THE ROLE OF TRANSPOSABLE ELEMENTS IN THE EVOLUTION OF FUNGAL ENDOPHYTE GENOMES

A thesis presented in partial fulfilment of the requirements for the degree of

Master of Science

in

Genetics

at Massey University, Palmerston North,

New Zealand.

Kelli Louise Fukumi Smith

2022

Abstract

Transposable elements (TEs) are mobile DNA sequences that can catalyse their own replication and movement within a genome. As a result of their proliferation, TEs have become a major constituent in almost all eukaryotic genomes. While TEs were historically overlooked or dismissed as 'junk' DNA, these elements have now been reappraised as important contributors to gene regulation and genome evolution. Markedly, in plant-associated fungi, it has been proposed that TEs regulate expression of genes that mediate the invasion of plants; the localisation of TEs proximal to invasion-mediating genes is proposed to create a niche for accelerated fungal evolution, extending their host-range and assisting in the antagonistic co-evolution with their host plants.

Epichloë is a genus of ascomycete fungi that live in close association with pasture grasses. This symbiosis can provide the host plants with profound bioprotective benefits such as increased resistance to drought, herbivory and pests. Hence, there is considerable interest in developing novel *Epichloë* strains with improved host-range. However, TEs in *Epichloë* genomes have been considerably inactivated by host genome defences, thus it was unclear whether active elements remain in this genus.

In this project, I have curated a high quality library of TEs in three closely related strains of the *Epichloë typhina* species complex. Using this data, I have demonstrated lineage-specific activity of TEs that have contributed to genome evolution in the recent history of this genus. Furthermore, I have identified sets of TEs that are enriched near virulence-related genes. The work produced here will serve as foundation for future studies to elucidate regulatory roles of TEs in *Epichloë*.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisors, Dr. David Winter and Professor Patrick Biggs. Thank you for your time, guidance, and invaluable wealth of knowledge. Thank you for teaching me more than I ever imagined I could learn, and for celebrating every milestone that paved the way to a complete thesis.

Dr. Julie Blommaert, your guidance and expertise during the TE library curations formed the key part to all my research. Thank you so much for making this project possible.

Thank you to the Cox lab group. Your near-inhuman level of productivity and expertise is incredibly inspiring and teaching. Thank you for all the feedback, and all the cake. To the Massey staff, my former lecturers, and the incredible administration team, you have all been an indispensable part of this journey.

Thank you to the Marsden Fund, managed by Royal Society Te Apārangi, for funding this project. Thank you also to the Joan Dingley Memorial Scholarship in Mycology, awarded by Massey University, for supporting my studies.

My Boffin Lounge officemates, you are the strangest, greatest group of people I have ever met. Thank you for the friendship, shared frustrations, potluck dinners, lockdown zooms, 10x daily tea runs, and of course our office mascot, Balthazar.

To my sister, Salli, and all my loved ones, thank you for all the times you made me laugh until I forgot that I was stressed.

Last but certainly not least, thank you to my mum, Masako, for the ceaseless encouragement. You kept telling everyone that I studied computer science, but you declared it with pride. I promise I'll translate my thesis to Japanese for you.

Contents

List of Figures

List of Tables

List of Acronyms

5mC 5-methylcytosine

AP Aspartic Protease

bp Base pair

CDD Conserved Domain Database

cDNA Complementary DNA

CNV Copy Number Variant

DNA Deoxyribonucleic Acid

DSB Double Strand Break

Ecl *Epichloë typhin*a subspecies *clarkii*

Ety *Epichloë typhina* subspecies *typhina*

Epo *Epichloë typhina* subspecies *poae*

INT Integrase

kb Kilobase

LINE Long Interspersed Nuclear Element

LTR Long Terminal Repat

MITE Miniature Inverted Repeat Transposable Element

MTase Methyl transferase

ORF Open Reading Frame

pol II RNA polymerase II

RIP Repeat Induced Point Mutations

RNA Ribonucleic acid

rRNA Ribosomal Ribonucleic Acid

RNAse Ribonuclease

TE Transposable Element

TIR Terminal Inverted Repeat

TSD Target Site Duplication

UTR Untranslated Region

List of Programmes and Software

Name	Version	Reference
Advanced Consensus Maker	web	[1]
AliView	v1.26	[2]
Antibiotics and Secondary Metabolite Analysis Shell (antiSMASH)	fungiSMASH	[3]
BedTOOLS	v2.27.1	[4]
Biopython	v1.79	[5]
Basic Local Alignment Search Tool (BLAST)	v2.0.0+, blastn	[6]
National Center for Biotechnology Conserved Domain Database (NCBI CDD)	web	[7]
CDhit	v4.8.1	[8]
CENSOR	web	[9]
EffectorP	v2.0	[10]
Funannotate	v1.6.0	[11]
LAST	web	[12]
Multiple Alignment using Fast Fourier Transformation (MAFFT)	web	[13]
minimap2	v2.2	[14]
Occultercut	v1.1	[15]
Python	3.8.10	[16]
R	v.4.0.0	[17]
R Studio	v1.2.5042	[18]
RepeatMasker	v4.0.6	[19]
RepeatModeler2	v2.0.2	[20]
RepeatModeler4.pl		
SignalP	v5.0	[21]
Tandem Repeat Finder	web	[22]

Chapter 1

Introduction

1.1 The widespread success of transposable elements in eukaryotic genomes

For many years, scientists were perplexed by the discordance between the size of an organism's genome and the complexity of the organism [23]. Why do some 'simple' organisms harbour genomes that are far larger than comparatively 'complex' organisms? Why do organisms with relatively few genes have such large genomes? [24] Today, it is widely accepted that genome size does not correspond to total gene number or organismal complexity. Large eukaryotic genomes harbour stable genic regions that are highly conserved by virtue of purifying selection, however, these conserved regions often account for only a portion of the total genome [25]. Much of the remaining genomic space has since been resolved to be non-coding DNA such as degenerative gene copies (pseudogenes), intergenic regions, or sequences of structural or unknown function [26–28]. Above all, it is widely accepted that much of the variation in eukaryotic genome size is owed to non-coding dynamic elements called Transposable Elements (TEs) [29–31].

TEs are mobile DNA sequences with the capacity to replicate and transpose from one chromosomal locus to another. These elements persist in the genome via 'selfish' means: many elements encode all the necessary components to independently recruit endogenous host machinery and facilitate their own replication and transposition [32]. As a result, TEs have shown remarkable success, particularly in eukaryotes, where they have invaded virtually all eukaryotic genomes with only a few exceptions (apicomplexan protists; [33]). The proportion of a genome occupied by TEs varies greatly between species, but most notably, half the human genome [34] and nearly 85% of the maize genome [35] are comprised of TEs. These TE populations consist of both actively transposing elements, and immobile, decaying relics of TEs.

The genetic variation induced by TEs differs to those caused by single nucleotide polymorphisms (SNPs), copy number variants (CNVs), and other common mutagenic mechanisms [36]. Most notably, TEs cause very large insertions ranging

from a few hundred to thousands of base pairs in length. Unlike CNVs, the sequence content of new TE insertions is non-random and encompasses the complete regulatory and coding sequences encoded within the element. Thus, each new TE copy has the potential to transpose again into a new genomic location.

The degree to which TEs are able to proliferate within a genome is governed by both properties intrinsic to the elements themselves, and properties of their host genome [32]. TE-intrinsic factors encompass the capacity of the TE to promote its own transposition and evade negative selection. Although TEs encode all or part of the machinery required to proliferate in a genome, it has been hypothesised that some elements may have the capacity to self-regulate their own copy numbers [37]. Thus the rate of proliferation may be balanced against negative effects on host fitness in a way that would ultimately decrease propagation of a given TE [38]. Genome-intrinsic factors refer to the restriction of TE propagation by the host genome; this occurs by virtue of negative selection against the phenotypes created by the action of TEs, and by specialised TE-defence mechanisms encoded by the host genome. A partial release of factors that constrict TE transposition is proposed to be a key characteristic in very large genomes [39]. TE-intrinsic and genome-intrinsic factors underpin the rate at which TEs can proliferate, the rate at which they are negatively selected against by the host genome, and the rate at which they can accumulate and reach fixation in the genome [32]. These factors in turn are affected by external factors. Notably, TE proliferation and TE-induced mutations are frequently associated with exposure to environmental stressors [40–43].

In addition to proliferating within a genome, TEs can spread between hosts in two ways. The first, vertical inheritance, passes TEs from parent to offspring by virtue of TE proliferation in the germline. The second, horizontal transfer, enables the transfer of TEs across mating barriers. Horizontal transposon transfer is widely reported within and between kingdoms and is widely accepted to occur with higher frequency than horizontal gene transfer in eukaryotes [44–48]. Taken together, the remarkable

persistence of TEs underpins a majority of the variation in eukaryotic genome size, and the dynamic nature of these elements plays a paramount role in eukaryotic genome evolution [39, 49].

1.2 TE Classification

While the proportion of genomes occupied by TEs varies significantly between species, there is also a great diversity in the population of TEs within a given genome. Transposable elements can be classified into subgroups based on structural and enzymatic features. These factors determine their mechanism of transposition and, by extension, their impact and success in host genomes [50]. As a result, TE-invaded species each have a TE repertoire consisting of numerous TEs of different classes, each with variation in the copy number, activity, and age. As these classes of TEs each have unique properties and evolutionary histories, the classification of TEs is a crucial step in studying the contribution of these elements to genome evolution.

In 1989, David Finnegan established that at the highest order, TEs can be classified by their transposition intermediate [51]. Class I TEs, or retrotransposons, utilize an RNA intermediate and transpose via a "copy and paste" mechanism. Here, the TE is transcribed as RNA, and the RNA intermediate is then reversed transcribed to cDNA and reintegrated into a new location. Class II TEs, or DNA transposons, employ a "cut and paste" transposition mechanism with a DNA intermediate. These elements are excised from their location, and reintegrated into a new location [32, 51]. Following this discovery, Class I and Class II TEs have been further divided into orders, superfamilies, families, and subfamilies to create a hierarchical classification system (Table 1.1) [50]. Within each subgroup, elements can be further regarded as autonomous or non-autonomous elements. Autonomous TEs are elements that encode the necessary enzymes to promote their own independent replication and transposition. In contrast, non-autonomous elements are unable to transpose independently, however

Class	Order	Superfamily
Class I (Retrotransposons)	**Long Terminal Repeat (LTR)**	Copia Gypsy Bel-Pao Retrovirus ERV
	Long Interspersed Nuclear Element (LINE)	R2 RTE Jockey L1 I
	Short Interspersed Nuclear Element (SINE)	tRNA 7SL 5S
	Penelope-like Element (PLE)	Penelope
	Dictyostelium Intermediate Repeat Sequence (DIRs)	DIRS Ngaro VIPER
Class II (DNA Transposons)	**Terminal Inverted Repeat (TIR)**	Tc1-mariner hAT Mutator Merlin Transib P PiggyBac PIF-Harbinger CACTA
	Crypton	Crypton
	Helitron	Helitron
	Maverick	Maverick

Table 1.1: TE Classification system proposed by Wicker *et. al* (2007). Bolded orders are the focus of this thesis. MITEs, not listed in this table, belong to Class II and are commonly deletion derivatives of TIR elements.

they carry *cis*-acting sequences that enable transposition if the required enzymatic functions are provided by a transposition-competent autonomous element in *trans* [52]. Some non-autonomous elements are deletion derivatives of autonomous elements, hence they retain features of their autonomous counterpart, but have lost domains that enable independent replication. In summary, a myriad classification determinants have paved the way to extensive groupings of TEs that are widely present in eukaryotes. In this review, I will focus primarily on the four groups of TEs that are present in the fungal genomes focal to this project: long-terminal repeat (LTR) retrotransposons, non-LTR long interspersed nuclear elements (LINEs), terminal inverted repeat (TIR)

DNA transposons, and miniature inverted repeat transposable elements (MITEs) (Figure 1.1).

1.2.1 Class I: LTR retrotransposons

Long terminal repeat (LTR) elements are a major subclass of class I retrotransposons. They are typically 5-7 kb long [53], however in cases they can be as short as 100 bp, or as long as 22 kb [50, 54]. These elements are found in almost all major eukaryotic taxonomic groups [55] and resemble retroviruses, with which they are believed to share an evolutionary relationship [53]. A defining feature of LTR elements is the identical long terminal repeat sequences ranging from 100 bp to more than 5 kb in length that flank the inner portion of the TE. The LTRs typically comprise of a canonical 5'-TG and CA-3' structure and create target site duplications (TSD) of 4-6 bp that arise as a result of a fixed-length staggered cut induced in the donor DNA during TE insertion (Figure 1.1) [56].

LTR retrotransposons are autonomous elements that are transcriptionally regulated by a RNA polymerase II (polII) promoter found within the LTR. These elements generally harbour two open reading frames (ORFs): a polyprotein (*pol*) and a *gag* viral structure protein [32]. The *pol* ORF encodes enzymes that facilitate the movement of the element: a reverse transcriptase (RT) that binds to a primer binding site downstream of the 5' LTR and transcribes the RNA copy into cDNA for transposition; integrase (INT) that integrates element into the host DNA; aspartic protease (AP) that cleaves the polyprotein to form smaller protein products from the large transposon transcripts; and RNAseH (RH) which cleaves the RNA/DNA hybrid formed during transposition. Additional ORFs of unknown function have also been reported. Copia and Gypsy are two superfamilies within the LTR retrotransposons order, and differ only in the arrangement of enzyme-encoding genes within their *pol* ORFs (Figure 1.1) [32, 57].

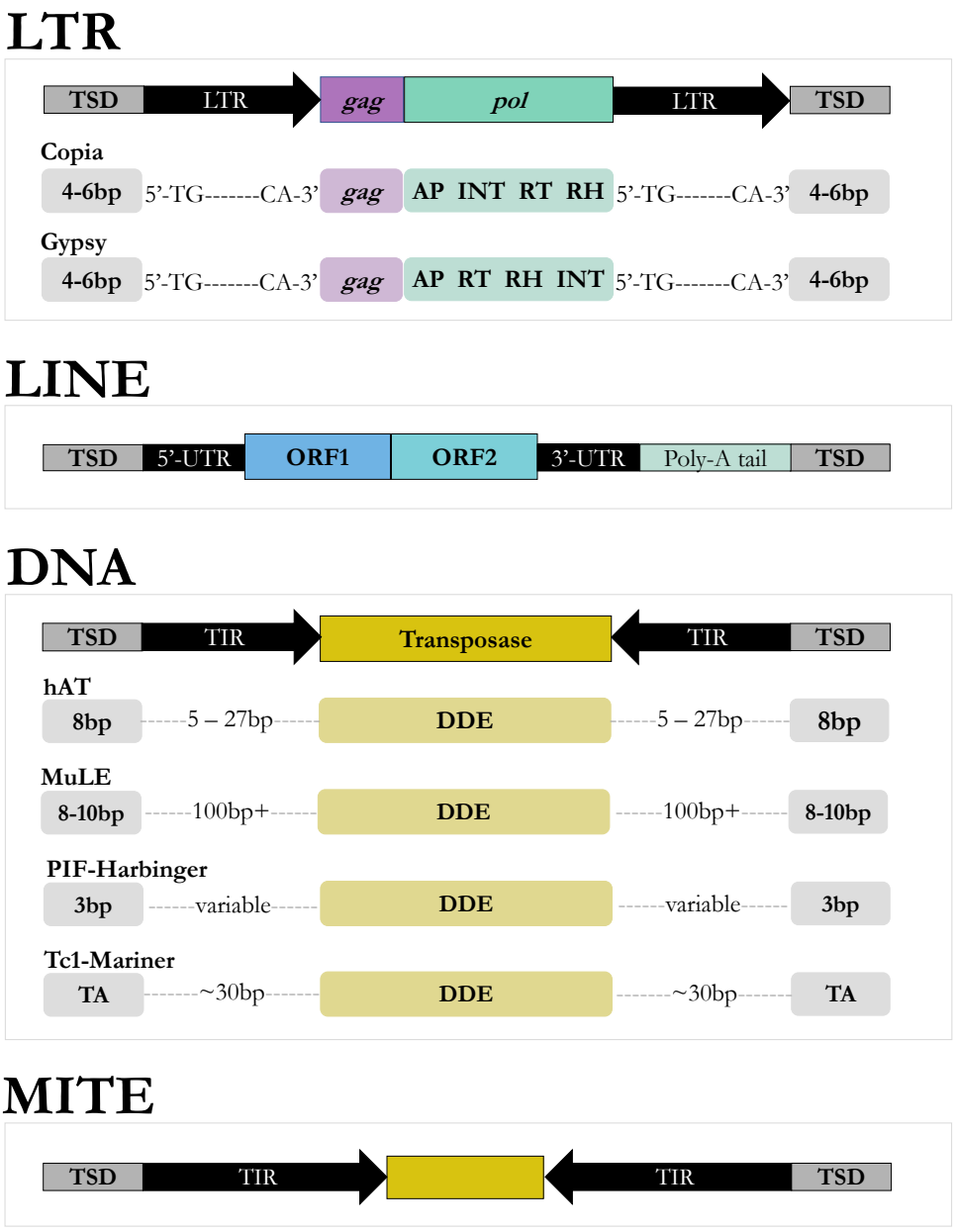

Figure 1.1: Overview of basic TE structure. Members of the LTR family harbour target site duplications (TSDs) of 4-6 bp, and characteristic long terminal repeat (LTR) sequences that show a 5'-TG CA-3' motif. The internal portion contains open reading frames (ORFs) for gag, and pol. Pol encoded enzymes differ in order between the two superfamilies. LINE elements harbour variable TSDs, two ORFs flanked by untranslated regions (UTRs), and a characteristic poly-A tail. DNA and MITE elements have TSDs of variable length and motifs which flank terminal inverted repeat sequences (TIRs) of varying lengths. Each focal DNA element of this project harbours a transposase characterised by a DDE amino acid triad. MITEs do not possess transposases.

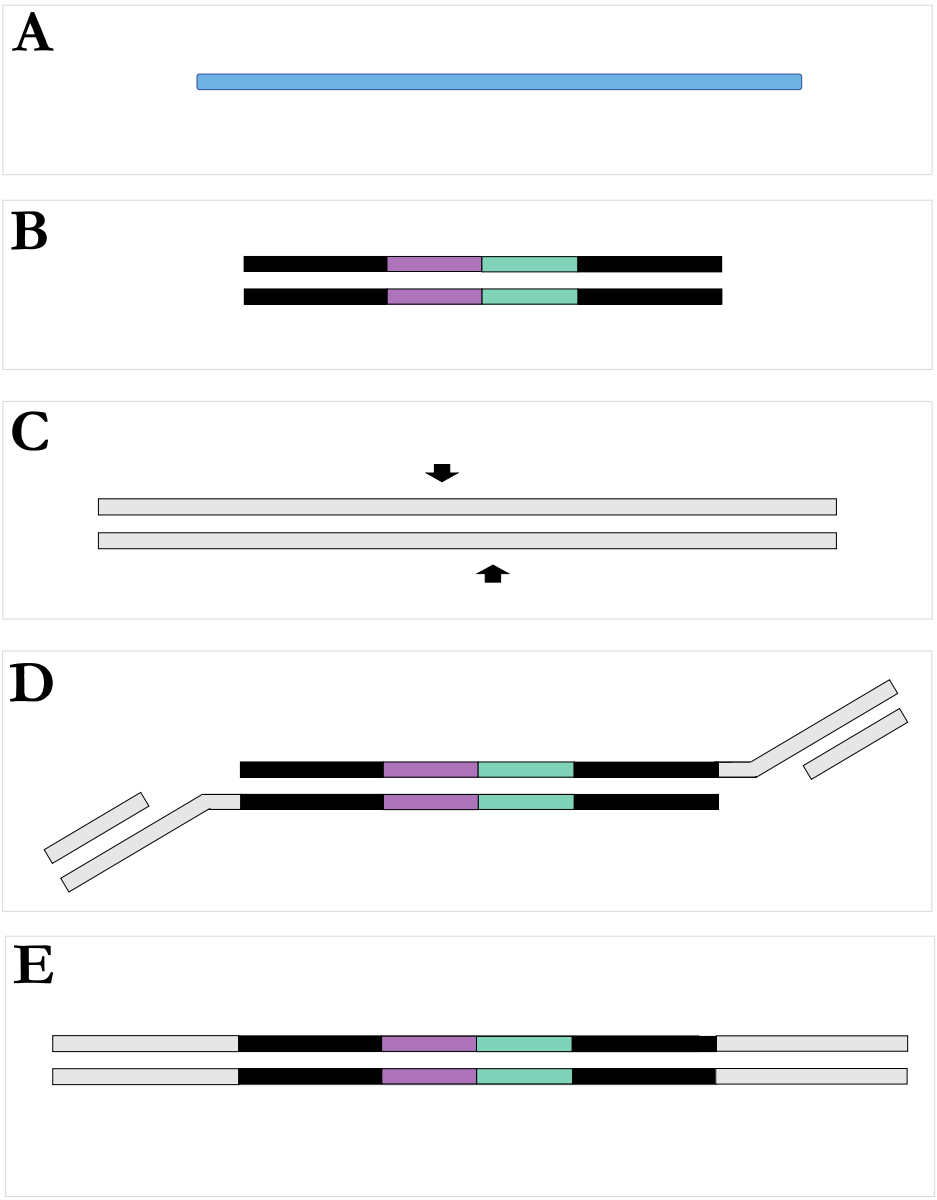

Figure 1.2: LTR retrotransposition. (A) An RNA intermediate of the LTR element is transcribed by endogenous pol II. The RNA harbours a primer binding site downstream of the 5' LTR that serves as a binding site for encoded RT (not shown). (B) the intermediate is reverse transcribed into cDNA in a multistep process resulting in a double stranded DNA copy of the element. This cDNA copy harbours the complete LTRs (black) and conserved domains (purple and green). (C-D) Encoded intergrase mediates the transposition; a fixed length staggered cut (black arrows) is induced in the host-DNA (grey), and the newly synthesised cDNA copy is integrated into host DNA. (E) Endogenous repair machinery repairs the breaks to complete the transposition process.

Transposition and integration of LTR elements occur via a cleavage and strand transfer method mediated by the integrase (Figure 1.2) [53, 58]. They are reported to preferentially insert within other LTRs and AT-rich regions [55, 59]. As LTR elements amplify through a copy and paste mechanism, the transposition events lead to increasing copy numbers.

1.2.2 Class I: LINEs

Long interspersed nuclear elements (LINEs) are a diverse group of autonomous Class I elements. These elements lack LTRs, and are thus referred as non-LTR retrotransposons. LINEs are the most common Class I elements in mammalian genomes, however they are relatively rare in plants and fungi in comparison to LTR elements. LINEs are distinguished into five superfamilies named L1, R2, RTE, I, and Jockey, which are each further divided into many subclasses [50]. Like LTR elements, LINEs encode a reverse transcriptase and in some cases, an RNAseH enzyme. However, they typically lack the protease and integrase seen in LTR retrotransposons [60].

Full-length LINES are 3-7 kb in length and harbour two open reading frames, ORF1 and ORF2. An RNA binding protein within ORF1 acts to stabilise RNA and assist in strand transfer during the reverse transcription process [53]. The second ORF (ORF2) encodes a nuclease to cleave genomic DNA, and a reverse transcriptase to convert the RNA intermediate to cDNA. In some cases (superfamily I), an RNAseH may also be in ORF2. One characteristic of LINEs is a poly-A tail or adenine-rich sequence at the 3' end of the element that assists in transposition (Figure 1.1) [53, 61].

Like LTR retrotransposons, LINE transposition directly correlates with copy number. LINEs transpose via target-primed reverse transcription, and are transcriptionally regulated by a promoter in the 5'-UTR [62, 63]. LINEs lack the primer binding site reported in LTRs. Instead, the single stranded break in host DNA that is induced by the encoded endonuclease frees a 3'-hydroxy group that then serves as a primer.

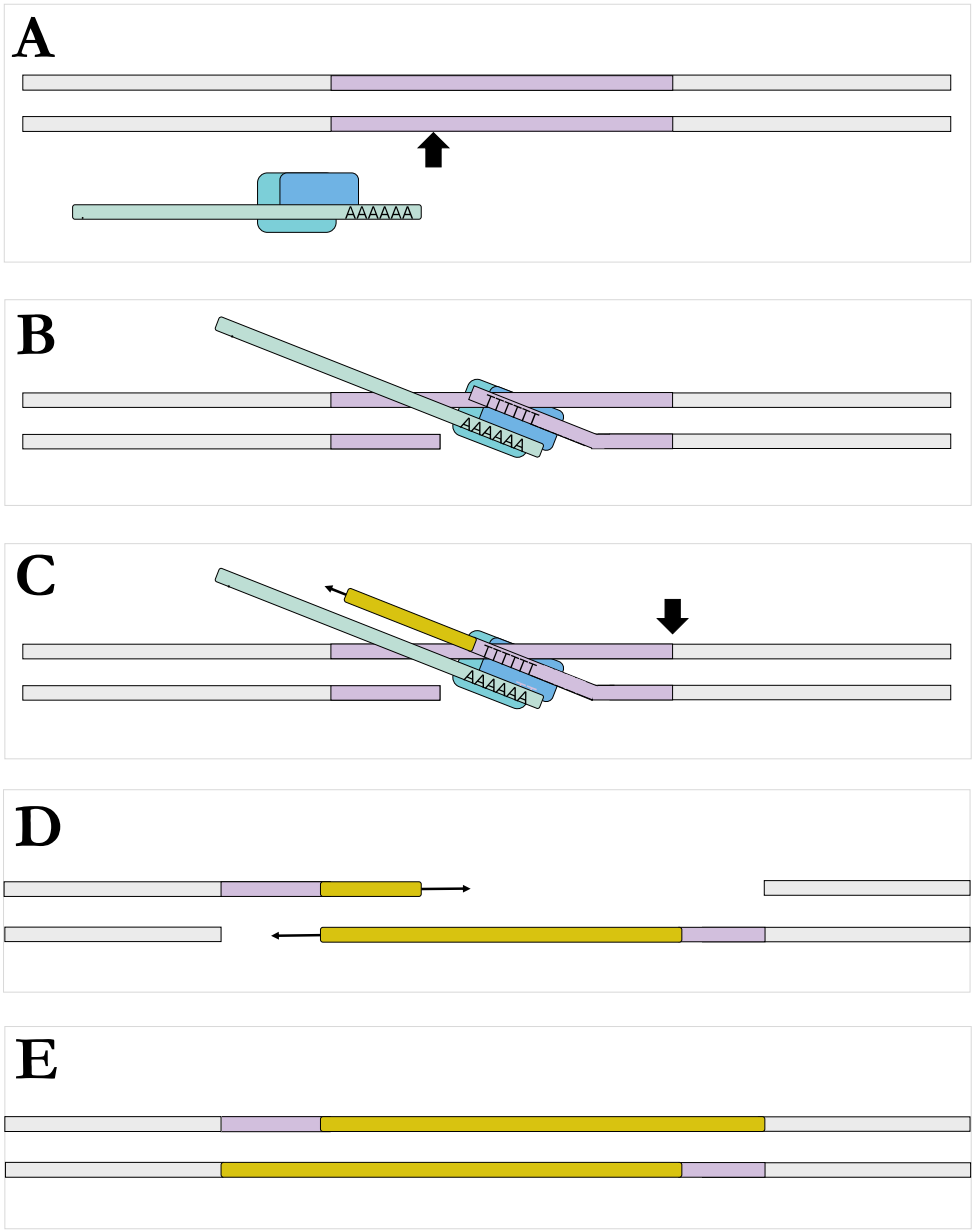

Figure 1.3: LINE retrotransposition. (A) The LINE-encoded nuclease (black arrow) forms a single stranded break in an AT-rich region of host DNA and frees a 3'-OH group that serves as a primer for RT (not shown). (B) The LINE RNA intermediate (green), stabilised by ORF1 and ORF2 proteins (blue), harbours a poly-A tail on the 3' end that can bind to the T-rich region of host DNA. (C) Encoded RT synthesises the first strand of DNA, and the second strand of host DNA is nicked by the encoded nuclease a few base pairs away from the first break. (D) The second strand is synthesised. (E) Endogenous repair machinery fixes the gaps.

The transposition process is then completed by LINE-encoded enzymes and endogenous host-repair machinery (Figure 1.3) [53]. The integration of LINEs into host DNA generates a truncated 5' region in the element, hence new insertions are often partial and non-functional.

1.2.3 Class II: DNA transposons

Class 2 elements, or DNA transposons, are present in almost all eukaryotes [50]. These elements mobilise via a DNA intermediate. Currently, Class II elements can be categorised into two main subclasses by virtue of the transposition mechanism (subclass I and subclass II), and are further divided into four major groups (orders) on a basis of enzymatic differences. As the focus of this thesis is fungal TEs, here I will focus on TIR DNA transposons, a member of the subclass I "cut-and-paste" elements. These elements are mobilised by DDE transposases, named after the DDD or DDE amino acid triad found within the protein domain [32]. As a result, they are often referred as DDE transposons, however, the classification of these elements by virtue of DDE transposases has been debated, as DDE transposases may be one of the most abundant and ubiquitous genes in nature [64], and have been found in non-DDE TE groups [50]. TIR elements are characterised by terminal inverted repeats (TIRs) that flank the coding sequence of the TE. Members of this order can be further distinguished into superfamilies by virtue of features such as TIR length, TSD length, and the TSD motif that is representative of target site preference (Figure 1.1).

Transposition of TIR elements typically involves a double stranded break at each end of the transposon, and at the new target site in host DNA. Transposition is mediated by the encoded transposase and endogenous repair machinery (Figure 1.4) [65]. In contrast to Class I elements, the "cut-and-paste" mechanism used by Class II elements does not itself lead to an increase in copy number. Nevertheless, these elements can proliferate within a genome. In particular, DNA elements often transpose

during DNA synthesis, copying themselves from replicated to as-yet-replicated regions of the replication-fork and creating new copies [50].

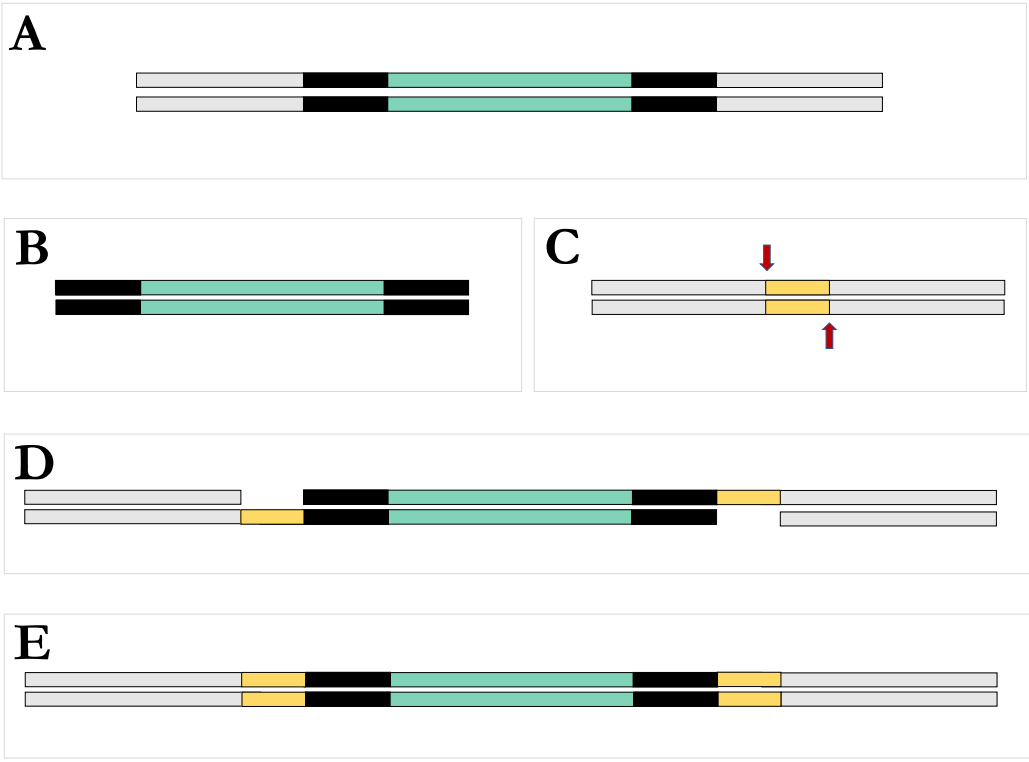

Figure 1.4: Basic transposition of Class II cut and paste elements. (A-B) Excision of the DNA element from the host DNA (grey) is catalysed by transposase. The excised element contains the complete TIRs (black) and internal sequence (green). (C) Both strands of the host DNA are cleaved by the encoded endonuclease at its sequence-specific target site (yellow). (D) The excised DNA intermediate is reintegrated into the new chromosomal location, and (E) the fixed staggered gaps are repaired by endogenous DNA polymerase and DNA ligase, forming target site duplications representative of the target site that flank the newly integrated element.

1.2.4 Class II: MITEs

Miniature inverted repeat transposable elements (MITEs) are non-autonomous deletion derivatives of DNA elements that can be structurally characterised by their terminal inverted repeats, high copy number, and small size (50-800 bp) [66]. They are abundant in plants and fungi and reported to show target site bias close to genic regions. Consistent with this, MITEs are proposed to play a role in gene regulation [66–69], and are over-represented near transcription start sites of upregulated genes [70]. Although MITEs lack the capacity to independently transpose, they are present in high copy numbers, due to a process termed cross mobilisation; as these elements are often deletion derivatives of Class II elements, they can retain transposase binding sites of autonomous DNA transposons and utilise this as a means for mobilisation [71, 72].

1.3 TEs as mutagens

Despite their ubiquity among eukaryotic genomes, TEs were historically dismissed as "junk DNA", or selfish genomic parasites: non-functional DNA sequences that serve no benefit to the host [73]. Pioneering work by Nobel laureate Barbara McClintock first discovered TEs and their impact on the host genome [74, 75]. McClintock, while studying the mosaic colour patterns in maize, noted that the frequency of kernel color changes were too rapid to attribute to classic genetic mutations. McClintock proposed genes were able to "jump" across or within chromosomes, consequently impacting the expression of genes, and turning physical traits "on" or "off." In this work, she identified two genetic loci responsible for these changes observed in maize: *Activator (Ac)*, and *Dissociation (Ds)*. This work lead to the discovery of what is now known as the *Ac/Ds* transposable element system [76]. *Ds* (now known to be a non-autonomous DNA transposon) was discovered first, and could mobilise when associated with *Ac* (a self-mobilising, autonomous DNA transposon). McClintock reported that both *Ac* and *Ds* could insert into genic regions and alter the kernel colour of maize. Taken together, her results suggested that the genome harboured dynamic loci that contributed to phenotypic variation. At the time of discovery, McClintock's hypothesis remained ignored; the theory of a dynamic genomic landscape deviated from the existing understanding of gene expression, and was not well-received. However, more than 30 years later in 1983, she was awarded a Nobel Prize in honour of her discoveries [77].

Today, it is recognised that in most cases, the effects of TEs are 'nearly neutral' with regard to their host's fitness. Thus, these elements are able to accumulate and reach fixation within the host genome [78, 79]. However, the extent of TE impact can span the continuum between beneficial and deleterious. Beneficial insertions may undergo molecular domestication to serve function in the genome, and deleterious mutations are rapidly selected against. Above all, as TEs are potent sources of polymorphisms, their activity can have a profound impact on the fitness of the host genome.

Their insertion, relocation, and repression can give rise to functional consequences such as alterations to gene expression and regulation, large scale structural rearrangements, and mutant phenotypes [80–82].

1.3.1 TE-induced mutant phenotypes

Mutant phenotypes and genetic diseases that arise as a consequence of TE activity are well documented in the human genome. LINEs, the only autonomous elements that remain active in man, are responsible for over 120 genetic diseases [81]. A majority of these diseases are a result of insertional mutagenesis that results in gene inactivation. This may occur when the element inserts into an exon, or may arise as a result of an intronic insertion that results in aberrant splicing [81]. Further, LINE-1 insertions in human and primate genomes have been reported to induce target-site deletions as a result of their target-primed reverse transcription [83, 84].

Due to the threat of TEs on host fitness, TEs are typically silenced by epigenetic modifications such as DNA methylation [85]. While it is unsurprising that new, active TEs can cause disease, epigenetic dysregulation and derepression of TEs are also reported to underpin multiple diseases such as cancers and autoimmune disorders in man [31, 86].

1.3.2 Rearrangements, gene acquisition, and transduction

The movement of TEs within a genome can induce novel chromosome restructuring, giving rise to karyotypic variations or acquisition of new genes. As TEs present the genome with multiple near-identical sequences, strand-transfer can occur between these homologous sequences [87]. This enables ectopic recombination between transposable elements and is reported to produce chromosomal rearrangements. Further, a high TE density has been reported to correlate with high recombination rates and chromosome

length polymorphisms [82, 88, 89]. It is important to note that immobile TEs can also serve as a substrate for ectopic recombination due to sequence similarity, thus deactivated TEs can still indirectly contribute to genome evolution. In addition to inactive TEs and newly inserted TEs, the host genome can be negatively impacted erroneous double strand break (DSB) repair following the excision of a DNA element during transposition [90].

The transposition of TEs can lead to the formation or acquisition of new genes. For example, during transposition, flanking nucleotide sequences can be mobilised alongside the TE. On rare occasions, captured fragments such as exons and promoters may be transduced into existing genes. In turn, this has reportedly given rise to gene fragment duplications [91], and the formation of new chimaeric genes [92]. Non-mendellian inheritance of genes via transposon-mediated horizontal gene transfer has also been reported [93].

1.3.3 Molecular domestication

TE-invaded genomes are replete with relics of TEs that have mutated and immobilised throughout evolution. It is widely accepted that among these ancient TEs, some have been co-opted to serve a cellular function in the host genome. Here, TE-derived sequences are captured by the host genome and repurposed to serve functions that are beneficial for host fitness [80, 94, 95]. These TE-derived sequences are widely reported to play a role in gene-regulatory networks. For example, TEs have been demonstrated to contain binding sites for transcription factors, and up to 40% of transcription factor binding sites in the mouse and human genome were derived from TEs [96]. Further, epigenetic studies have revealed TE-derived transcription start sites and TE-derived promoters to act as tissue-specific promoters. [97, 98]. Other functions of co-opted TEs include domestication of TE nucleases, and TE-derived non-coding regulatory RNA transcripts that modulate transcriptional and post-transcriptional alterations to

host gene expression [99, 100].

