durrens, P., Lepingle, A., Malpertuy, A., Neuveglise, C., Ozier-Kalogeropoulos, O., Potier, S., Saurin, W., Termier, M., Wesolowski-Louvel, M., Wincker, P., Souciet, J. and Weissenbach, J. (2000), 'Genomic exploration of the hemiascomycetous yeasts: 21. comparative functional classification of genes', *FEBS Lett* **487**(1), 134–149.

Galtier, N., Roux, C., Rousselle, M., Romiguier, J., Figuet, E., Glemin, S., Bierne, N. and Duret, L. (2018), 'Codon usage bias in animals: Disentangling the effects of natural selection, effective population size, and gc-biased gene conversion', *Mol Biol Evol* **35**(5), 1092–1103.

Genolevures, C., Souciet, J. L., Dujon, B., Gaillardin, C., Johnston, M., Baret, P. V., Cliften, P., Sherman, D. J., Weissenbach, J., Westhof, E., Wincker, P., Jubin, C., Poulain, J., Barbe, V., Segurens, B., Artiguenave, F., Anthouard, V., Vacherie, B., Val, M. E., Fulton, R. S., Minx, P., Wilson, R., Durrens, P., Jean, G., Marck, C., Martin, T., Nikolski, M., Rolland, T., Seret, M. L., Casaregola, S., Despons, L., Fairhead, C., Fischer, G., Lafontaine, I., Leh, V., Lemaire, M., de Montigny, J., Neuveglise, C., Thierry, A., Blanc-Lenfle, I., Bleykasten, C., Diffels, J., Fritsch, E., Frangeul, L., Goeffon, A., Jauniaux, N., Kachouri-Lafond, R., Payen, C., Potier, S., Pribylova, L., Ozanne, C., Richard, G. F., Sacerdot, C., Straub, M. L. and Talla, E. (2009), 'Comparative genomics of protoploid saccharomycetaceae', *Genome Res* **19**(10), 1696–709.

Gerber, A., Grosjean, H., Melcher, T. and Keller, W. (1998), 'Tad1p, a yeast trna-specific adenosine deaminase, is related to the mammalian pre-mrna editing enzymes adar1 and adar2', *The EMBO Journal* **17**(16), 4780 – 4789.

Geurts, A. M., Hackett, C. S., Bell, J. B., Bergemann, T. L., Collier, L. S., Carlson, C. M., Largaespada, D. A. and Hackett, P. B. (2006), 'Structure-based prediction of insertion-site preferences of transposons into chromosomes', *Nucleic Acids Res* **34**(9), 2803–2811.

Gilbert, C. and Cordaux, R. (2013), 'Horizontal transfer and evolution of prokaryote transposable elements in eukaryotes', *Genome Biol Evol* **5**(5), 822–832.

GIRI (2016).
URL: *http://www.girinst.org*

Gladieux, P., Ropars, J., Badouin, H., Branca, A., Aguileta, G., de Vienne, D. M., Rodriguez de la Vega, R. C., Branco, S. and Giraud, T. (2014), 'Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes', *Mol Ecol* **23**(4), 753–773.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S. G. (1996), 'Life with 6000 genes', *Science* **274**(5287), 546, 563–567.

Goodier, J. L. and Kazazian, H. H., J. (2008), 'Retrotransposons revisited: the restraint and rehabilitation of parasites', *Cell* **135**(1), 23–35.

Gorinsek, B., Gubensek, F. and Kordis, D. (2004), 'Evolutionary genomics of chromoviruses in eukaryotes', *Mol Biol Evol* **21**(5), 781–798.

Grabundzija, I., Irgang, M., Mates, L., Belay, E., Matrai, J., Gogol-Doring, A., Kawakami, K., Chen, W., Ruiz, P., Chuah, M. K., VandenDriessche, T., Izsvak, Z. and Ivics, Z. (2010), 'Comparative analysis of transposable element vector systems in human cells', *Mol Ther* **18**(6), 1200–1209.

Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pav'e, A. (1980), 'Codon catalog usage and the genome hypothesis', *Nucleic Acids Res* **8**(1), 49 – 62.

Gypsy Database 2.0 (2010).
URL: *http://gydb.org/index.php/Ty3/Gypsy*

Han, J. S. (2010), 'Non-long terminal repeat (non-ltr) retrotransposons: mechanisms, recent developments, and unanswered questions', *Mob DNA* **1**(15), 1 –12.

Harrison, R. J. and Charlesworth, B. (2011), 'Biased gene conversion affects patterns of codon usage and amino acid usage in the saccharomyces sensu stricto group of yeasts', *Mol Biol Evol* **28**(1), 117–129.

Hauk, G., McKnight, J. N., Nodelman, I. M. and Bowman, G. D. (2010), 'The chromodomains of the chd1 chromatin remodeler regulate dna access to the atpase motor', *Mol Cell* **39**(5), 711–723.

Havecker, E. R., Gao, X. and Voytas, D. F. (2004), 'The diversity of ltr retrotransposons', *Genome Biol* **5**(225), 1–6.

Hayward, A. (2017), 'Origin of the retroviruses: when, where, and how?', *Curr Opin Virol* **25**, 23–27.

Hehenberger, E., Tikhonenkov, D. V., Kolisko, M., Del Campo, J., Esaulov, A. S., Mylnikov, A. P. and Keeling, P. J. (2017), 'Novel predators reshape holozoan phylogeny and reveal the presence of a two-component signaling system in the ancestor of animals', *Curr Biol* **27**(13), 2043–2050.

Hess, J., Skrede, I., Wolfe, B. E., LaButti, K., Ohm, R. A., Grigoriev, I. V. and Pringle, A. (2014), 'Transposable element dynamics among asymbiotic and ectomycorrhizal amanita fungi', *Genome Biol Evol* **6**(7), 1564–1578.

Hillis, D. M. and Bull, J. J. (1993), 'An empirical test of bootstrapping as a method of assessing confidence in phylogenetic analysis', *Syst Biol* **42**(2), 182–192.

Hinegardner, R. and Engelberg, J. (1963), 'Rationale for a universal genetic code', *Science* **142**(3595), 1083 – 1085.

Hoffmeyer, T. T. and Burkhardt, P. (2016), 'Choanoflagellate models - monosiga brevicollis and salpingoeca rosetta', *Curr Opin Genet Dev* **39**, 42–47.

Huang, C. R. L., Burns, K. H. and Boeke, J. D. (2012), 'Active transposition in genomes', *Annual Review of Genetics* **46**(1), 651–675.

Huda, A. and Jordan, I. K. (2009), 'Analysis of transposable element sequences using censor and repeatmasker', *Methods Mol Biol* **537**, 323–336.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C. and Bork, P. (2017), 'Fast genome-wide functional annotation through orthology assignment by eggnog-mapper', *Mol Biol Evol* **34**(8), 2115–2122.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., Jensen, L. J., von Mering, C. and Bork, P. (2016), 'eggnog 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences', *Nucleic Acids Res* **44**(D1), 286–293.

Ikemura, T. (1981), 'Correlation between the abundance of escherichia coli transfer rnas and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the e. coli translational system', *Journal of Molecular Biology* **151**(3), 389–409.

Ikemura, T. (1985), 'Codon usage and trna content in unicellular and multicellular organisms', *Mol Bio Evol* **2**(1), 13 – 34.

Ivics, Z., Hackett, P. B., Plasterk, R. H. and Isvak, Z. (1997), 'Molecular reconstruction of sleeping beauty, a tc1-like transposon from fish, and its transposition in human cells', *Cell* **91**, 501 – 510.

James-Clark, H. (1868), *On the Spongiae Ciliatae as Infusoria Flagellata: or observations on the structure, animality, and relationship of Leucosolenia botryoides, Bowerbank.*, Nat. Hist, Memoirs Boston Soc.

James, T. Y., Kauff, F., Schoch, C. L., Matheny, P. B., Hofstetter, V., Cox, C. J., Celio, G., Gueidan, C., Fraker, E., Miadlikowska, J., Lumbsch, H. T., Rauhut, A., Reeb, V., Arnold, A. E., Amtoft, A., Stajich, J. E., Hosaka, K., Sung, G. H., Johnson, D., O'Rourke, B., Crockett, M., Binder, M., Curtis, J. M., Slot, J. C., Wang, Z., Wilson, A. W., Schussler, A., Longcore, J. E., O'Donnell, K., Mozley-Standridge, S., Porter, D., Letcher, P. M., Powell, M. J., Taylor, J. W., White, M. M., Griffith, G. W., Davies, D. R., Humber, R. A., Morton, J. B., Sugiyama, J., Rossman, A. Y., Rogers, J. D., Pfister, D. H., Hewitt, D., Hansen, K., Hambleton, S., Shoemaker, R. A., Kohlmeyer, J., Volkmann-Kohlmeyer, B., Spotts, R. A., Serdani, M., Crous, P. W., Hughes, K. W., Matsuura, K., Langer, E., Langer, G., Untereiner, W. A., Lucking, R., Budel, B., Geiser, D. M., Aptroot, A., Diederich, P., Schmitt, I., Schultz, M., Yahr, R., Hibbett, D. S., Lutzoni, F., McLaughlin, D. J., Spatafora, J. W. and Vilgalys, R. (2006), 'Reconstructing the early evolution of fungi using a six-gene phylogeny', *Nature* **443**(7113), 818–822.

Jangam, D., Feschotte, C. and Betran, E. (2017), 'Transposable element domestication as an adaptation to evolutionary conflicts', *Trends Genet* **33**(11), 817–831.

Jeuck, A., Arndt, H. and Nitsche, F. (2014), 'Extended phylogeny of the craspedida (choanomonada)', *Eur J Protistol* **50**(4), 430–443.