1.4 TEs in plant-associated fungi

Plant-associated fungi are ideal systems to study the contribution of TEs to genome evolution. These species typically have relatively compact genomes, which nevertheless house a diversity of TEs. In addition, it has become increasing clear that these TEs play an important role in the co-evolution between a given fungus and its plant host(s) [67]. There are already considerable genomic resources built around plant-associated fungi. Fungal phytopathogens are well studied due to their economic and agricultural importance, and the threat they represent to global food security [101]. Conversely, plant-symbiotic fungi can provide their hosts with increased tolerance to abiotic stressors, thus there is motivation to exploit these factors for agricultural applications [102, 103].

In many characterised fungal species, TEs play an important role in genome structure and plasticity. In particular, the activity of TEs is proposed to accelerate the evolution of fungal genes that mediate the colonisation of plants. By extension, this allows the fungus to outpace detection by its host plant, assists in the host-microbe co-evolution, and potentially extends the host range of the fungus [104]. This phenomenon is believed to occur by virtue of both TE-intrinsic and genome-intrinsic factors that act to expand and restrain TE populations.

1.4.1 Fungal defences against TEs

The potentially deleterious impacts of TE integration have given rise to adaptive defence mechanisms in eukaryotic genomes that act to prevent further transcription and proliferation of TEs. These defence mechanisms include transposon methylation [105], meiotic silencing and quelling [106], and repeat-induced point mutations (RIP; [107]).

RIP, a homology-dependent genome defence system, is a defence mechanism unique to fungi that targets repeat sequences for hypermutation, and can facilitate epigenetic silencing of targeted sequences via methylation [107]. RIP promotes irreversible C-to-T transitions in large repeat sequences (>400 bp) that share >80% identity [108, 109]. This is believed to deactivate and immobilise TEs over the course of one or several generations [110].

RIP targets cytosines in a preferred nucleotide context, often CpA dinucleotides with the favoured 3' adjacent nucleotide differing between species, leading to an accrual of AT-rich genomes with characteristic dinucleotide patterns [109, 111]. The current understanding of RIP was pioneered by work in *Neurospora crassa.* Selker and colleagues first observed a tandem duplication in 5S rRNA with unusual methylation of cytosines, and robust C-to-T mutations throughout the genome [112]. In following work, Selker *et al.* [113] defined that RIP involves detection of duplications in chromosomal DNA, specifically in haploid germline nuclei in pre-meiotic mitosis events. This detection occurs regardless of factors such as the origin of duplication, transcriptional state, and relative position in the genome. Cytosines on both strands of the duplex are mutated, and occasional mutation of cytosines in neighbouring non-repetitive regions may occur [114]. The process of RIP leads an accrual of C-to-T transitions, hence AT-rich regions are biochemical hallmarks of RIP-deactivated repeat sequences [114].

The underlying molecular mechanism of RIP remains poorly understood in most fungal species [115]. In *N. crassa*, concomitant methylation of RIP-detected cytosines was reported to be catalysed by RID, a putative C5-cytosine methyltransferase (5mC MTase). The putative model suggested the signal for cytosine methylation is the DNA duplication itself [116]. The C5-cytosine methylation (5mC) seen in *N. crassa* is conserved in all domains of life [117], and the methylation mediated-deamination of cytosines is known to promote C-to-T transitions, consistent with the action of RIP [112, 118, 119]. However, DNA and chromatin modifications of the fungal epigenome remain poorly investigated beyond the findings that 5mC is enriched in regions of repetitive

DNA and TEs. In addition, further investigation into the phylogenetic distribution of 5mC MTases determined unique combinations of 5mC MTases possessed between fungal clades; the gain and loss of genes encoding these enzymes across fungi have been shaped by gene duplications and losses, and RIP-associated MTases such as RID are more recently derived in fungi from ancestral maintenance enzymes [117]. Consistent with the non-uniform composition of MTases across fungal taxa, the methylation associated with RIP is suggested to occur by virtue of functionally distinct methyltransferases across different fungal species. For example, there is an absence of RID in several RIP-competent species [114]. In these instances, combinations of unique non-RID MTases can mediate RIP [114]. Taken together, these findings make apparent that while the mechanisms of RIP require further investigation, there is heterogeneity of the RIP mechanism across fungal species.

In addition to differences in the RIP mechanism between fungi, the extent to which RIP affects the genome also varies between species. Interestingly, the abundance of MTases does not positively correlate with levels of 5mC across fungal genomes [117]. A near complete absence of active TEs by virtue of RIP is one notable feature of the *N. crassa* genome, however, evidence of RIP remains undetected in its close relative *Sordaria macrospora* [120]. It must also be noted that RIP can be costly to the host, and the RIP-mediated defense against TEs can be lost via the loss of MTases [121].

While the primary role of RIP is to immobilise TEs, the impact of RIP can extend to non-TE genetic elements. For example, as RIP targets regions of high sequence similarity, it may prevent new genes that could otherwise arise as a result of duplications. In addition, "leaky RIP" may induce C-to-T transitions in genes proximal to TEs, or promote TE-mediated silencing of genes proximal to TE clusters [122].

1.4.2 TE organisation in fungal genomes

Fungal pathogens and symbionts are highly diverse in both lifestyle and in the interactions they share with their host plant. However, a common characteristic in plant-fungi interplay is the recognition of the fungus by the plant innate immune system. In order to dampen a host defence response and successfully colonise the plant, plant-associated fungi must deploy secreted proteins, collectively referred to as effectors. Upon delivery into the plant, effectors modulate plant immunity and govern the invasion process by altering host cell structure and function [123–125]. Thus, effectors inhibit plant defence and are used similarly by both pathogenic and mutalistic fungi [126, 127]. In addition to effector genes, fungal secondary metabolites, which perform an array of functions, are also crucial players in interactions with other organisms [128]. Comparative analyses of effectors and secondary metabolites between fungal genomes have demonstrated aberrant phylogenetic distribution, extreme sequence diversity, and very few homologues between fungal species [126]. These observations substantiate the theory that genes involved in the colonisation of plants continually evolve to retain and improve their capacity for plant invasion. Interestingly, effectors and secondary metabolite genes are non-randomly distributed throughout the fungal genome, and have been observed to localise with TEs [129].

TEs are reported to play a profound role in the genome organisation of the pathogenic fungi of plants. While the genomes of fungal pathogens are highly diverse, many species are often described to have 'core' chromosomes, and 'accessory' or 'dispensable' chromosomes. Here, the core chromosomes comprise of highly conserved genes that are essential for the survival of the organism. Conversely, accessory chromosomes are not required for survival, but may confer an adaptive advantage in plant invasion. Accessory chromosomes, often cited in the literature as "lineage-specific", "type B" or "supernumerary" chromosomes, are typically dense in TEs and other repetitive sequences, and are largely devoid of genes [130–132]. However, the genes present in these gene-poor regions often encode effectors and other virulence-specific genes. As TEs

are potent sources of polymorphisms that can induce structural variations and regulate gene expression, these accessory chromosomes and the genes within them are observed to evolve faster than core chromosomes [104, 133]. In addition to TE activity, this rapid evolution may be attributed to a relaxation of selection against mutations in these regions due to low gene density. On occasion, the evolution of accessory chromosomes is so great that they show no detectable synteny even between closely related species that share conserved core genomes [131]. Further, the accessory chromosomes may be present or absent in select individuals within the same population. Thus, although the organism does not rely on accessory chromosomes for survival, it is widely accepted that these accessory regions can be important determinants for virulence and host specificity [134]. This phenomenon is well supported in *Zymoseptoria triciti*, a fungal wheat pathogen, in which deletion of accessory chromosomes negatively impacted the fitness of the fungus during infection of its host [134].

Associations between plants and fungal pathogens are underpinned by an antagonistic co-evolution in which the fungus evolves to evade plant immunity whilst the plant evolves to recognise fungal secreted proteins. Hence, the accelerated evolution of virulence-related genes presents new means of overcoming plant-recognition and resistance [135] and, by extension, may contribute to host-specificity.

The distinct separation of core and accessory chromosomes is not the only way in which conserved genes and virulence-related genes may be organised. Instead, in several characterised species, this same phenomenon is observed as what could be described as a "patchwork" genome in which highly conserved core sequences are interspersed with blocks of divergent accessory regions (Figure 1.5). This structure, termed the "two-speed genome" is well documented in several fungal lineages [67, 82, 136–138]. The accessory regions in these genomes are typically AT-rich as a result of RIP-deactivated TEs [67, 139]. As a result, these two-speed genomes harbour a distinct bipartite structure consisting of alternating regions of AT-rich blocks enriched with TEs and effectors, and gene-rich regions of relatively equal nucleotide composition [104]. It

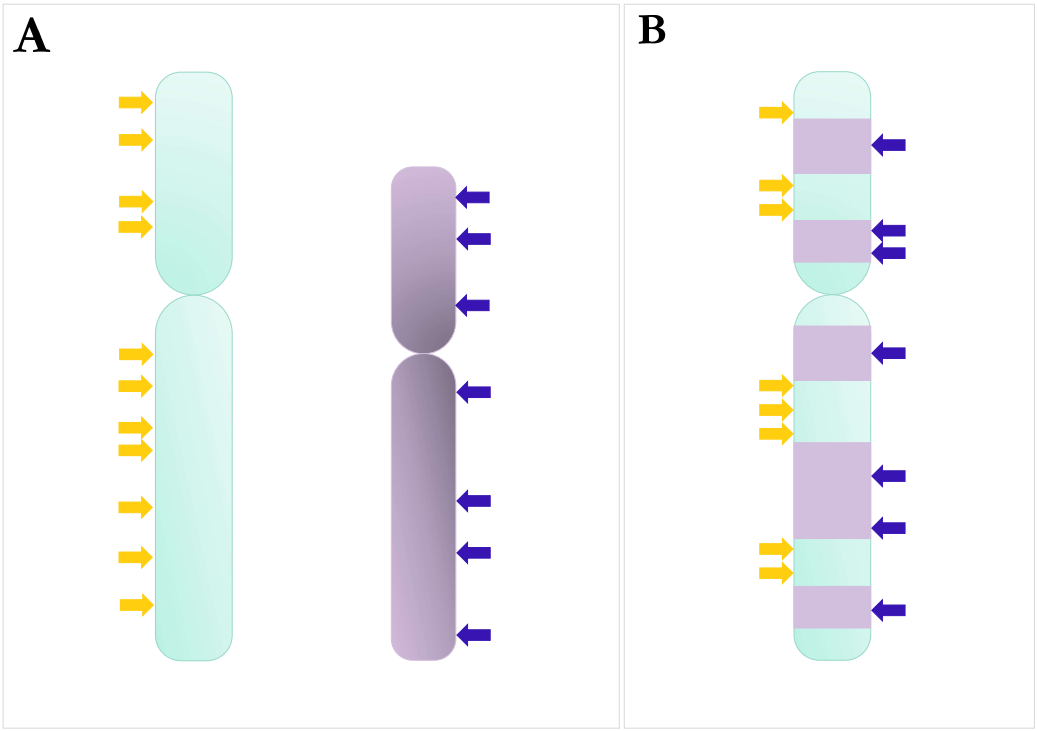

Figure 1.5: Two observed genome structures in plant-pathogenic fungi. (A) The core chromosome (green) is enriched with non-virulence-related genes (yellow) with low or no TE content. The accessory chromosome (purple) is enriched with TEs and genes that mediate virulence and plant-invasion (blue). (B) The two speed genome structure consists of alternating blocks of gene-rich regions (green) enriched with non-virulence genes (yellow), and TE-rich regions (purple) enriched with virulence-related genes (blue).

is important to note that while several unrelated species have diverged to have two-speed genomes [82, 137, 140, 141], "one-speed" genomes with no compartmentalisation [142], and multi-compartmented "multi-speed" genomes have also been reported [143]. Thus, co-evolutionary selection pressures do not result in one chromosomal structure. The generation of more high-quality genomes across fungal species may provide further insight into genome architectures and elucidate the origin of these compartmentalised structures.

1.5 *Epichloë*

The study of plant-colonising fungi has predominantly been restricted to fungi of agricultural and economic importance. As a result, far less is known about genome organisation of fungi in natural ecosystems. This is largely due to fewer telomere-to-telomere chromosomal assemblies of these populations which are highly valuable when elucidating genomic structure and evolution [130]. An exception to this is *Epichloë*, a genus of filamentous fungi that lives in close association with grasses in the subfamily *Pooideae* [144]. Species within this genus can span the continuum between parasitic and mutualistic; the asexual life cycle exhibits asymptomatic infection that provides the host plant with protective benefits, and the virulent sexual life cycle encloses the host inflorescence in the fungal fruiting structure and can sterilize the host plant [145–148]. The focus of this project is predominantly asexual *Epichloë* species, which are considered to act as mutualists. The systemic infection of these asexual *Epichloë* species can provide the host plant with profound bioprotective benefits such as increased resistance to drought, pathogens, and pests [149–151]. These benefits prompt significant interest in extending the host range of these species for agricultural applications [152]. To further understand how *Epichloë* associates with its host plants, *Epichloë* secondary metabolites have been extensively studied, with fewer studies currently conducted on effector genes [153–157].

Recent studies have characterised TE populations within select species of *Epichloë*. Preliminary research in *Epichloë festucae* prior to this project found that TEs have a profound influence on the genome structure. In addition to creating the distinct, compartmentalised two-speed structure, the AT-rich tracts in the *E. festucae* genome accounted for almost half of all inter-chromosomal contacts, and thus act as key determininants to the 3D structure of the genome. Further, MITEs within *E. festucae* appear to have been transcriptionally active in the recent history of the genus and are proposed to be regulators of gene expression that may contribute to host-adaptation [67, 158]. This hypothesis is strengthened by the overrepresentation of MITEs near

transcription start sites of near genes that demonstrate the largest differential expression in axenic culture and *in planta* growth conditions. However, beyond these findings, the role of TEs in *Epichloë* remains obscure. This is largely due to the fact that most DNA transposons and retrotransposon families have been extensively affected by RIP, and it is unclear whether active elements remain in this genome [67].

1.6 Research objectives

Epichloë species provide an excellent system in which to study the impact of transposable elements on gene regulation and genome evolution, particularly in the context of plant colonisation. The relatively small genomes, considerable repeat content, and diversity of species within the *Epichloë* genus allows whole-genome approaches to elucidate the roles of TEs in the evolution of this genus. In this work, I focus on three sub-species belonging to the *Epichloë typhina* species complex: *E. typhina* subsp. *clarkii* (Ecl), *E. typhina* subp. *typhina* (Ety), and *E. typhina* subsp. *poae* (Epo). These subspecies are genetically differentiated in natural populations and have adapted to different grass species across a short evolutionary timeline [159, 160]. While phylogenetic analyses have been conducted in *Epichloë* [161], a precise timeline of the divergence of *Epichloë* lineages is yet to be established. Thus, in this work, any reference to the "recent history" of this genus refers to the time lapsed since the divergence between members of this species complex. A currently unpublished phylogentic study (I. Dumville, personal communication) has established the relationships between my focal subspecies, and suggests the divergence has occured within the last 10 million years (Figure 1.6).

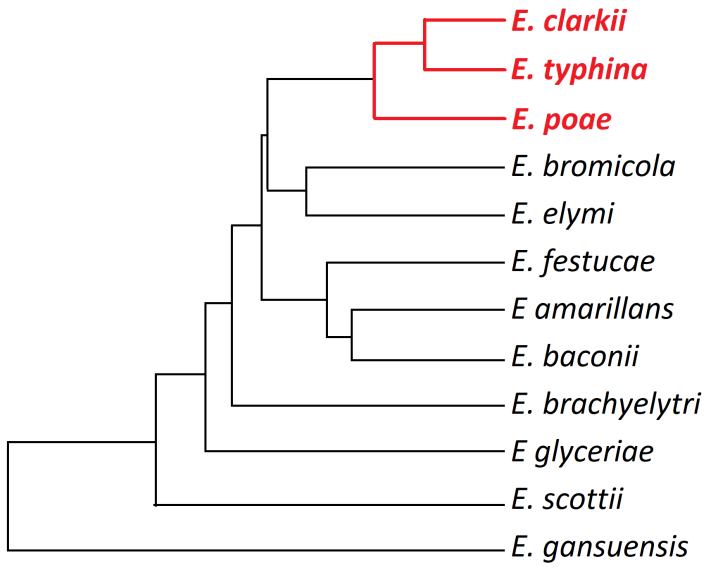

Figure 1.6: A representative model of the phylogenetic relationship shared between the *E.typhina* species complex and its relatives. The focal sub-species of this project are denoted in red. Figure adapted from work by Quenu *et al.* [161]

The objectives of this project were therefore to:

1. Identify TEs in three sub-species of the *Epichloë typhina* species complex and create a high quality custom TE library for these genomes.

2. Determine the TE populations in all three focal lineages, and investigate factors such as the class composition, copy numbers, estimated age of transposition, and distribution of TEs.

3. Infer the association between TEs and genes predicted to play a role in host invasion.

Chapter 2

Manual Curation of a custom TE library for *Epichloë typhina*

2.1 Abstract

Transposable elements (TEs) are major genomic constituents in allmost all eukaryotic genomes. However, the study of these elements presents myriad analytical challenges that require extensive labour to resolve. As a result, current TE analyses cannot keep pace with the ever-growing production of reference genomes, and the study of TEs remain subordinated to that of non-TE genomic components.

An accurate classification of TEs requires high quality databases of known TEs. However, the lack of representative TEs for most non-model organisms substantially contributes to the generation of poorly-annotated or unclassified TEs. As such, it is widely accepted that in order to obtain robust TE annotations, one must use a combination of *de novo*- and homology-based annotations coupled with an extensive manual curation step in which all TEs within a given genome are characterised via conserved structural and enzymatic features.

In this chapter, I have completed the manual curation of an *Epichloë typhina* TE library using three sub-species of the *Epichloë typhina* species complex: (sub-species *clarkii*, *poae*, and *typhina*). This process achieved a substantial decrease in the proportion of unknown elements, precise reclassification of previously incorrectly annotated TEs, and the characterisation of several TEs that can not be identified by automatic-identification software. This work resulted in a high quality library tailored for *Epichloë* genomes, and will serve as a robust database for future TE studies in *Epichloë*.

2.2 Introduction

The development of long-read sequencing technologies has revolutionised the study of TEs. Previously, the repetitive, complex, and nested structure of these elements required reconstruction of full-length TEs from short reads. Further, the TEs present in short-read assemblies were often excluded from analyses, as these reads could not be reliably mapped back to the genome [162, 163]. Long read sequencing technologies have largely overcome these technological hurdle; the increased read-length and quality, alongside the reduced cost of sequencing technologies now sees high quality assemblies being routinely produced for almost any organism. Despite the wealth of TE data produced by long read sequencing, the analyses of these elements still present myriad analytical challenges. Currently available software tools are able to identify repetitive regions in assembled genomes and match structural elements of reference TEs to genome assemblies. However, these programs are not typically able to accurately define the boundaries of TE-encoding sequences, accurately identify the key sequence properties necessary for robust classification, or confidently identify families and sub-families of TEs within a genome. Thus, careful analysis of the TE content of a genome requires both automated detection of putative TE sequences and manual curation of the resulting libraries [164].

TE discovery begins with the automated detection of putative TEs from an unannotated genome. Currently, TE annotation approaches can be split into two categories. The first is repository- or homology- based annotation, in which genomic sequences are compared to a database of known TEs. The second, *de novo* annotation, searches for over-represented sequences in genome assemblies or raw-read data. Homology-based annotation is most powerful when a database of known TEs from a closely related species is available. *De novo* annotation typically identifies common repeats, but may miss low-copy number TEs. The most robust automated annotation will therefore combine these approaches. Nevertheless, the combination of *de novo* and

homology-based annotation is unable to accurately annotate lineage-specific, heavily mutated, or low-frequency TEs, particularly those in non-model species [164, 165]. For example, it has been demonstrated that when using sub-optimal annotation approaches, an increase in phylogenetic distance between a queried genome and the source of reference TEs leads to decreased ability to accurately identify TEs [164]. Taken together, these restrictions demonstrate that no single software can detect all repeats [163, 165], and robust TE prediction requires manual curation of a custom library in addition to *de novo*- and homology-based annotation [164, 166].

High quality TE libraries tailored for the species of interest is the most powerful way to prevent inaccuracies in downstream analyses. However, manual curation of TEs is time-consuming, labour intensive, and therefore not performed in a large number of TE studies. Furthermore, manual curation is a specialised and exclusive skill often passed down within TE research groups [166]. To ensure these skills are accessible to those outside of the field, Platt *et al.* [164] and Goubert *et al.* [166] outlined current approaches to hand-curation that have been demonstrated in many studies [167–171]. In this chapter, I have followed these protocols to generate a high quality TE library for three closely related strains of the *Epichloë typhina* species complex: *E. typhina* subsp. *clarkii* (Ecl), *E. typhina* subsp. *typhina* (Ety), and *E. typhina* subsp. *poae* (Epo). For each lineage, I generated an initial TE annotation using automated methods. Alignments of resulting TE families were then extracted from their respective reference genomes and used to classify each family according to current TE taxonomy. A representative model sequence was generated for each TE family. This hand curation process resulted in the first high-quality TE library for *Epichloë*.

2.3 Methods

2.3.1 Genome Sequences

Reference genomes for *Epichloë typhina* subsp. *clarkii* strain Ecl_1605_22 (BioProject ID **PRJNA533212**), *Epichloë typhina* subsp. *typhina* strain Ety_1756 (BioProject ID **PRJNA533210**), and *Epichloë typhina* subsp. *poae* strain Etp76 were already prepared for this work [148]. Each genome was resolved with 7 telomere-to-telomere chromosomes and no gaps, making these genomes ideal for the analysis of difficult-to-assemble regions like TEs.

2.3.2 Automatic TE annotations

An iterative approach was used to automatically identify TEs in each reference genome. First, *de novo* TE libraries were estimated from each reference genome using RepeatModeler2 (v.2.0.1) [20]. This suite of programs uses RepeatScout to identify multi-copy regions of a genome [172], and RECON to discover sequences that contain motifs associated with known transposable element families [173]. This pipeline also makes use of LTR_retriever and LTR_Harvest to identify LTR retrotransposon sequences [174, 175]. All putative repetitive elements identified in this way are then clustered to form repeat families and classified into repeat classes following DFAM specifications [176]. Following this, the locations of specific transposable elements were then discovered by running RepeatMasker (v4.0.6) [19], a homology-based annotation method that also uses the default DFAM database to perform searches against known TEs.

2.3.3 Library preparation

The discovery algorithms of automatic identification softwares utilise databases of known TE sequences sourced from a wide range of organisms. The phylogenetic distance between the *Epichloë* and the source of the known TE can lead to inaccurate annotations; many sequences returned this way may be denoted as unknown repeat elements, or be represented by poorly-matched model consensus sequences that are not specific to the *Epichloë* genomes. To create a library of consensus sequences specific to *Epichloë*, the RepeatModeler4 pipeline was used. This pipeline is a specialised script provided for this work by specialists in the field (J. Blommaert, personal communication). The script queries automatic RepeatMasker-predicted consensus sequences in each lineage against their respective reference genome. Following this, up to the top 20 best BLAST hits for each consensus sequence is retrieved [6] from the genome, and the retrieved sequences are aligned using MAFFT under EINSI parameters (recommended for aligning <200 sequences containing large, unalignable regions) [13]. All returned alignments harbour 2 kb flanking sequence on the 5' and 3' end of each repeat element to ensure the entire element is retrieved.

2.3.4 Consensus generation

All sequence alignments produced by the RepeatModeler4 pipeline were manually inspected in AliView (v. 1.26) [2]. TE classes and superfamilies may have canonical structures, such as 5' - TG...CA - 3' motifs in LTR elements, or they may harbour TSDs of a specific length or motif (Figure 1.1). Thus, each alignment was searched for known TE start and end signatures in accordance to the classification system proposed by Wicker *et al.* [50]. In the absence of known TE boundaries, the well conserved portion of the alignment was used to generate the new consensus sequence, and the sequence was assigned the suffix '.inc' to denote an incomplete sequence. After determining the element boundaries, a new consensus was generated from the aligned

sequences using Advanced Consensus Maker [1] with the majority threshold set to 0.75. By default, if the most common nucleotide was below a minimum fraction of 0.5, the consensus nucleotide was recorded as '?'. All bases denoted as '?' were later defined using a custom python script that applies a strict logical tie-breaking algorithm to resolve ambiguities. The final consensus sequences were then aligned to the original sequence alignment in AliView (default aligner MUSCLE v3.8.425 [177]) and manually inspected to ensure they sufficiently represented the sequence alignment.

2.3.5 Higher order classification

Self-similar domains such as LTRs and TIRs are an important feature of many TE classes. These were identified in each new consensus sequence by performing a self-alignment with MAFFT under default parameters [13]. The dotplot generated from self alignment was then assessed for structural features characteristic of different TE orders (Figure 2.1). If the sequences were determined to have LTRs or TIRs (Figure 2.1A-C), these sequences were extracted using LAST web servers under default parameters [12]. Satellite sequences are often identified by RepeatMasker as long sequences of unknown classification. As these sequences comprise of multiple repeating monomers, any consensus sequence that appeared to be a satellite (Figure 2.1E), was submitted to Tandem Repeat Finder under default parameters [22], and a monomer was determined based on a combination of copy number and length. Following advice from experts, each monomer was duplicated to form a dimer. This was done to increase coverage as it is often impossible to determine the beginning and end of a monomer (J. Blommaert, personal communication). To confirm that the dimer was representative of the satellite, the dimer was aligned to the full-length consensus sequence using MAFFT. The dimer was considered to be representative if the partial dotplot generated this way retained the overall pattern that was observed in the self-alignment of the full-length sequence (Figure 2.1E-F).

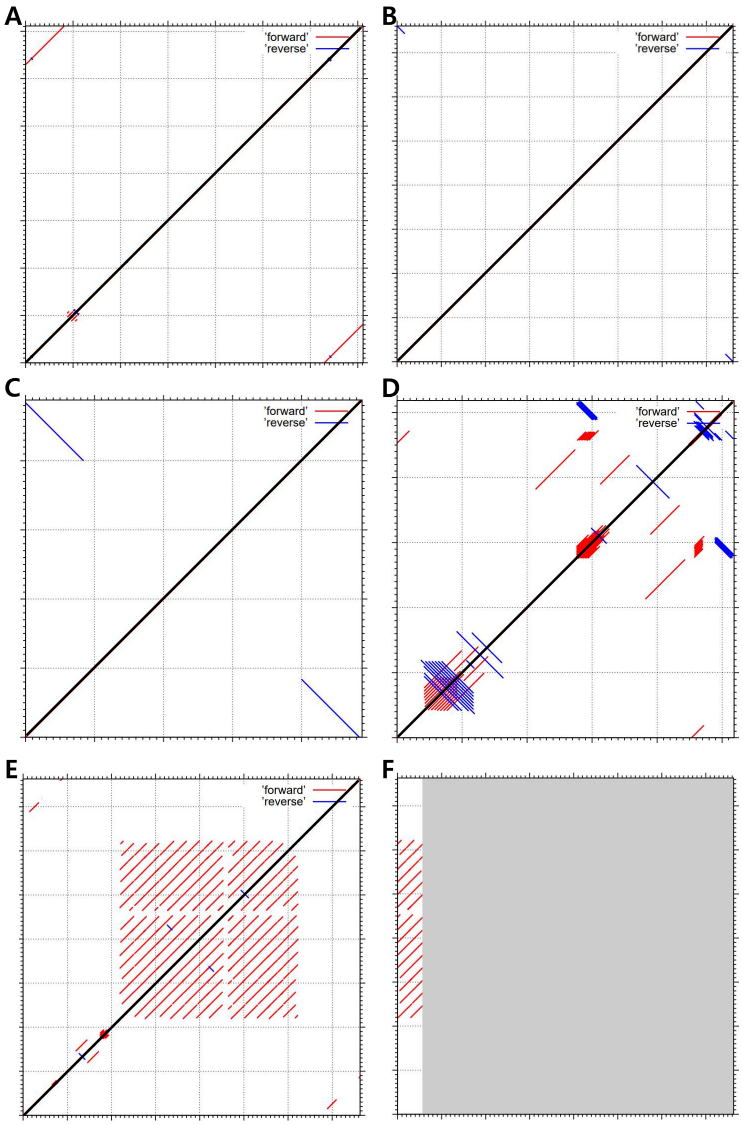

Figure 2.1: Self-similar domains for order classification. Black central line represents internal TE sequence and arises from self-alignment of consensus sequences. (A) Terminal repeats (red) characteristic of LTR elements are shown in the upper left and lower right corner. (B) Small TIRs (blue) can be observed in DNA elements in the upper left and lower right corners. (C) Proportionally large TIR sequences are representative of MITEs (blue). (D) An example of a simple repeat sequence. Self alignment reveals multiple short repeats throughout the consensus sequence (red and blue). (E) Satellite sequences form a distinct pattern of short, repetitive sequences (red). (F) Alignment of a single satellite dimer to its full-length consensus sequence shows that the pattern exhibited in E is represented by the dimer (red).

2.3.6 Conserved Domains

Many TEs encode characteristic protein domains (Figure 1.1). Identifying conserved domains therefore allows for confident annotation of TE classes and superfamilies. To identify conserved domains, all sequences were submitted to NCBI Conserved Domain Search under default parameters [7].

2.3.7 Comparison to known TEs from Repbase

To query each new consensus sequence against known fungal TEs, all consensus sequences were submitted to the Repbase database using the CENSOR tool [9]. A custom R script was written to identify TE-consensus sequences where the best hit to a Repbase sequence had an alignment score $\geqq$ 2.5 times the mean of all other hits for that consensus. This threshold was determined by inspecting alignment scores for TEs with existing accurate annotations. The CENSOR results were also manually inspected to identify elements that may have been removed by this threshold as a result of having a a number of high scoring alignments, or scored marginally below the the elected threshold but nevertheless offered insight to potential classifications.

2.3.8 Consensus clustering

Genomes often harbour fragmented TE sequences. These may arise as a result of TEs transposing within other TEs in a nested structure [178], or due to mutations that leave partial length remnants of TEs [179]. These partial sequences can generate multiple TE-models for a single TE-family, with each model representing only a portion of the true consensus sequence. Similarly, the three TE libraries produced to this point will contain considerable redundancy, as many TE families will represented in all three genomes. I used an all-against-all search approach to identify and merge redundant consensus models across all three libraries. All new consensus sequences were queried against

one another using BLAST (Blast 2.0.0+, blastn default parameters [6]). In addition, the consensus sequences from all three genomes were clustered using CDhit (v4.8.1, 90% identity, local alignment [8]). Local alignment was favored as DNA elements and MITEs may share TIRs, but harbour divergent internal sequences.

2.3.9 Reannotation of TEs

The final classifications were determined via a combination of structure, TSD length, canonical motifs, conserved domains, and sequence homology to known fungal TEs. Putative classifications that offered moderate but not extensive evidence for group-membership were denoted with a '?' suffix to assist with downstream analyses. The consensus sequences from all three genomes were combined to create a single master library which was provided to RepeatMasker for re-annotation of the focal genomes.

2.3.10 Data availability

A repository containing analyses and scripts developed for this project are available at `https://github.com/KelliSmith17/Masters`

2.4 Results and Discussion

2.4.1 Manual curation greatly improves classification of TEs

An automated annotation of the three focal genomes using RepeatMasker identified 559 unique TE consensus sequences in the *E. typhina* species complex. The sequences identified this way prior to manual-curation will henceforth be referred to as the "automatic library." The number of consensus sequences identified in the automatic library was approximately equal among fungal lineages, with 206, 195, and 158 sequences derived from Ecl, Ety, and Epo respectively. As expected, the majority (72%) of these consensus sequences were of 'unknown' classification.

De novo annotation softwares are able to predict heavily degraded, novel, fragmented, or lineage specific repeat sequences. However, once identified, such elements typically lack sequence homology to known TEs, making it impossible to successfully classify them to an order or superfamily during homology-based annotation. The lack of representative TEs for most non-model organisms substantially contributes to the generation of poorly-annotated or unclassified TEs.

Currently, manual inspection of sequences is the most reliable method of reducing the unknown elements in annotations [166]. By focusing on the structural properties of TE orders and superfamilies, these approaches can classify sequences that lack homology to any known TE. All 559 sequences identified in the automatic library underwent a complete manual-classification process. 77 sequences were excluded from manual curation due to poor alignments or insufficient sequences in their respective alignments that gave rise to a new consensus sequence with extensive ambiguity. To account for within- and between-lineage TE redundancy, all curated consensus sequences were clustered using CDhit. Following clustering, the final hand-curated TE library was reduced to 288 unique consensus sequences across all lineages.

To examine the changes in classification between the two libraries, I compared

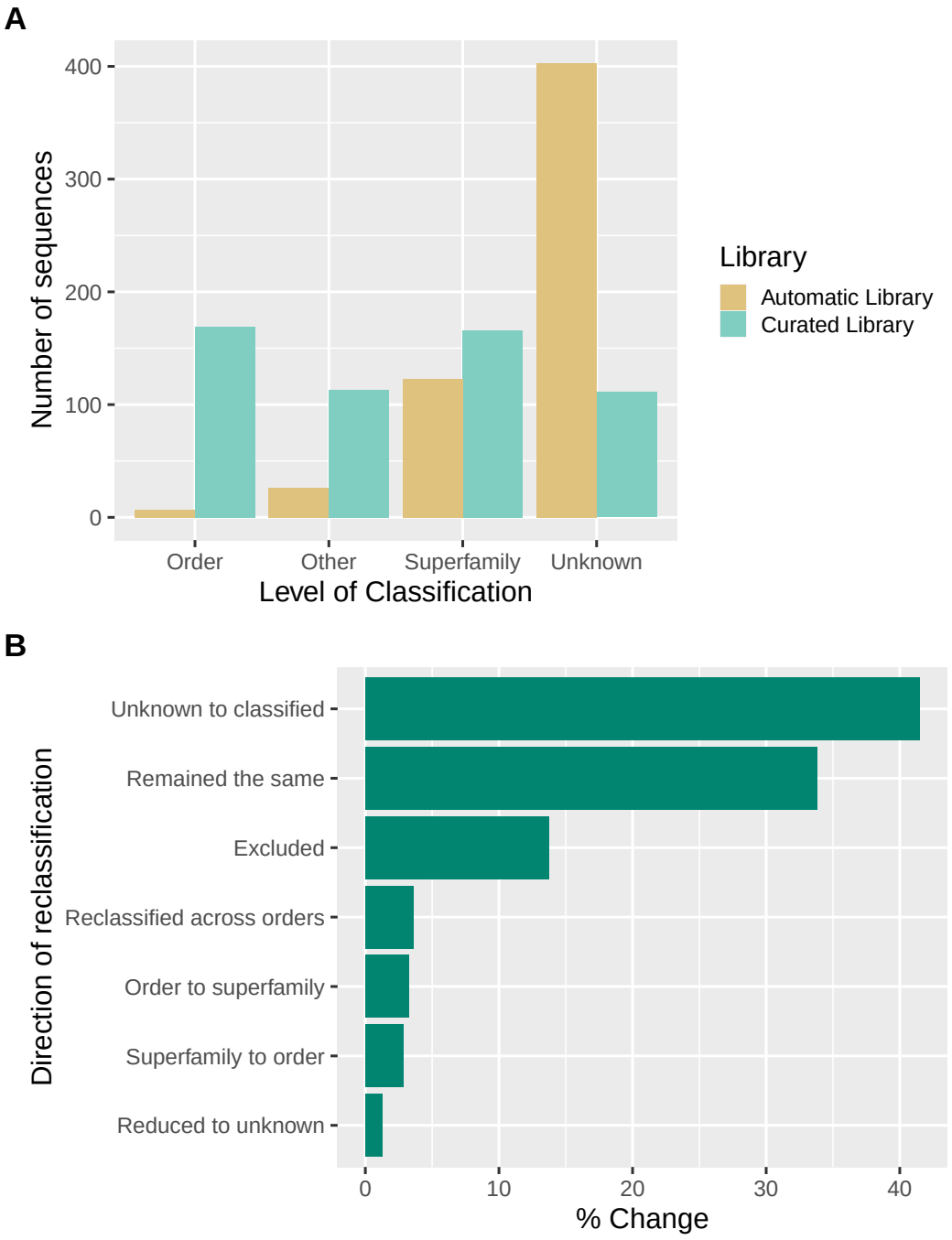

Figure 2.2: Change in classification level across the automatic and curated library. "Other" refers to other interspersed repeats such as simple repeats and satellites

each automatically identified consensus sequence to its manually-curated counterpart in the pre-clustered library. Manual-curation resulted in an increase in classification at the level of order and superfamily, and a dramatic reduction in unknown elements (Figure 2.2A). As a result, 41.5% of the new curated library comprised of consensus sequences that were previously annotated as unknown repeat elements. In addition, 3.22% of the library consisted of order-level annotations that were successfully refined to a superfamily. In contrast, 2.86% of sequences previously classified to a superfamily were demoted to an order due to insufficient evidence during manual curation. Lastly, 1.25% of sequences were reclassified as unknown and 3.58% of the initial library was entirely reclassified across orders. Only 33.81% of automated annotations were not altered by the hand-curation exercise (Figure 2.2B).

2.4.2 Manual curation allows accurate determination of the TE content of each genome

As is typical for a non-model organism, my initial TE libraries had a high proportion of unclassified elements. A high non-classification rate can lead to biased estimates of the TE content of genomes, in turn giving rise to incorrect conclusions about the roles of different classes in genome evolution. It is striking that the manual-curation process improved the identification of some TE classes more than others. In particular, manual curation identified many MITEs, short or complex LTR elements, and DNA/hAT elements.