Jia, J. and Xue, Q. (2009), 'Codon usage biases of transposable elements and host nuclear genes in arabidopsis thaliana and oryza sativa', *Genomics, Proteomics and Bioinformatics* **7**(4), 175–184.

Jiang, R. H. Y., Tyler, B. M. and Govers, F. (2006), 'Comparative analysis of phytophthora genes encoding secreted proteins reveals conserved synteny and lineage-specific gene duplications and deletions', *The American Phytopathological Society* **19**(12), 1311–1321.

Joly-Lopez, Z., Forczek, E., Hoen, D. R., Juretic, N. and Bureau, T. E. (2012), 'A gene family derived from transposable elements during early angiosperm evolution has reproductive fitness benefits in arabidopsis thaliana', *PLoS Genet* **8**(9), 1–10.

Joly-Lopez, Z., Hoen, D. R., Blanchette, M. and Bureau, T. E. (2016), 'Phylogenetic and genomic analyses resolve the origin of important plant genes derived from transposable elements', *Mol Biol Evol* **33**(8), 1937–1956.

Jones, D. T. (1999), 'Protein secondary structure prediction based on position-specific scoring matrices', *J Mol Biol* **292**, 195 – 202.

Jordan, I. K. and McDonald, J. F. (1998), 'Tempo and mode of ty element evolution in saccharomyces cerevisiae', *Genetics* **151**, 1341 – 1351.

Jordan, I. K. and McDonald, J. F. (1999), 'Comparative genomics and evolutionary dynamics of saccharomyces cerevisiae ty elements', *Genetica* **107**, 3 – 13.

Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y. and Ikemura, T. (2001), 'Codon usage and trna genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with cg-dinucleotide usage as assessed by multivariate analysis', *J Mol Evol* **53**(4-5), 290–8.

Kaneko, Y. and Banno, I. (1991), 'Re-examination of *Saccharomyces bayanus* strains by DNA-hybridization and electrophoretic karyotyping', *IFO Res Comm* **15**, 30–41.

Katoh, K., Misawa, K., Kuma, K. and Miyata, T. (2002), 'Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform', *Nucleic Acids Res* **30**(14), 3059 – 3066.

Kazazian, H. H., J. (2004), 'Mobile elements: drivers of genome evolution', *Science* **303**(5664), 1626–1632.

Kidwell, M. G. (2002), 'Transposable elements and the evolution of genome size in eukaryotes', *Genetica* **115**, 49 – 63.

Kidwell, M. G. and Lisch, D. (1997), 'Transposable elements as sources of variation in animals and plants', *Proc Natl Acad Sci U S A* **94**(15), 7704–7711.

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

King, N. (2005), 'Choanoflagellates', *Curr Biol* **15**(4), R113–R114.

King, N., Westbrook, M. J., Young, S. L., Kuo, A., Abedin, M., Chapman, J., Fairclough, S., Hellsten, U., Isogai, Y., Letunic, I., Marr, M., Pincus, D., Putnam, N., Rokas, A., Wright, K. J., Zuzow, R., Dirks, W., Good, M., Goodstein, D., Lemons, D., Li, W., Lyons, J. B., Morris, A., Nichols, S., Richter, D. J., Salamov, A., Sequencing, J. G., Bork, P., Lim, W. A., Manning, G., Miller, W. T., McGinnis, W., Shapiro, H., Tjian, R., Grigoriev, I. V. and Rokhsar, D. (2008), 'The genome of the choanoflagellate monosiga brevicollis and the origin of metazoans', *Nature* **451**(7180), 783–8.

King, N., Young, S. L., Abedin, M., Carr, M. and Leadbeater, B. S. (2009), 'The choanoflagellates: heterotrophic nanoflagellates and sister group of the metazoa', *Cold Spring Harb Protoc* **2009**(2), pdb emo116.

Kordis, D. (2005), 'A genomic perspective on the chromodomain-containing retrotransposons: Chromoviruses', *Gene* **347**(2), 161–73.

Kupiec, M. and Petes, T. D. (1988), 'Allelic and ectopic recombination betweenty elements in yeast', *Genetics Society of America* **119**(1), 549 – 559.

Kurtzman, C. (2003), 'Phylogenetic circumscription of *Saccharomyces, Kluyveromyces* and other members of the Saccharomycetaceae, and the proposal of the new genera *Lachancea, Nakaseomyces, Naumovia, Vanderwaltozyma* and *Zygotorulaspora*.', *FEMS Yeast Research* **4**(3), 233–245.

Kurtzman, C. P. (2011), *The Yeasts: A Taxonomic Study*, 2 Saccharomyces, Kluyveromyces and Other Genera of the Saccharomycetaceae, 5th edn, Elsevier Science.

Kurtzman, C. and Robnett, C. (2003), 'Phylogenetic relationships among yeasts of the *Saccharomyces* complex determined from multigene sequence analyses', *FEMS Yeast Research* **3**(4), 417–432.

LaBella, A. L., Opulente, D. A., Steenwyk, J. L., Hittinger, C. T. and Rokas, A. (2019), 'Variation and selection on codon usage bias across an entire subphylum', *PLoS Genet* **15**(7), 1–25.