MITEs are of interest in plant-associated fungi in general, and *Epichloë* in particular. However, these non-autonomous Class II TEs are often overlooked in TE analyses. Indeed, RepeatMasker, the most widely used tool for TE-annotation, does not predict these elements. Multiple softwares dedicated to detecting MITEs have been developed, however these pipelines are limited by insufficient databases of reference MITEs and technical difficulties in detecting these elements. Current softwares often

come with high error rates, false positive MITE annotation for LTR and DNA elements, struggle to process large-scale genomes, and fail to detect MITEs that harbor indel mutations within the TIRs [66, 180–182]. In *Epichloë* genomes, MITEs have been associated with the upstream regions of important genes [67, 68, 158] and may play a regulatory role in the expression of these genes. Thus, it was of considerable interest to characterise the MITE population in *Epichloë* genomes. Initial hand curation identified 73 MITE consensus sequences that were reduced to 41 unique consensus sequences. These elements were previously annotated as simple repeats and unknown elements. (Figure 2.3). The MITEs had a mean length of 350 bp, and mean TIR length of 67 bp. TSDs were identified for all but 6 elements, with the most common TSD being a 2 bp TA motif (Table S2.1). The curated MITEs were highly conserved: all retained full TIRs and only two MITEs harbored divergent internal structures. The 41 unique MITE families identified in this work is a dramatic increase from the 13 MITE families previously characterised in *Epichloë* [68].

In addition to MITEs, automatic identification by RepeatMasker struggled to resolve many DNA elements in the focal genomes. In particular, DNA/hAT elements were most frequently unclassified in the automatically library, and almost all elements initially identified as 'buffer' elements were shown to be DNA/hAT elements. These hAT elements were often difficult to identify during hand-curation as many were AT-rich, had poorly defined boundaries, and indiviual elements within a given family could be highly divergent from each other. Despite these challenges, hand-curation identified 19 hAT consensus sequences that retained conserved domains, structural features such as 8 bp TSDs, and/or shared high similarity with known fungal hAT elements reported on the RepBase database.

Although the automatic TE annotation accurately identified full-length LTR elements, many short or incomplete LTR sequences could only be identified by hand-curation. The missed LTRs were often short, complete LTR elements (<650 bp), or were partial LTR copies that harbored the canonical 5'TG start or 3'CA end, and retained

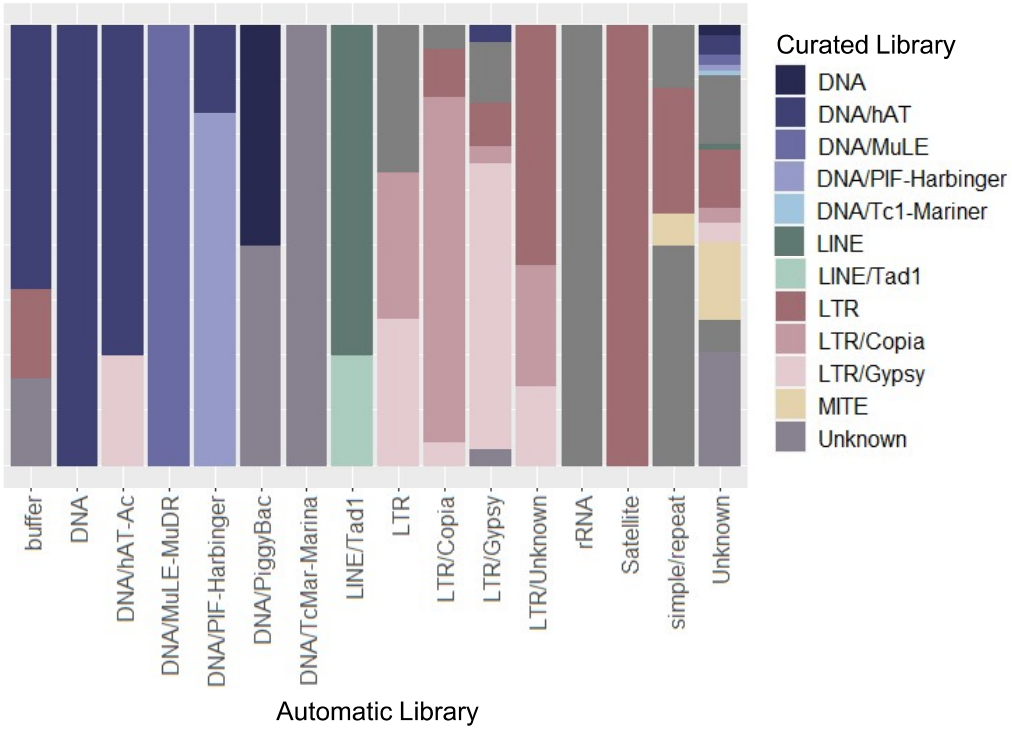

Figure 2.3: Change in classification across the automatic and curated library. Classifications determined by automatic identification are represented across the x-axis and are proportionally shaded by the new classification in the curated library.

conserved domains or sequence homology with other known LTR elements. Some of these elements may be solo-LTRs (short elements that arise as a result of unequal recombination between two LTR elements) that are often missed by RepeatMasker [183]. In addition, many of the longer LTRs elements harboured satellite-like structures between their LTRs. These satellite-like structures may result from nested insertion of TEs, with transposition or recombination within a compound-element generating very complex repetitive structures. Reclassification of these elements was determined by the presence of LTRs and TSDs consistent with LTR elements. Overall, 90 short, complex, or nested elements that were previously annotated as unknown, simple repeats, or satellites were successfully reclassified as LTR, LTR Copia, and LTR Gypsy elements.

While manual curation substantially resolved a number of MITEs, LTRs, and DNA elements, LINE elements annotations were disadvantaged during manual curation. LINE elements get truncated upon insertion due to their transposition mechanism [53], hence new insertions gradually decrease in length. Putative LINEs were classified on a basis of sequence homology to known fungal LINE elements, and truncated 5' sequences. In the absence of robust matches to LINEs in the RepBase database, these elements were not further classified into superfamilies.

2.4.3 Conserved domains within the curated TE library

Some TE classes have characteristic protein domains that mediate their transposition (Figure 1.1), thus the identification of these domains assists with the classification of elements. TE populations within a host genome may harbor no conserved domains, all conserved domains, or retain partial conserved domains such as in the case of degraded autonomous elements. Consistent with this, in the curated library 54 of the 288 consensus sequences harbored at least one conserved domain associated with a TE class. Of these, only 10 sequences harboured two characteristic domains. In addition, two elements contained domains of an unknown relevance (Table 2.1).

Most notably, more than half of DNA/hAT consensus sequences retained a highly conserved C-terminal dimerization domain found widely in transposase of hAT superfamily members [184]. A small number of DNA elements had putative DDE transposases, characteristic of these classes of TEs [50]. LTRs elements harbored a Ribonuclease H enzyme or a reverse transcriptase, consistent with previous LTR studies [185]. Finally, two LINE elements harboured reverse transcriptases. MITEs, as expected, harboured no conserved domains as they are non-autonomous elements with no functional domains. Other elements lacking conserved domains may include degraded copies of once-functional elements.

Conserved Domain	**TE group**	**Count**
DDE_Tnp_ISL3 super family	DNA	1
DDE_Tnp_ISL3 super family	DNA/MuLE	4
Dimer_Tnp_hAT	DNA/hAT	11
Glycosyltransferase_GTB-type super family	Unknown	1
RNase_H_like super family	LTR/Copia	3
RNase_H_like super family	LTR/Gypsy	4
RNase_H_like super family, RT_LTR	LTR/Gypsy	1
RNase_H_like super family, RVT_2 superfamily	LTR/Copia	1
RNase_HI_RT_Ty1, RVT_2 super family	LTR/Copia	2
RT_like super family	LINE	2
RT_like super family	LTR	1
RT_like super family	LTR/Gypsy	9
RT_like super family, RT_RNaseH_2 super family	LTR	1
RT_like super family, RT_RNaseH_2 super family	LTR/Gypsy	2
RT_LTR	LTR/Gypsy	4
RT_LTR, RNase_H_like super family	LTR/Gypsy	1
RT_LTR, RT_RNaseH_2 super family	LTR/Gypsy	1
RT_RNaseH_2 super family	LTR	1
RVT_2 super family	LTR/Copia	2
RVT_2 super family, RNase_H_like super family	LTR/Copia	1
SEN1 N terminal	Unknown	1
Total		**54**

Listed domain names are NCBI assigned IDs for each unique domain
Tnp: Transposase; RT & RVT: reverse transcriptase; RNase: ribonuclease

Table 2.1: Conserved domains within the curated library.

2.4.4 Manual curation leads to decreased TE content

To assess the effect of manual curation on total TE content, I reannotated each of the three focal genomes with the new TE library. I then compared the total length and copy number of all primary TE annotations across the pre- and post-curated library. Interestingly, the new library reduced genome coverage by a sum of 35.7Mb across all lineages, and the total copy number of TEs reduced by 38,515 (Table 2.2 - 2.3). Some reduction in total length was expected due to the refinement of element boundaries during manual-curation. For example, satellites in the original library were often annotated as long unknown elements many kilobases in length. Manual curation resolved these satellites, and identified them to have repeat-units of 54-166 bp in length. Correct annotation of these satellites reduced the total length covered by satellites while increasing copy number of individual satellites. My manual-curation approach also removed many non-TE repeats, including low complexity sequences and microsatellites, short AT- or GC- rich sequences, and poly- purine or pyrimidine stretches that are masked by RepeatMasker. These non-TE sequences were present in high copy numbers in the automatic library (n = 20 859). In addition to these sequences, my hand-curated library excluded 77 consensus sequences identified by RepeatMaker but for which I was unable to generate a majority rule consensus sequence. Given the highly-divergent nature of the sequences identified in these excluded sequences, it is unlikely they represent a true TE family. Thus excluding these elements from my manually-curated library improved the over-all accuracy of my TE annotations.

Interestingly, the most dramatic reduction in TE coverage was seen in LTR elements, despite the fact that the curated library contains 77 *more* LTR consensus sequences than the automatic library. The total length of LTR elements across all three genomes decreased by 42.32%. This reduction can be partly explained by the exclusion of 11 LTR consensus sequences and reclassification of four others. In addition, my manually curated library revealed that many LTR consensus sequences in the automated library had overly-large flanking sequences. Thus my refined LTR models decreased

the size of many LTR elements. The decrease in size of many individual LTR consensus sequences combined with the identification of 77 new LTR models in my hand-curated library means the copy number of these elements was reduced by only 5.12% . Despite these reductions, LTR remain both the most common elements (>12,000 copies) and the largest component of the TE-set (>50 Mb) when the curated library is used to annotate genomes.

The only TE classes to increase in genome coverage following the manual curation process were LINEs and DNA elements. This result can be explained by the number of new consensus sequences that were identified for these elements during hand curation (Table S2.2). In particular, the new curation identified 50 new consensus sequences for DNA elements that were previously annotated as unknown or buffer elements. The total length of DNA elements increased by 5Mb (289.59%) with an increase in copy number of 2008 elements (214.3%). LINEs increased by 710 kb (40.42%) in length and 179 (38.49%) in copy number. In addition, MITEs, which were previously not predicted by RepeatMasker, occupied a total of 583.75 kb with a copy number of 2795.

	Total length (Mb)		Change	
Order	Automatic library	Curated library	Length (Mb)	%
DNA	1.73	6.75	5.01	289.59
LINE	1.76	2.47	0.71	40.42
Low_complexity	0.10	0.00	-0.10	-100.00
LTR	91.03	52.51	-38.53	-42.32
rRNA	0.01	0.01	-0.00	-32.85
Satellite	0.05	0.27	0.22	404.07
Simple_repeat	2.01	0.83	-1.18	-58.65
Unknown	7.71	5.35	-2.36	-30.65

% denotes percentage difference in total length

Table 2.2: Comparison of total genome length covered by each library.

	Copy number		Change	
TE order	Automatic library	Curated library	Copy number	%
DNA	937	2945	2008	214.30
LINE	465	644	179	38.49
Low_complexity	2187	0	-2187	-100.00
LTR	12922	12262	-660	-5.11
rRNA	15	76	61	406.67
Satellite	50	284	234	468.00
Simple_repeat	36184	713	-35471	-98.03
Unknown	9186	3712	-5474	-59.59

% denotes percentage change in copy number

Table 2.3: Comparison of total copy numbers detected by each library

2.4.5 A TE library for *Epichloë typhina*

The manual curation process resulted in the first TE library tailored for *E. typhina*. Consensus redundancy between the three focal lineages resulted in a compact library of 288 unique consensus sequences. Of these, only 3 consensus sequences were lineage specific, and the remaining TEs were shared between two or more lineages, demonstrating a high similarity in TE repetoire between the genomes. Of the 288 unique repeat sequences, 91.67% were TEs and 8.33% were other interspersed repeats (satellites, rRNA, simple repeats). 194 (67.36%) of all consensus sequences provided conserved structural and/or enzymatic features of the respective classes. 26 sequences (9.03%) retained partial features and were denoted with a "?' suffix, and 67 sequences (23.26%) were incomplete sequences denoted with an ".inc" suffix. One sequence (0.35%) was both partial and incomplete (.inc? suffix). Reannotation of the three genomes revealed that the TEs predicted in this new curated library account for 67.04%, 46.1%, and 57.44% of the total genome length in Ecl, Ety, and Epo, respectively (Table S3.1). Other interspersed repeats such as satellites, rRNA, and simple repeats account for 0.37%, 0.36%, and 0.26% of total genome length in Ecl, Ety, and Epo, respectively.

TE class	Total Number	%.inc	%.?
DNA	6	16.67	33.33
DNA/hAT	19	73.68	21.05
DNA/MuLE	8	0	50
DNA/PIF-Harbinger	4	0	25
DNA/Tc1-Mariner	4	0	100
LINE	8	0	25
LINE/Tad1	1	0	0
LTR	45	6	0
LTR/Copia	23	8.7	0
LTR/Gypsy	31	22.58	0
MITE	41	0	9.76
rRNA	1	0	0
Satellite.dimer	11	0	0
Simple_repeat	12	0	0
Unknown	74	50	0

%.inc denotes percentage of consensus sequences that are incomplete
%.? denotes percentage of consensus sequences that are putative

Table 2.4: Overview of final consensus sequences in the curated library

2.5 Conclusion

This work created the first manually-curated library for *Epichloë typhina*. By combining *de novo*- and homology-based annotation alongside in-depth and current approaches to TE classifications, this library successfully tailored a set of TE consensus sequences specific to these genomes. This process greatly improved the classification of TEs, strengthening the theory that manual curation is essential for robust TE annotation, particularly in non-model species. Most notably, 72% of sequences in the pre-curated library were annotated as "unknown" repeat sequences. During manual curation, over 40% of these sequences were successfully classified, reducing the total percentage of unknown elements to 25.69%. Further, manual curation was able to refine several TE consensus' from the level of order to the level of superfamily. In total, only 33.81% of all annotations remained the same across the pre- and post-curated library.

During reclassification, the new library was able to resolve several under-represented TE classes that were commonly annotated as unknown elements, buffers,

satellites, or simple repeats. In particular, these underrepresented elements were reclassified as small LTR elements, DNA/hAT elements, and MITEs. The new library characterised 41 unique MITE consensus sequences, providing valuable insight to a class of interest in fungi that are currently not detected by RepeatMasker. These 41 new consensus sequences are a stark increase from the 13 MITEs previously characterised in *Epichloë* [68]. Further, manual curation successfully identified the presence of conserved domains present in 18.75% of sequences. These domains are not detected using automatic annotation methods, but are crucial determinants in classification.

Interestingly, despite the merit of performing manual-curation, the new library showed an overall decrease in genome coverage, both in copy number and in length. This may be due to limitations caused by computational restrictions and properties intrinsic to TEs. In particular, 77 automatically-identified TE consensus sequence were excluded from manual curation. This is due excessive sequence divergence in the alignments that meant a majority rules consensus sequence could not be reliably generated. Such degraded TEs would likely add little to downstream analyses of TEs.

The final TE library comprises of 288 unique consensus sequences, of which 264 are TEs and 24 are other interspersed repeats (satellites, simple repeats, short rRNA). Almost 70% of all sequences retained conserved domains and structural features that paved the way to robust reclassification. The remaining sequences were a combination of putative sequences that retained partial features, and incomplete, fragmented sequences that retained just a portion of the full element. Despite limitations, this library substantiates the belief that manual curation is essential for accurate TE annotation, and demonstrates how high quality reference genomes can assist in generating high quality TE libraries, even for non-model species that are underrepresented in homology-based TE databases. Above all, the data produced in this work will serve as a valuable resource for future fungal TE studies.

2.6 Supplementary Information

Name	Class	Length	TSD length	TSD motif	TIR length
Ecl_rnd-1_family-42	MITE	292	2	TA	36
Ecl_rnd-1_family-47	MITE	244	2-3	TA?	42
Ecl_rnd-1_family-115	MITE	269	2	TA	38
Ecl_rnd-1_family-127	MITE	368	6-9	unique	111
Ecl_rnd-1_family-130	MITE?	78	2	TA	39
Ecl_rnd-1_family-138	MITE	421	2	TA?	51
Ecl_rnd-1_family-142	MITE	318	2	TA	61
Ecl_rnd-1_family-155	MITE	466	6-8	unique	57
Ecl_rnd-1_family-165	MITE	107	5	unique	109
Ecl_rnd-4_family-324	MITE	403	7-8	unique	57
Ety_rnd-1_family-39	MITE	376	9	unique	116
Ety_rnd-1_family-44	MITE	429	6-8	unique	59
Ety_rnd-1_family-53	MITE	243	2	TA	30
Ety_rnd-1_family-57	MITE	398	8-9	unique	85
Ety_rnd-1_family-58	MITE	442	8-9	unique	84
Ety_rnd-1_family-59	MITE	430	8-9	unique	74
Ety_rnd-1_family-63	MITE	501	NA	NA	32
Ety_rnd-1_family-107	MITE	497	2	TA	154
Ety_rnd-1_family-117	MITE	357	2	TA	60
Ety_rnd-1_family-136	MITE	357	8-9	unique	91
Ety_rnd-1_family-138	MITE	332	2	TA	61
Ety_rnd-1_family-139	MITE	327	2	TA	61
Ety_rnd-1_family-141	MITE	344	2	TA	61
Ety_rnd-1_family-145	MITE	492	7-8	unique	61
Ety_rnd-1_family-148	MITE	518	8	unique	57
Ety_rnd-1_family-163	MITE?	775	NA	NA	124
Ety_rnd-1_family-176	MITE	453	8-9	unique	73
Ety_rnd-4_family-223	MITE	249	6-9	unique	112
Epo_rnd-1_family-12	MITE	586	2	TA	118
Epo_rnd-1_family-38	MITE	265	2-4	unique	32
Epo_rnd-1_family-41	MITE	135	2	TA	56
Epo_rnd-1_family-43	MITE	266	2	TA	35
Epo_rnd-1_family-67	MITE	194	2	TA	35
Epo_rnd-1_family-68	MITE	251	2-4	unique	40
Epo_rnd-1_family-71	MITE	340	NA	NA	113
Epo_rnd-1_family-93	MITE?	389	NA	NA	59
Epo_rnd-1_family-94	MITE	271	NA	NA	35
Epo_rnd-1_family-97	MITE	270	2	TA	36
Epo_rnd-1_family-120	MITE	192	2	TA	88
Epo_rnd-2_family-40	MITE	280	2	TA	45
Epo_rnd-3_family-47	MITE?	442	NA	NA	52

Table S2.1: MITEs in *E. typhina*

Automatic Class	**Curated Class**	**Number of sequences**
buffer	DNA/hAT	6
	LTR	2
	Unknown	2
DNA	DNA/hAT	1
DNA/hAT-Ac	DNA/hAT	3
	LTR/Gypsy	1
DNA/MuLE-MuDR	DNA/MuLE	3
DNA/PIF-Harbinger	DNA/PIF-Harbinger	4
	DNA/hAT	1
DNA/PiggyBac	DNA	1
	Unknown	1
DNA/TcMar-Marina	Unknown	2
LINE/Tad1	LINE	6
	LINE/Tad1	2
LTR	excluded	2
	LTR/Copia	2
	LTR/Gypsy	2
LTR/Copia	LTR/Copia	29
	LTR	4
	excluded	2
	LTR/Gypsy	2
LTR/Gypsy	LTR/Gypsy	33
	excluded	7
	LTR	5
	DNA/hAT	2
	LTR/Copia	2

	Unknown	2
LTR/Unknown	LTR	6
	LTR/Copia	3
	LTR/Gypsy	2
rRNA	excluded	1
Satellite	LTR	1
Simple repeat	Simple repeat	5
	LTR	4
	excluded	2
	Satellite	2
	MITE	1
Unknown	Unknown	104
	MITE	72
	excluded	63
	LTR	53
	Satellite	19
	DNA/hAT	18
	LTR/Gypsy	17
	LTR/Copia	14
	DNA	9
	DNA/MuLE	9
	Simple repeat	9
	DNA/PIF-Harbinger	6
	LINE	5
	DNA/Tc1-Mariner	4
	rRNA	1

Table S2.2: Change in classification between libraries

2.6.1 Curation table and annotation data

The following section contains the data collected during consensus generation and classification of TEs. For brevity, some columns have been removed from the pages below, however, the full spreadsheet and the guide to interpretations are available at **curation_records.xlsx** at `https://github.com/KelliSmith17/Masters`.

The curation spreadsheet contains extensive detail on each consensus sequence created during manual-curation and records features such as: consensus name and TE class; consensus length and sequence; TSD length and motif; the internal structure of the TE, i.e., satellite-like or simple-repeat-like; the status of the TE with regards to truncated ends; conserved domains and Repbase results; general comments; the previous RepeatMasker classification using the automatic library; and information on the consensus clusters formed by CDhit.

In addition to the consensus spreadsheet, the RepeatMasker outputs generated upon reannotating all three genomes with the new curated library are also available on the repository under **RepeatMasker_outputs**. This raw data contains information such as the genomic locations of every TE copy within each genome, and the divergence of each copy from its respective consensus sequence.

Name	TE class	Consensus	Length	TSD len	TSD motif	TIR len	LTR len	SR	Sat	.inc	Conserved domain	Repbase	Automatic Classification	Cluster n	Cluster spp.
Ecl_rnd-1_family-1	LTR/Copia	TGTTAATTC	6744	5	unique	NA	49	no	no	no	NA	LTR/Copia	rnd-1_family_1#LTR_Copia	8	Ecl
Ecl_rnd-1_family-2	DNA/hAT	TGTTTGTA	7358	7-8	unique	NA	NA	yes	yes	no	Dimer_Tnp_hAT	DNA/hAT	rnd-1_family_2#buffer	1	Ecl
Ecl_rnd-1_family-3	LTR	TGTTAGAAC	6242	4-6	unique	NA	286	no	no	no	NA	NA	rnd-1_family-3#LTR_Copia	1	Ecl
Ecl_rnd-1_family-4	LTR/Copia	TGTTAAAAA	7113	4-6	unique	NA	816	no	no	no	RNase_HI_RT_Ty1, I	LTR/Copia	rnd-1_family-4#LTR_Copia	2	Ecl, Ety
Ecl_rnd-1_family-5	LTR/Copia	TGTTAATTC	6503	5	unique	NA	249	no	no	no	NA	LTR/Copia	rnd-1_family-5#LTR_Copia	1	Ecl
Ecl_rnd-1_family-6	LTR/Gypsy	TGTTAGGAC	7766	5	unique	NA	256	no	no	no	NA	LTR/Gypsy	rnd-1_family-6#LTR_Gypsy	1	Ecl
Ecl_rnd-1_family-14	Unknown	TATCTAGGT	9206	NA	NA	NA	NA	yes	no	both	SEN1 N terminal	NA	rnd-1_family-14#Unknown	1	Ecl
Ecl_rnd-1_family-17	Satellite.dimer	TATATAGTT	134	NA	NA	NA	NA	no	yes	no	NA	NA	rnd-1_family-17#Unknown	1	Ecl
Ecl_rnd-1_family-18	LTR/Copia	TGTCAGGAA	5429	5	unique	NA	224	no	no	no	NA	LTR/Copia	rnd-1_family-18#LTR_Copia	3	Ecl, Ety, Epo
Ecl_rnd-1_family-19	Unknown	TTAATATAC	7653	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-19#Unknown	1	Ecl
Ecl_rnd-1_family-21	Unknown.inc	TTATATACC	10042	NA	NA	NA	NA	yes?	no	both	NA	NA	rnd-1_family-21#Unknown	1	Ecl
Ecl_rnd-1_family-23	LTR/Gypsy.inc	TGTGAGAAC	3905	NA	NA	NA	NA	no	no	3'	RNase_H_like super f	LTR/Gypsy	rnd-1_family-23#Unknown	1	Ecl
Ecl_rnd-1_family-24	LTR/Copia	TGTTGCACC	5467	4-5	unique	NA	234	no	no	no	RVT_2 super family, I	LTR/Copia	rnd-1_family-24#LTR_Copia	2	Ecl, Ety
Ecl_rnd-1_family-27	Unknown.inc	GCTCTTAAA	10760	NA	NA	NA	NA	yes?	no	both	Glycosyltransferase_G	NA	rnd-1_family-27#buffer	2	Ecl
Ecl_rnd-1_family-29	LTR	TGTTACGAT	2158	4-5	unique	NA	152	no	yes	no	NA	NA	rnd-1_family-29#Simple_repeat	2	Ecl
Ecl_rnd-1_family-30	Unknown	TATCCCCTC	793	2	TA	NA	NA	no	no	no	NA	NA	rnd-1_family-30#Unknown	2	Ecl
Ecl_rnd-1_familiy-31	LTR	TGTTACGAT	2252	4-5	unique	NA	152	no	yes	no	NA	NA	rnd-1_family-31#Simple_repeat	1	Ecl
Ecl_rnd-1_family-32	Satellite.dimer	TTTTATTTTA	200	NA	NA	NA	NA	no	yes	no	NA	NA	rnd-1_family-32#Unknown	1	Ecl
Ecl_rnd-1_family-34	LTR/Copia	TGTTGATTC	2300	5	unique	NA	249	no	no?	no	NA	LTR/Copia	rnd-1_family-34#LTR_Copia	1	Ecl
Ecl_rnd-1_family-35	LTR/Gypsy	TGTTACGTG	10223	4	unique	NA	204	no	no	no	NA	LTR/Gypsy	rnd-1_family-35#LTR_Gypsy	1	Ecl
Ecl_rnd-1_family-36	Satellite.dimer	TACTTTTAA	202	NA	NA	NA	NA	no	yes	no	NA	NA	rnd-1_family-36#Simple_repeat	3	Ecl, Ety
Ecl_rnd-1_family-38	LTR	TGTTACGTG	2165	4-5	unique	NA	152	no	yes	no	NA	NA	rnd-1_family-38#Simple_repeat	1	Ecl
Ecl_rnd-1_family-39	LTR	TGTTACGAT	2191	5	unique	NA	152	no	yes	no	NA	NA	rnd-1_family-39#Simple_repeat	1	Ecl
Ecl_rnd-1_family-40	DNA/MuLE	GAGGTAGTA	2916	8	unique	84	NA	no	no	no	DDE_Tnp_ISL3 supe	NA	rnd-1_family-40#DNA_MuLE-M	1	Ecl
Ecl_rnd-1_family-42	MITE	CAATACACC	292	2	TA	36	NA	no	no	no	NA	NA	rnd-1_family-42#Unknown	10	Ecl, Ety, Epo
Ecl_rnd-1_family-43	LTR/Copia	TGGACCAAA	3060	NA	NA	NA	357	no	no	no	NA	NA	rnd-1_family-43#Unknown	1	Ecl
Ecl_rnd-1_family-44	Unknown.inc	TAATAATAG	7903	NA	NA	NA	NA	yes?	no	both	NA	NA	rnd-1_family-44#Unknown	1	Ecl
Ecl_rnd-1_family-45	LTR/Copia	TGTTGAGAA	5474	NA	NA	NA	251	no	no	no	RVT_2 super family	NA	rnd-1_family-45#LTR_Copia	1	Ecl
Ecl_rnd-1_family-46	Unknown.inc	ATGTAAAGC	7032	NA	NA	NA	NA	yes?	no	both	NA	NA	rnd-1_family-46#buffer	2	Ecl
Ecl_rnd-1_family-47	MITE	TACAGTCTA	244	2-3	TA?	42	NA	no	no	no	NA	NA	rnd-1_family-47#Unknown	6	Ecl, Ety
Ecl_rnd-1_family-48	LINE	TTAAAAAAA	2145	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-48#LINE_Tad1	1	Ecl
Ecl_rnd-1_family-49	DNA/hAT.inc	AAAAAATTT	4154	NA	NA	NA	NA	yes	no	both	Dimer_Tnp_hAT	DNA/hAT	rnd-1_family-49#Unknown	1	Ecl
Ecl_rnd-1_family-50	Unknown.inc	TATAAACTT	2875	NA	NA	NA	NA	no	no	both	NA	NA	rnd-1_family-50#Unknown	1	Ecl
Ecl_rnd-1_family-51	Unknown.inc	TTTAGCTAA	6171	NA	NA	NA	NA	yes?	no	both	NA	NA	rnd-1_family-51#Unknown	1	Ecl
Ecl_rnd-1_family-53	Unknown	TTGATTTTT	745	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-53#Unknown	1	Ecl
Ecl_rnd-1_family-57	Unknown	ATAATATTA	5542	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-57#Unknown	1	Ecl
Ecl_rnd-1_family-60	DNA/hAT.inc	GTAAAAATA	5189	NA	NA	NA	NA	yes	no	both	Dimer_Tnp_hAT	DNA/hAT	rnd-1_family-60#Unknown	6	Ecl, Ety, Epo
Ecl_rnd-1_family-62	LTR	TGTAAGAGA	607	5-6	unique	NA	174	no	no	no	NA	NA	rnd-1_family-62#Unknown	7	Ecl, Ety
Ecl_rnd-1_family-63	LTR	TGTAAGAGA	561	5-6	unique	NA	164	no	no	no	NA	NA	rnd-1_family-63#Unknown	2	Ecl, Epo
Ecl_rnd-1_family-67	LTR	TGTTACGAT	2224	4-6	unique	NA	152	no	yes	no	NA	NA	rnd-1_family-67#Unknown	1	Ecl
Ecl_rnd-1_family-71	DNA/hAT.inc	TAATCAGAG	4539	NA	NA	NA	NA	yes?	yes?	no	Dimer_Tnp_hAT	DNA/hAT	rnd-1_family-71#Unknown	1	Ecl
Ecl_rnd-1_family-72	LTR/Gypsy	TGTGTGCCT	11318	NA	NA	NA	200	no	no	no	NA	LTR/Gypsy	rnd-1_family-72#LTR_Gypsy	1	Ecl
Ecl_rnd-1_family-73	LTR/Gypsy	TGTGTGCCT	11623	4	unique	NA	NA	no	no	no	RNase_H_like super f	LTR/Gypsy	rnd-1_family-73#LTR_Gypsy	1	Ecl
Ecl_rnd-1_family-76	LTR/Copia	TGTGGAGTC	4445	5	unique	NA	254	no	no	no	RNase_HI_RT_Ty1, I	NA	rnd-1_family-76#Unknown	1	Ecl
Ecl_rnd-1_family-77	LTR	TGTGTGGCT	3283	4	unique	NA	444	no	yes	no	NA	NA	rnd-1_family-77#Unknown	1	Ecl
Ecl_rnd-1_family-78	Unknown	GGCCTAGTT	5651	NA	NA	NA	112	yes?	no	no	NA	NA	rnd-1_family-78#Unknown	1	Ecl

Name	TE class	Consensus	Length	TSD len	TSD motif	TIR len	LTR len	SR	Sat	.inc	Conserved domain	Repbase	Automatic Classification	Cluster n	Cluster spp.
Ecl_rnd-1_family-79	Unknown.inc	AACTATAAT	5983	NA	NA	NA	NA	yes?	no	both	NA	NA	rnd-1_family-79#Unknown	1	Ecl
Ecl_rnd-1_family-80	Unknown.inc	TACGCGTTC	1901	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-80#DNA_PiggyBac	1	Ecl
Ecl_rnd-1_family-81	DNA/PIF-Harbi	GGGCGTGA	3053	3	TAA or TT	35	NA	no	no	no	NA	NA	rnd-1_family-81#DNA_PIF-Harb	1	Ecl
Ecl_rnd-1_family-83	Unknown.inc	TCTTCCCTA	4242	NA	NA	NA	NA	no	no	5'	NA	NA	rnd-1_family-83#Unknown	1	Ecl
Ecl_rnd-1_family-84	Satellite.dimer	ATAGCGTAA	278	NA	NA	NA	NA	no	yes	no	NA	NA	rnd-1_family-84#Unknown	1	Ecl
Ecl_rnd-1_family-85	Unknown	CAAGATCCT	7924	2	TA?	NA	NA	no	no	no	NA	NA	rnd-1_family-85#Unknown	7	Ecl, Ety, Epo
Ecl_rnd-1_family-87	Unknown	CACACCTGC	171	2	TA	NA	NA	no	no	no	NA	NA	rnd-1_family-87#Unknown	3	Ecl, Ety
Ecl_rnd-1_family-88	LINE	NATATATAT	6571	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-88#LINE_Tad1	2	Ecl, Ety
Ecl_rnd-1_family-93	LINE	ATTCCTGTC	7561	NA	NA	NA	NA	no	no	no	RT_like super family	NA	rnd-1_family-93#LINE_Tad1	2	Ecl, Ety
Ecl_rnd-1_family-94	LTR?	TTTTGTCGG	2713	4-5	unique	NA	506	no	yes	no	NA	NA	rnd-1_family-94#Unknown	5	Ecl
Ecl_rnd-1_family-97	Unknown	CACACCGAC	1473	2	TA	NA	NA	no	no	no	NA	NA	rnd-1_family-97#DNA_TcMar-M	2	Ecl
Ecl_rnd-1_family-98	DNA/hAT.inc	CAGAGCTTT	4500	NA	NA	NA	NA	yes?	yes?	5'	Dimer_Tnp_hAT	DNA/hAT	rnd-1_family-98#Unknown	1	Ecl
Ecl_rnd-1_family-102	DNA?	TATCTCCTTA	1410	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-102#DNA_PiggyBa	1	Ecl
Ecl_rnd-1_family-105	LINE	CAAAAAACT	6912	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-105#LINE_Tad1	1	Ecl
Ecl_rnd-1_family-106	DNA/PIF-Harbi	GGGTCTGAT	3051	3	TAA	15	NA	no	no	no	NA	NA	rnd-1_family-106#DNA_PIF-Ha	1	Ecl
Ecl_rnd-1_family-109	Unknown.inc	TATTGCTAT	256	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-109#Unknown	1	Ecl
Ecl_rnd-1_family-113	Unknown	AGTGGCAG	324	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-113#Unknown	1	Ecl
Ecl_rnd-1_family-114	DNA/MuLE	GGGCAACTA	2325	9	unique	211	NA	no	no	no	NA	NA	rnd-1_family-114#Unknown	1	Ecl
Ecl_rnd-1_family-115	MITE	TCTCTTTGA	269	2	TA	38	NA	no	no	no	NA	NA	rnd-1_family-115#Unknown	1	Ecl
Ecl_rnd-1_family-119	LTR	TGTCATGGT	522	5-6	unique	NA	174	no	no	no	NA	NA	rnd-1_family-119#Unknown	5	Ecl, Ety, Epo
Ecl_rnd-1_family-121	Unknown.inc	TAATAGTAC	8050	NA	NA	NA	NA	no	no	both	NA	NA	rnd-1_family-121#Unknown	1	Ecl
Ecl_rnd-1_family-123	DNA/MuLE	GAGTTCTGC	2779	9	unique	34	NA	no	no	no	DDE_Tnp_ISL3 supe	NA	rnd-1_family-123#DNA_MuLE-I	2	Ecl
Ecl_rnd-1_family-124	LTR	ACCCTGAAG	508	NA	NA	NA	85	no	no	no	NA	NA	rnd-1_family-124#Unknown	1	Ecl
Ecl_rnd-1_family-125	LTR.inc	CTATTATTA	7229	NA	NA	NA	NA	no	no	5'	RT_like super family, I	NA	rnd-1_family-125#LTR_Gypsy	1	Ecl
Ecl_rnd-1_family-127	MITE	GAGTACGTA	368	6-9	unique	111	NA	no	no	no	NA	NA	rnd-1_family-127#Unknown	1	Ecl
Ecl_rnd-1_family-128	Satellite.dimer	TTTATTTAT	322	NA	NA	NA	NA	no	yes	no	NA	NA	rnd-1_family-128#Unknown	1	Ecl
Ecl_rnd-1_family-129	Unknown	AAAAAAGAC	5696	NA	NA	NA	NA	no	no	3'?	NA	NA	rnd-1_family-129#Unknown	4	Ecl
Ecl_rnd-1_family-130	MITE?	CAGTGGGTC	78	2	TA	39	NA	no	no	no	NA	NA	rnd-1_family-130#Unknown	3	Ecl, Ety, Epo
Ecl_rnd-1_family-132	LTR.inc	TGTCACGGC	6392	NA	NA	NA	NA	no	no	3'	RT_like super family	NA	rnd-1_family-132#LTR_Gypsy	2	Ecl
Ecl_rnd-1_family-133	DNA/MuLE?	GGGCGTGG	2799	8-9	unique	210	NA	no	no	no	DDE_Tnp_ISL3 supe	NA	rnd-1_family-133#Unknown	2	Ecl
Ecl_rnd-1_family-134	Unknown.inc	TGCGAGGC	7590	NA	NA	NA	NA	no	no	both	NA	NA	rnd-1_family-134#LTR_Gypsy	2	Ecl
Ecl_rnd-1_family-135	LTR/Gypsy.inc	ATTAAGGG	4136	NA	NA	NA	NA	no	no	both	RT_LTR, RT_RNaseH	LTR/Gypsy	rnd-1_family-135#Unknown	1	Ecl
Ecl_rnd-1_family-136	LTR/Copia	TGTTGATTC	2266	5	unique	NA	251	no	yes?	no	NA	LTR/Copia	rnd-1_family-136#Unknown	1	Ecl
Ecl_rnd-1_family-137	LTR/Copia	TGTTGATTC	2295	4-5	unique	NA	252	no	yes?	no	RNase_H_like super f	LTR/Copia	rnd-1_family-137#Unknown	1	Ecl
Ecl_rnd-1_family-138	MITE	GTTACAGTA	421	2	TA?	51	NA	no	no	no	NA	NA	rnd-1_family-138#Unknown	1	Ecl
Ecl_rnd-1_family-141	Unknown.inc	GGATAAGTA	6392	NA	NA	NA	NA	no	no	both	NA	NA	rnd-1_family-141#Unknown	1	Ecl
Ecl_rnd-1_family-142	MITE	AGTTAGGAC	318	2	TA	61	NA	no	no	no	NA	NA	rnd-1_family-142#Unknown	3	Ecl
Ecl_rnd-1_family-143	LTR.inc	TGTAAGGG	4270	NA	NA	NA	NA	no	no	3'	RT_RNaseH_2 super	NA	rnd-1_family-143#Unknown	1	Ecl
Ecl_rnd-1_family-146	Unknown.inc	AGTAACTAT	7313	NA	NA	NA	NA	no	no	both	NA	NA	rnd-1_family-146#LTR_Gypsy	1	Ecl
Ecl_rnd-1_family-148	LTR	TGTTGAAAT	341	NA	NA	NA	56	no	no	no	NA	NA	rnd-1_family-148#Unknown	2	Ecl
Ecl_rnd-1_family-149	Unknown.inc	CTAATCCTT	5426	NA	NA	NA	NA	no	no	both	NA	NA	rnd-1_family-149#Unknown	1	Ecl
Ecl_rnd-1_family-150	DNA/hAT?	GGCTCGTGT	2695	8	unique	64	NA	no	no	no	NA	NA	rnd-1_family-150#Unknown	1	Ecl
Ecl_rnd-1_family-151	DNA	TAATCGTGC	1749	NA	NA	61	59	no	no	no	NA	NA	rnd-1_family-151#Unknown	1	Ecl
Ecl_rnd-1_family-153	DNA/Tc1-Marin	ATCGTGCCA	1704	2	TA	117	NA	no	no	no	NA	NA	rnd-1_family-153#Unknown	1	Ecl
Ecl_rnd-1_family-155	MITE	GGGGTTAC	466	6-8	unique	57	NA	no	yes	no	NA	NA	rnd-1_family-155#Unknown	1	Ecl
Ecl_rnd-1_family-156	DNA/MuLE?	TGAATTGCT	1528	9	unique	124	NA	no	no	no	NA	NA	rnd-1_family-156#Unknown	1	Ecl