Lang, B. F., O'Kelly, C., Nerad, T., Gray, M. W. and Burger, G. (2002), 'The closest unicellular relatives of animals', *Curr Biol* **12**, 1773–1778.

Lee, H. E., Ayarpadikannan, S. and Kim, H. S. (2015), 'Role of transposable elements in genomic rearrangement, evolution, gene regulation and epigenetics in primates', *Genes Genet Syst* **90**(5), 245–257.

Lee, I. and Harshey, R. M. (2003), 'Patterns of sequence conservation at termini of long terminal repeat (ltr) retrotransposons and dna transposons in the human genome: lessons from phage mu', *Nucleic Acids Res* **31**(15), 4531–4540.

Lee, S. I. and Kim, N. S. (2014), 'Transposable elements and genome size variations in plants', *Genomics Inform* **12**(3), 87–97.

Legras, J. L., Galeote, V., Bigey, F., Camarasa, C., Marsit, S., Nidelet, T., Sanchez, I., Couloux, A., Guy, J., Franco-Duarte, R., Marcet-Houben, M., Gabaldon, T., Schuller, D., Sampaio, J. P. and Dequin, S. (2018), 'Adaptation of s. cerevisiae to fermented food environments reveals remarkable genome plasticity and the footprints of domestication', *Mol Biol Evol* **35**(7), 1712–1727.

Lerat, E., Capy, P. and Biémont, C. (2002), 'Codon usage by transposable elements and their host genes in five species', *J Mol Evol* **54**(5), 625–637.

Lerat, E., Rizzon, C. and Biemont, C. (2003), 'Sequence divergence within transposable element families in the *Drosophila melanogaster* genome', *Genome Res* **13**(8), 1889–96.

Levin, H. L. and Moran, J. V. (2011), 'Dynamic interactions between transposable elements and their hosts', *Nat Rev Genet* **12**(9), 615–627.

Li, W., Cowley, A., Uludag, M., Gur, T., McWilliam, H., Squizzato, S., Park, Y. M., Buso, N. and Lopez, R. (2015), 'The embl-ebi bioinformatics web and programmatic tools framework', *Nucleic Acids Res* **43**(W1), W580–584.

Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., Davey, R. P., Roberts, I. N., Burt, A., Koufopanou, V., Tsai, I. J., Bergman, C. M., Bensasson, D., O'Kelly, M. J., van Oudenaarden, A., Barton, D. B., Bailes, E., Nguyen, A. N., Jones, M., Quail, M. A., Goodhead, I., Sims, S., Smith, F., Blomberg, A., Durbin, R. and Louis, E. J. (2009), 'Population genomics of domestic and wild yeasts', *Nature* **458**(7236), 337–341.

Liti, G., Peruffo, A., James, S. A., Roberts, I. N. and Louis, E. J. (2005), 'Inferences of evolutionary relationships from a population survey of LTR-retrotransposons and telomeric-associated sequences in the *Saccharomyces sensu stricto* complex', *Yeast* **22**(3), 177–92.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. and Law, M. (2012), 'Comparison of next-generation sequencing systems', *J Biomed Biotechnol* **2012**, 1–11.

Liu, Y., Steenkamp, E. T., Brinkmann, H., Forget, L., Philippe, H. and Lang, B. F. (2009), 'Phylogenomic analyses predict sistergroup relationship of nucleariids and fungi and paraphyly of zygomycetes with significant support', *BMC Evol Biol* **9**(272), 1–11.

Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J. M., Tamarit, D., Aguilar-Rodríguez, J., Vicente-Ripolles, M., Fuster, G., Bernet, G., Maumus, F., Munoz-Pomer, A., Sempere, J. M., Latorre, A. and Moya, A. (2010), 'The gypsy database (gydb) of mobile genetic elements: release 2.0', *Nucleic Acids Res* **39**(1), 70–74.

Lloyd, A. T. and Sharp, P. M. (1992), 'Evolution of codon usage patterns: the extent and nature of divergence between candida albicans and saccharomyces cerevisiae', *Nucleic Acids Res* **20**(20), 5289–5295.

Looke, M., Kristjuhan, K. and Kristjuhan, A. (2011), 'Extraction of genomic dna from yeasts for pcr-based applications', *Biotechniques* **50**(5), 325–328.

Lowe, T. M. and Chan, P. P. (1997), 'trnascan-se: a program for improved detection of transfer rna genes in genomic sequence.', *Nucleic Acids Res* **25**(5), 955–964.

Lu, L., Chen, J., Robb, S. M. C., Okumoto, Y., Stajich, J. E. and Wessler, S. R. (2017), 'Tracking the genome-wide outcomes of a transposable element burst over decades of amplification', *Proc Natl Acad Sci U S A* **114**(49), E10550–E10559.