Name	TE class	Consensus	Length	TSD len	TSD motif	TIR len	LTR len	SR	Sat	.inc	Conserved domain	Repbase	Automatic Classification	Cluster n	Cluster spp.
Ecl_rnd-1_family-161	Unknown.inc	TTTAAGGTT	4196	NA	NA	NA	NA	yes?	no	3'	NA	NA	rnd-1_family-161#Unknown	1	Ecl
Ecl_rnd-1_family-164	rRNA	CACATACGA	126	NA	NA	NA	NA	no	no	no	NA	multicopy_gene/r	rnd-1_family-164#Unknown	1	Ecl
Ecl_rnd-1_family-165	MITE	GGGGACGT/	107	5	unique	109	NA	no	no	no	NA	NA	rnd-1_family-165#Unknown	1	Ecl
Ecl_rnd-1_family-166	DNA.inc	AGGACGCCT	2280	NA	NA	NA	NA	no	no	3'	DDE_Tnp_ISL3 supe	NA	rnd-1_family-166#Unknown	2	Ecl
Ecl_rnd-1_family-167	DNA/MuLE?	TGAATTGCT	2234	6-9	unique	123	NA	no	no	no	NA	NA	rnd-1_family-167#Unknown	1	Ecl
Ecl_rnd-1_family-169	Unknown	TATCGTGCG	251	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-169#Unknown	1	Ecl
Ecl_rnd-4_family-3	Unknown.inc	ACTACTAGA	2574	NA	NA	NA	NA	no	no	5'	NA	NA	rnd-4_family-3#Unknown	1	Ecl
Ecl_rnd-4_family-28	Simple_repeat	TAANCAGAC	1993	NA	NA	NA	NA	no	yes	no	NA	NA	rnd-4_family-28#Simple_repeat	1	Ecl
Ecl_rnd-4_family-31	Simple_repeat	CTAATATTT/	2869	NA	NA	NA	NA	yes	no	no	NA	NA	rnd-4_family-31#Simple_repeat	1	Ecl
Ecl_rnd-4_family-32	LTR/Copia	TGTTGGAAC	6641	5	unique	NA	281	no	no	no	NA	LTR/Copia	rnd-4_family-32#LTR_Copia	1	Ecl
Ecl_rnd-4_family-53	LTR/Gypsy	TGTTATGGA	6850	4-5	unique	NA	232	no	no	no	RT_like super family, I	LTR/Gypsy	rnd-4_family-53#LTR_Gypsy	3	Ecl
Ecl_rnd-4_family-104	LTR/Gypsy	TGTTAATTC	6356	5	unique	NA	247	no	no	no	NA	NA	rnd-4_family-104#LTR_Copia	1	Ecl
Ecl_rnd-4_family-311	Unknown	CGGGGTTAC	8254	NA	NA	NA	NA	yes?	no	no	NA	NA	rnd-4_family-311#Unknown	2	Ecl, Ety
Ecl_rnd-4_family-322	Simple_repeat	ACAATACAA	2107	NA	NA	NA	NA	yes	yes	no	NA	NA	rnd-4_family-322#Simple_repeat	2	Ecl, Ety
Ecl_rnd-4_family-324	MITE	GGGGTTAC/	403	7-8	unique	57	NA	no	no	no	NA	NA	rnd-4_family-342#Simple_repeat	1	Ecl
Ecl_rnd-4_family-642	Unknown	TAGGGAAAC	7417	NA	NA	NA	NA	no	no	no	NA	NA	rnd-4_family-642#Unknown	1	Ecl
Ecl_rnd-4_family-669	LTR/Copia	TGTTAATTC	6689	5	unique	NA	245	no	no	no	NA	LTR/Copia	rnd-4_family-669#LTR_Copia	1	Ecl
Ety_rnd-1_family-0	LTR/Copia	TGTTAGAAC	6714	4	unique	NA	284	no	no	no	NA	LTR/Copia	rnd-1_family-0#LTR_Copia	3	Ecl, Ety
Ety_rnd-1_family-2	LTR/Copia	TGTTAATTC	6742	4-5	unique	NA	259	yes?	no	no	NA	LTR/Copia	rnd-1_family-2#LTR_Copia	5	Ety
Ety_rnd-1_family-4	LTR/Gypsy	TGTTAGGAC	7676	5	unique	NA	236	no	no	no	RT_like super family	LTR/Gypsy	rnd-1_family-4#LTR_Gypsy	1	Ety
Ety_rnd-1_family-5	LTR/Copia.inc	TATAAATAA.	4596	NA	NA	NA	NA	no	no	5'	NA	LTR/Copia	rnd-1_family-5#LTR_Gypsy	1	Ety
Ety_rnd-1_family-9	LTR/Gypsy	TGTTATGGA	6782	5	unique	NA	226	no	no	no	RT_like super family	LTR/Gypsy	rnd-1_family-9#LTR_Gypsy	1	Ety
Ety_rnd-1_family-10	LTR/Gypsy	TGTGAGGCC	6586	4-5	unique	NA	193	no	no	no	RT_LTR, RNase_H_l	LTR/Gypsy	rnd-1_family-10#LTR_Gypsy	15	Ecl, Ety
Ety_rnd-1_family-11	LTR/Gypsy	TGTAAGGGC	7043	5	unique	NA	204	no	no	no	RT_like super family	LTR/Gypsy	rnd-1_family-11#LTR_Gypsy	1	Ety
Ety_rnd-1_family-15	Simple_repeat	CTAACAATC	7793	NA	NA	NA	NA	yes	no	no	NA	NA	rnd-1_family-15#Unknown	3	Ety
Ety_rnd-1_family-16	LTR/Gypsy	TGTTACGTG	10733	4-5	unique	NA	194	no	no	no	RT_like super family	LTR/Gypsy	rnd-1_family-16#LTR_Gypsy	1	Ety
Ety_rnd-1_family-19	DNA/hAT.inc	CACTACATA	4527	NA	NA	NA	NA	yes	no	3'	Dimer_Tnp_hAT	DNA/hAT	rnd-1_family-19#buffer	1	Ety
Ety_rnd-1_family-20	Unknown.inc	AGTTAGTAA	13032	NA	NA	NA	NA	no	no	both	NA	NA	rnd-1_family-20#Unknown	2	Ety
Ety_rnd-1_family-23	Unknown.inc	AAAATAATT	4378	NA	NA	NA	NA	yes?	yes?	both	NA	NA	rnd-1_family-23#Unknown	1	Ety
Ety_rnd-1_family-32	LINE/Tad1	AATAGATAT	6790	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-32#LINE_Tad1	2	Ety
Ety_rnd-1_family-33	DNA/hAT.inc	ATAATCAGA	3686	NA	NA	NA	NA	yes	no	both	NA	DNA/hAT	rnd-1_family-33#buffer	2	Ety
Ety_rnd-1_family-34	Unknown.inc	ATTAATTAC	11189	NA	NA	NA	NA	no	yes?	both	NA	NA	rnd-1_family-34#Unknown	1	Ety
Ety_rnd-1_family-35	Unknown.inc	ACTATTAAG	6655	NA	NA	NA	NA	yes?	yes?	both	NA	NA	rnd-1_family-35#Unknown	1	Ety
Ety_rnd-1_family-39	MITE	GGACATCTA	376	9	unique	116	NA	no	no	no	NA	NA	rnd-1_family-39#Unknown	1	Ety
Ety_rnd-1_family-44	MITE	GGACATCTA	429	6-8	unique	59	NA	no	no	no	NA	NA	rnd-1_family-44#Unknown	3	Ety
Ety_rnd-1_family-45	LTR/Copia	TGTTGAGAA	5830	5	unique	NA	248	no	no	no	RNase_H_like super f	NA	rnd-1_family-45#LTR_Copia	2	Ety
Ety_rnd-1_family-46	DNA/MuLE	GGACCTCTA	2941	8-9	unique	84	NA	no	no	no	NA	NA	rnd-1_family-46#DNA_MuLE-M	3	Ecl, Ety
Ety_rnd-1_family-47	LTR	TGTGAGGCC	583	4-5	unique	NA	172	no	no	no	NA	NA	rnd-1_family-47#Unknown	6	Ety, Epo
Ety_rnd-1_family-49	Satellite.dimer	GCCTCGACC	264	NA	NA	NA	NA	no	yes	no	NA	NA	rnd-1_family-49#Unknown	1	Ety
Ety_rnd-1-family-52	Simple_repeat	CAGAGCTTT	1825	NA	NA	NA	NA	yes	no	no	NA	NA	rnd-1_family-52#Unknown	1	Ety
Ety_rnd-1_family-53	MITE	TATCTCCTTA	243	2	TA	30	NA	no	no	no	NA	NA	rnd-1_family-53#Unknown	7	Ecl, Ety
Ety_rnd-1_family-54	LTR	TGTCAGGTC	631	5	unique	NA	106	no	no	no	NA	NA	rnd-1_family-54#Unknown	5	Ecl, Ety
Ety_rnd-1_family-56	Unknown.inc	ATAAAGATA	6793	NA	NA	NA	NA	yes?	yes?	both	NA	NA	rnd-1_family-56#Unknown	1	Ety
Ety_rnd-1_family-57	MITE	GGACATCTA	398	8-9	unique	85	NA	no	no	no	NA	NA	rnd-1_family-57#Unknown	2	Ety
Ety_rnd-1_family-58	MITE	GGGCATCTA	442	8-9	unique	84	NA	no	no	no	NA	NA	rnd-1_family-58#Unknown	1	Ety
Ety_rnd-1_family-59	MITE	GGGCAACTA	430	8-9	unique	74	NA	no	no	no	NA	NA	rnd-1_family-59#Unknown	1	Ety

Name	TE class	Consensus	Length	TSD len	TSD motif	TIR len	LTR len	SR	Sat	.inc	Conserved domain	Repbase	Automatic Classification	Cluster n	Cluster spp.
Ety_rnd-1_family-60	LTR/Gypsy	TGTGTGCCG	11881	5	unique	NA	185	no	no	no	RT_like super family	LTR/Gypsy	rnd-1_family-60#LTR_Gypsy	4	Ecl, Ety, Epo
Ety_rnd-1_family-62	LTR	TGTTACGAT	2237	5	unique	NA	152	no	yes	no	NA	NA	rnd-1_family-62#buffer	1	Ety
Ety_rnd-1_family-63	MITE	ACAGTAAAT	501	NA	NA	32	NA	no	no	no	NA	NA	rnd-1_family-63#Unknown	1	Ety
Ety_rnd-1_family-65	DNA?	AGAGCGTTC	2607	NA	NA	60	NA	no	no	no	NA	NA	rnd-1_family-65#Unknown	1	Ety
Ety_rnd-1_family-66	Unknown	CAGAGCTTT	414	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-66#Unknown	3	Ety, Epo
Ety_rnd-1_family-67	LTR/Gypsy.inc	TATTGTGTG	4896	NA	NA	NA	NA	no	no	3'	RT_like super family	LTR/Gypsy	rnd-1_family-67#Unknown	1	Ety
Ety_rnd-1_family-68	LTR	TGTGGAGTC	5585	5	unique	NA	241	no	no	no	NA	NA	rnd-1_family-68#Unknown	1	Ety
Ety_rnd-1_family-71	Simple_repeat	CTATAACAC	6108	NA	NA	NA	NA	yes	yes?	no	NA	NA	rnd-1_family-71#Unknown	1	Ety
Ety_rnd-1_family-73	Unknown	TTAAAATAG	2924	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-73#Unknown	1	Ety
Ety_rnd-1_family-74	Unknown	TAGTTATTG	774	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-74#Unknown	2	Ety, Epo
Ety_rnd-1_family-75	Unknown	TGTAAAACC	3769	NA	NA	NA	NA	yes?	no	no	NA	NA	rnd-1_family-75#Unknown	1	Ety
Ety_rnd-1_family-78	Unknown.inc	CGTCCCCCT	8631	NA	NA	NA	NA	yes?	no	no	NA	NA	rnd-1_family-78#Unknown	2	Ety
Ety_rnd-1_family-79	Unknown.inc	AGATTACTA	6616	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-79#Unknown	1	Ety
Ety_rnd-1_family-80	LTR	TGTGGCTAC	2577	4	unique	NA	379	no	yes	no	NA	NA	rnd-1_family-80#Unknown	2	Ety
Ety_rnd-1_family-81	Unknown	CAGTAGGTC	99	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-81#Unknown	2	Ecl, Ety
Ety_rnd-1_family-82	Unknown	TGTCGCAAG	75	4-5	unique	NA	NA	no	no	no	NA	NA	rnd-1_family-82#Unknown	1	Ety
Ety_rnd-1_family-87	Unknown.inc	TCAGGGAAA	5111	NA	NA	NA	NA	no	no	3'	NA	NA	rnd-1_family-87#Unknown	2	Ety, Epo
Ety_rnd-1_family-88	Simple_repeat	CAGAGCTTT	1902	NA	NA	NA	NA	no	yes	no	NA	NA	rnd-1_family-88#Unknown	1	Ety
Ety_rnd-1_family-89	DNA/hAT.inc	TACACTACA	5089	NA	NA	NA	NA	yes	no	no	Dimer_Tnp_hAT	DNA/hAT	rnd-1_family-89#DNA_hAT-Ac	4	Ety, Epo
Ety_rnd-1_family-95	DNA/PIF-Harbi	GGGTGTGAT	3270	3	unique	46	NA	no	no	no	NA	NA	rnd-1_family-95#Unknown	4	Ety, Epo
Ety_rnd-1_family-96	Unknown	TGTCGGGGC	505	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-96#Unknown	1	Ety
Ety_rnd-1_family-97	Unknown	GTACTATAT	9585	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-97#Unknown	1	Ety
Ety_rnd-1_family-99	Unknown.inc	TGTATATAA	4574	NA	NA	NA	NA	no	no	both	NA	NA	rnd-1_family-99#Unknown	1	Ety
Ety_rnd-1_family-100	Unknown.inc	TGTTGAGTG	6614	NA	NA	NA	NA	no	no	3'	NA	NA	rnd-1_family-100#Unknown	5	Ety, Epo
Ety_rnd-1_family-102	DNA/hAT.inc	TATAAGCAG	4486	NA	NA	NA	NA	yes?	no	3'	Dimer_Tnp_hAT	DNA/hAT	rnd-1_family-102#Unknown	1	Ety
Ety_rnd-1_family-103	Unknown	GAGCCTATC	2083	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-103#Unknown	1	Ety
Ety_rnd-1_family-105	Unknown.inc	TTCATTCTTA	4469	NA	NA	NA	NA	no	no	both	NA	NA	rnd-1_family-105#Unknown	1	Ety
Ety_rnd-1_family-106	LTR	TGTTACGTG	2250	5	unique	NA	150	no	yes	no	NA	NA	rnd-1_family-106#Unknown	2	Ety
Ety_rnd-1_family-107	MITE	CGGATGTCC	497	2	TA	154	NA	no	no	no	NA	NA	rnd-1_family-107#Unknown	2	Ety
Ety_rnd-1_family-110	Unknown.inc	TTAATATGC	5077	NA	NA	NA	NA	no	no	both	NA	NA	rnd-1_family-110#Unknown	1	Ety
Ety_rnd-1_family-111	Unknown	TACAGATTT	1449	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-111#Unknown	2	Ety
Ety_rnd-1_family-113	LTR?	TGTTACGTG	2154	5	unique	NA	152	no	yes	no	NA	NA	rnd-1_family-113#Unknown	1	Ety
Ety_rnd-1_family-115	DNA	GAGCGTTGC	2881	8-9	unique	52	NA	no	no	no	NA	NA	rnd-1_family-115#Unknown	3	Ecl, Ety
Ety_rnd-1_family-116	DNA/hAT.inc	TTTTTATTTA	4216	NA	NA	NA	NA	yes	no	no	Dimer_Tnp_hAT	DNA/hAT	rnd-1_family-116#Unknown	1	Ety
Ety_rnd-1_family-117	MITE	AGTTAGGCC	357	2	TA	60	NA	no	no	no	NA	NA	rnd-1_family-117#Unknown	1	Ety
Ety_rnd-1_family-120	Simple_repeat	TAAAAGTTA	8130	NA	NA	NA	NA	yes	yes?	no	NA	NA	rnd-1_family-120#Unknown	1	Ety
Ety_rnd-1_family-123	LTR	TGTGATGAT	502	NA	NA	NA	170	no	no	no	NA	NA	rnd-1_family-123#Unknown	2	Ety
Ety_rnd-1_family-125	LINE?	ATAAATTAT	5826	NA	NA	NA	NA	no	no	3'	NA	NA	rnd-1_family-125#Unknown	1	Ety
Ety_rnd-1_family-131	DNA/hAT.inc	TTAAGAATA	3076	NA	NA	NA	NA	yes?	no	5'	NA	DNA/hAT	rnd-1_family-131#Unknown	1	Ety
Ety_rnd-1_family-132	Satellite.dimer	ATAACCACA	206	NA	NA	NA	NA	no	yes	no	NA	NA	rnd-1_family-132#Unknown	1	Ety
Ety_rnd-1_family-136	MITE	GGGGACGT	357	8-9	unique	91	NA	no	no	no	NA	NA	rnd-1_family-136#Unknown	1	Ety
Ety_rnd-1_family-137	Unknown.inc	ATTAGAGCT	9569	NA	NA	NA	NA	no	no	both	NA	NA	rnd-1_family-137#Unknown	1	Ety
Ety_rnd-1_family-138	MITE	AGTTAGCCN	332	2	TA	61	NA	no	no	no	NA	NA	rnd-1_family-138#Unknown	1	Ety
Ety_rnd-1_family-139	MITE	AGCTAGCCG	327	2	TA	61	NA	no	no	no	NA	NA	rnd-1_family-139#Unknown	1	Ety
Ety_rnd-1_family-141	MITE	AGTTAGCAC	344	2	TA	61	NA	no	no	no	NA	NA	rnd-1_family-141#Unknown	1	Ety
Ety_rnd-1_family-143	DNA/Tc1-Marin	ATCGTGCCA	1815	2	TA	80	NA	no	no	no	NA	NA	rnd-1_family-143#Unknown	2	Ety, Epo

Name	TE class	Consensus	Length	TSD len	TSD motif	TIR len	LTR len	SR	Sat	.inc	Conserved domain	Repbase	Automatic Classification	Cluster n	Cluster spp.
Ety_rnd-1_family-144	DNA/hAT.inc	TATAAACAG	3770	NA	NA	NA	NA	yes?	no	5'	NA	DNA/hAT	rnd-1_family-144#Unknown	1	Ety
Ety_rnd-1_family-145	MITE	GGGGATAC/	492	7-8	unique	61	NA	no	no	no	NA	NA	rnd-1_family-145#Unknown	1	Ety
Ety_rnd-1_family-146	Unknown	AGGGTGTG/	1473	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-146#Unknown	1	Ety
Ety_rnd-1_family-148	MITE	GGGCATACA	518	8	unique	57	NA	no	NA	no	NA	NA	rnd-1_family-148#Unknown	1	Ety
Ety_rnd-1_family-149	Unknown.inc	GATATTCTA	7213	NA	NA	NA	NA	no	no	both	NA	NA	rnd-1_family-149#Unknown	1	Ety
Ety_rnd-1_family-150	LTR	TGTTAAAAT	327	4	unique	NA	68	no	no	no	NA	NA	rnd-1_family-150#Unknown	2	Ety
Ety_rnd-1_family-151	LTR	TGTAAGGGT	7824	5	unique	NA	96	no	no	no	NA	NA	rnd-1_family-151#Unknown	1	Ety
Ety_rnd-1_family-154	Unknown.inc	AACTATAAC	5590	NA	NA	NA	NA	yes?	no	5'	NA	NA	rnd-1_family-154#Unknown	1	Ety
Ety_rnd-1_family-158	DNA/Tc1-Marin	CGGTGGCAC	3089	2	TA	52	NA	no	no	no	NA	NA	rnd-1_family-158#Unknown	1	Ety
Ety_rnd-1_family-161	DNA	TGGAGGTA/	1446	NA	NA	37	NA	no	no	no	NA	NA	rnd-1_family-161#Unknown	1	Ety
Ety_rnd-1_family-163	MITE?	TGAATTGCT	775	NA	NA	124	NA	no	no	no	NA	NA	rnd-1_family-163#Unknown	1	Ety
Ety_rnd-1_family-165	LTR	TGTAAGGGT	8127	5	unique	NA	94	no	no	no	NA	NA	rnd-1_family-165#Unknown	1	Ety
Ety_rnd-1_family-166	LTR/Copia	TGTTGGATT	5697	5	unique	NA	139	no	no	no	RVT_2 super family	NA	rnd-1_family-166#Unknown	4	Ecl, Ety
Ety_rnd-1_family-173	Unknown.inc	CGGTGGCAC	4202	NA	NA	NA	NA	no	yes?	3'	NA	NA	rnd-1_family-173#Unknown	1	Ety
Ety_rnd-1_family-174	Unknown.inc	TTACTATCCT	3742	NA	NA	NA	NA	no	no	5'	NA	NA	rnd-1_family-174 #Unknown	1	Ety
Ety_rnd-1_family-175	DNA/hAT.inc	TGTAATGTT	3573	NA	NA	NA	NA	yes?	no	3'	NA	DNA/hAT	rnd-1_family-175#Unknown	1	Ety
Ety_rnd-1_family-176	MITE	GGCTCCATG	453	8-9	unique	73	NA	no	no	no	NA	NA	rnd-1_family-176#Unknown	1	Ety
Ety_rnd-3_family-14	Simple_repeat	TAATTAGAG	1756	d	NA	NA	NA	yes	yes	no	NA	NA	rnd-3_family-14#Simple_repeat	1	Ety
Ety_rnd-3_family-15	Unknown	TATTGATTT	2932	NA	NA	NA	NA	no	no	no	NA	NA	rnd-3_family-15#Unknown	1	Ety
Ety_rnd-3_family-18	Satellite.dimer	AAAAGTGG/	202	NA	NA	NA	NA	no	yes	no	NA	NA	rnd-3_family-18#Simple_repeat	1	Ety
Ety_rnd-3_family-40	LTR/Gypsy	TGTTAGGAC	7811	5	unique	NA	263	no	no	no	RT_LTR	LTR/Gypsy	rnd-3_family-40#LTR	2	Ety
Ety_rnd-4_family-57	LTR	TGTTGATTC	6651	5	unique	NA	255	no	no	no	NA	NA	rnd-4_family-57#LTR_Copia	1	Ety
Ety_rnd-4_family-116	LTR/Copia.inc	TTATTGAAA	9377	NA	NA	NA	NA	no	no	no	RNase_H_like super f	LTR/Copia	rnd-4_family-116#LTR_Copia	1	Ety
Ety_rnd-4_family-122	LTR/Gypsy.inc	TTTAGGAGC	7379	NA	NA	NA	NA	yes?	yes	no	RT_like super family	LTR/Gypsy	rnd-4_family-122#LTR_Copia	1	Ety
Ety_rnd-4_family-179	Simple_repeat	AGAGCTTTG	1874	NA	NA	NA	NA	yes	no	no	NA	NA	rnd-4_family-179#Unknown	1	Ety
Ety_rnd-4_family-223	MITE	CAACACGTA	249	6-9	unique	112	NA	no	no	no	NA	NA	rnd-4_family-233#Unknown	2	Ety
Ety_rnd-4_family-495	Unknown	TGTGAGACC	183	5	unique	NA	NA	no	no	no	NA	NA	rnd-4_family-495#Unknown	1	Ety
Epo_ltr-1_family-1	LTR	TGTTACGTG	2015	4-5	unique	NA	175	no	no	no	NA	NA	ltr-1_family-1#LTR_Unknown	1	Epo
Epo_ltr-1_family-2	LTR	TGTTACGTG	1984	5	unique	NA	151	no	no	no	NA	NA	ltr-1_family-2#Satellite	1	Epo
Epo_ltr-1_family-3	LTR	TGTTACGAT	1750	4-5	unique	NA	154	no	no	no	NA	NA	ltr-1_family-3#LTR_Unknown	1	Epo
Epo_ltr-1_family-4	LTR	TGTTACGAT	2117	5	unique	NA	154	no	yes?	no	NA	NA	ltr-1_family-4#LTR_Unknown	1	Epo
Epo_ltr-1_family-6	LTR/Copia	TGTTGCCCG	7157	5	unique	NA	287	no	no	no	RNase_H_like super f	LTR/Copia	ltr-1_family-6#LTR_Copia	5	Ety, Epo
Epo_ltr-1_family-7	LTR/Gypsy.inc	TAATAAGAT	2162	NA	NA	NA	NA	no	no	3'	NA	NA	ltr-1_family-7#LTR_Unknown	1	Epo
Epo_ltr-1_family-8	LTR/Gypsy	TGTGAGGCC	11832	NA	NA	NA	NA	no	no	no	RNase_H_like super f	LTR/Gypsy	ltr-1_family-8#LTR_Gypsy	1	Epo
Epo_ltr-1_family-10	LTR	AGCTTTGGT	8178	NA	NA	NA	NA	no	no	no	NA	NA	ltr-1_family-10#LTR_Gypsy	2	Ety, Epo
Epo_ltr-1_family-11	LTR	TTTGACCGC	4096	NA	NA	NA	184	no	no	no	NA	NA	ltr-1_family-11#LTR_Unknown	1	Epo
Epo_ltr-1_family-12	LTR/Copia	TGTTGACCG	5374	4	unique	NA	173	no	no	no	NA	NA	ltr-1_family-12#LTR_Copia	1	Epo
Epo_ltr-1_family-13	LTR/Copia	TGTTAGATC	6739	NA	NA	NA	258	no	no	no	NA	LTR/Copia	ltr-1_family-13#LTR_Unknown	3	Epo
Epo_ltr-1_family-16	LTR/Copia	TGTTAGAAC	6660	5	unique	NA	288	no	no	no	NA	LTR/Copia	ltr-1_family-16#LTR_Copia	2	Epo
Epo_ltr-1_family-17	LTR	TGTCAGGAA	5259	4	unique	NA	216	no	no	no	NA	NA	ltr-1_family-17#LTR_Unknown	2	Epo
Epo_ltr-1_family-19	LTR/Gypsy.inc	CATAGAGAT	3455	NA	NA	NA	NA	no	no	5'	NA	LTR/Gypsy	ltr-1_family-19#LTR_Unknown	1	Epo
Epo_rnd-1_family-0	LTR/Gypsy	TGTTAGGAC	10218	5-6	unique	NA	262	no	no	no	RT_LTR	LTR/Gypsy	rnd-1_family-0#LTR_Gypsy	1	Epo
Epo_rnd-1_family-1	LTR/Gypsy	TGTGAGGCC	6494	5-6	unique	NA	158	no	no	no	RNase_H_like super f	LTR/Gypsy	rnd-1_family-1#LTR_Gypsy	1	Epo
Epo_rnd-1_family-6	LTR/Gypsy	TGTTATGGA	6792	4-5	unique	NA	221	no	no	no	RT_like super family,	LTR/Gypsy	rnd-1_family-6#LTR_Gypsy	2	Epo
Epo_rnd-1_family-7	LTR/Gypsy	TGTAAGGGC	7114	4-6	unique	NA	215	no	no	no	RT_LTR	LTR/Gypsy	rnd-1_family-7#LTR_Gypsy	2	Epo
Epo_rnd-1_family-12	MITE	TATTGATTT	586	2	TA	118	NA	no	no	no	NA	NA	rnd-1_family-12#Unknown	1	Epo

Name	TE class	Consensus	Length	TSD len	TSD motif	TIR len	LTR len	SR	Sat	.inc	Conserved domain	Repbase	Automatic Classification	Cluster n	Cluster spp.
Epo_rnd-1_family-14	LTR?	AGAGCTTTG	1739	NA	NA	NA	NA	yes?	no	no	NA	NA	rnd-1_family-14#Unknown	1	Epo
Epo_rnd-1_family-15	LTR/Gypsy	TGTTACGTG	10280	5	unique	NA	200	no	no	no	NA	LTR/Gypsy	rnd-1_family-15#LTR_Gypsy	2	Epo
Epo_rnd-1_family-17	Satellite.dimer	ATCAGAGCA	188	NA	NA	NA	NA	no	yes	no	NA	NA	rnd-1_family-17#Unknown	1	Epo
Epo_rnd-1_family-20	Unknown	TATTAATAA	796	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-20#Unknown	2	Epo
Epo_rnd-1_family-21	DNA/hAT.inc	ATACTATAN	4560	NA	NA	NA	NA	yes?	no	no	Dimer_Tnp_hAT	DNA/hAT	rnd-1_family-21#LTR_Gypsy	4	Epo
Epo_rnd-1_family-22	LTR	TGTTACGAT	1822	4-5	unique	NA	157	no	no	no	NA	NA	rnd-1_family-22#Unknown	1	Epo
Epo_rnd-1_family-26	LINE	AAAACCTTC	6510	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-26#LINE_Tad1	2	Epo
Epo_rnd-1_family-28	DNA/PIF-Harbi	GGGTCTGAT	3074	3	unique	34	NA	no	no	no	NA	NA	rnd-1_family-28#DNA_PIF-Harl	4	Ecl, Epo
Epo_rnd-1_family-31	LTR/Gypsy	TGTGTGCCG	10447	NA	NA	NA	NA	no	no	no	NA	LTR/Gypsy	rnd-1_family-31#LTR_Gypsy	1	Epo
Epo_rnd-1_family-35	LTR?	TGTGTGGCT	2449	4-6	unique	NA	393	no	yes	no	NA	LTR/Gypsy	rnd-1_family-35#Unknown	1	Epo
Epo_rnd-1_family-38	MITE	GTCCACCCC	265	2-4	unique	32	NA	no	no	no	NA	NA	rnd-1_family-38#Unknown	2	Epo
Epo_rnd-1_family-40	LTR?	TGTGTGGCT	2989	4-6	unique	NA	216	no	yes	no	NA	NA	rnd-1_family-40#Unknown	1	Epo
Epo_rnd-1_family-41	MITE	TCTCCTTGA	135	2	TA	56	NA	no	no	no	NA	NA	rnd-1_family-41#Unknown	2	Epo
Epo_rnd-1_family-43	MITE	CAATACACC	266	2	TA	35	NA	no	no	no	NA	NA	rnd-1_family-43#Unknown	1	Epo
Epo_rnd-1_family-46	Satellite.dimer	GTAAAAAGC	108	NA	NA	NA	NA	no	yes	no	NA	NA	rnd-1_family-46#Unknown	3	Epo
Epo_rnd-1_family-51	LTR/Gypsy	AGTTGGAGC	10927	5	unique	NA	157	no	no	no	RNase_H_like super f	NA	rnd-1_family-51#LTR_Gypsy	2	Epo
Epo_rnd-1_family-61	DNA/hAT.inc	TAACTTAAA	3470	NA	NA	NA	NA	yes?	no	no	Dimer_Tnp_hAT	DNA/hAT	rnd-1_family-61#DNA_hAT-Ac	1	Epo
Epo_rnd-1_family-66	Unknown	CCCCTCCCA	837	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-66#Unknown	1	Epo
Epo_rnd-1_family-67	MITE	CAGTCTAAC	194	2	TA	35	NA	no	no	no	NA	NA	rnd-1_family-67#Unknown	1	Epo
Epo_rnd-1_family-68	MITE	CAGTCTAAC	251	2-4	unique	40	NA	no	no	no	NA	NA	rnd-1_family-68#Unknown	1	Epo
Epo_rnd-1_family-71	MITE	TCAACACGT	340	NA	NA	113	NA	no	no	no	NA	NA	rnd-1_family-71#Unknown	1	Epo
Epo_rnd-1_family-72	Unknown.inc	ATCCTAATA	2515	NA	NA	NA	NA	no	no	5'	NA	NA	rnd-1_family-72#Unknown	1	Epo
Epo_rnd-1_family-78	Satellite.dimer	TATATACGT	120	NA	NA	NA	NA	no	yes	no	NA	NA	rnd-1_family-78#Unknown	1	Epo
Epo_rnd-1_family-80	Unknown	TATTGTGTG	2130	NA	NA	NA	NA	no	yes	no	NA	NA	rnd-1_family-80#Unknown	1	Epo
Epo_rnd-1_family-83	DNA/Tc1-Marin	CAGTGGGTC	1396	2	TA	20	NA	no	no	no	NA	NA	rnd-1_family-83#Unknown	1	Epo
Epo_rnd-1_family-85	LTR/Gypsy.inc	ATAACTAAG	4894	NA	NA	NA	NA	no	no	5'	NA	NA	rnd-1_family-85#Unknown	3	Ecl, Epo
Epo_rnd-1_family-86	Unknown	TGTGGAGTC	253	3	unique	NA	NA	no	no	no	NA	NA	rnd-1_family-86#Unknown	1	Epo
Epo_rnd-1_family-89	Unknown.inc	TCTTAGAGT	6974	NA	NA	NA	NA	no	no	3'	NA	NA	rnd-1_family-89#Unknown	2	Epo
Epo_rnd-1_family-91	LTR.inc?	ATAAATAGA	6089	NA	NA	NA	NA	no	no	3'	NA	NA	rnd-1_family-91#Unknown	2	Epo
Epo_rnd-1_family-93	MITE?	GGACCTCTA	389	NA	NA	59	NA	no	no	no	NA	NA	rnd-1_family-93#Unknown	1	Epo
Epo_rnd-1_family-94	MITE	TATCTCTTTC	271	NA	NA	35	NA	no	no	no	NA	NA	rnd-1_family-94#Unknown	2	Epo
Epo_rnd-1_family-96	Unknown	TATTTTGTC	2166	NA	NA	NA	NA	no	yes	no	NA	NA	rnd-1_family-96#Unknown	2	Epo
Epo_rnd-1_family-97	MITE	TCCCTTTGA	270	2	TA	36	NA	no	no	no	NA	NA	rnd-1_family-97#Unknown	1	Epo
Epo_rnd-1_family-98	LTR	TGTCATGGT	496	5	unique	NA	168	no	no	no	NA	NA	rnd-1_family-98#Unknown	1	Epo
Epo_rnd-1_family-109	Unknown.inc	TGTCACAGA	4886	NA	NA	NA	NA	no	no	3'	NA	NA	rnd-1_family-109#Unknown	1	Epo
Epo_rnd-1_family-115	DNA/hAT?	GGGCGTGG	2531	8	unique	54	NA	no	no	no	NA	NA	rnd-1_family-115#Unknown	1	Epo
Epo_rnd-1_family-117	DNA/hAT?	GGGCGTGG	2483	8	unique	60	NA	no	no	no	NA	NA	rnd-1_family-117#Unknown	1	Epo
Epo_rnd-1_family-118	LTR	TGTTAGAAT	1792	NA	NA	NA	81	no	no	no	NA	NA	rnd-1_family-118#LTR_Copia	1	Epo
Epo_rnd-1_family-120	MITE	GCCTAGCAG	192	2	TA	88	NA	no	no	no	NA	NA	rnd-1_family-120#Unknown	1	Epo
Epo_rnd-1_family-126	DNA/hAT?	TAGGGCGTC	2647	7-8	unique	57	NA	no	no	no	NA	NA	rnd-1_family-126#Unknown	3	Ety, Epo
Epo_rnd-1_family-127	LTR.inc	TGTTAAGGT	2091	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-127#Unknown	1	Epo
Epo_rnd-1_family-128	LTR.inc	GTGAAATAC	5715	NA	NA	NA	NA	no	no	5'	NA	NA	rnd-1_family-128#LTR_Gypsy	1	Epo
Epo_rnd-1_family-129	LTR.inc	TGTCTTATT	6629	NA	NA	NA	NA	no	no	both	NA	NA	rnd-1_family-129#LTR_Gypsy	1	Epo
Epo_rnd-1_family-130	Unknown	ACAGATCTT	86	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-130#Unknown	2	Epo
Epo_rnd-1_family-132	DNA/MuLE?	GTGAATTGC	2639	7-8	unique	125	NA	no	no	no	DDE_Tnp_ISL3 supe	NA	rnd-1_family-132#Unknown	1	Epo
Epo_rnd-1_family-134	LINE?	AAGAATTAC	5480	NA	NA	NA	NA	no	no	no	NA	NA	rnd-1_family-134#Unknown	1	Epo

Name	TE class	Consensus	Length	TSD len	TSD motif	TIR len	LTR len	SR	Sat	.inc	Conserved domain	Repbase	Automatic Classification	Cluster n	Cluster spp.
Epo_rnd-1_family-135	LINE	AAAGCTCTA	3536	NA	NA	NA	NA	no	no	no	RT_like super family	NA	rnd-1_family-135#LINE_Tad1	1	Epo
Epo_rnd-2_family-40	MITE	CAATACACC	280	2	TA	45	NA	no	no	no	NA	NA	rnd-2_family-40#Unknown	1	Epo
Epo_rnd-2_family-53	Unknown	TAATGCTTA	1502	NA	NA	NA	NA	no	no	no	NA	NA	rnd-2_family-53#Unknown	1	Epo
Epo_rnd-2_family-54	Simple_repeat	AGAGCATTG	2152	NA	NA	NA	NA	yes	yes?	no	NA	NA	rnd-2_family-54#Unknown	1	Epo
Epo_rnd-3_family-26	LTR/Gypsy	AAGCCTAGC	13686	NA	NA	NA	NA	no	no	no	RT_LTR	LTR/Gypsy	rnd-3_family-26#LTR_Gypsy	1	Epo
Epo_rnd-3_family-47	MITE?	AGTTGNGCT	442	NA	NA	52	NA	no	yes	no	NA	NA	rnd-3_family-47#Unknown	1	Epo
Epo_rnd-3_family-55	Unknown	AANTGTATC	236	NA	NA	NA	NA	no	no	no	NA	NA	rnd-3_family-55#Unknown	1	Epo
Epo_rnd-3_family-170	LTR/Gypsy	TGTCACGAC	7832	5	unique	NA	259	no	no	no	RT_like super family	LTR/Gypsy	rnd-3_family-170#LTR	2	Epo
Epo_rnd-4_family-186	Unknown	TGTTGGAAT	2044	NA	NA	NA	NA	no	no	5'	NA	NA	rnd-4_family-186#Unknown	1	Epo
Epo_rnd-4_family-200	Unknown	GGCACACAC	8133	NA	NA	NA	NA	no	no	no	NA	NA	rnd-4_family-200#Unknown	1	Epo
Epo_rnd-4_family-214	Simple_repeat	TATACAGAG	2060	NA	NA	NA	NA	yes	no	no	NA	NA	rnd-4_family-214#Unknown	1	Epo
Epo_rnd-4_family-298	LTR/Gypsy	GAACCCTTA	6498	NA	NA	NA	NA	no	no	no	RT_like super family	LTR/Gypsy	rnd-4_family-298#LTR_Gypsy	1	Epo

Chapter 3

Transposable elements contribute to evolution in the *Epichloë typhina* species complex

3.1 Abstract

In recent years, there has been growing enthusiasm around the roles of transposable elements (TEs) in gene regulation and genome evolution. Currently, a large body of evidence suggests TEs accelerate the evolution of virulence-related genes in plant-associated fungi, aiding in the antagonistic co-evolution with their host plants. However, the potentially deleterious impact of TEs have given rise to adaptive genome defence systems that prevent proliferation of these elements. One defence mechanism unique to fungi is repeat-induced point mutations (RIP). Previous studies reported that the extent of RIP is so great in *Epichloë* that it is unclear if active elements remain in these genomes. As a result, it is unknown if TEs have contributed to evolution and gene regulation in the recent history of this genus.

In this chapter, I investigate the contribution of TEs to the evolution of three sub-species of the *Epichloë typhina* species complex. All three genomes showed diversity in TE composition and frequency, and the invasion and subsequent inactivation of TEs largely explained the between-lineage variation in genome size. In addition, many sequences within the TE repertoire of each species remained unaffected by RIP. Taken together, these results suggest that TEs drive the evolution of *Epichloë typhina*, and active element remain in these genomes. Above all, a striking example of TE-mediated lineage-specific expansion was observed in *Epichloë typhina* subsp. *clarkii*, evident in the correlation between genome size and recent expansion of LTR Copia elements. In addition to identifying recent activity of TEs within this species complex, this work also demonstrated that MITEs are overrepresented near two classes of genes involved in the invasion of plants. These results will provide a foundation for future functional studies to elucidate regulatory roles of TEs in *Epichloë.*

3.2 Introduction

In recent years, TEs have been reappraised in pathogenic fungi of plants as important contributors to evolution. Although fungal genomes are highly diverse, comparative analyses have revealed a number of common concepts. Namely, a distinct, compartmentalised genome structure has been reported in several fungal species as a result of non-random TE distribution [67, 82, 138]. This structure, termed "two speed genomes," is an archetypal model of genome plasticity in plant-associated fungi. Here, physical compartmentalisation of the genome gives rise to a discrete bipartite architecture consisting of gene-dense/TE-poor regions that alternate with with gene-poor/TE-dense regions [136]. The two-speed genome presents a number of adaptive advantages for plant-associated fungi. In many characterized species, conserved housekeeping genes are present in the gene-dense regions, and the gene-sparse/TE-rich region harbour an enriched repertoire of effector genes (Figure 1.5) [136, 186]. Effectors, a collective term for a suite of secreted proteins, mediate the invasion of plants by modulating the plants immune response. Markedly, the gene-sparse/TE-rich regions that harbour the effectors appear to evolve faster than gene dense regions as a result of TE-mediated instability. This prompts uneven rates – or two speeds – of evolution across the genome [104, 136]. The localisation of effector genes proximal to TEs is proposed to affect the pathogenicity and host range of fungi. By consensus, it is agreed that these two-speed genomes create a niche for accelerated fungal evolution, allowing the fungi to undergo rapid evolution of virulence-related genes and outpace detection by its host plant [136].

Epichloë is an agriculturally important genus of filamentous fungi that can form intimate mutualistic relationships with cool season grasses [144]. Due to the bioprotective benefits received by *Epichloë*-infected hosts, there is considerable interest in developing novel strains with extended host ranges for agricultural purposes. *Epichloë* genomes harbour the compartmentalised two-speed genomic structure, and insight on

two sibling strains within the *Epichloë typhina* species complex determined that the TE-dense regions underpin the divergence between two otherwise syntenic sibling strains. Further, recent studies in *E. festucae* demonstrate the genome is replete with TEs that play a profound role in the 3D organisation of the genome, and some elements (MITEs) are associated with the regulation of effector candidate genes [67, 158]. However, in these *Epichloë* strains, the extent of RIP is so great, that it is unclear whether active elements still contribute to the genome evolution of this genus. Hence, the dynamics of TEs in *Epichloë* remains obscure.

In this study, I present a complete genome for *Epichloë typhina* subsp. *poae*, and focus on three closely related sub-species within the *E. typhina* complex: *E. typhina* subsp. *typhina* (Ety), *E. typhina* subsp. *poea* (Epo), and *E. typhina* subsp. *clarkii* (Ecl). These sub-species will henceforth be referred to as *lineages*, with *sibling strain* referring specifically to the relationship between Ecl and Ety. The *E. typhina* species complex is genetically differentiated in natural populations, and gene flow between lineages is restricted or prevented by virtue of host-specificity [159, 160]. The selected strains of these sub-species therefore provide an example of ecologically distinct lineages that have adapted to different grass hosts across a short evolutionary timeline (estimated to be within the last 10 million years; Section 1.6)). Here I use the first TE library manually curated for *Epichloë* to test the hypotheses that TEs still contribute to genome evolution over a time-scale at which they may contribute to host adaptation, and investigate the association between TEs and genes that underpin plant-colonisation.

3.3 Methods

3.3.1 Data availability

A repository containing analyses and scripts developed for this project are available at `https://github.com/KelliSmith17/Masters`

3.3.2 Genomic Sequences and TE annotation

The work here used the TE annotations for each of the focal genomes developed in Chapter 2.

3.3.3 Gene annotations

Initial gene prediction and functional annotations for each genome were achieved using Funannotate v1.6.0 [11], a pipeline widely used and designed for the annotation of fungal genomes. I additionally identified several classes of proteins known to be of particular importance in plant-associated fungi. Secreted proteins for each genome were predicted using SignalP (v5.0)[21], a deep neural network-based approach that predicts signal peptides and cleavage sites characteristic of secreted proteins. Candidate effector proteins (small secreted proteins that interact with the fungi's plant-host) were identified using EffectorP (v2.0)[10]. This software implements a machine learning approach that uses the composition and chemical properties of known effectors to calculate the probability that a given protein is an effector. Candidate effectors were defined as proteins with fewer than 200 amino acid residues, containing a signal peptide motif and having an EffectorP probability ≥ 0.5 [158]. Fungal secondary metabolites, proteins associated with the production of chemicals that interact with the fungi's host or environment, were predicted using Antibiotics and Secondary Metabolite Analysis Shell (antiSMASH)[3].

3.3.4 Synteny

Synteny between the three genomic sequences was analysed from pairwise alignments of all genomes produced using minimap2 (v2.2)[14] with the command line argument '-x asm20'. The resulting alignments were filtered and visualised using pafR [187].

3.3.5 Genome compartmentalisation

AT-rich compartments of each reference genome were identified using Occultercut (v.1.1)[15] This program identifies regions of distinct nucleotide composition within a given genome without requiring a threshold that defines an AT-rich region. This software assumes that the genome comprises of AT- and non-AT-rich blocks, and the provided reference genome is recursively split at genomic regions where the Jensen-Shannon divergence statistic is maximised. Nucleotide sequences of AT- and non-AT-rich regions identified by OcculterCut were extracted from reference genomes using Bedtools getfasta [4].

3.3.6 TE localisation and dating

The localisation of TE classes and orders within AT- or non-AT-rich regions of each genome was determined using Bedtools intersect [4].

The relative timing of of TE mobilisation was estimated using RepeatR [188], an R package developed as part of this work. The timeline was estimated by examining the percent deviation between a TE copy and its respective consensus sequence. This divergence performs as an indirect measure of the time lapsed since transposition.

3.3.7 RIP

A custom python script that takes advantage of BioPython [5, 16] was developed to calculate the RIP index of Margolin et al [189]. The RIP indices measure the depletion of RIP-targeted nucleotides by calculating the frequency known pre- and post-RIP dinucleotides, namely (CpA+TpG)/(ApC+GpT). By measuring these frequencies, the presence of RIP can be determined in a given sequence.

3.3.8 TE/gene associations

The hypothesis that certain TE classes are over-represented near classes of functional genes was tested using a permutation approach. The specific test used for effector proteins serves to demonstrate this approach. The number of TEs of a given class occurring within 1 kb of an effector (i.e. being close that the TE may plausibly contribute to regulation of this gene [122]) was calculated for each genome. To test whether the observed number of TEs occurring close to these genes was more than should be expected, a null distribution of this statistic was generated. This null distribution was produced by generating 1000 size-matched samples from all non-effector genes, and calculating the number of TEs neighbouring within 1 kb of these randomly selected control genes. TEs were considered to be under- or over-represented within 1 kb of an effector when they fell outside of the $0.025 - 0.975$ quantiles of the null distribution (i.e the observed value was more extreme than $> 95\%$ of samples in the null distribution).

This permutation approach was used to test the hypotheses that effector or secondary metabolite genes were (A) more likely to occur within 1 kb of a an AT-rich region of the genome and (B) more likely to have TEs of different classes within 1 kb of their coding region.

3.3.9 Analyses

Analyses were conducted using R v. 4.0.0 [17] in RStudio v 1.2.5042 [18].

3.4 Results and Discussion

3.4.1 The *E. poae* genome has undergone extensive rearrangement

The reference genome used for *E. poae* (Epo) has not appeared in a published paper. Although the sequencing and assembly of this genome was not part of this project, and will be described in an upcoming paper, I report a brief summary of this genome here. The complete Epo genome comprises of seven nuclear chromosomes and a complete mitochondrial chromosome with a total length of 38.3Mb. The genome harbours 7 620 predicted genes, of which 7 504 are predicted protein-coding genes. Of these protein coding genes, 585 (7.98%) were predicted encode secreted proteins and 130 were predicted to be putative small secreted proteins. Among these, 82 were likely to be effectors, 43 predicted to be non-effectors, and 5 unlikely effectors. In addition, 300 secondary metabolite genes were identified.

The *E. typhina* (Ety) and *E. clarkii* (Ecl) genomes are known to be highly syntenic, sharing a highly conserved core genome that is interspersed with AT-rich regions comprised of differing repeat content [148]. My results confirmed this finding, with a high degree of synteny between Ecl and Ety at the chromosome level and no evidence of large rearrangements between non-homologous chromosomes (Figure 3.1A). However, it is important to note that chromosome 3 in Ecl is homologous to chromosome 1 in Ety. This is due to the fact that chromosomes are numbered from largest to smallest in *Epichloë*. My results further confirm that the near-perfect synteny is largely restricted to the non-AT-rich regions of the Ecl and Ety genomes (Figure 3.2) [148].

In contrast, the Epo genome has undergone extensive recombination. Pairwise alignments between Epo and Ecl show each Epo chromosome is a patchwork of sequences homologous to multiple different chromosomes from Ecl and Ety (3.1B). Further, the distinct conservation of non-AT-rich regions observed between Ecl and Ety is less pronounced in Epo (Figure 3.2). The contrasting patterns of synteny revealed by

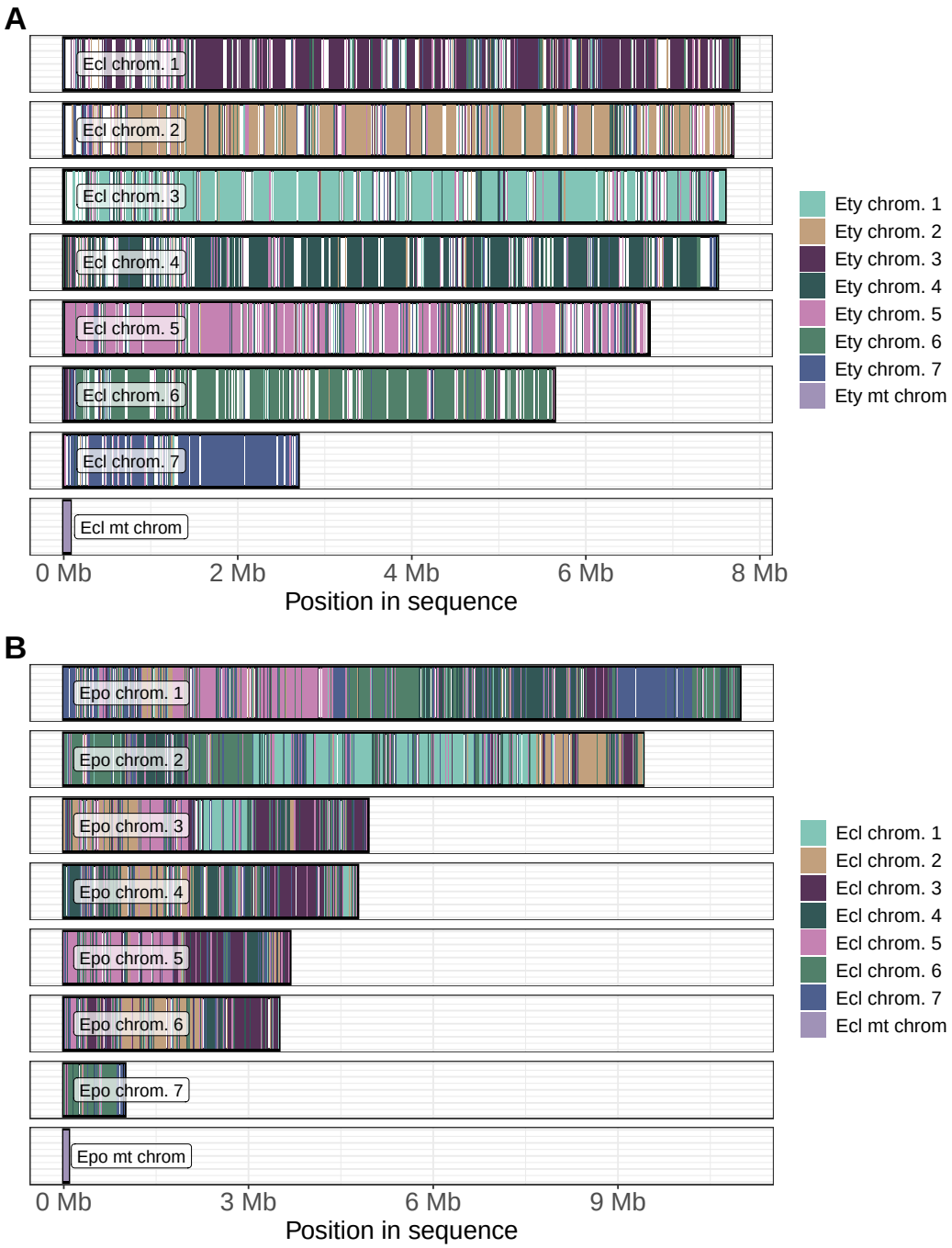

Figure 3.1: Synteny between Ecl, Epo and Ety genomes. Each subfigure represents a pairwise alignment between a reference and query genome. Each chromosome of the reference genome is shown as a box, with sections shaded according chromosome that section is homologous to in the query genome. (A) Ecl reference genome shaded by homology to Ety query genome. (B) Epo reference genome shaded by homology to Ecl query genome

this comparison suggests the Epo genome has been the subject of large scale genome rearrangements following the divergence these *E. typhina* sub-species.

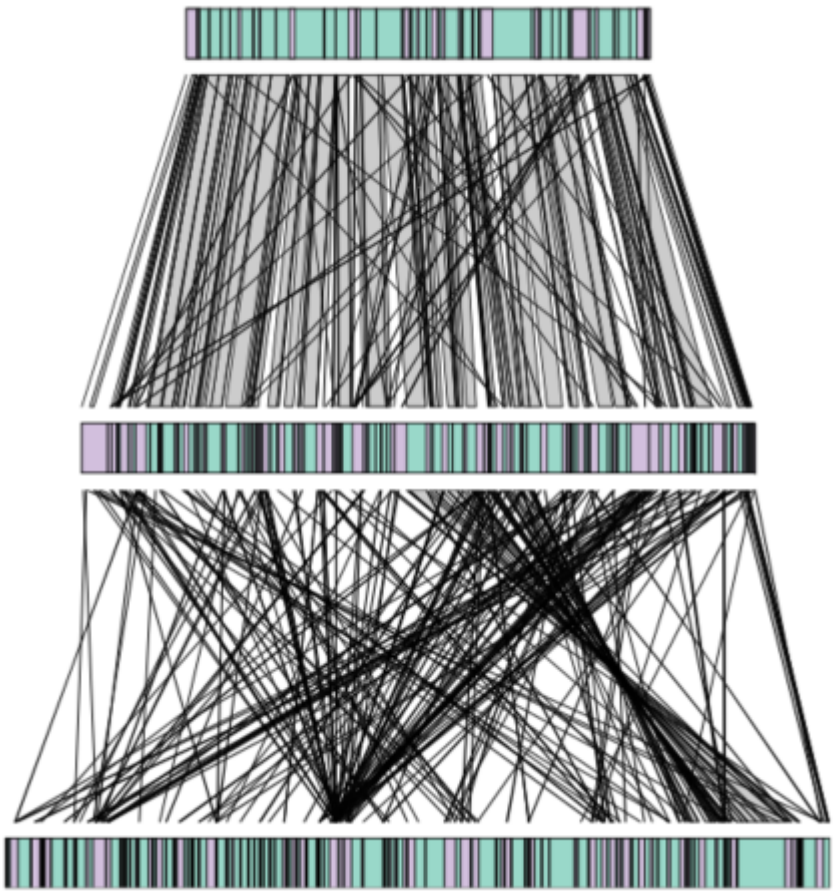

Figure 3.2: Synteny of chromosome 2. Ety (top) and Ecl (middle) demonstrate high conservation of chromosome 2. This synteny is largely confined to non-AT-rich regions (green) in comparison to AT-rich regions (purple). Each region is defined by a black border. The conservation of non-AT-rich regions is less pronounced in Epo (bottom).

Although reconstructing evolutionary events occurring in Epo is beyond the scope of this work, the large number of rearrangements in this genome is a striking and unexpected result. A recent paper examining the evolution of genome structure in 15 *Epichloë* lineages (including the previously published Ety and Ecl genomes) showed genome structure is conserved across many *Epichloë* species [161]. Although some rearrangements are observed between *Epichloë* chromosomes, the degree of rearrangement between the closely related lineages described here is unusual.

Previous studies report distinct boundaries in the genome organisation of *Epichloë* with AT-rich regions comprised mostly of repetitive DNA sequences such as TEs, and non-AT-rich regions of relatively equal nucleotide composition [67, 148]. These equilibrated non-AT-rich regions will henceforth be referred as gene-rich regions. Consistent with previous reports in *Epichloë*, the Epo genome was highly partitioned with alternating gene-rich and AT-rich blocks. The gene-rich region contained 99.71% of known genes, and 16% of known TEs. Contrarily, the AT-rich region was almost devoid of genes (0.29% of known genes) and harboured 84% of known TEs. These results are very similar to those already published for Ecl and Ety: between 99.88% and 99.93% of genes in these genomes cross a gene-rich region, and TEs localise within the AT-rich blocks [148].

3.4.2 Genome size differences are explained by AT-rich regions

Despite their recent divergence, the genomes of the three lineages vary considerably in size, with Ecl the largest (45.6Mb), its sister strain Ety the smallest size (33.8Mb), and Epo an intermediate size (38.2Mb). This between-strain variation in genome size is almost entirely explained by the AT-rich portion of each genome (Figure 3.3A), with respect to the total length, number, and genomic proportion of AT-rich blocks (Table 3.1). In line with previous reports in *Epichloë* [148], the AT-rich regions of all focal lineages were remarkably rich in A/T nucleotides (Figure 3.3B). The mean AT contents of all AT-rich blocks were 72.61%, 75.39%, and 74.51% in Ecl, Ety, and Epo, respectively. Contrarily, gene-rich regions had relatively equal nucleotide content, with a mean G/C content of 51.12% (Ecl), 51.18% (Ety), and 51.11% (Epo).

AT-rich regions in fungal genomes are typically the result of RIP inducing C-to-T transitions in repetitive sequences [114]. To confirm the contribution of RIP to the abundance of AT-rich regions in all focal genomes, I calculated the RIP indices according to Margolin *et al.* [189]. As RIP tends to be targeted on CpA and TpG

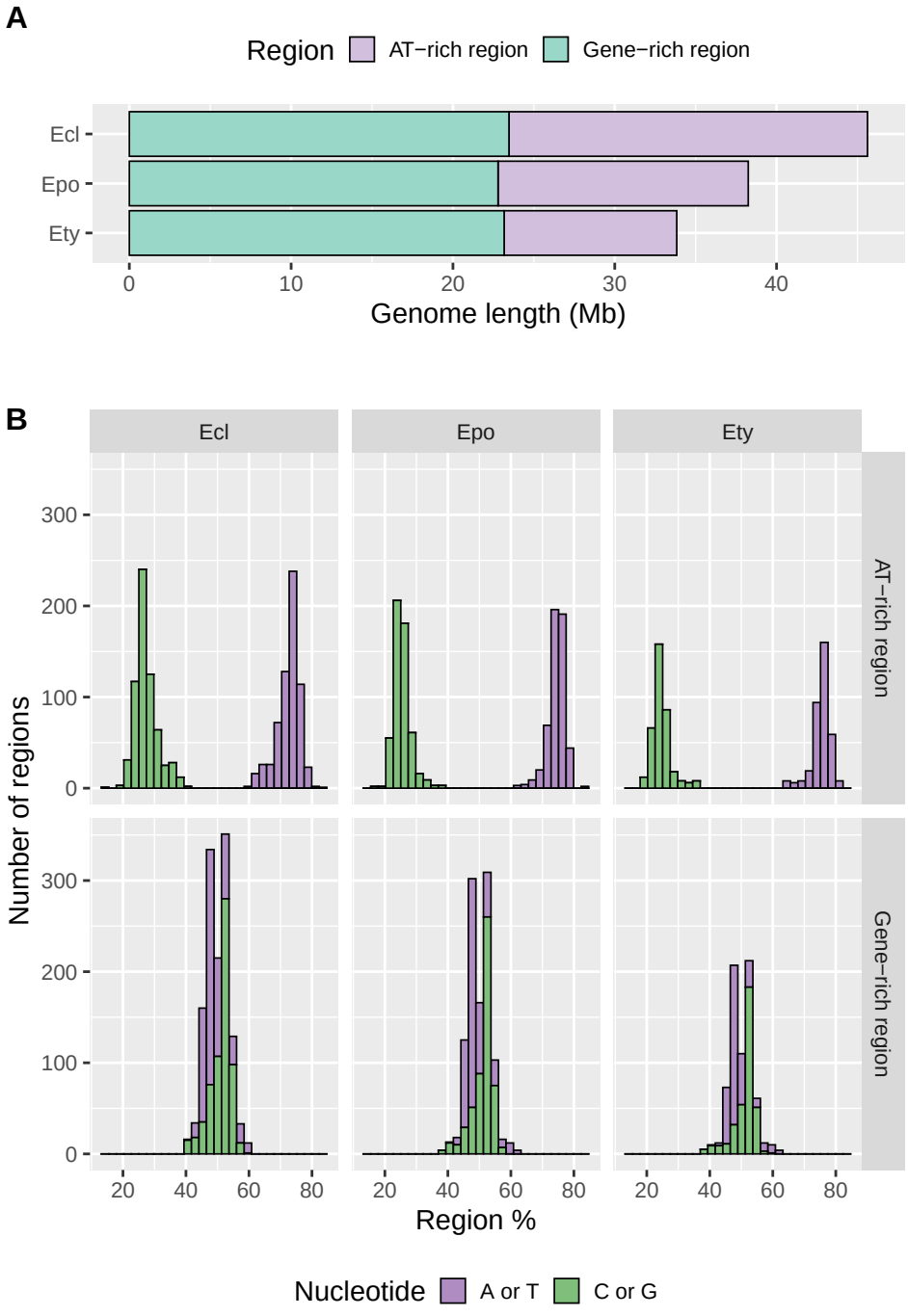

Figure 3.3: AT- and gene-rich regions. (A) the total length of AT- and gene-rich regions within each genome. (B) Nucleotide composition of AT- and gene-rich regions. The x-axis represents the percentage of A/T or C/G nucleotides, with the y-axis representing each AT-rich or gene-rich region in the genome.

Lineage	ref len (Mb)	AT n	AT len (Mb)	proportion
Ecl	45.6	648	22.16	48.57
Ety	33.8	362	10.67	31.53
Etp	38.2	538	15.46	40.43

ref len (Mb) = total length of reference genome in Mb
AT n = total number of AT-rich regions
AT len = total length of AT-rich regions in Mb
proportion = total proportion of AT-rich blocks in the genome

Table 3.1: AT rich regions

dinucleotides, thus causing a corresponding increase in TpA dinucleotides, the relative frequency of each dinucleotide can be used to calculate the action of RIP in a sequence [139]. Using (CpA+TpG)/(ApC+GpT), a RIP index of <0.9 suggests a presence of RIP [67]. The AT rich regions of our focal species were all observed to have low RIP index values, demonstrating the remarkable extent of RIP in these regions (Table 3.2). Further, the presence of RIP is substantially limited to these AT-rich regions, with very little RIP indicated in the gene-rich regions in all three genomes (Figure S3.1). Taken together, these findings confirm that the accrual of AT-rich tracts in these genomes is a consequence of RIP-recognition and deactivation following TE invasion. Thus the observed differences in genome size among the *E. typhina* lineages are underpinned by differing TE populations within each genome.

	Mean RIP index	
Lineage	AT-rich region	Gene-rich region
Ecl	0.37	1.24
Ety	0.26	1.24
Epo	0.29	1.25

Table 3.2: Mean RIP index of AT-rich and gene-rich regions of each genome

3.4.3 Distribution of TEs among genomic components

The results reported so far demonstrate that *Epichloë typhina* sub-species have the typical compartmentalised structure of an *Epichloë* genome, with distinct AT-rich regions

dominated by inactivated TEs and contrasting gene-rich regions with approximately equal nucleotide content. Nevertheless certain TE classes demonstrate insertion preferences for specific motifs, nucleotide content or, in the case of MITEs, gene-rich regions [68]. To determine if specific TE classes are present outside of the AT-rich regions, I examined the localisation of TE classes within AT- and gene-rich regions in each focal genome.

As expected, the AT-rich regions of all genomes are dominated by TEs (Figure 3.4). The AT-rich component of the each genome is mostly comprised of LTR elements, with a substantial minority of DNA and LINE elements in all species. The gene-rich regions are densely packed, with the majority of this component being occupied by gene-sequences. Interestingly, all three genomes contain a considerable number of LTR elements in the gene-rich regions. This was an unusual discovery as LTR elements are known to preferentially insert into AT-rich sequences or into other LTR elements [190]. The detection of these elements in this work may be as a result of the curated TE consensus sequences used for this TE annotation, which identified several shorter LTR families.

MITEs make up only a small proportion of each genome, and are almost entirely restricted to the gene-rich regions, consistent with reports that these elements preferentially insert into genic regions [67]. In addition to MITEs, a small proportion of DNA elements were also present in the gene-rich regions. With the exception of some LTR elements, TEs within the gene-rich region were typically very short (<500 bp; Figure S3.2). Many of these sequences are only partial fragments of their respective consensus sequences (Table S3.1), thus they may be fragmented relics of inactive TEs that are often evenly distributed throughout the genome [191].

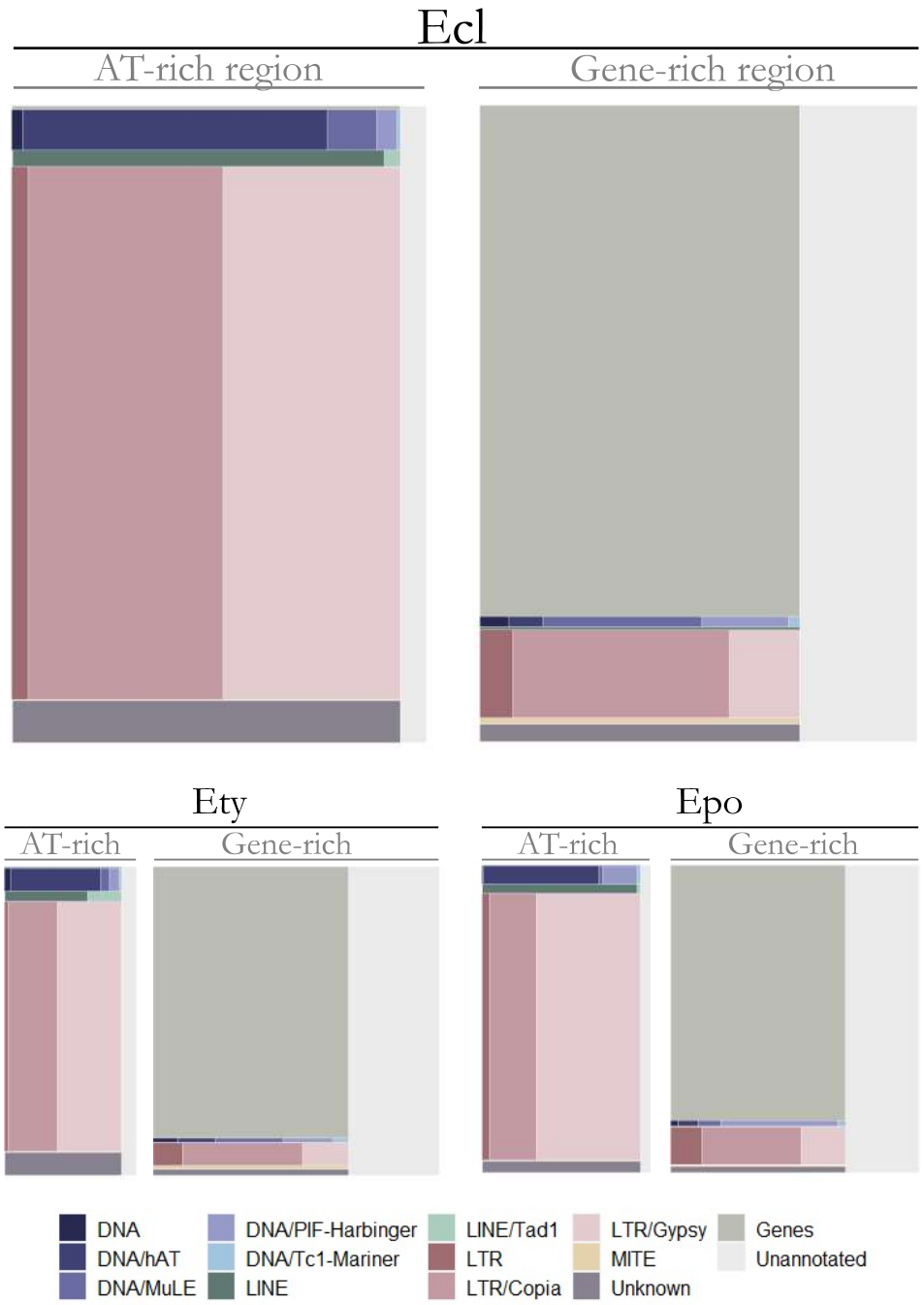

Figure 3.4: Mosaic plots demonstrating the composition of *Epichloë* genomes. Each sub-figure represents one genome, with the AT-rich and gene-rich regions of each divided into rectangles proportional to the size of each component (x axis). Each rectangle is then further subdivided according to proportion of sequences assigned to the TE classes identified in the legend (y axis).

3.4.4 Some TE sequences show no evidence of RIP

The presence of non-MITE TEs in gene-rich regions is a surprising result. Previous work on TEs in *Epichloë* [67, 148] concluded that the majority of LTR and DNA elements had been deactivated by RIP, and were thus restricted to the AT-rich regions of the genome. To better understand how RIP has acted on different classes of TEs, I calculated the RIP index for each individual TE sequence in each genome and compared the distribution of this score between TE classes. Each TE class has a characteristic distribution for the RIP index, which is reflected in each genome (Figure 3.5). It is notable that LTR Copia, LTR Gypsy, and all LINE elements have very low RIP scores, suggesting the majority of these elements have been deactivated. However, overall, TE classes in each genome show considerable variation in RIP extent.

Interestingly, several elements within each TE class exceeded the typical length threshold for RIP, however showed no evidence of RIP (Figure S3.3). It is possible that RIP is not completely effective at targeting all repetitive sequences in these lineages. These non-mutated TE sequences may represent recent transposition or copying events; *Epichloë* species that are able to reproduce asexually and rarely undergo meiosis likely demonstrate less RIP activity, as RIP predominantly occurs in the pre-meoitic mitosis events. Therefore, newer insertions in the focal lineages may have still evaded RIP. An additional explanation for these non-targeted sequences may be due to the relaxation of RIP deactivation for neutral TE insertions such as those within AT-rich regions that are therefore not proximal to genes. I tested these hypothesis by isolating TE copies that exceed the 400 bp RIP threshold and examining the presence of RIP. No notable pattern of RIP could be detected when the age and localisation of TEs was considered for these longer elements (Figure S3.4 - S3.5). Taken together, this suggests that the RIP mechanisms in the three focal genomes do not target all TEs, and active elements may remain.

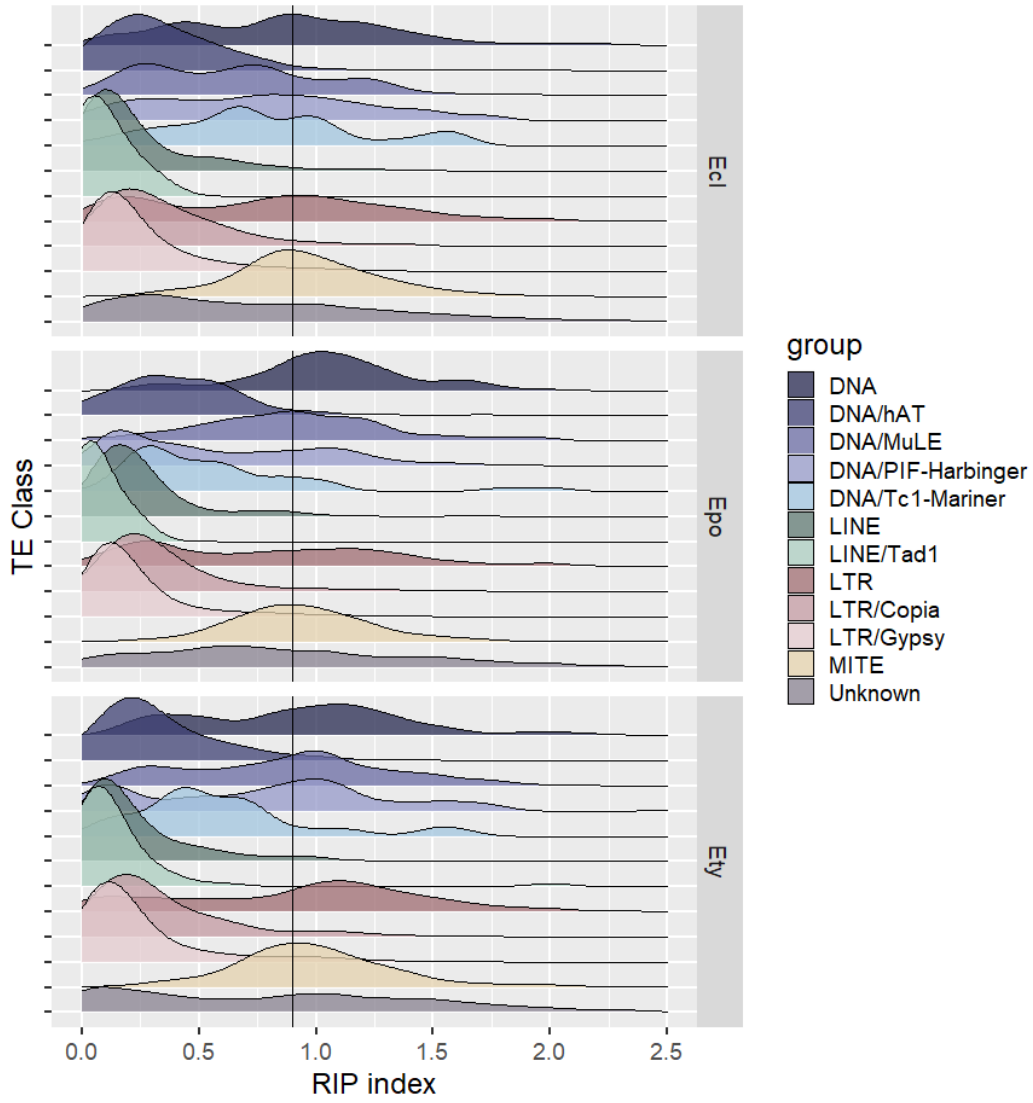

Figure 3.5: Density plot depicting distribution of RIP index scores across all TE copies per genome. Vertical line represents RIP cut-off of 0.9, where values lower than this line suggest a presence of RIP.