Macrogen (2018).
URL: *https://dna.macrogen.com/eng/*

Malik, H. S. and Eickbush, T. H. (1999), 'Modular evolution of the integrase domain in the ty3/gypsy class of ltr retrotransposons', *Journal of Virology* **73**(6), 5186 – 5190.

Marin, I. and Llore´ns, C. (2000), 'Ty3/gypsy retrotransposons: Description of new arabidopsis thaliana elements and evolutionary perspectives derived from comparative genomic data', *Mol Bio Evol* **17**(7), 1040 – 1049.

Marinoni, G., Manuel, M., Petersen, R. F., Hvidtfeldt, J., Sulo, P. and Piskur, J. (1999), 'Horizontal transfer of genetic material among *Saccharomyces* yeasts', *J Bacteriol* **181**(20), 6488–96.

Marsit, S., Mena, A., Bigey, F., Sauvage, F. X., Couloux, A., Guy, J., Legras, J. L., Barrio, E., Dequin, S. and Galeote, V. (2015), 'Evolutionary advantage conferred by an eukaryote-to-eukaryote gene transfer event in wine yeasts', *Mol Biol Evol* **32**(7), 1695–1707.

Martienssen, R. and Colot, V. (2001), 'Dna methylation and epigenetic inheritance in plants and filamentous fungi', *Science* **293**(5532), 1070 – 1074.

McClintock, B. (1950), 'The origin and behavior of mutable loci in maize', *PNAS* **36**, 344 – 355.

McClintock, B. (1984), 'The significance of responses of the genome to challenge', *Science* **226**(4676), 792–801.

Melek, M., Gellert, M. and van Gent, D. C. (1998), 'Rejoining of dna by the rag1 and rag2 proteins', *Science* **280**(5361), 301 – 303.

Mikata, K., Ueda-Nishimura, K. and Hisatomi, T. (2001), 'Three new species of saccharomyces sensu lato van der walt from yaku island in japan: Saccharomyces naganishii sp. nov., saccharomyces humaticus sp. nov. and saccharomyces yakushimaensis sp. nov.', *International Journal of Systematic and Evolutionary Microbiology* **51**, 2189 – 2198.

Miller, M. A., Pfeifer, W. and Schwartz, T. (2010), 'Creating the cipres science gateway for inference of large phylogenetic trees', pp. 1 – 7.

Miller, W. J., McDonald, J. F., Nouand, D. and D., A. (1999), 'Molecular domestication – more than a sporadic episode in evolution', *Genetica* **107**(1), 197–207.

Milne, I., Stephen, G., Bayer, M., Cock, P. J., Pritchard, L., Cardle, L., Shaw, P. D. and Marshall, D. (2013), 'Using tablet for visual exploration of second-generation sequencing data', *Brief Bioinform* **14**(2), 193–202.

Muszewska, A., Hoffman-Sommer, M. and Grynberg, M. (2011), 'LTR retrotransposons in fungi', *PLoS One* **6**(12), 1–10.

Muñoz-López, M. and García-Pérez, J. L. (2010), 'Dna transposons: Nature and applications in genomics', *Current Genomics* **11**, 115 – 128.

NCYC (2016).
URL: *https://catalogue.ncyc.co.uk/strains/*

Neuvéglise, C., Feldmann, H., Bon, E., Gaillardin, C. and Casaregola, S. (2002), 'Genomic evolution of the long terminal repeat retrotransposons in hemiascomycetous yeasts', *Genome Res* **12**(6), 930–943.

Nisiotou, A. A. and Nychas, G. J. (2008), 'Kazachstania hellenica sp. nov., a novel ascomycetous yeast from a botrytis-affected grape must fermentation', *Int J Syst Evol Microbiol* **58**(Pt 5), 1263–1267.

Nitsche, F., Carr, M., Arndt, H. and Leadbeater, B. S. (2011), 'Higher level taxonomy and molecular phylogenetics of the choanoflagellatea', *J Eukaryot Microbiol* **58**(5), 452–62.

Norris, R. (1965), 'Neustonic marine craspedomonadales (choanofiagellates) from washington and california', *The Journal of Protozoology* **12**(4).

Novikova, O. S. and Blinov, A. G. (2009), '[origin, evolution, and distribution of different groups of non-ltr retrotransposons among eukaryotes]', *Genetika* **45**(2), 149–59.

Novoa, E. M., Pavon-Eternod, M., Pan, T. and Ribas de Pouplana, L. (2012), 'A role for trna modifications in genome structure and codon usage', *Cell* **149**(1), 202–213.

Nwokeoji, A. O., Kilby, P. M., Portwood, D. E. and Dickman, M. J. (2016), 'Rnaswift: A rapid, versatile rna extraction method free from phenol and chloroform', *Anal Biochem* **512**, 36–46.

O'Donnell, K. A. and Burns, K. H. (2010), 'Mobilizing diversity: transposable element insertions in genetic variation and disease', *Mobile DNA* **1**(21), 1– 10.

Oliver, K. R. and Greene, W. K. (2009), 'Transposable elements: powerful facilitators of evolution', *Bioessays* **31**(7), 703–714.