3.4.5 The TE repertoire of each focal genome differs

The results so far presented have demonstrated similarities in the TE content of all three focal genomes. Each genome has an AT-rich component predominantly comprised of deactivated TEs and a number TE sequences in an otherwise gene-rich component. Nevertheless, these genomes differ markedly in size as a result of the total amount of deactivated TEs present in the AT-rich component. This raises that question of

whether these genome size differences are the result of active expansion and subsequent deactivation of TE families in each lineage, or the differential loss of deactivated TEs in some lineages.

If active TEs have contributed to the genome evolution of *E. typhina* post-divergence, the frequency of repeat classes is expected to differ between each genome. To investigate this, I calculated the frequency of each TE class per genome to investigate the contribution of individual classes to each genome. I found a considerable diversity in the composition and frequency of TE families in each genome (Figure 3.6). LTR retrotransposons were the predominant constituent across all three lineages; the total length of LTR retrotransposons accounted for 78.09%, 72.66%, and 81.34% of the total length of all TEs in Ecl, Ety, and Epo, respectively (52.35%, 33.5%, and 46.72% of the total genome size in Ecl, Ety, and Epo, respectively; Table S3.1). In addition to greatest total length, LTR elements also demonstrated the highest copy numbers, particularly in Gypsy elements (Epo), and Copia (Ecl and Epo). Secondary to LTR elements, all genomes also harboured high copy numbers of MITEs, consistent with previous studies in *Epichloë* [67].

The differences in the composition of TEs in each genome described above could be explained by the loss of TE classes present in a common ancestor of all three lineages, i.e., elements that have undergone such great mutation that they are no longer detected on a basis of sequence homology. Alternatively, the difference may be underpinned by lineage specific expansions of TEs in select lineages. I tested these alternative hypotheses by investigating the age of transposition events in each genome. This was achieved by investigating the percentage differences between a given copy of each repeat in comparison to its respective consensus sequence. Elements that have recently undergone transposition will show little deviation from the consensus sequence, with older TE copies showing higher sequence divergence by virtue of natural processes such as genetic drift. All three lineages demonstrated a recent expansion of TEs (<10% divergence; Figure 3.7), demonstrating TE activity that post-dates the divergence of

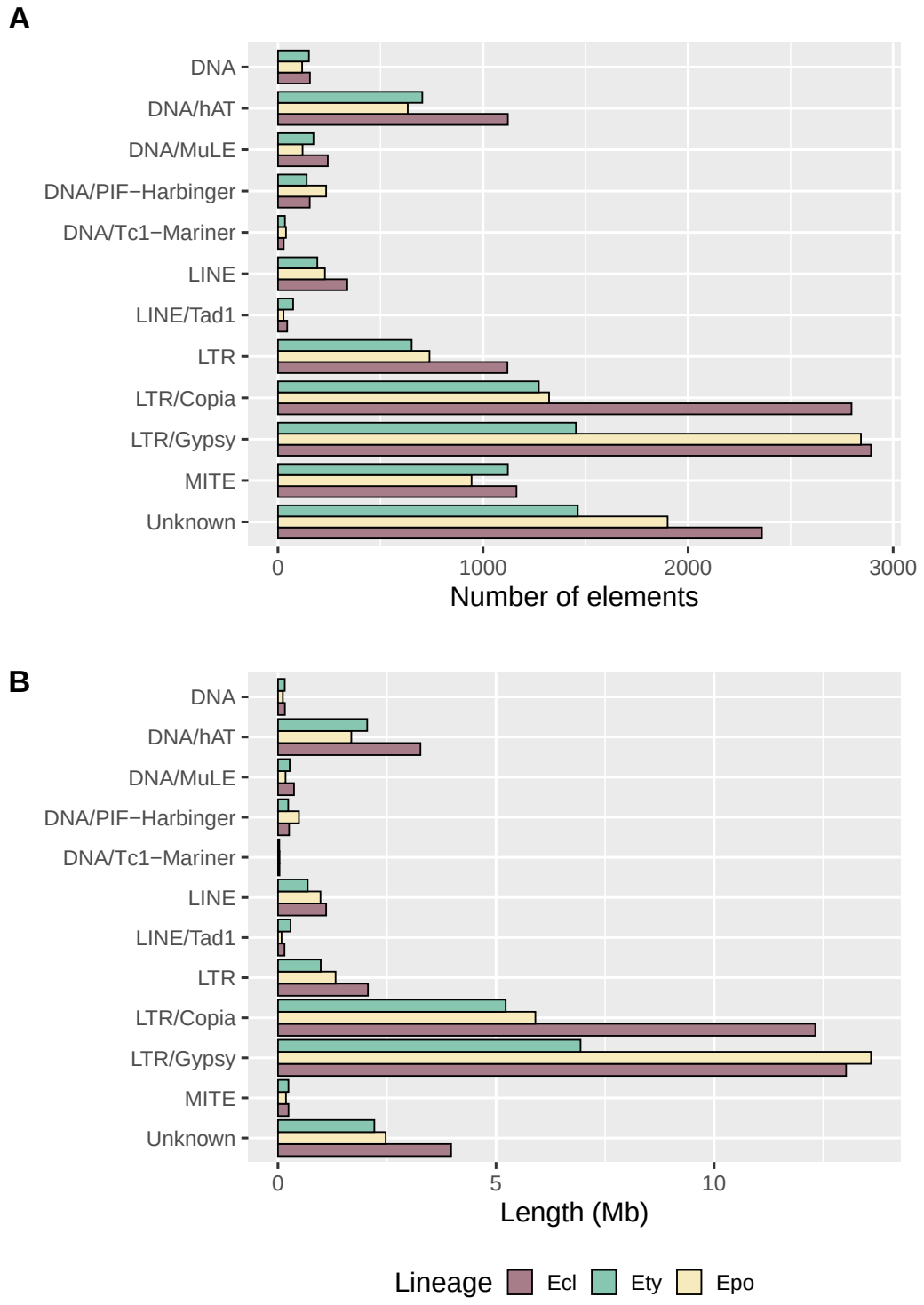

Figure 3.6: TE composition of Ecl, Ety, and Epo. (A) Total copy number of each TE class. (B) Total length of each TE class

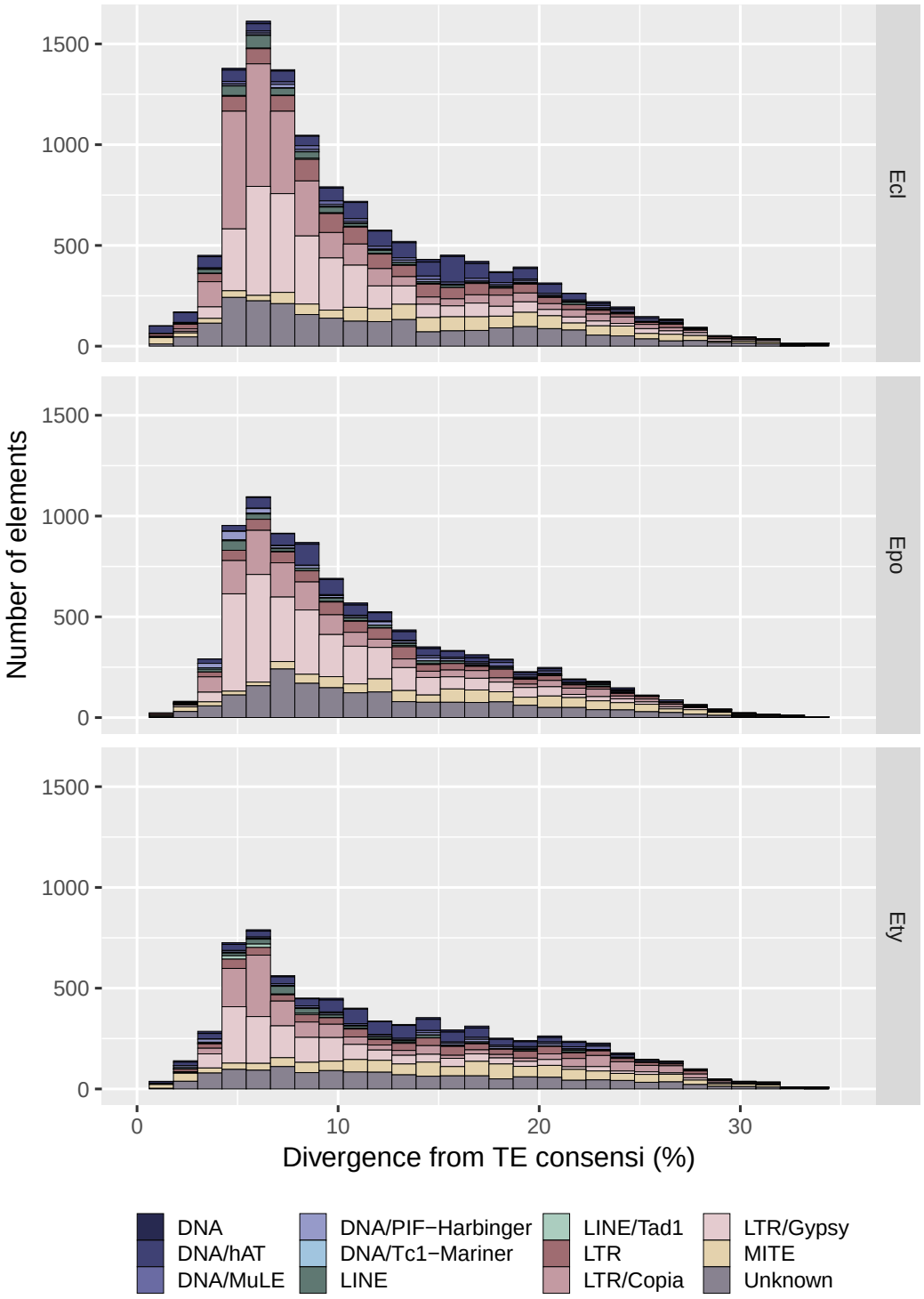

Figure 3.7: Divergence of TE copies from their consensus sequence. Newer insertions are expected to be centered closer to 0%, with older TE copies showing greater divergence

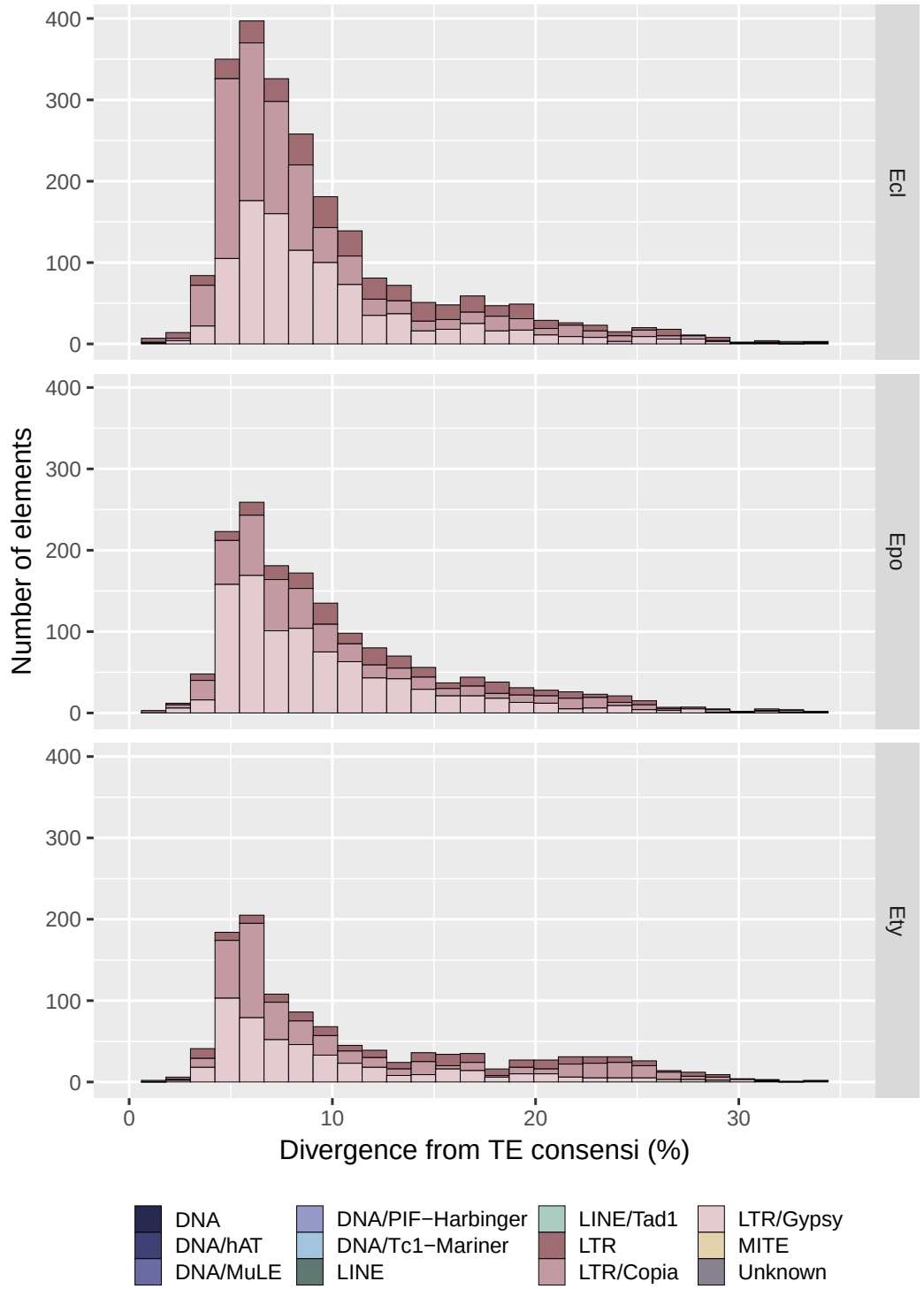

Figure 3.8: Divergence of LTR copies from their consensus sequence estimates the timeline of TE insertions. Newer insertions are expected to be centered closer to 0%, with older TE copies showing greater divergence

this species complex. The <10% divergence measure can not precisely predict the timeline of transposition events, however, active TEs can mutate at rates 10^3 times greater than nuclear genes and fixed, inactive TEs [192, 193], thus divergence below 10% suggest an insertion has occurred relatively recently in the history of TE activity. Ecl has undergone the greatest recent TE invasion, and its sister strain Ety appears to be the least impacted. LTR elements were the largest contributor to this recent TE expansion, and one striking example of this is seen in LTR Copia elements in Ecl (Figure 3.8). Interestingly, Ecl and Ety share a closer phylogenetic relationship with one another than with Epo (Figure 1.6), however Ety does not exhibit TE invasion at the same scale as Ecl and Epo. This suggests that the proliferation of TEs within each genome post-dates the divergence of Epo from Ecl and Ety, as well as the divergence of Ecl and Ety. Further, despite all three genomes sharing the same repertoire of TE sequences with only 3 lineage specific TE sequences present in these genomes (Section 2.4.5), the expansion of LTR elements in Ecl and Epo is underpinned by different LTR sequences in each genome (Figure S3.6). Taken together, this recent expansion of TEs indicates lineage-specific expansion of TEs in this species complex.

3.4.6 Are TEs associated with important genes?

Having established that TEs have likely been active in the recent history of *E. typhina* genomes, I next considered whether these elements may have contributed to phenotypic evolution in these lineages. Classes of genes that are known to play a particularly important role in the evolution of plant-associated fungi include effectors and secondary metabolite genes. The former are small secreted proteins that leave the fungal cell and interact directly with the host plant, and the latter encode proteins that synthesise small molecules that underpin interaction with the host and environment [124, 128]. Previous studies in *E. festucae* show that the most highly up-regulated genes harboured an overrepresentation of MITEs near their transcription start sites [158]. This may be due to positive selection of MITEs following integration events that confer an adaptive

advantage for plant-colonisation. In addition, MITEs may contribute to gene expression responsible for other growth conditions in *Epichloë* [67].

In many plant-associated fungi, important genes fall within or near to a repeat-heavy genomic component [104]. Consistent with this, if TEs contribute to the regulation of gene expression in *E. typhina*, they are expected to more frequently occur near function specific genes. Although there are very few genes in AT-rich regions of *E. typhina* genomes (<1% in all three genomes), it has been suggested these AT-rich regions contribute to the regulation of neighbouring genes that localise near to these regions [67, 191]. I found neither effector genes or secondary metabolite genes are more likely to fall within 1 kb of an AT-rich region than the sampled control genes (Figure 3.9). The result suggests the large AT-rich tracts created by RIP do not contribute directly to the regulation of these important genes.

Having demonstrated no general relationship between specific AT-rich regions of the genome and functionally important genes, I next considered whether specific TE classes might be associated with these gene classes. The sequence motifs present in TEs can act as regulatory sequences, which either drive or modify the expression of nearby genes. My results show MITEs are significantly more likely to occur in the 1 kb upstream region of both effector genes (Figure 3.10) and secondary metabolite genes (Figure 3.11). This result is consistent across species, but most striking in Ecl where MITEs are three-fold over-represented within 1 kb upstream of effectors. While some LTR elements appear to be overrepresented in individual lineages, MITEs are the only class of TEs that are consistently over-represented across all lineages for both classes of functional genes.

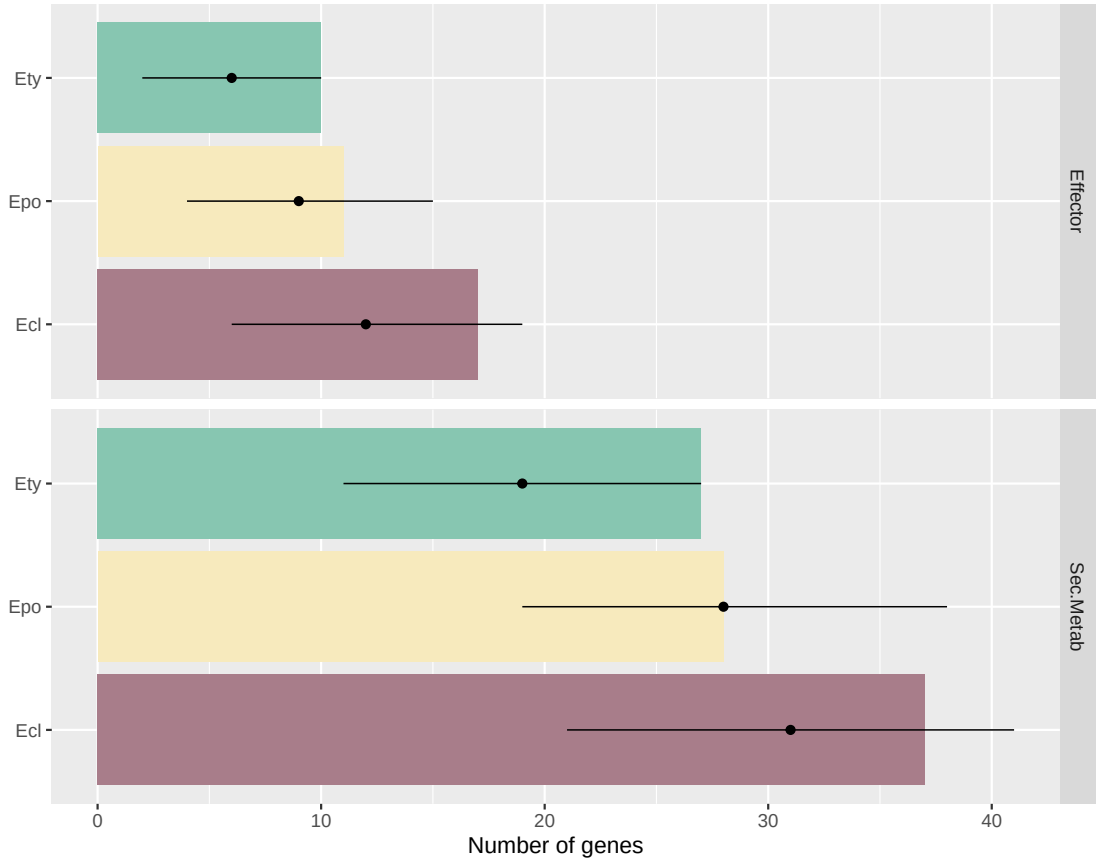

Figure 3.9: Association between AT-rich regions and functional genes. Each bar represents the total number of effector or secondary metabolite genes falling within 1 kb of an AT-rich region. The overlaid error bars show the 0.025 and 0.975 quantiles of the a null distribution for the same statistic, generated through a permutation process (1000 iterations). A gene-class is significantly associated with AT-rich regions if the shaded bar for that class extends beyond the corresponding error bar.

The analyses above focus only on those genes known to be associated with important biological functions in *Epichloë*. It might be argued that TEs other than MITEs still contribute to gene regulation in this genus, despite not being associated with the relatively small number of genes considered here. I examined this possibility by comparing the distribution of distance to nearest gene for all TEs in all genomes. For all TE classes, the majority of sequences did not localise near any gene, however, DNA elements and MITEs localised within 1 kb upstream or downstream of a gene more frequently than other elements (Figure 3.12).

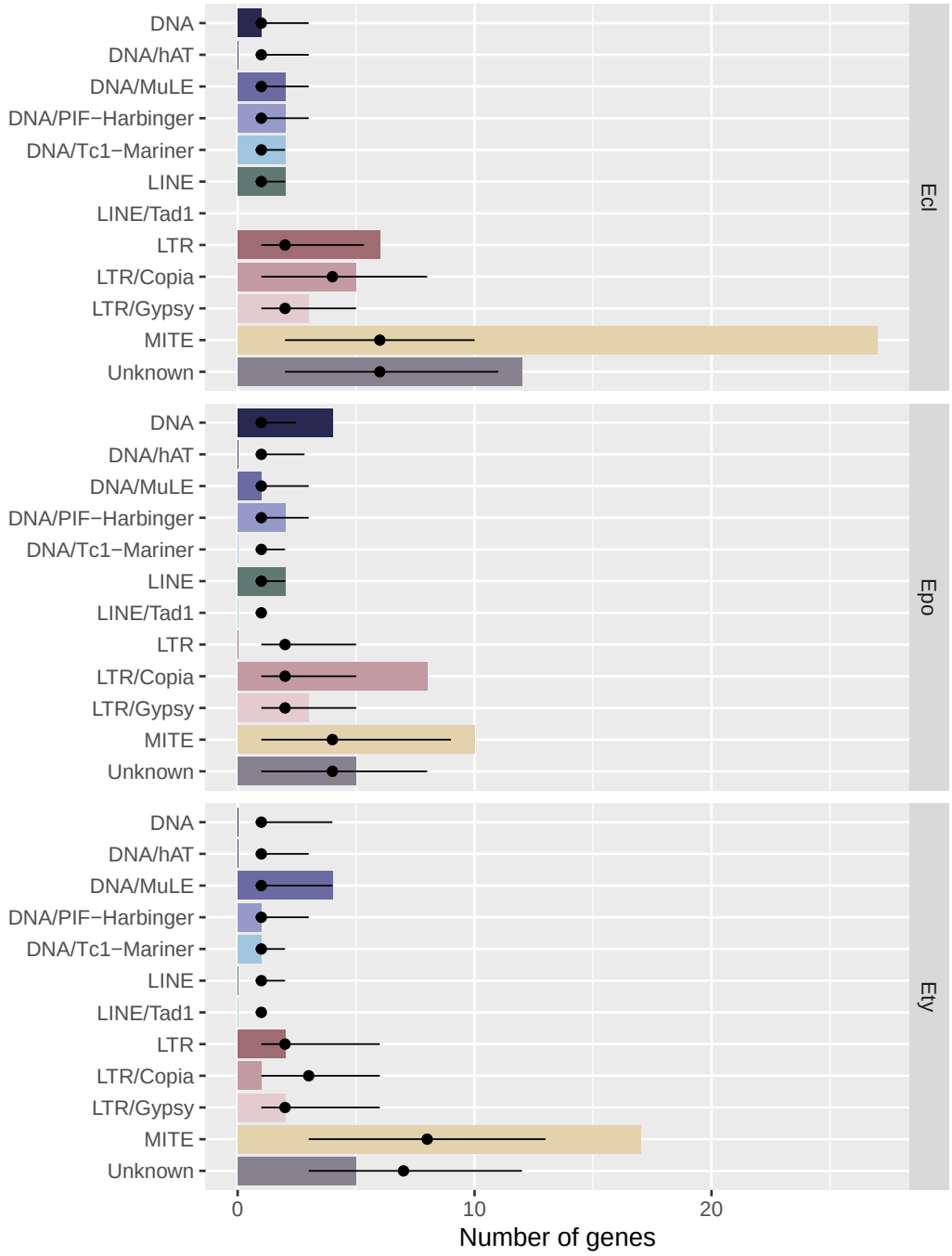

Figure 3.10: Association between TE classes and effectors. Each bar represents the total number of effector genes with a TE of a given class in its 1 kb upstream region. The overlaid error bars show the 0.025 and 0.975 quantiles of the a null distribution for the same statistic, generated through a permutation process. A TE class is significantly associated with effector-encoding genes if the shaded bar for that class extends beyond the corresponding error bar.

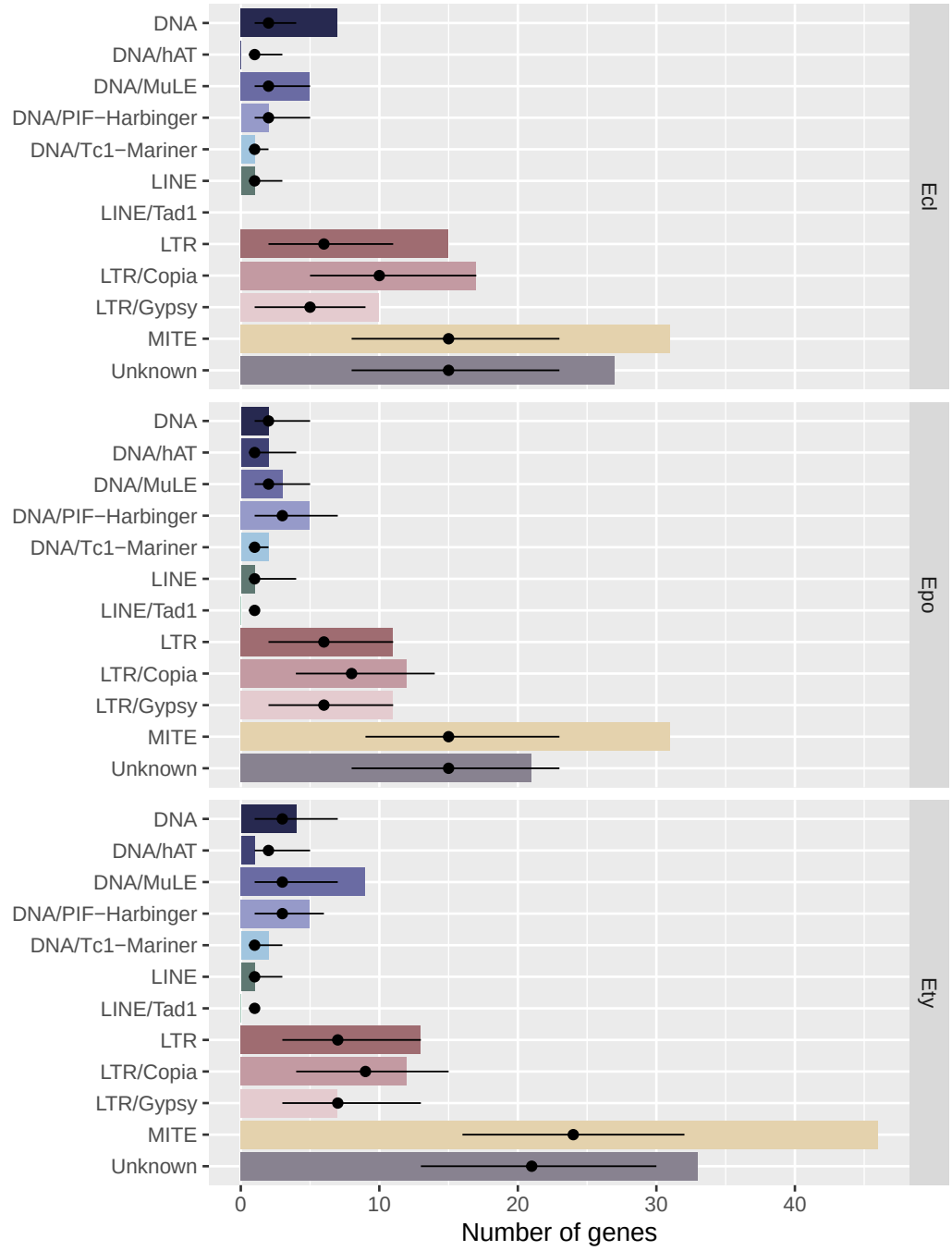

Figure 3.11: Association between TE classes and secondary metabolite genes. Each bar represents the total number of secondary metabolite genes with a TE of a given class in its 1 kb upstream region. The overlaid error bars show the 0.025 and 0.975 quantiles of the a null distribution for the same statistic, generated through a permutation process. A TE class is significantly associated with SM-encoding genes if the shaded bar for that class extends beyond the corresponding error bar.

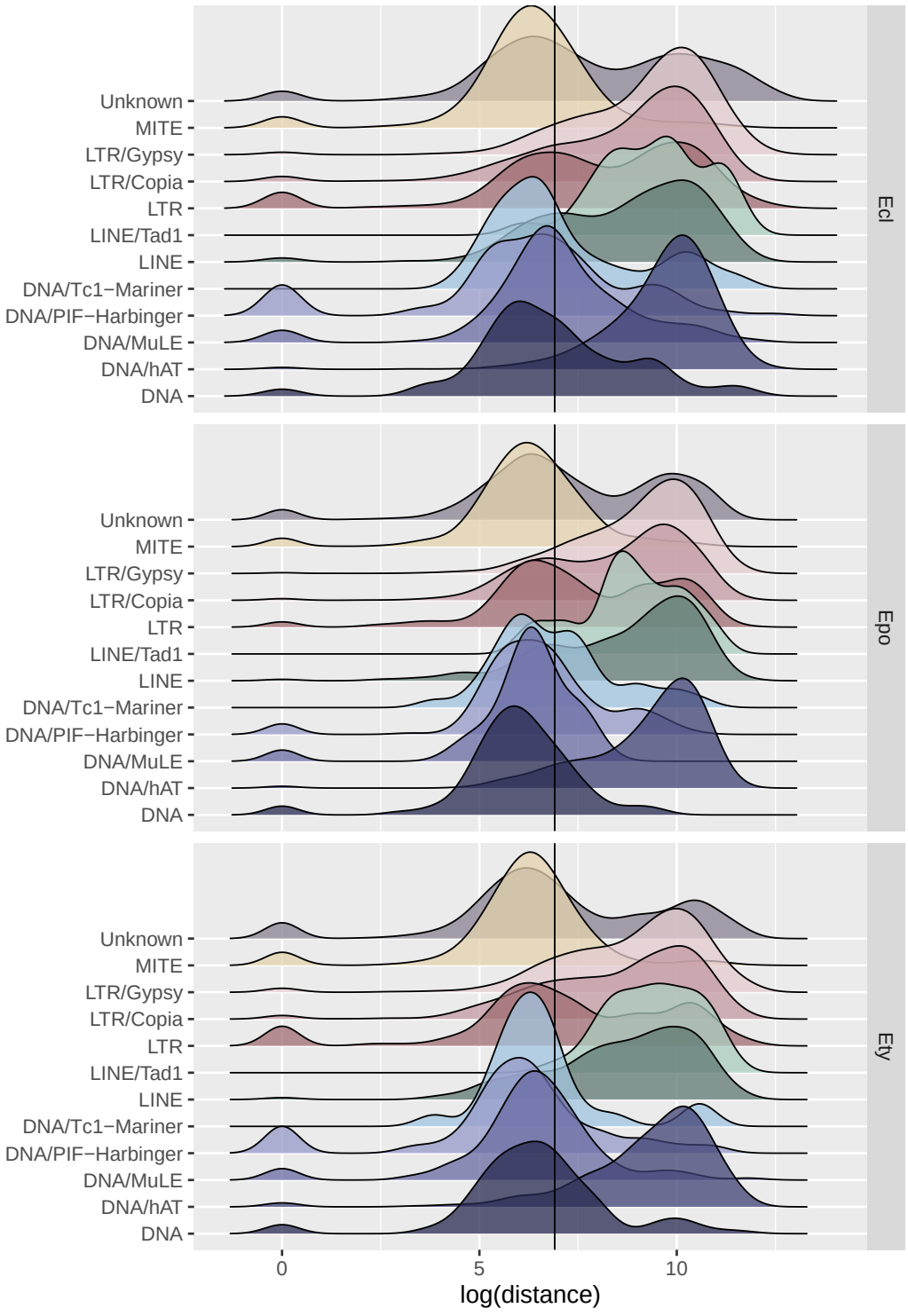

Figure 3.12: Distribution of distance between TEs and their nearest upstream or downstream gene. Vertical line represents 1 kb in distance.

3.5 Conclusions

In this chapter I have used the high quality TE annotations produced in Chapter 2 to analyse the role of these elements in the evolution of three closely related fungal lineages. The results I have produced suggest that TEs have actively contributed to the evolution of *Epichloë* genomes in recent timescales. Although it is clear that these elements have influenced the evolution of genome structure, their contribution to phenotypic evolution is much less certain.

Comparative analyses of all three lineages demonstrated that the differences in genome size are almost entirely explained by AT-rich regions. Consistent with previous reports [67, 148], these AT rich regions were significantly affected by RIP in comparison to gene-rich regions, suggesting that these genome size differences are underpinned by TE invasion followed by deactivation. However, despite the high signals of RIP within the genome, not all TEs showed signals of RIP, suggesting that active elements remain in the genomes. The sequences that have evaded RIP appear to be relatively random; RIP is established to target sequences above a length threshold (>400 bp), however TEs that measure above the 400 bp RIP threshold remain unaffected by RIP, while others measuring below the threshold show high levels of RIP. As the focal lineages of this work undergo asexual reproduction, it is possible that newer insertions have not yet been detected by RIP, as RIP predominantly occurs in pre-meiotic mitosis events. However, when testing this hypothesis, the age of transposition did not exhibit clear patterns of RIP. In addition, proximity to genes (ie TEs in AT-rich or gene-rich regions) did not appear to explain the absence of RIP in these long TE copies. Thus, it is possible that RIP in *Epichloë* do not target all TEs.

All three genomes had unique TE repertoires comprising of different copy numbers of select TE classes. These results suggest lineage specific expansion of TEs, thus to investigate the timeline of insertions, I examined the divergence of each TE copy with its consensus sequence. Newer insertions are expected to show very little

sequence divergence from its respective consensus, while older integrations and ancient TE copies are expected to show high levels of divergence by virtue of mutational processes. An enrichment of TEs with <10% divergence substantiated the hypothesis that lineage-specific expansion of TEs has occurred post-dating the divergence of the three lineages. These expansions were predominantly underpinned by LTR retrotransposons, and one striking example of this phenomenon was observed in Ecl, where a considerable expansion of LTR Copia elements was observed.

Though it is clear that TEs have contributed to genome evolution in *E. typhina*, it is not certain that they have made a major contribution to phenotypic evolution during that time. It has been suggested that the AT-rich regions of *Epichloë* genomes contribute to gene regulation through the 3D structure of DNA in the nucleus [67]. However, I found no evidence that secondary metabolite and effector genes, crucial contributors to host-interaction in *Epichloë*, are associated with these regions. However, there is clear evidence that MITEs are associated with effectors and secondary metabolite genes, suggesting the possibility that these elements act as regulators for the expression of functionally important genes. However, this pattern could also reflect the insertion bias that sees MITEs preferentially inserting into genic regions [68].

The results presented here suggest it is unlikely that the contribution of TEs to gene expression is pervasive across these genes. Nevertheless, it is possible that specific TE insertions are of importance for gene regulation. Notably, I have demonstrated that the gene-rich component of these genomes is not entirely devoid of TEs. It has long been known that MITEs are common in genic regions [68], but I have shown elements of all classes occur in proximity to genes. Future work building from my results could include identifying specific novel TE insertions in gene-rich regions of genome. In addition, if functional genomic data such as RNAseq data was generated for these species, it might be possible to link novel gene expression traits occurring in a fungal lineage with specific TE insertions near to that gene.

In addition to the TE analysis presented here, this chapter represents the first description of a full-length telomere-to-telomere genome for Epo. Although this genome will be described in more detail in forthcoming publications, it is interesting to note how unusual the Epo genome is among *Epichloë* sequences. A recent study has demonstrated a high degree of conservation of genome structure across the entirety of this genus [161]. However, the Epo genome showed extensive chromosomal rearrangements when compared to its sibling taxa. This genome has not been more subject to TE-expansion than Ecl or Ety, so this marked difference cannot be simply tied to the action of TEs. Determining whether TEs may have contributed to these rearrangements would require a complete reconstruction of the chromosomal evolution in *Epichloë typhina*, a task that is well beyond the scope of this work. However, the high quality TE annotations produced here will be an important resource for any investigation into genome evolution in Epo.

3.6 Supplementary Information

lineage	TE group	tot. len. (Mb)	mean. len. (kb)	n	%TE	%genome	%complete
Ecl	DNA	0.14	1.04	130	0.44	0.30	7.69
Ecl	DNA/hAT	2.28	2.81	811	7.44	4.99	7.64
Ecl	DNA/MuLE	0.33	1.49	220	1.07	0.72	19.55
Ecl	DNA/PIF-Harbinger	0.24	1.62	146	0.77	0.52	22.60
Ecl	DNA/Tc1-Mariner	0.03	1.19	26	0.10	0.07	19.23
Ecl	LINE	0.92	3.07	299	3.00	2.01	9.03
Ecl	LINE/Tad1	0.09	3.66	25	0.30	0.20	0.00
Ecl	LTR	1.40	1.67	837	4.57	3.06	15.29
Ecl	LTR/Copia	11.17	4.00	2792	36.54	24.50	24.25
Ecl	LTR/Gypsy	11.31	4.15	2724	36.98	24.79	20.56
Ecl	MITE	0.21	0.21	1027	0.69	0.46	18.40
Ecl	Unknown	2.47	1.50	1651	8.08	5.42	14.48
Ety	DNA	0.13	1.01	125	0.81	0.37	11.20
Ety	DNA/hAT	1.30	2.59	502	8.35	3.85	5.38
Ety	DNA/MuLE	0.23	1.53	149	1.46	0.67	14.09
Ety	DNA/PIF-Harbinger	0.20	1.58	126	1.27	0.59	23.02
Ety	DNA/Tc1-Mariner	0.02	0.93	25	0.15	0.07	16.00
Ety	LINE	0.51	3.43	149	3.28	1.51	7.38
Ety	LINE/Tad1	0.18	3.82	46	1.13	0.52	0.00
Ety	LTR	0.71	1.44	490	4.53	2.09	16.12
Ety	LTR/Copia	4.58	3.60	1272	29.37	13.54	14.39
Ety	LTR/Gypsy	6.04	4.41	1372	38.76	17.87	22.38
Ety	MITE	0.22	0.21	1037	1.39	0.64	15.91
Ety	Unknown	1.48	1.29	1146	9.50	4.38	13.70
Epo	DNA	0.09	0.92	98	0.41	0.24	13.27
Epo	DNA/hAT	1.19	2.48	479	5.41	3.11	8.56
Epo	DNA/MuLE	0.13	1.25	102	0.58	0.33	15.69
Epo	DNA/PIF-Harbinger	0.42	1.93	216	1.90	1.09	19.91
Epo	DNA/Tc1-Mariner	0.03	0.87	38	0.15	0.09	13.16
Epo	LINE	0.72	3.55	204	3.30	1.89	13.24
Epo	LINE/Tad1	0.03	3.23	10	0.15	0.08	10.00
Epo	LTR	1.06	1.76	600	4.82	2.77	13.33
Epo	LTR/Copia	5.31	3.72	1426	24.17	13.88	22.37
Epo	LTR/Gypsy	11.50	4.26	2698	52.35	30.07	21.83
Epo	MITE	0.16	0.19	818	0.71	0.41	13.33
Epo	Unknown	1.33	1.14	1162	6.05	3.47	15.92

tot. len = Total length of elements in Mb
mean. len = mean length of elements in kb
n = total copy number
%TE = total percent of each TE order accounted for by the element denoted in "TE group"
%genome = total percentage of the genome accounted for by the TE class
%complete = total percentage of genomic TE copies that encompassed the entire consensus sequence

Table S3.1: Summary statistics for TE annotation

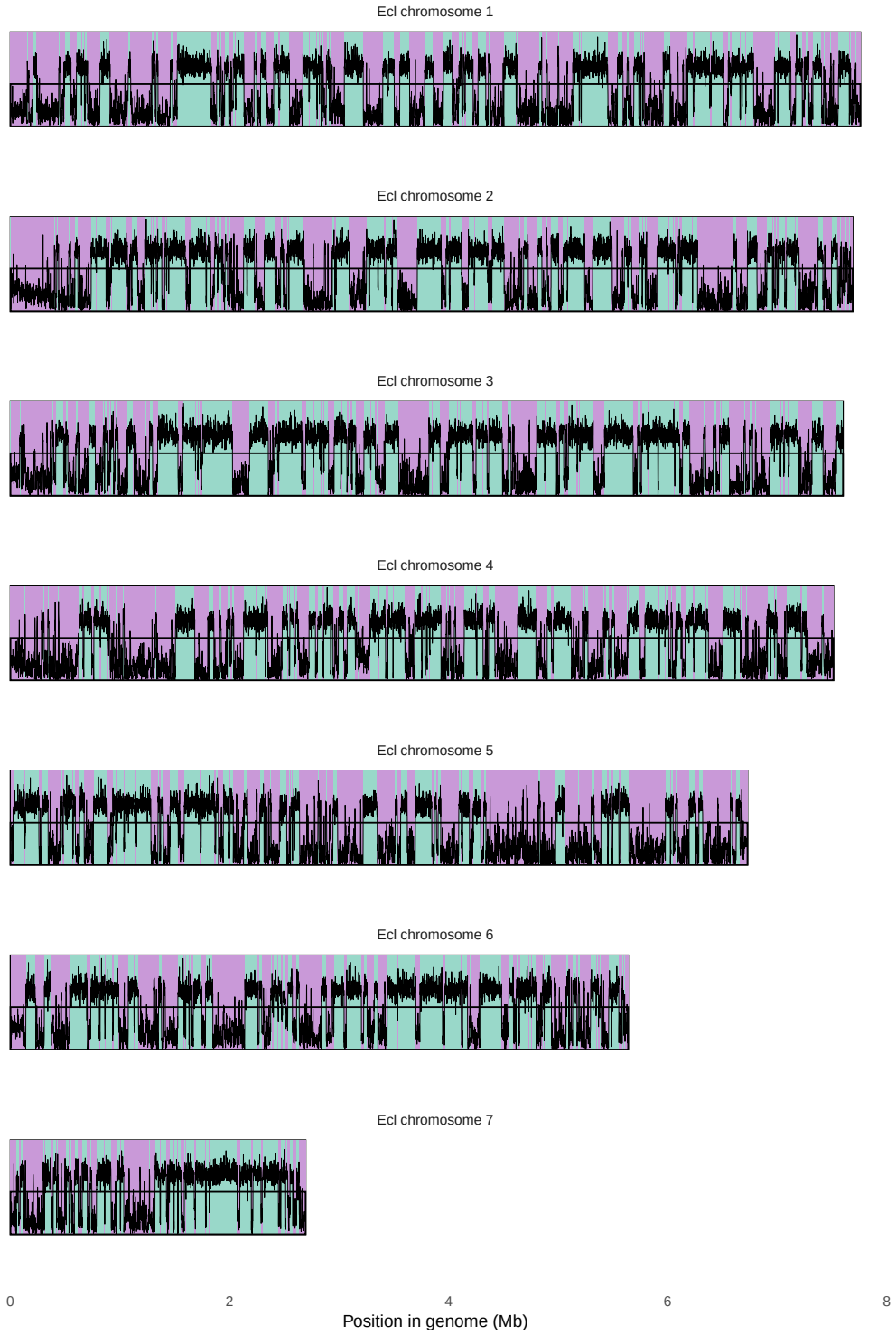

Figure S3.1: RIP across genomic compartments in Ecl. The RIP index was measured in 1 kb stretches across the genome and is represented as a line graph. The horizontal line represents 0.9 RIP threshold, with lower scoring values indicative of RIP deactivation. AT-rich regions are represented in purple. The patterns of RIP across AT and gene-rich regions is almost identical in Ety and Epo (not shown).

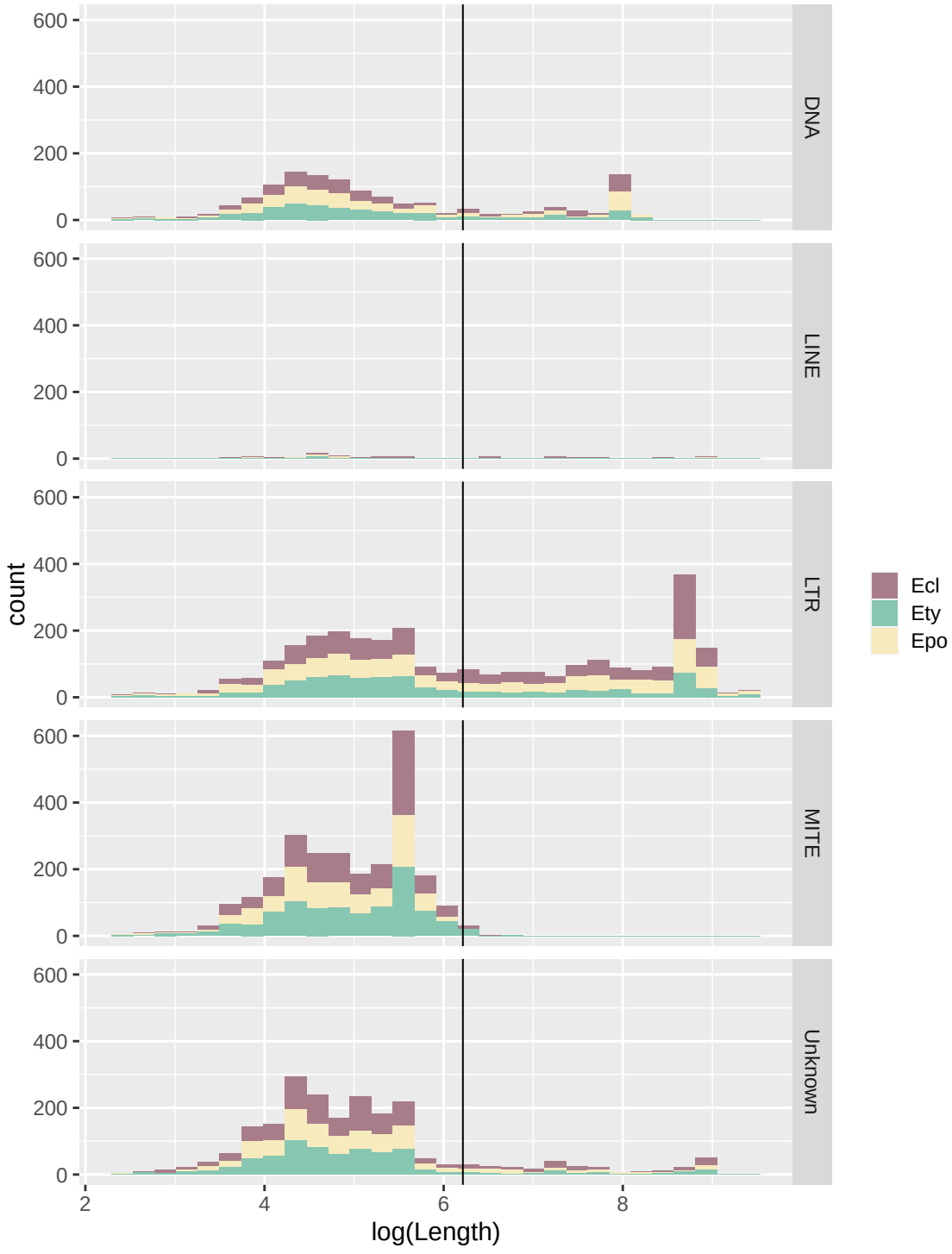

Figure S3.2: Length of TE elements residing in the gene-rich region. Vertical line indicates 500 bp in length.

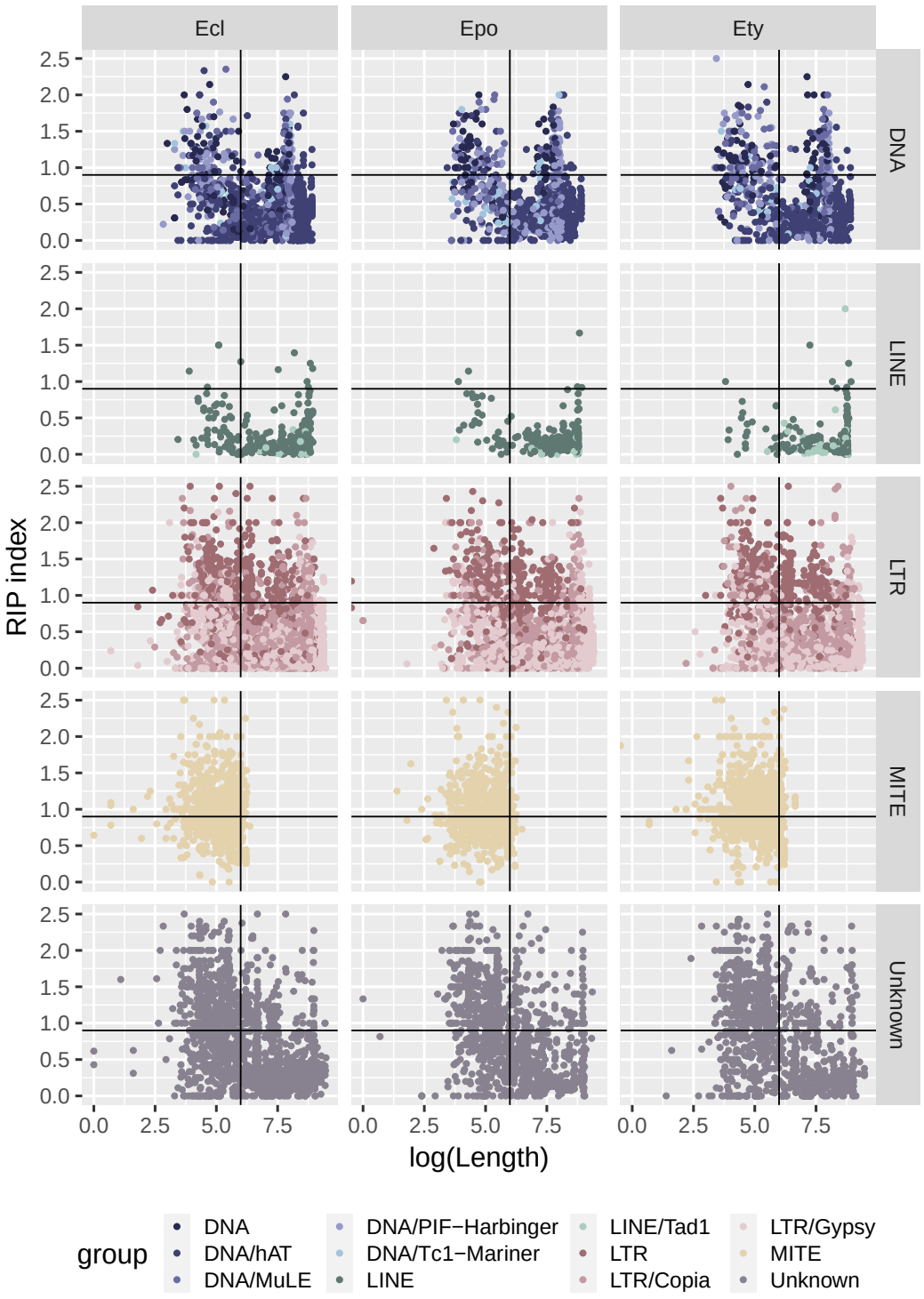

Figure S3.3: RIP vs length of element. Horizontal line depicts RIP length threshold, with lower scoring values indicating targeting by RIP. Vertical line denotes 400 bp and TEs measuring higher than this value are expected to be targeted by RIP. Thus, TEs are expected to appear only in the the lower left quarter and upper right quarter of each plot are if RIP targets all sequences >400 bp.

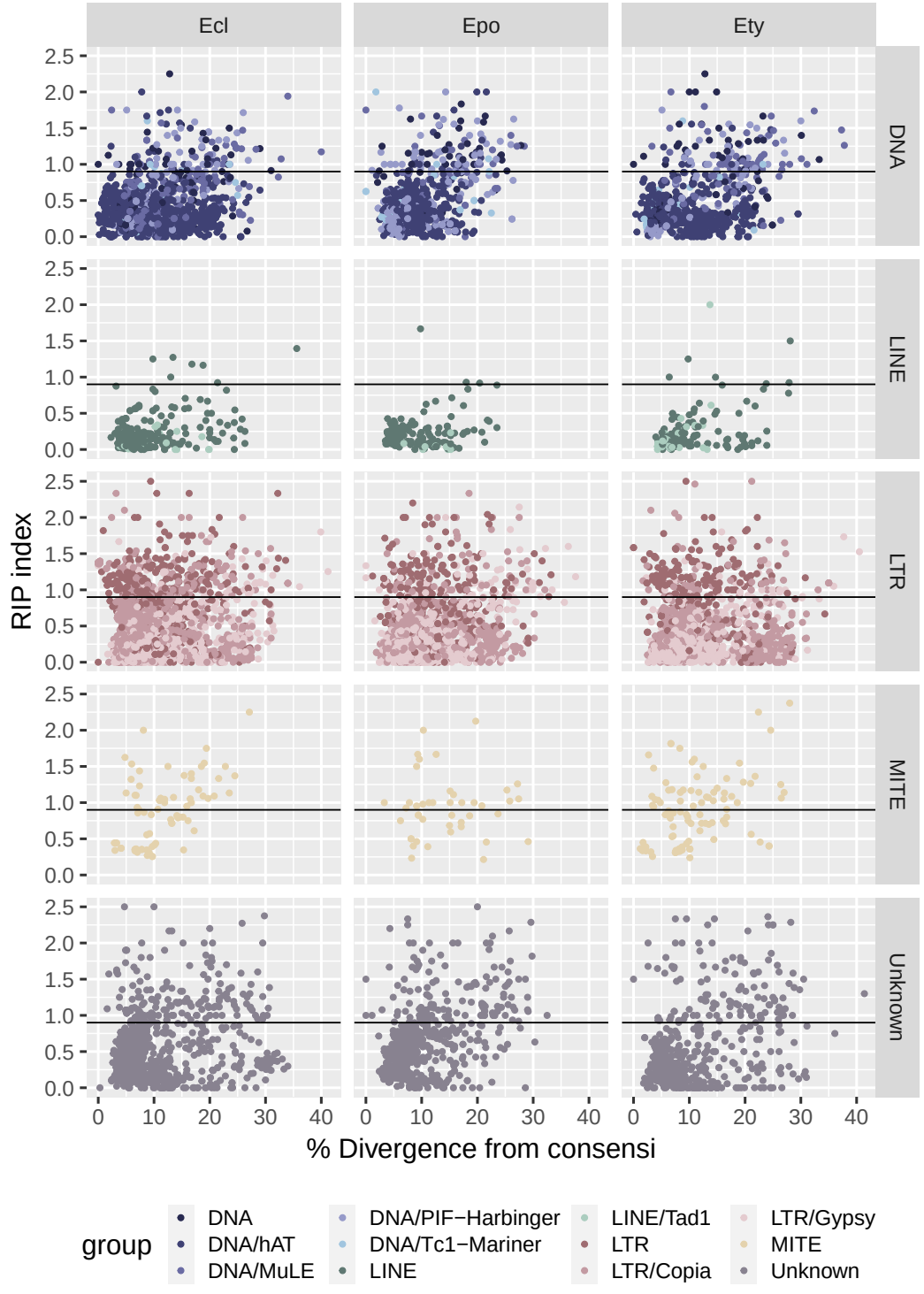

Figure S3.4: RIP vs divergence of long TE copies (>400 bp) from their respective consensus sequence. This divergence can be used to estimate the relative age of the element, with older elements showing higher divergence. Vertical line denotes a RIP score of 0.9. Values below this line suggest the sequence has been targeted by RIP.

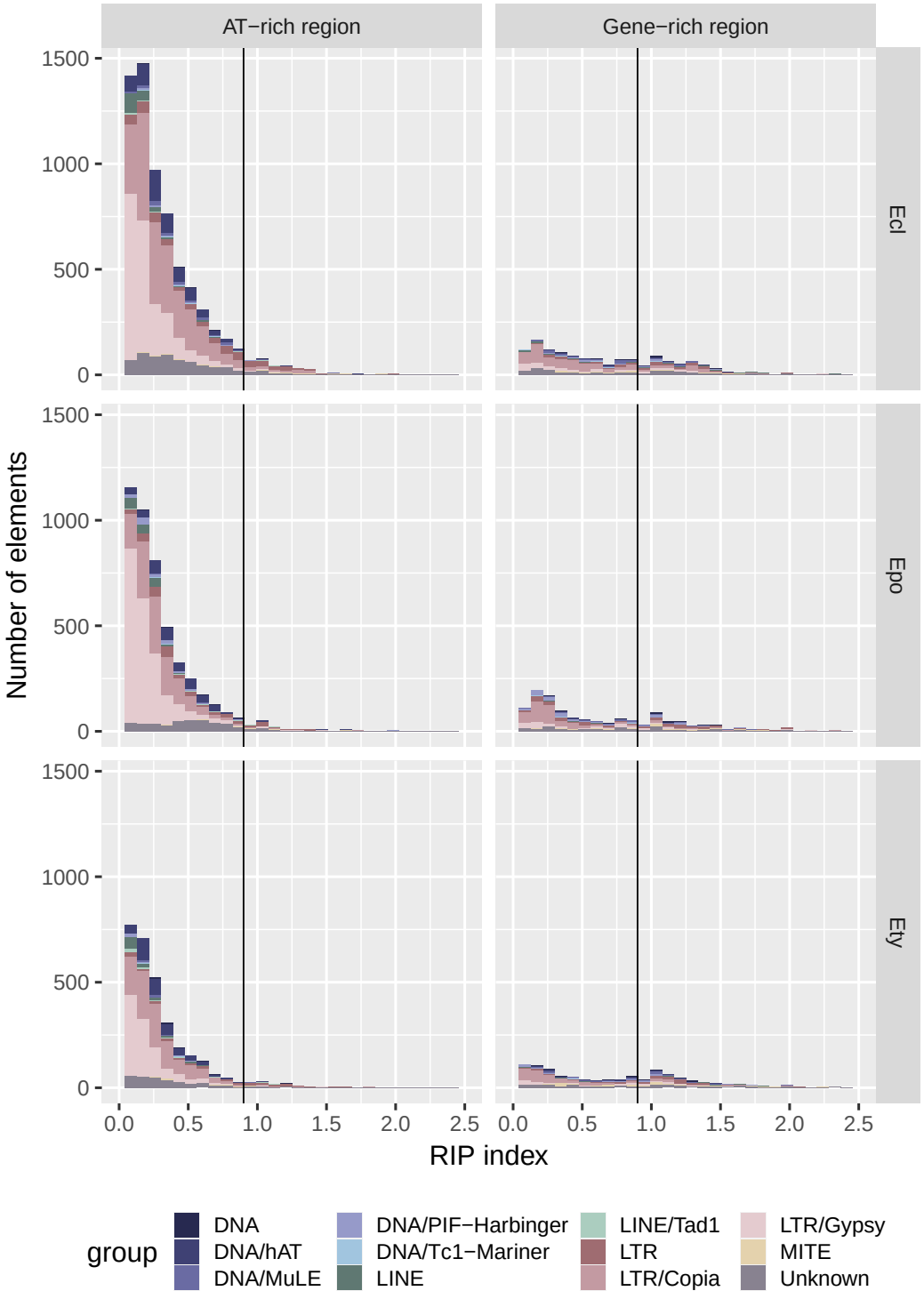

Figure S3.5: RIP vs genomic compartment for long elements (>400 bp). Horizontal line depicts RIP threshold of 0.9, with lower scoring values indicating the presence of RIP.

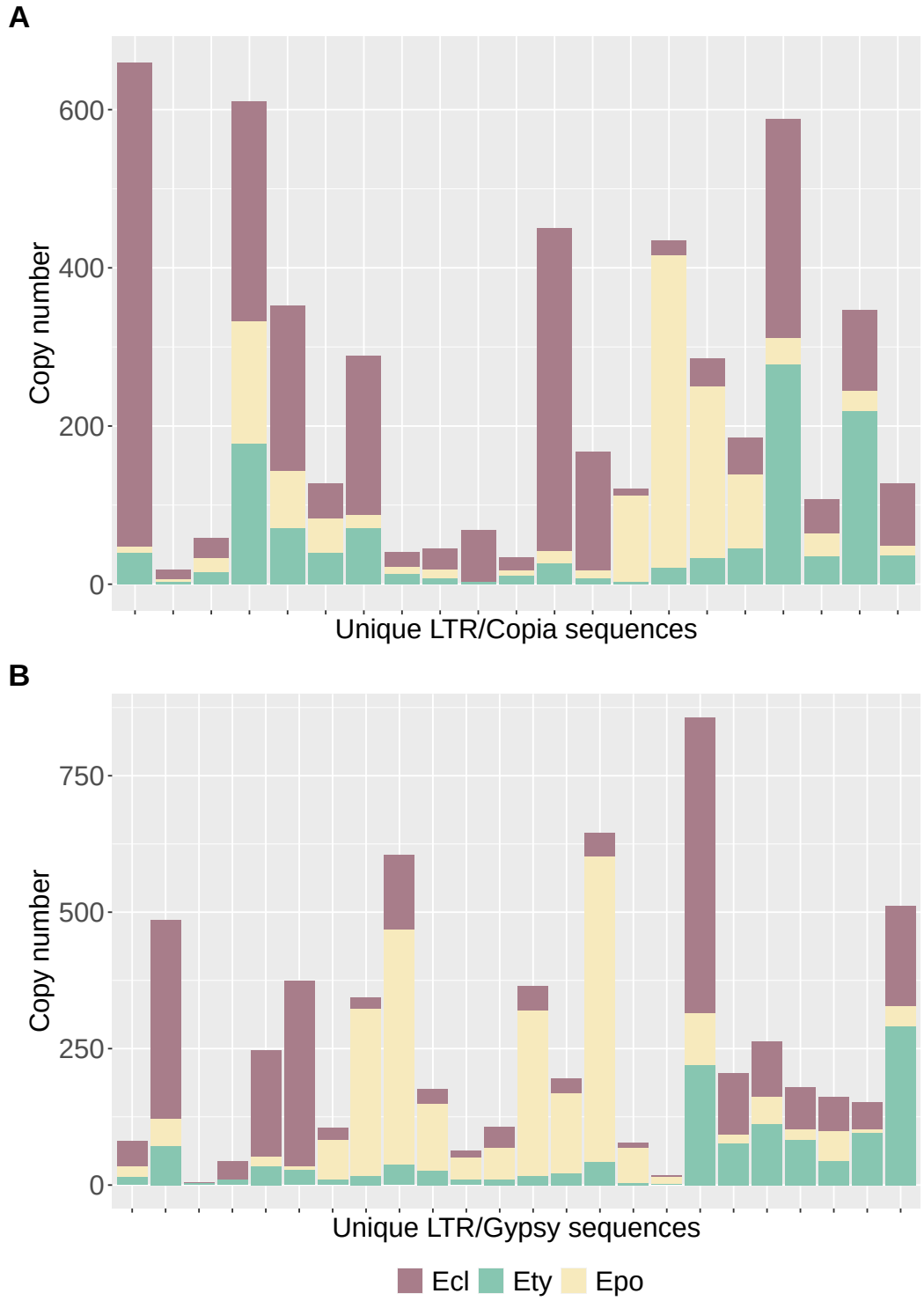

Figure S3.6: Copy number of LTR families. Each bar refers to a unique sequence belonging to (A) the LTR/Copia superfamily, and (B) the LTR/Gypsy superfamily.

Chapter 4

General Discussion

4.1 General Discussion

The remarkable success of modern sequencing technology has revolutionised the way we generate, study, and understand genomes. Despite the explosion of research in genomics, TEs remain subordinated to non-TE genomic components despite the fact that they are a major genomic constituent in most eukaryotic genomes.

Generating a comprehensive understanding of TEs within a genome is hindered by factors intrinsic to TEs themselves, coupled with restrictions in the softwares and methods currently employed for the study of these elements. For example, TEs are dynamic genomic units that can rapidly proliferate, degrade, and accrue mutations. These factors give rise to lineage-specific TE populations that are recalcitrant to TE annotation softwares. Even so, there is a large motivation to study TEs due to their impact on the host genome; TEs are becoming increasingly recognised as important contributors to gene regulation and genome structure.

The focus of this project was *Epichloë*, a genus of symbiotic fungi that live in close association with pasture grasses. Preliminary results prior to this project proposed that TEs profoundly impact the structure of *Epichloë* genomes, and regulate the expression of specific genes that allow *Epichloë* species to adapt to specific plant hosts. However, due to the extent of RIP, it was unclear whether actively transposing elements remained in this genus. As a result, the overall impact of TEs in the evolution of *Epichloë* genomes remained obscure [67]. *Epichloë* genomes are relatively small with rich TE content, making them an excellent model for the study of TEs. By studying TEs across a range of *Epichloë* species that are ecologically distinct and adapted to different grass hosts, I was able to test the hypothesis that TEs contribute to the evolution *Epichloë* genomes and, in particular, contribute to the evolution of functionally important genes involved in host-specificity.

The objectives of this project were completed:

1. Complete the manual curation of a custom TE library for three *Epichloë* genomes. This work produced the first high quality TE library to ensure robust downstream TE analyses.
2. Study the TE populations present in the three focal genomes. This included identifying whether TEs have been active within these genomes post-dating the divergence of the three lineages, as well as characterising the TE populations and extent of RIP within each genome.
3. Identify host-association genes from each *Epichloë* genomes. By identifying putative effectors and secondary metabolites, I could test for associations between novel TE integration and host-association genes.

4.2 Findings

4.2.1 Objective One

An *Epichloë typhina* master TE library was curated using three strains of the *E. typhina* species complex (sub-species *clarkii* (Ecl), *typhina* (Ety), and *poae* (Epo)). During manual curation, 66.19% of all TE consensus sequences were reclassified. A large portion of these sequences were previously classified as unknown repeat sequences. In total, the process of manual curation reduced the number of unknown elements from 72% to 25.69%, emphasising the need for manual curation in non-model species (Section 2.4.1).

Further, a number of repeat elements previously annotated as interspersed repeats such as satellites and simple repeats were found to be nested, complex, or degraded TE copies. In particular, these elements were resolved to be LTR elements, DNA/hAT elements, and MITEs (Section 2.4.2). MITEs are a class of interest due to

their abundance in fungal genomes, and potential role in gene regulation [67, 68, 158]. These elements are not currently identified by RepeatMasker, the most widely used software for TE annotation. Further, programs designed to identify MITEs routinely fail to do so in fungal genomes. In this work, I characterised 41 unique MITE consensus sequences that provided a valuable resource for downstream MITE analyses (Table S2.1). In addition to MITEs, automatic identification does not currently identify conserved domains within TEs. My manual curation process identified conserved protein domains in 54 of the 288 consensus sequences. As conserved domains can serve as key determinants in TE classification, the identification of these domains ensures robust classification of TEs (Section 2.4.3).

Interestingly the curated library improved the quality of TE annotations (assigning more TEs to a class and identifying conserved domains), but reduced the coverage of TEs within each genome when compared to the automatic library. This result is partly due to the exclusion of 77 automatically-identified consensus sequences during manual curation. These excluded sequences harboured excessive ambiguity upon generation of a new consensus sequence. Such highly degraded (or possibly paralagous) sequences add little to downstream analyses of TEs, so their exclusion has little impact on conclusions drawn in this report. Another reason for the lower total coverage was refining of the boundaries of TE elements (Section 2.4.4). The manual curation process identified many TE consensus sequences with poorly-defined boundary sequences. Excluding these flanking regions from the consensus models decreased the total coverage of TEs, but improved the accuracy of each TE annotation.

The final TE library for *E. typhina* presents 288 unique consensus sequences, of which 67% retain characteristic structural and/or enzymatic features, 9.06% retain partial features, 23.26% were incomplete copies retaining only a portion of the TE, and 0.35% that were both partial and incomplete. The three *E. typhina* genomes show a high level of TE sharing, with only three lineage-specific TEs identified within the library. In total, the sequences predicted in this new curated library account for 67.03%,

46.1%, and 57.44% of the total genome length in Ecl, Ety, and Epo, respectively. (Section 2.4.5).

4.2.2 Objective Two

Previous work on *Epichloë* has established this genus has a rich TE repertoire, and suggested some elements may contribute to gene regulation. However, because these studies focused on a single genome [67] or a single class of TE [68], an overview of the action of these elements in this genus has not yet been presented. In Chapter 3, I present the first multi-lineage study on the dynamics of TEs in *Epichloë*.

This work includes the first summary of a complete genome for Epo. The contents of this genome are typical for the genus, a patchwork of AT-rich regions that are almost devoid of genes and gene-rich regions with approximately equal GC content. However, this genome has undergone a remarkable number of rearrangements, with each Epo chromosome appearing as a patchwork of homologous Ety and Ecl chromosomes. While previous studies have observed large chromosomal rearrangements between members of the *Epichloë* genus, the degree of rearrangement between the three closely related sub-species studied in this work is unusual (Section 3.4.1).

Upon investigating the divergent AT-rich regions of each genome, I found that the between-lineage difference in genome size was almost entirely explained by these AT-rich regions. As AT-rich regions are often biochemical hallmarks of RIP deactivated sequences [139], I calculated the extent of RIP across the genome. I found the AT-rich regions of all genomes indeed showed very high signals of RIP. The action of RIP was distinctly limited to the AT-rich regions, with very little RIP detected in the gene-rich regions of each genome. Taken together, this suggests the between-lineage variation in genome size is underpinned by the proliferation and subsequent deactivation of TEs in each genome (Section 3.4.2).

To determine the distribution of TEs across these AT- and gene-rich regions, I investigated the localisation of TEs and genes within these distinct blocks. While TEs did indeed reside predominantly within the AT-rich regions, some exceptions included the localisation of MITEs, a small portion of LTR elements, DNA elements, and unknown repeat elements within the GC-rich regions. These elements were typically <500 bp in length, and may be fragmented relics of deactivated TEs that are often evenly distributed throughout fungal genomes (Section 3.4.3) [191].

Having established the presence of TEs in the gene-rich regions, I examined the extent of RIP for every type of TE present in each genome. I found not all TEs had been deactivated by RIP, however, the presence of RIP appeared to be relatively random. For example, some long TEs expected to be targeted by RIP showed no signals of RIP, while other TEs that were not expected to be affected by RIP showed extensive hallmarks of RIP (Section 3.4.4). Previous studies hypothesised that all long repeat elements had been detected by RIP [67], however the work here has demonstrated that some long sequences have evaded RIP, and suggests that active elements may remain in these genomes.

To investigate potential TE activity post-dating the divergence of the three genomes, I investigated the TE repertoire of each genome. Consistent with other studies in *Epichloë*, the predominant constituent in all three genomes were LTR elements and MITEs, however, each genome showed considerable variation in the frequency of each TE class, suggesting lineage-specific expansion of specific TE classes following the divergence of this species complex. To confirm this TE activity, a timeline of insertions was estimated by examining the percentage deviation of any given TE from its respective consensus sequences. All three genomes demonstrated a recent expansion of TEs with <10% divergence, substantiating the theory that each genome has undergone lineage-specific expansion. One striking example of this recent expansion was seen in LTR Copia elements in Ecl. The expansion of LTR elements, particular in Ecl and Ety were underpinned by unique TE families, suggesting that unique TE families are

responsible for the expansion of each genome (Section 3.4.5).

4.2.3 Objective Three

Having confirmed recent activity of TEs among the three genomes, it was of considerable interest to investigate the association between TEs and genes known to mediate host-plant invasion. In particular, the genes of interest were effector genes and secondary metabolite genes. In many plant-associated fungi, virulence-related genes fall within or near to repeat-heavy genomic regions. TEs that localise within 1 kb upstream of a gene are hypothetically capable of impacting gene regulation [122]. I tested whether these virulence-related genes are more likely to fall within an AT-rich region or within 1 kb of these regions, and found that they are not more likely to localise in these areas. This suggests that the TE-dense AT-rich regions do not directly contribute to the regulation of these genes. I next considered associations between TEs and virulence-related genes, and found that MITEs are significantly more likely to localise within 1 kb upstream of both effector and secondary metabolite genes than within 1 kb of non-virulence related genes. This phenomenon is most pronounced in Ecl, and is consistent with previous findings in *E. festucae* (Section 3.4.6) [67].

Similar results have also been observed beyond the *Epichloë* genus. For example, in *Fusarium oxysporum*, MITEs are closely associated with effector genes, and partial copies of MITEs are always present in the gene promoter regions of effectors [194]. Further, in the filamentous fungus *Zymoseptoria*, genes that are potentially involved in plant infection were identified via presence/absence variations between 26 strains of *Z. triciti*. Upon examining an association between TEs and the genes identified this way, it was established that MITEs are located significantly closer to these genes than what would be expected from a random distribution [115]. Interestingly, in this study, Lorrain *et al* [115] also established that non-MITE TEs are associated with effector genes in each *Z. triciti* genome, however the TE class associated with these

genes differ between the selected strains. This phenomenon was also observed in my results, where some non-MITE elements appeared to be over-represented within 1 kb upstream of an effector or secondary metabolite genes, however, unlike MITEs, this over-representation was not consistent across all three genomes (Figure 3.10 - 3.11). Future lineage-specific TE studies may elucidate roles of these specific TEs in the regulation of each genome.

4.3 Limitations

4.3.1 Objective One

Fragmented sequences: TE copies are very often nested or partial [178, 195–197]. This presents challenges during manual curation if the fragmented sequence has not retained sufficient length to identify structural features or conserved domains. When a sequence appeared to be a fragment of a TE, the consensus sequence was rerun through the RepeatModeler4 pipeline with extended flanking sequences. Despite this, on many occasions, the boundaries of the element could not be retrieved, and the sequence could not be classified. Partial sequences were also queried against each consensus sequence in the new library I created in an attempt to obtain a full length copy. While clustering of the library and reciprocal BLAST resolved some incomplete sequences, a number obtained no matches and could not be classified into the appropriate TE class. Unfortunately, this is a current limitation intrinsic to the nature of TEs. In the future, the availability of new high-quality fungal TE libraries may help elucidate the classification of these unknown sequences.

Classification of lineage-specific TEs: During propagation, active TEs can mutate at rates 10^3 times higher than nuclear genes and fixed, inactive TEs [192, 193]. This presents a number of difficulties when creating a unified classification system for TEs as the accumulation of mutations may see TE classes diverging from their characteristic features. This phenomenon is seen in TEs that have evolved across a long evolutionary timescale in multiple taxa [192, 198].

In this work, hAT elements in *Epichloë* were one example of this phenomenon as they were not always consistent with the unified classification system outlined by Wicker et al [50]. In Chapter 2, classification of hATs was determined by a combination of conserved domain, TSD length and motif, TIR length, and sequence similarity to

known hATs. More than half of all hAT consensus sequences retained part of a transposase domain characteristic of hATs, and almost all of the consensus sequences (93%) had high sequence similarity with other fungal hATs reported on RepBase. They also harboured 8 bp TSDs characteristic of hATs. However, on occasion, the TIR length and total length of the element were not consistent with previous literature. Members of the hAT superfamily are reported to typically have short TIRs of 5 to 27 bp in length with limited sequence similarity between family members. The total length is typically less than 4 kb [50, 199]. However, the aforementioned hATs in *Epichloë* often harboured longer TIRs, and exceeded 4 kb in length, with the longest complete element measuring greater than 7 kb in length. In the presence of conserved domains or TSDs consistent with hAT elements, these atypical TIR lengths and total lengths were not considered in classification. This decision was based on previous reports of hAT elements that exceed 4 kb in length, such as a previously published 12 kb hAT element in *Chlamydomas reinhardtii* [200]. In addition, TIR lengths are reported to differ between species.