Oliver, K. R. and Greene, W. K. (2012), 'Transposable elements and viruses as factors in adaptation and evolution: an expansion and strengthening of the te-thrust hypothesis', *Ecol Evol* **2**(11), 2912–33.

Orgel, L. E. and Crick, F. H. (1980), 'Selfish DNA: the ultimate parasite', *Nature* **284**(5757), 604–607.

Parfrey, L. W., Lahr, D. J. G., Knolf, A. H. and Katz, L. A. (2011), 'Estimating the timing of early eukaryotic diversification with multigene molecular clocks', *PNAS* **108**(33), 13624–13629.

Peden, J. F. (1999), Analysis of codon usage, Thesis.

Piednoël, M., Gonçalves, I. R., Higuet, D. and Bonnivard, E. (2011), 'Eukaryote dirs1-like retrotransposons: an overview', *BMC Genomics* **12**(621), 1–18.

Ponstingl, H. (2014), 'Smalt manual', (0.7.4), 1–10.

Pritham, E. J. (2009), 'Transposable elements and factors influencing their success in eukaryotes', *J Hered* **100**(5), 648–655.

Qiagen (2017).
URL: *https://www.qiagen.com/us/*

Rafels-Ybern, A., Torres, A. G., Grau-Bove, X., Ruiz-Trillo, I. and Ribas de Pouplana, L. (2017), 'Codon adaptation to trnas with inosine modification at position 34 is widespread among eukaryotes and present in two bacterial phyla', *RNA Biology* **15**(4-5), 500–507.

Ran, W. and Higgs, P. G. (2012), 'Contributions of speed and accuracy to translational selection in bacteria', *PLoS Biol* **7**(12), 1–7.

Rannala, B. and Yang, Z. (1996), 'Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference', *Molecular Evolution* **43**, 304–311.

Reilly, M. T., Faulkner, G. J., Dubnau, J., Ponomarev, I. and Gage, F. H. (2013), 'The role of transposable elements in health and diseases of the central nervous system', *J Neurosci* **33**(45), 17577–17586.

RepeatMasker (1996).
URL: *http://repeatmasker.org/*

Roche (2018).
URL: *https://sequencing.roche.com/en/products-solutions/by-category/library-preparation/library-quantification*

Ronquist, F. and Huelsenbeck, J. P. (2003), 'MrBayes 3: Bayesian phylogenetic inference under mixed models', *Bioinformatics* **19**(12), 1572–4.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M. A. and Huelsenbeck, J. P. (2012), 'Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space', *Syst Biol* **61**(3), 539–542.

Rozas, J., Sanchez-DelBarrio, J. C., Messeguer, X. and Rozas, R. (2003), 'Dnasp, dna polymorphism analyses by the coalescent and other methods', *Bioinformatics* **19**(18), 2496–2497.

Sandmeyer, S., Patterson, K. and Bilanchone, V. (2015), 'Ty3, a position-specific retrotransposon in budding yeast', *Microbiol Spectr* **3**(2), 1 – 29.

Sarilar, V., Bleykasten-Grosshans, C. and Neuvéglise, C. (2015), 'Evolutionary dynamics of *hAT* transposon families in Saccharomycetaceae', *Genome Biol Evol* **7**(1), 172–190.

Sarilar, V., Sterck, L., Matsumoto, S., Jacques, N., Neuveglise, C., Tinsley, C. R., Sicard, D. and Casaregola, S. (2017), 'Genome sequence of the type strain clib 1764t (= cbs 14374t) of the yeast species kazachstania saulgeensis isolated from french organic sourdough', *Genom Data* **13**, 41–43.

Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Yaschenko, E. and Ye, J. (2009), 'Database

resources of the national center for biotechnology information', *Nucleic Acids Res* **37**(Database issue), D5–15.

Schaack, S., Gilbert, C. and Feschotte, C. (2010), 'Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution', *Trends Ecol Evol* **25**(9), 537–46.

Schrodinger (2017), 'The pymol molecular graphics system', *Nucleic Acids Res* .

Shalchian-Tabrizi, K., Minge, M. A., Espelund, M., Orr, R., Ruden, T., Jakobsen, K. S. and Cavalier-Smith, T. (2008), 'Multigene phylogeny of choanozoa and the origin of animals', *PLoS One* **3**(5), 1–7.

Sharp, P. and Cowe, E. (1991), 'Synonymous codon usage in saccharomyces cerevisiae', *Yeast* **7**, 657 – 678.

Sharp, P. M., Cowe, E., Higgins, D. G., Shields, D. C., Wolfe, K. H. and Wright, F. (1988), 'Codon usage patterns in escherichia coli, bacillus subtilis, saccharomyces cerevisiae, schizosaccharomyces pombe, drosophila melanogaster and homo sapiens; a review of the considerable within-species diversity.', *Nucleic Acids Res* **16**(17), 8207–8211.

Sharp, P. M., Tuchy, T. M. F. and R., M. K. (1986), 'Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes', *Nucleic Acids Res* **14**(13), 5125–5143.

Shen, J. J., Dushoff, J., Bewick, A. J., Chain, F. J. and Evans, B. J. (2013), 'Genomic dynamics of transposable elements in the western clawed frog (silurana tropicalis)', *Genome Biol Evol* **5**(5), 998–1009.

Shields, D. C. and Sharp, P. M. (1989), 'Evidence that mutation patterns vary among drosophila transposable elements', *J Mol Biol* **207**, 843–846.

Shirasu, K., Schulman, A. H., Lahaye, T. and Schulze-Lefert, P. (2000), 'A contiguous 66-kb barley dna sequence provides evidence for reversible genome expansion', *Genome Res* **10**(1), 908 – 915.

Silva, J., Bastida, F., Bidwell, S. L., Johnson, P. J. and Carlton, J. M. (2005), 'A potentially functional mariner transposable element in the protist trichomonas vaginalis', *Mol Biol Evol* **22**(1), 126 – 134.

Silvestro, D. and Michalak, I. (2011), 'raxmlgui: a graphical front-end for raxml', *Organisms Diversity and Evolution* **12**(4), 335–337.

Sinzelle, L., Kapitonov, V. V., Grzela, D. P., Jursch, T., Jurka, J., Izsvak, Z. and Ivics, Z. (2008), 'Transposition of a reconstructed harbinger element in human cells and functional homology with two transposon-derived cellular genes', *Proc Natl Acad Sci U S A* **105**(12), 4715–4720.

Slotkin, R. K. and Martienssen, R. (2007), 'Transposable elements and the epigenetic regulation of the genome', *Nat Rev Genet* **8**(4), 272–85.

Smith, N. G. C. and Eyre-Walker, A. (2001), 'Synonymous codon bias is not caused by mutation bias in g1c-rich genes in humans', *Mol Bio Evol* **18**(6), 982–986.

Snoke, M. S., Berendonk, T. U., Barth, D. and Lynch, M. (2006), 'Large global effective population sizes in paramecium', *Mol Biol Evol* **23**(12), 2474–2479.

Soderlund, C., Bomhoff, M. and Nelson, W. M. (2011), 'Symap v3.4: a turnkey synteny system with application to plant genomes', *Nucleic Acids Res* **39**(10), 1–9.

Soderlund, C., Nelson, W., Shoemaker, A. and Paterson, A. (2006), 'Symap: A system for discovering and viewing syntenic regions of fpc maps', *Genome Res* **16**(9), 1159–68.

Song, G., Dickins, B. J., Demeter, J., Engel, S., Gallagher, J., Choe, K., Dunn, B., Snyder, M. and Cherry, J. M. (2015), 'Agape (automated genome analysis pipeline) for pan-genome analysis of saccharomyces cerevisiae', *PLoS One* **10**(3), 1–19.

Southworth, J., Armitage, P., Fallon, B., Dawson, H., Bryk, J. and Carr, M. (2018), 'Patterns of ancestral animal codon usage bias revealed through holozoan protists', *Molecular Biology and Evolution* **35**(10), 2499 – 2511.

Southworth, J., Grace, C. A., Marron, A. O., Fatima, N. and Carr, M. (2019), 'A genomic survey of transposable elements in the choanoflagellate salpingoeca rosetta reveals selection on codon usage', *Mob DNA* **10**, 44.

Stamatakis, A. (2014), 'Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies', *Bioinformatics* **30**(9), 1312–1313.

Stoletzki, N. and Eyre-Walker, A. (2007), 'Synonymous codon usage in escherichia coli: selection for translational accuracy', *Mol Biol Evol* **24**(2), 374–81.

Suga, H., Chen, Z., de Mendoza, A., Sebe-Pedros, A., Brown, M. W., Kramer, E., Carr, M., Kerner, P., Vervoort, M., Sanchez-Pons, N., Torruella, G., Derelle, R., Manning, G., Lang, B. F., Russ, C., Haas, B. J., Roger, A. J., Nusbaum, C. and Ruiz-Trillo, I. (2013), 'The capsaspora genome reveals a complex unicellular prehistory of animals', *Nat Commun* **4**(2325), 1–9.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/23942320*

Suh, S. O. and Zhou, J. J. (2011), 'Kazachstania intestinalis sp. nov., an ascosporogenous yeast from the gut of passalid beetle odontotaenius disjunctus', *Antonie Van Leeuwenhoek* **100**(1), 109–115.

Tempel, S. (2012), 'Using and understanding repeatmasker', *Methods Mol Biol* **859**, 29–51.

Thermofisher (2017).
URL: *https://www.thermofisher.com/uk/en/home.html*

Thompson, J. D., Gibson, T. J. and Higgins, D. G. (2002), 'Multiple sequence alignment using clustalw and clustalx', *Curr Protoc Bioinformatics* **Chapter 2**, Unit 2 3.