However, after acknowledging that TIRs in hAT elements may be longer than expected, it became increasingly difficult to discern between hATs and MuLEs. This is due to the fact that MuLEs, close relatives to hATs, also often harbour 8 bp TSDs (typically 8-10 bp), but are often distinguished by longer TIRs [201]. In the absence of conserved domains or similarities to other known fungal TEs, many putative hATs or MuLEs were denoted with a '?' suffix, or demoted simply to DNA elements. Further characterisation of fungal TEs may elucidate the atypical structures of some TEs in fungi.

4.3.2 Objectives Two and Three

Secondary alignments: During annotation, RepeatMasker identifies TEs to be a primary alignment, or a secondary alignment. Matches are considered to be secondary

when there is a higher scoring match that partly includes the domain of the match (>80%)[19]. Hence, the risk of including secondary matches is an over-representation of TEs in the genome due to overlapping boundaries between the primary and secondary match. In addition, secondary alignments are difficult to manage during downstream analyses as RepeatMasker does not record pairings between secondary alignments and their primary partner. I determined these associations based on overlapping boundaries between the primary and secondary copies. Many secondary alignments belonged to the same family as their primary match, hence their inclusion would not serve any purpose in downstream analyses, and it would increase the copy number and perceived length of TE content simply due to sequence similarity between two TEs of the same class. Contrarily, primary DNA annotations often had secondary MITEs. This may be due to the MITE elements being deletion derivatives of the DNA element, i.e., RepeatMasker assigns a fragment of a DNA element (such as the TIR) as a primary match, and identifies the derivative MITE as a secondary match. If MITEs are a class of interest in downstream analyses, the exclusion of secondary alignments may risk under-representation of this class in the genome, or the inclusion of secondary alignments may hinder the analyses due to partial fragments of DNA elements being perceived as MITEs. Due to the difficult nature of managing secondary alignments, for the purpose of this project, secondary alignments were excluded from the analyses.

RIP-affected sequences In chapter 3, I measured the divergence of a TE copy from its respective consensus sequence to estimate the relative age of the TE. It is important to note here that divergence values can only serve as rough estimates not only due to the rapid evolution of TEs, but also due to the accrual of RIP-induced mutations. In addition to potentially obscuring the estimated timeline of insertions, detecting RIP in itself comes with limitations. The presence of RIP is detected via dinucleotide frequency variations and point-mutations. As RIP targets nucleotides in a preferred dinucleotide context (CpA dinucleotides), known pre- and post- RIP dinucleotide frequencies can be used to measure the extent of RIP. However, it must be

appreciated that C-to-T transitions in the CpA dinucleotides can also occur as a result of spontaneous mutation. As a result, it can be difficult to discern whether mutations that are typical of RIP have indeed occurred purely due to a targeted defence against TEs. Some studies utilise methods to potentially differentiate between RIP-induced mutations and random mutations by reconstructing trees from TE alignments with masked CpA dinucleotides (as described by Winter et al. [67]). This, however, was beyond the scope of this project. However, in all three focal genomes of this study, the extent of RIP was so great, and largely consistent with other *Epichloë* studies that it is very unlikely that natural mutational processes had the capacity to obscure these results.

Sequence similarity between DNA transposons and MITEs MITEs are typically deletion derivatives of DNA elements. As a result, the TIR of MITEs may retain very high sequence similarity to the DNA element from which it was derived. This risks the incorrect annotation of MITEs as DNA elements when using homology-based annotation software such as RepeatMasker. In this work, some short DNA elements resided within genic regions which is an insertion preference typical of MITE elements. However, it was impossible to discern whether these elements were indeed MITEs that had been incorrectly annotated as DNA elements. This is due to the fact that while the total length of a TE serves as a classification determinant, many of these short elements were only partial copies. Thus, it was not possible to determine whether these TEs were truncated copies of DNA elements or true MITEs. Nevertheless, the potential for MITEs to be incorrectly annotated as DNA elements is an important feature to consider when interpreting results.

4.4 Future Work

Transposable elements are a greatly understudied genomic component due to their dynamic and repetitive nature. However, the the work produced in this thesis may serve as a resource for future TE studies. Within and beyond the *Epichloë* genus, the curation of a high quality TE library will provide a resource for future fungal studies that wish to conduct homology-based annotations of TEs. As current homology-based TE databases remain dominated by animal- and plant-derived TEs, the fungal TEs characterised in this work may assist in more robust fungal annotations due to the decreased phylogenetic distance between a given fungal genome and *Epichloë*.

Within the *Epichloë* genus, the identification of long TE elements that have not been targeted by RIP suggests that TEs may indeed remain active in these genomes, a result that was not determined prior to this project [67]. This hypothesis is strengthened by the TE activity that has been shown here to post-date the divergence of members of the *E. typhina* species complex. This activity prompts motivation to continue studying the dynamics of TEs within the *Epichloë* genome, particularly in their association with genes.

While this research demonstrated that effector and secondary metabolite genes are not significantly enriched near TE-dense regions of this genome, the data did identify that MITEs reside in gene-rich regions and are consistently overrepresented within 1 kb of effector and secondary metabolite genes across all three genomes. This is a distance at which a TE can hypothetically influence gene expression [191]. Thus, future studies may benefit from conducting gene expression studies using functional genomic data such as RNAseq data for the three focal genomes. This may elucidate novel gene expression traits such as up-regulation or down-regulation of these genes proximal to MITEs. In addition, the genes of interest may be extended beyond effectors and secondary metabolites, for example to genes involved in fungal growth.

Above all, the curation of a high quality TE library for *Epichloë*, characterisation of recently active TE classes, and the large body of literature that appraises the role of TEs in fungal genome evolution presents reason to continue the study of TEs within the *Epichloë* genus.

Bibliography

1. Los Almost National Laboratory. *Advanced Consensus Maker*. `https://www.hiv.lanl.gov/content/sequence/CONSENSUS/AdvCon.html`.

2. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30,** 3276–3278 (2014).

3. Blin, K. *et al.* antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research* **49,** 29–35 (2021).

4. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

5. Cock, P. J. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25,** 1422–1423 (2009).

6. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215,** 403–410 (1990).

7. Lu, S. *et al.* CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Research* **48,** D265–D268 (2020).

8. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).

9. Kohany, O., Gentles, A. J., Hankus, L. & Jurka, J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* **7,** 1–7 (2006).

10. Sperschneider, J. *et al.* EffectorP: predicting fungal effector proteins from secretomes using machine learning. *New Phytologist* **210,** 743–761 (2016).

11. Palmer, J. & Stajich, J. *Funannotate: a fungal genome annotation and comparative genomics pipeline* 2016.

12. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Research* **21,** 487–493 (2011).

13. Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics* **9,** 286–298 (2008).

14. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34,** 3094–3100 (2018).

15. Testa, A. C., Oliver, R. P. & Hane, J. K. OcculterCut: a comprehensive survey of AT-rich regions in fungal genomes. *Genome Biology and Evolution* **8,** 2044–2064 (2016).

16. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual* ISBN: 1441412697 (Scotts Valley, CA, 2009).

17. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2020). `https://www.R-project.org/`.

18. RStudio Team. *RStudio: Integrated Development Environment for R* RStudio, Inc. (Boston, MA, 2020). `http://www.rstudio.com/`.

19. Smit, A., Hubley, R. & Green, P. *RepeatMasker Open-4.0. 2013–2015* 2015.

20. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117,** 9451–9457 (2020).

21. Petersen, T. N., Brunak, S., Von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* **8,** 785–786 (2011).

22. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27,** 573–580 (1999).

23. Gregory, T. R. Synergy between sequence and size in large-scale genomics. *Nature Communications Genetics* **6,** 699–708 (2005).

24. Eddy, S. R. The C-value paradox, junk DNA and ENCODE. *Current Biology* **22,** 898–899 (2012).

25. Cvijović, I., Good, B. H. & Desai, M. M. The effect of strong purifying selection on genetic diversity. *Genetics* **209,** 1235–1278 (2018).

26. Gilbert, W. Why genes in pieces? *Nature* **271,** 501–501 (1978).

27. Balakirev, E. S. & Ayala, F. J. Pseudogenes: are they" junk" or functional DNA? *Annual Review of Genetics* **37,** 123–151 (2003).

28. Cavalier-Smith, T. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. *Journal of Cell Science* **34,** 247–278 (1978).

29. Pritham, E. J. Transposable elements and factors influencing their success in eukaryotes. *Journal of Heredity* **100,** 648–655 (2009).

30. Sotero-Caio, C. G., Platt, R. N., Suh, A. & Ray, D. A. Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biology and Evolution* **9,** 161–177 (2017).

31. Enriquez-Gasca, R., Gould, P. A. & Rowe, H. M. Host gene regulation by transposable elements: the new, the old and the ugly. *Viruses* **12,** 1089 (2020).

32. Wells, J. N. & Feschotte, C. A field guide to eukaryotic transposable elements. *Annual Review of Genetics* **54,** 539–561 (2020).

33. Kissinger, J. C. & DeBarry, J. Genome cartography: charting the apicomplexan genome. *Trends in Parasitology* **27,** 345–354 (2011).

34. Tang, W., Mun, S., Joshi, A., Han, K. & Liang, P. Mobile elements contribute to the uniqueness of human genome with 15,000 human-specific insertions and 14 Mbp sequence increase. *DNA Research* **25,** 521–533 (2018).

35. Schnable, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326,** 1112–1115 (2009).

36. Payer, L. M. & Burns, K. H. Transposable elements in human genetic disease. *Nature Communications Genetics* **20,** 760–772 (2019).

37. Lohe, A. R. & Hartl, D. L. Autoregulation of mariner transposase activity by overproduction and dominant-negative complementation. *Molecular Biology and Evolution* **13,** 549–555 (1996).

38. Bourque, G. *et al.* Ten things you should know about transposable elements. *Genome Biology* **19,** 1–12 (2018).

39. Kidwell, M. G. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115,** 49–63 (2002).

40. Krishnan, P. *et al.* Transposable element insertions shape gene regulation and melanin production in a fungal pathogen of wheat. *BMC Biology* **16,** 1–18 (2018).

41. Chen, F. *et al.* Fungicide-induced transposon movement in *Monilinia fructicola*. *Fungal Genetics and Biology* **85,** 38–44 (2015).

42. Casacuberta, E. & González, J. The impact of transposable elements in environmental adaptation. *Molecular Ecology* **22,** 1503–1517 (2013).

43. Ogasawara, H., Obata, H., Hata, Y., Takahashi, S. & Gomi, K. Crawler, a novel Tc1/mariner-type transposable element in *Aspergillus oryzae* transposes under stress conditions. *Fungal Genetics and Biology* **46,** 441–449 (2009).

44. Park, M., Christin, P.-A. & Bennetzen, J. L. Sample sequence analysis uncovers recurrent horizontal transfers of transposable elements among grasses. *Molecular Biology and Evolution* **38,** 3664–3675 (2021).

45. Venner, S. *et al.* Ecological networks to unravel the routes to horizontal transposon transfers. *PLoS biology* **15,** e2001536 (2017).

46. Roulin, A. *et al.* Whole genome surveys of rice, maize and sorghum reveal multiple horizontal transfers of the LTR-retrotransposon Route66 in *Poaceae*. *BMC Evolutionary Biology* **9,** 1–10 (2009).

47. Gilbert, C. & Feschotte, C. Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. *Current Opinion in Genetics & Development* **49,** 15–24 (2018).

48. Suh, A. Horizontal transfer of transposons as genomic fossils of host-parasite interactions. *The Evolution and Fossil Record of Parasitism,* 451–463 (2021).

49. Feschotte, C. & Pritham, E. J. DNA transposons and the evolution of eukaryotic genomes. *Annual Review of Genetics* **41,** 331 (2007).

50. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nature Communications Genetics* **8,** 973–982 (2007).

51. Finnegan, D. J. Eukaryotic transposable elements and genome evolution. *Trends in Genetics* **5,** 103–107 (1989).

52. Arkhipova, I. R. & Yushenova, I. A. Giant transposons in eukaryotes: is bigger better? *Genome Biology and Evolution* **11,** 906–918 (2019).

53. Finnegan, D. J. Retrotransposons. *Current Biology* **22,** 432–437 (2012).

54. Neumann, P., Požárková, D. & Macas, J. Highly abundant pea LTR retrotransposon Ogre is constitutively transcribed and partially spliced. *Plant Molecular Biology* **53,** 399–410 (2003).

55. Jedlicka, P., Lexa, M. & Kejnovsky, E. What can long terminal repeats tell us about the age of LTR retrotransposons, gene conversion and ectopic recombination? *Frontiers in Plant Science* **11,** 644 (2020).

56. Linheiro, R. S. & Bergman, C. M. Whole Genome Resequencing Reveals Natural Target Site Preferences of Transposable Elements in Drosophila melanogaster. *PloS One* **7,** 1–12 (Feb. 2012).

57. Muszewska, A., Hoffman-Sommer, M. & Grynberg, M. LTR retrotransposons in fungi. *PloS One* **6,** e29425 (2011).

58. Zhang, L. *et al.* The structure and retrotransposition mechanism of LTR-retrotransposons in the asexual yeast *Candida albicans*. *Virulence* **5,** 655–664 (2014).

59. Wei, L. *et al.* New insights into nested long terminal repeat retrotransposons in Brassica species. *Molecular Plant* **6,** 470–482 (2013).

60. Levin, H. L. It's prime time for reverse transcriptase. *Cell* **88,** 5–8 (1997).

61. Finnegan, D. Transposable elements: how non-LTR retrotransposons do it. *Current Biology* **7,** R245–R248 (1997).

62. Cost, G. J., Feng, Q., Jacquier, A. & Boeke, J. D. Human L1 element target-primed reverse transcription in vitro. *The EMBO Journal* **21,** 5899–5910 (2002).

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

64. Aziz, R. K., Breitbart, M. & Edwards, R. A. Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Research* **38,** 4207–4217. ISSN: 0305-1048. `https://doi.org/10.1093/nar/gkq140` (Mar. 2010).

65. Hickman, A. B. & Dyda, F. DNA transposition at work. *Chemical Reviews* **116,** 12758–12784 (2016).

66. Fattash, I. *et al.* Miniature inverted-repeat transposable elements: discovery, distribution, and activity. *Genome* **56,** 475–486 (2013).

67. Winter, D. J. *et al.* Repeat elements organise 3D genome structure and mediate transcription in the filamentous fungus *Epichloë festucae*. *PLoS Genetics* **14,** e1007467 (2018).

68. Fleetwood, D. J. *et al.* Abundant degenerate miniature inverted-repeat transposable elements in genomes of epichloid fungal endophytes of grasses. *Genome Biology and Evolution* **3,** 1253–1264 (2011).

69. Keidar-Friedman, D., Bariah, I. & Kashkush, K. Genome-wide analyses of miniature inverted-repeat transposable elements reveals new insights into the evolution of the *Triticum-Aegilops* group. *PloS One* **13,** e0204972 (2018).

70. Winter, D. J. *et al.* Repeat elements organise 3D genome structure and mediate transcription in the filamentous fungus *Epichloë festucae*. *PLoS Genetics* **14,** e1007467 (2018).

71. Jiang, N., Feschotte, C., Zhang, X. & Wessler, S. R. Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). *Current Opinion in Plant Biology* **7,** 115–119 (2004).

72. Feschotte, C. & Mouches, C. Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a pogo-like DNA transposon. *Molecular Biology and Evolution* **17,** 730–737 (2000).

73. Orgel, L. E. & Crick, F. H. Selfish DNA: the ultimate parasite. *Nature* **284,** 604–607 (1980).

74. McClintock, B. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences* **36,** 344–355 (1950).

75. McClintock, B. Induction of instability at selected loci in maize. *Genetics* **38,** 579 (1953).

76. Du, C., Hoffman, A., He, L., Caronna, J. & Dooner, H. K. The complete Ac/Ds transposon family of maize. *BMC Genomics* **12,** 1–12 (2011).

77. Nanjundiah, V. Barbara McClintock and the discovery of jumping genes. *Resonance* **1,** 56–62 (1996).

78. Arkhipova, I. R. Neutral theory, transposable elements, and eukaryotic genome evolution. *Molecular Biology and Evolution* **35,** 1332–1337 (2018).

79. Iranzo, J., Gómez, M. J., Lopez de Saro, F. J. & Manrubia, S. Large-scale genomic analysis suggests a neutral punctuated dynamics of transposable elements in bacterial genomes. *PLoS Computational Biology* **10,** e1003680 (2014).

80. Jangam, D., Feschotte, C. & Betrán, E. Transposable element domestication as an adaptation to evolutionary conflicts. *Trends in Genetics* **33,** 817–831 (2017).

81. Hancks, D. C. & Kazazian, H. H. Roles for retrotransposon insertions in human disease. *Mobile DNA* **7,** 1–28 (2016).

82. Tsushima, A. *et al.* Genomic plasticity mediated by transposable elements in the plant pathogenic fungus *Colletotrichum higginsianum. Genome Biology and Evolution* **11,** 1487–1500 (2019).

83. Gilbert, N., Lutz-Prigge, S. & Moran, J. V. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110,** 315–325 (2002).

84. Narita, N. *et al.* Insertion of a 5'truncated L1 element into the 3'end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy. *The Journal of Clinical Investigation* **91,** 1862–1867 (1993).

85. Choi, J., Lyons, D. B., Kim, M. Y., Moore, J. D. & Zilberman, D. DNA methylation and histone H1 jointly repress transposable elements and aberrant intragenic transcripts. *Molecular Cell* **77,** 310–323 (2020).

86. Rolland, A. *et al.* The envelope protein of a human endogenous retrovirus-W family activates innate immunity through CD14/TLR4 and promotes Th1-like responses. *The Journal of Immunology* **176,** 7636–7644 (2006).

87. Gray, Y. H. It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends in Genetics* **16,** 461–468 (2000).

88. Davière, J.-M., Langin, T. & Daboussi, M.-J. Potential role of transposable elements in the rapid reorganization of the *Fusarium oxysporum* genome. *Fungal Genetics and Biology* **34,** 177–192 (2001).

89. Thon, M. R. *et al.* The role of transposable element clusters in genome evolution and loss of synteny in the rice blast fungus *Magnaporthe oryzae*. *Genome Biology* **7,** 1–9 (2006).

90. Hedges, D. & Deininger, P. Inviting instability: transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutation Research* **616,** 46–59 (2007).

91. Morgante, M. *et al.* Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nature Genetics* **37,** 997–1002 (2005).

92. Elrouby, N. & Bureau, T. E. Bs1, a new chimeric gene formed by retrotransposon-mediated exon shuffling in maize. *Plant Physiology* **153,** 1413–1424 (2010).

93. McDonald, M. C. *et al.* Transposon-mediated horizontal transfer of the host-specific virulence protein ToxA between three fungal wheat pathogens. *MBio* **10,** e01515–19 (2019).

94. Brosius, J. Retroposons—seeds of evolution. *Science* **251,** 753–753 (1991).

95. Miller, W. J., McDonald, J. F., Nouaud, D. & Anxolabéhère, D. in *Transposable Elements and Genome Evolution* 197–207 (Springer, 2000).

96. Sundaram, V. *et al.* Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Research* **24,** 1963–1976 (2014).

97. Lee, H. J. *et al.* Epigenomic analysis reveals prevalent contribution of transposable elements to cis-regulatory elements, tissue-specific expression, and alternative promoters in zebrafish. *Genome Research,* gr–276052 (2022).

98. Miao, B. *et al.* Tissue-specific usage of transposable element-derived promoters in mouse development. *Genome Biology* **21,** 1–25 (2020).

99. Liang, S. C. *et al.* Kicking against the PRCs–a domesticated transposase antagonises silencing mediated by Polycomb group proteins and is an accessory component of Polycomb repressive complex 2. *PLoS Genetics* **11,** e1005660 (2015).

100. Ramsay, L. *et al.* Conserved expression of transposon-derived non-coding transcripts in primate stem cells. *BMC Genomics* **18,** 1–13 (2017).

101. Bebber, D. P. & Gurr, S. J. Crop-destroying fungal and oomycete pathogens challenge food security. *Fungal Genetics and Biology* **74,** 62–64 (2015).

102. Singh, L. P., Gill, S. S. & Tuteja, N. Unraveling the role of fungal symbionts in plant abiotic stress tolerance. *Plant Signaling & Behavior* **6,** 175–191 (2011).

103. Kivlin, S. N., Emery, S. M. & Rudgers, J. A. Fungal symbionts alter plant responses to global change. *American Journal of Botany* **100,** 1445–1457 (2013).

104. Frantzeskakis, L., Kusch, S. & Panstruga, R. The need for speed: compartmentalized genome evolution in filamentous phytopathogens. *Molecular Plant Pathology* **20,** 3 (2019).

105. Jeon, J. *et al.* Genome-wide profiling of DNA methylation provides insights into epigenetic regulation of fungal development in a plant pathogenic fungus, *Magnaporthe oryzae*. *Scientific Reports* **5,** 1–11 (2015).

106. Nakayashiki, H. RNA silencing in fungi: mechanisms and applications. *FEBS Letters* **579,** 5950–5957 (2005).

107. Galagan, J. E. & Selker, E. U. RIP: the evolutionary cost of genome defense. *Trends in Genetics* **20,** 417–423 (2004).

108. Hane, J. K. & Oliver, R. P. RIPCAL: a tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences. *BMC Bioinformatics* **9,** 1–12 (2008).

109. Cambareri, E. B., Jensen, B. C., Schabtach, E. & Selker, E. U. Repeat-induced GC to AT mutations in *Neurospora*. *Science* **244,** 1571–1575 (1989).

110. Pereira, D., Oggenfuss, U., McDonald, B. A. & Croll, D. Population genomics of transposable element activation in the highly repressive genome of an agricultural pathogen. *Microbial Genomics* **7** (2021).

111. Clutterbuck, A. J. Genomic evidence of repeat-induced point mutation (RIP) in filamentous ascomycetes. *Fungal Genetics and Biology* **48,** 306–326 (2011).

112. Selker, E. U. & Stevens, J. N. DNA methylation at asymmetric sites is associated with numerous transition mutations. *Proceedings of the National Academy of Sciences* **82,** 8114–8118 (1985).

113. Selker, E. U. & Stevens, J. Signal for DNA methylation associated with tandem duplication in *Neurospora crassa*. *Molecular and Cellular Biology* **7,** 1032–1038 (1987).

114. Gladyshev, E. & Kleckner, N. DNA sequence homology induces cytosine-to-thymine mutation by a heterochromatin-related pathway in *Neurospora*. *Nature Genetics* **49,** 887–894 (2017).

115. Lorrain, C., Feurtey, A., Möller, M., Haueisen, J. & Stukenbrock, E. Dynamics of transposable elements in recently diverged fungal pathogens: lineage-specific transposable element content and efficiency of genome defenses. *G3* **11,** jkab068 (2021).

116. Selker, E. U. Repeat-induced gene silencing in fungi. *Advances in Genetics* **46,** 439–450 (2002).

117. Bewick, A. J. *et al.* Diversity of cytosine methylation across the fungal tree of life. *Nature Ecology & Evolution* **3,** 479–490 (2019).

118. Cooper, D. N., Mort, M., Stenson, P. D., Ball, E. V. & Chuzhanova, N. A. Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. *Human Genomics* **4,** 1–5 (2010).

119. Lieb, M. & Bhagwat, A. Repair of G–T Mismatches by *Escherichia coli* Vsr and Eukaryotic DNA Glycosylases (2013).

120. Nowrousian, M. *et al.* De novo assembly of a 40 Mb eukaryotic genome from short sequence reads: *Sordaria macrospora*, a model organism for fungal morphogenesis. *PLoS Genetics* **6,** e1000891 (2010).

121. Möller, M. *et al.* Recent loss of the Dim2 DNA methyltransferase decreases mutation rate in repeats and changes evolutionary trajectory in a fungal pathogen. *PLoS Genetics* **17,** e1009448 (2021).

122. Castanera, R. *et al.* Transposable elements versus the fungal genome: impact on whole-genome architecture and transcriptional profiles. *PLoS Genetics* **12,** e1006108 (2016).

123. Oliva, R. *et al.* Recent developments in effector biology of filamentous plant pathogens. *Cellular Microbiology* **12,** 705–715 (2010).

124. Selin, C., De Kievit, T. R., Belmonte, M. F. & Fernando, W. Elucidating the role of effectors in plant-fungal interactions: progress and challenges. *Frontiers in Microbiology* **7,** 600 (2016).

125. Lo Presti, L. *et al.* Fungal effectors and plant susceptibility. *Annual Review of Plant Biology* **66,** 513–545 (2015).

126. Grandaubert, J., Balesdent, M.-H. & Rouxel, T. Evolutionary and adaptive role of transposable elements in fungal genomes. *Advances in Botanical Research* **70,** 79–107 (2014).

127. Zamioudis, C. & Pieterse, C. M. Modulation of host immunity by beneficial microbes. *Molecular Plant-Microbe Interactions* **25,** 139–150 (2012).

128. Keller, N. P. Fungal secondary metabolism: regulation, function and drug discovery. *Nature Reviews Microbiology* **17,** 167–180 (2019).

129. Dallery, J.-F. *et al.* Gapless genome assembly of *Colletotrichum higginsianum* reveals chromosome structure and association of transposable elements with secondary metabolite gene clusters. *BMC Genomics* **18,** 1–22 (2017).

130. Witte, T. E., Villeneuve, N., Boddy, C. N. & Overy, D. P. Accessory Chromosome-Acquired Secondary Metabolism in Plant Pathogenic Fungi: The Evolution of Biotrophs Into Host-Specific Pathogens. *Frontiers in Microbiology* **12** (2021).

131. Hane, J. K. *et al.* A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. *Genome Biology* **12,** 1–16 (2011).

132. Bertazzoni, S. *et al.* Accessories make the outfit: accessory chromosomes and other dispensable DNA regions in plant-pathogenic fungi. *Molecular Plant-Microbe Interactions* **31,** 779–788 (2018).

133. Habig, M., Lorrain, C., Feurtey, A., Komluski, J. & Stukenbrock, E. H. Epigenetic modifications affect the rate of spontaneous mutations in a pathogenic fungus. *Nature Communications* **12,** 1–13 (2021).

134. Habig, M., Quade, J. & Stukenbrock, E. H. Forward genetics approach reveals host genotype-dependent importance of accessory chromosomes in the fungal wheat pathogen *Zymoseptoria tritici*. *MBio* **8,** e01919–17 (2017).

135. Rao, S., Sharda, S., Oddi, V. & Nandineni, M. R. The landscape of repetitive elements in the refined genome of chilli anthracnose fungus *Colletotrichum truncatum*. *Frontiers in Microbiology,* 2367 (2018).

136. Dong, S., Raffaele, S. & Kamoun, S. The two-speed genomes of filamentous pathogens: waltz with plants. *Current Opinion in Genetics & Development* **35,** 57–65 (2015).

137. Faino, L. *et al.* Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen. *Genome Research* **26,** 1091–1100 (2016).

138. Laurent, B. *et al.* High-resolution mapping of the recombination landscape of the phytopathogen *Fusarium graminearum* suggests two-speed genome evolution. *Molecular Plant Pathology* **19,** 341–354 (2018).

139. Gladyshev, E. Repeat-induced point mutation and other genome defense mechanisms in fungi. *Microbiology Spectrum* **5,** 5–4 (2017).

140. Haas, B. J. *et al.* Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* **461,** 393–398 (2009).

141. Rouxel, T. *et al.* Effector diversification within compartments of the Leptosphaeria maculans genome affected by Repeat-Induced Point mutations. *Nature Communications* **2,** 1–10 (2011).

142. Frantzeskakis, L. *et al.* Signatures of host specialization and a recent transposable element burst in the dynamic one-speed genome of the fungal barley powdery mildew pathogen. *BMC Genomics* **19,** 1–23 (2018).

143. Fokkens, L. *et al.* The multi-speed genome of *Fusarium oxysporum* reveals association of histone modifications with sequence divergence and footprints of past horizontal chromosome transfer events. *BioRxiv,* 465070 (2018).

144. Leuchtmann, A., Bacon, C. W., Schardl, C. L., White Jr, J. F. & Tadych, M. Nomenclatural realignment of *Neotyphodium* species with genus *Epichloë*. *Mycologia* **106,** 202–215 (2014).

145. Clay, K. & Schardl, C. Evolutionary origins and ecological consequences of endophyte symbiosis with grasses. *The American Naturalist* **160,** 99–127 (2002).

146. Schardl, C. L. The *Epichloae*, symbionts of the grass subfamily *Poöideae*. *Annals of the Missouri Botanical Garden,* 646–665 (2010).

147. Berry, D. *et al.* Cross-species transcriptomics identifies core regulatory changes differentiating the asymptomatic asexual and virulent sexual life cycles of grass-symbiotic *Epichloe* fungi. *G3* **12,** jkac043 (2022).

148. Treindl, A. D., Stapley, J., Winter, D. J., Cox, M. P. & Leuchtmann, A. Chromosome-level genomes provide insights into genome evolution, organization and size in *Epichloe* fungi. *Genomics* **113,** 4267–4275 (2021).

149. Malinowski, D. P. & Belesky, D. P. Adaptations of endophyte-infected cool-season grasses to environmental stresses: mechanisms of drought and mineral stress tolerance. *Crop Science* **40,** 923–940 (2000).

150. Bayat, F., Mirlohi, A. & Khodambashi, M. Effects of endophytic fungi on some drought tolerance mechanisms of tall fescue in a hydroponics culture. *Russian Journal of Plant Physiology* **56,** 510–516 (2009).

151. Tian, Z., Wang, R., Ambrose, K. V., Clarke, B. B. & Belanger, F. C. The *Epichloë festucae* antifungal protein has activity against the plant pathogen *Sclerotinia homoeocarpa*, the causal agent of dollar spot disease. *Scientific Reports* **7,** 1–15 (2017).

152. Kauppinen, M., Saikkonen, K., Helander, M., Pirttilä, A. M. & Wäli, P. R. *Epichloë* grass endophytes in sustainable agriculture. *Nature Plants* **2,** 1–7 (2016).

153. Young, C. *et al.* Molecular cloning and genetic analysis of a symbiosis-expressed gene cluster for lolitrem biosynthesis from a mutualistic endophyte of perennial ryegrass. *Molecular Genetics and Genomics* **274,** 13–29 (2005).

154. Young, C. A. *et al.* A complex gene cluster for indole-diterpene biosynthesis in the grass endophyte *Neotyphodium lolii*. *Fungal Genetics and Biology* **43,** 679–693 (2006).

155. Young, C. A. *et al.* Genetics, genomics and evolution of ergot alkaloid diversity. *Toxins* **7,** 1273–1302 (2015).

156. Eaton, C. J. *et al.* A core gene set describes the molecular basis of mutualism and antagonism in *Epichloë* spp. *Molecular Plant-Microbe Interactions* **28,** 218–231 (2015).

157. Scott, B., Green, K. & Berry, D. The fine balance between mutualism and antagonism in the *Epichloë festucae*–grass symbiotic interaction. *Current Opinion in Plant Biology* **44,** 32–38 (2018).

158. Hassing, B. *et al.* Analysis of *Epichloë festucae* small secreted proteins in the interaction with *Lolium perenne*. *PloS One* **14,** e0209463 (2019).

159. Schirrmann, M. K. & Leuchtmann, A. The role of host-specificity in the reproductive isolation of *Epichloë* endophytes revealed by reciprocal infections. *Fungal Ecology* **15,** 29–38 (2015).

160. Schirrmann, M. K., Zoller, S., Fior, S. & Leuchtmann, A. Genetic evidence for reproductive isolation among sympatric *Epichloë* endophytes as inferred from newly developed microsatellite markers. *Microbial Ecology* **70,** 51–60 (2015).

161. Quenu, M. *et al.* Telomere-to-Telomere Genome Sequences across a Single Genus Reveal Highly Variable Chromosome Rearrangement Rates but Absolute Stasis of Chromosome Number. *Journal of Fungi* **8,** 670 (2022).

162. Goerner-Potvin, P. & Bourque, G. Computational tools to unmask transposable elements. *Nature Communications Genetics* **19,** 688–704 (2018).

163. Lerat, E. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity* **104,** 520–533 (2010).

164. Platt, R. N., Blanco-Berdugo, L. & Ray, D. A. Accurate transposable element annotation is vital when analyzing new genome assemblies. *Genome Biology and Evolution* **8,** 403–410 (2016).

165. Bergman, C. M. & Quesneville, H. Discovering and detecting transposable elements in genome sequences. *Briefings in Bioinformatics* **8,** 382–392 (2007).

166. Goubert, C. *et al.* A beginner's guide to manual curation of transposable elements. *Mobile DNA* **13,** 1–19 (2022).

167. Suh, A. *et al.* Multiple lineages of ancient CR1 retroposons shaped the early genome evolution of amniotes. *Genome Biology and Evolution* **7,** 205–217 (2015).

168. Suh, A., Smeds, L. & Ellegren, H. Abundant recent activity of retrovirus-like retrotransposons within and among flycatcher species implies a rich source of structural variation in songbird genomes. *Molecular Ecology* **27,** 99–111 (2018).

169. Peona, V. *et al.* Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Molecular Ecology Resources* **21,** 263–286 (2021).

170. Jebb, D. *et al.* Six reference-quality genomes reveal evolution of bat adaptations. *Nature* **583,** 578–584 (2020).

171. Louha, S., Ray, D. A., Winker, K. & Glenn, T. C. A high-quality genome assembly of the North American Song Sparrow, *Melospiza melodia*. *G3: Genes, Genomes, Genetics* **10,** 1159–1166 (2020).

172. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21,** i351–i358 (2005).

173. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research* **12,** 1269–1276 (2002).

174. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiology* **176,** 1410–1422 (2018).

175. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9,** 1–14 (2008).

176. Hubley, R. *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Research* **44,** D81–D89 (2016).

177. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32,** 1792–1797 (2004).

178. Gao, C. *et al.* Characterization and functional annotation of nested transposable elements in eukaryotic genomes. *Genomics* **100,** 222–230 (2012).

179. Banuelos, M. & Sindi, S. Modeling transposable element dynamics with fragmentation equations. *Mathematical Biosciences* **302,** 46–66 (2018).

180. Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research* **38,** e199–e199 (2010).

181. Crescente, J. M., Zavallo, D., Helguera, M. & Vanzetti, L. S. MITE Tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes. *BMC Bioinformatics* **19,** 1–10 (2018).

182. Ye, C., Ji, G. & Liang, C. detectMITE: a novel approach to detect miniature inverted repeat transposable elements in genomes. *Scientific Reports* **6,** 1–11 (2016).

183. Hughes, J. F. & Coffin, J. M. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. *Proceedings of the National Academy of Sciences* **101,** 1668–1672 (2004).

184. Essers, L., Adolphs, R. H. & Kunze, R. A highly conserved domain of the maize activator transposase is involved in dimerization. *The Plant Cell* **12,** 211–223 (2000).

185. Kanaya, S. Ribonuclease H. *The FEBS Journal* **276,** 1481–1481 (2009).

186. Croll, D. & McDonald, B. A. The accessory genome as a cradle for adaptive evolution in pathogens. *PLoS Pathogens* **8,** e1002608 (2012).

187. Winter, D. *pafr: Read, Manipulate and Visualize 'Pairwise mApping Format' Data* (2020). `https://cran.r-project.org/web/packages/pafr/index.html`.

188. Winter, D. *repeatR: Parse and analyse RepeatMasker output* (2020). `https://github.com/dwinter/repeatR`.

189. Margolin, B. S. *et al.* A methylated *Neurospora* 5S rRNA pseudogene contains a transposable element inactivated by repeat-induced point mutation. *Genetics* **149,** 1787–1797 (1998).

190. Jedlicka, P., Lexa, M., Vanat, I., Hobza, R. & Kejnovsky, E. Nested plant LTR retrotransposons target specific regions of other elements, while all LTR retrotransposons often target palindromes and nucleosome-occupied regions: in silico study. *Mobile DNA* **10,** 1–14 (2019).

191. Muszewska, A., Steczkiewicz, K., Stepniewska-Dziubinska, M. & Ginalski, K. Transposable elements contribute to fungal genes and impact fungal lifestyle. *Scientific Reports* **9,** 1–10 (2019).

192. Wicker, T. *et al.* Reply: A unified classification system for eukaryotic transposable elements should reflect their phylogeny. *Nature Reviews Genetics* **10,** 276–276 (2009).

193. Gabriel, A., Willems, M., Mules, E. H. & Boeke, J. D. Replication infidelity during a single cycle of Ty1 retrotransposition. *Proceedings of the National Academy of Sciences* **93,** 7767–7771 (1996).

194. Schmidt, S. M. *et al.* MITEs in the promoters of effector genes allow prediction of novel virulence genes in *Fusarium oxysporum*. *BMC Genomics* **14,** 1–21 (2013).

195. Daron, J. *et al.* Organization and evolution of transposable elements along the bread wheat chromosome 3B. *Genome Biology* **15,** 1–15 (2014).

196. Kronmiller, B. A. & Wise, R. P. TEnest: automated chronological annotation and visualization of nested plant transposable elements. *Plant Physiology* **146,** 45–59 (2008).

197. Jiang, N. & Wessler, S. R. Insertion preference of maize and rice miniature inverted repeat transposable elements as revealed by the analysis of nested elements. *The Plant Cell* **13,** 2553–2564 (2001).

198. Seberg, O. & Petersen, G. A unified classification system for eukaryotic transposable elements should reflect their phylogeny. *Nature Reviews Genetics* **10,** 276–276 (2009).

199. Kempken, F. & Windhofer, F. The hAT family: a versatile transposon group common to plants, fungi, animals, and man. *Chromosoma* **110,** 1–9 (2001).

200. Ferris, P. J. Characterization of a Chlamydomonas transposon, Gulliver, resembling those in higher plants. *Genetics* **122,** 363–377 (1989).

201. Liu, K. & Wessler, S. R. Transposition of Mutator–like transposable elements (MULEs) resembles hAT and Transib elements and V (D) J recombination. *Nucleic Acids Research* **45,** 6644–6655 (2017).