Tsai, I. J., Bensasson, D., Burt, A. and Koufopanou, V. (2008), 'Population genomics of the wild yeast *Saccharomyces paradoxus*: Quantifying the life cycle', *Proc Natl Acad Sci U S A* **105**(12), 4957–62.

Tucker, J. M., Larango, M. E., Wachsmuth, L. P., Kannan, N. and Garfinkel, D. J. (2015), 'The *Ty1* retrotransposon restriction factor p22 targets *Gag*', *PLoS Genet* **11**(10), 1–29.

Tucker, R. P. (2013), 'Horizontal gene transfer in choanoflagellates', *J Exp Zool B Mol Dev Evol* **320**(1), 1–9.

Tuller, T., Waldman, Y. Y., Kupiec, M. and Ruppin, E. (2010), 'Translation efficiency is determined by both codon bias and folding energy', *Proc Natl Acad Sci U S A* **107**(8), 3645–3650.

van Luenen, H. G. A. M. and Plasterk, R. H. (1994), 'Target site choice of the related transposable elements tcl and tc3 of caenorhabditis elegans', *Nucleic Acids Res* **22**(3), 262–269.

Van't Hof, A. E., Campagne, P., Rigden, D. J., Yung, C. J., Lingley, J., Quail, M. A., Hall, N., Darby, A. C. and Saccheri, I. J. (2016), 'The industrial melanism mutation in british peppered moths is a transposable element', *Nature* **534**(7605), 102–105.

Vaughan-Martini, A., Lachance, M. A. and Kurtzman, C. (2011), *Kazachstania Zubkova (1971)*, 5th edn, Elsevier Science, Saint Louis, book section 34, pp. 439 – 470.

Visher, W. (1945), 'Über einen pilzähnlichen, autotrophen mikroorganismus, chlorochytridion, einige neue protococcales und die systematische bedeutung der chloroplasten.', *Verhandlungen der Naturforschenden Gessellschaft in Basel* **56**, 41– 49.

Walsh, A., Kortschaka, R., Gardner, M., Bertozzia, T. and Adelsona, D. (2013), 'Widespread horizontal transfer of retrotransposons', *PNAS* **110**(3), 1012 – 1016.

Wang, X. and Liu, X. (2016), 'Close ecological relationship among species facilitated horizontal transfer of retrotransposons', *BMC Evol Biol* **16**(201), 1–16.

Warren, I. A., Naville, M., Chalopin, D., Levin, P., Berger, C. S., Galiana, D. and Volff, J. N. (2015), 'Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates', *Chromosome Res* **23**(3), 505–31.

Wessler, S. R. (2006), 'Transposable elements and the evolution of eukaryotic genomes', *PNAS* **103**(47), 17600 – 17601.

Wilson, M. H., Coates, C. J. and George, A. L. J. (2007), 'Piggybac transposon-mediated gene transfer in human cells', *Molecular Therapy* **15**(1), 139 – 145.

Wolfe, K. H., Armisen, D., Proux-Wera, E., OhEigeartaigh, S. S., Azam, H., Gordon, J. L. and Byrne, K. P. (2015), 'Clade- and species-specific features of genome evolution in the Saccharomycetaceae', *FEMS Yeast Res* **15**(5), 1–12.

Wolfe, K. H. and Shields, D. C. (1997), 'Molecular evidence for an ancient duplication of the entire yeast genome', *Nature* **387**(6634), 708–13.

Wright, F. (1990), 'The 'effective number of codons' used in a gene', *Gene* **87**(1), 23–29.

Xiao, W. (2006), *Yeast Protocols*, Vol. 313, 2 edn, Humana Press, Totowa, New Jersey.

Xiong, Y. and Eickbush, T. H. (1990), 'Origin and evolution of retroelements based upon their reverse transcriptase sequences', *The EMBO Journal* **9**(10), 3353–3362.

Yannai, A., Katz, S. and Hershberg, R. (2018), 'The codon usage of lowly expressed genes is subject to natural selection', *Genome Biol Evol* **10**(5), 1237–1246.

Yue, J., Sun, G., Hu, X. and Huang, J. (2013), 'The scale and evolutionary significance of horizontal gene transfer in the choanoflagellate *Monosiga brevicollis*', *BMC Genomics* **14**(729), 1 – 10.

Zhang, L., Yan, L., Jiang, J., Wang, Y., Jiang, Y., Yan, T. and Cao, Y. (2014), 'The structure and retrotransposition mechanism of LTR-retrotransposons in the asexual yeast *Candida albicans*', *Virulence* **5**(6), 655–664.

Zhu, Y., Zou, S., Wright, D. A. and Voytas, D. F. (1999), 'Tagging chromatin with retrotransposons: target specificity of the saccharomyces ty5 retrotransposon changes with the chromosomal localization of sir3p and sir4p', *Genes Dev* **13**(20), 2738–2749.

Zubkova, R. (1971), 'Genus novum saccharomycetacearum e kazachstania', *Botanicheskie Materialy- Gerbariia Instituta Botaniki Akademia Nauk Kazahskoa SSR.* **7**, 53 – 56.