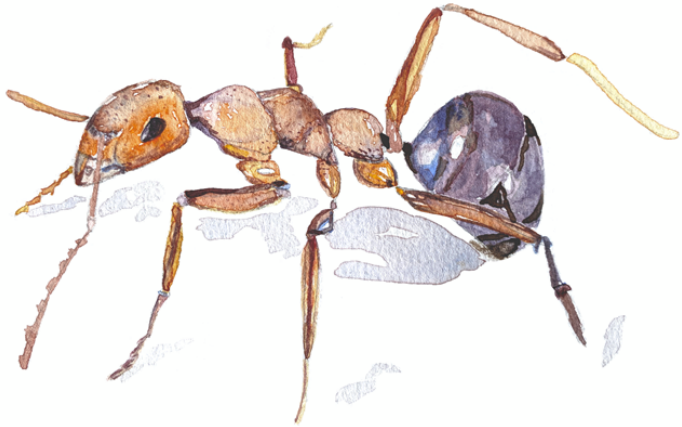
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Repeated increase of transposable elements and satellite DNA in three hybrid (*Formica aquilonia* × *F. polyctena*) wood ant populations



Patrick Heidbreder

Master's Programme in Ecology and Evolutionary Biology

Faculty of Biological and Environmental Sciences

University of Helsinki

April 2022

Abstract

Faculty: Faculty of Biological and Environmental Sciences

Degree programme: Master's Programme in Ecology and Evolutionary Biology

Study track: Ecology and Evolutionary Biology

Author: Patrick Heidbreder

Title: Repeated increase of transposable elements and satellite DNA in three hybrid (*Formica aquilonia* × *F. polyctena*) wood ant populations

Level: Master's Thesis

Month and year: April 2022

Number of pages: 51 (47 + Appendix)

Keywords: Transposons, genomic shock, genome expansion

Supervisor or supervisors: Dr. Pierre Nouhaud and Dr. Jonna Kulmuni

Where deposited: HELDA - Digital Repository of the University of Helsinki.

Additional information:

Abstract:

Hybridization between species is widespread across the tree of life and plays a role in adaptation, speciation and evolution. A critical component of hybridization is the compatibility of the combining genomes. Mechanisms that create incompatibilities, such as transposable element (TE) activity, are thus central to understanding and predicting the evolutionary effects of hybridization. The genomic shock hypothesis posits a burst of TE activity in hybrid genomes due to the uncoupling of TEs and their silencers. While many studies on this topic have focused on laboratory hybrids, there is relatively little data for wild hybrid populations, especially in non-model species. Here, I take advantage of a recent (< 50 generations ago), natural, and replicated hybridization events between two wood ant species, *Formica aquilonia* and *F. polyctena*, to test for increased TE abundance in hybrids. Analyses of whole genomes (N_{total} = 99) from both parental species and three hybrid populations revealed significantly more total TE copies in all hybrid populations compared to each parental species, and this partly remained after controlling for genome size, suggesting TE reactivation in the hybrids. LINE, DNA, and LTR elements, as well as multiple new and unclassified repeats, contributed most to the observed increase. However, I also found concurrent increases in satellite DNA copies in hybrids, suggesting heterochromatin expansion after hybridization. So while the observed TE-copy number increase I have detected is consistent with the genomic shock hypothesis, additional work is required to demonstrate and fully characterize TE reactivation in hybrids. Overall my work contributes to our understanding of the effects of hybridization on TE reactivation, satellite DNA abundance, and genome size evolution in natural populations.

i. Contribution Statement

The following people contributed to this thesis:

PH - Patrick Heidbreder
PN - Pierre Nouhaud
JK - Jonna Kulmuni

P.N. conceived the initial project idea. P.N and J.K collected the data. P.N., J.K., and P.H., contributed to planning study questions and analyses. P.N. trimmed the raw sequence data. P.H. proposed the hypotheses and carried out the TE abundance analysis, genome size estimates, and statistical analysis. P.H wrote the thesis.

1.1 Background

1.1.1 What is hybridization

The evolution of species is often visualized as a bifurcating tree, with the base as a common ancestor and a successive splitting of branches representing the divergence of populations through time. This picture is powerful in its ability to show the relationship of species by shared ancestry. However, it is a simplified depiction. Instead of the neatly diverging lines commonly depicted in the tree of life (Fig. 1A) there is a web of connections between branches (Fig. 1C) (Abbott et al. 2013). This web is the result of hybridization (Fig. 1 B) the reproduction between genetically distinct populations producing offspring of mixed ancestry (Barton and Hewitt 1985). Hybridization is widespread; it plays a role in nearly all speciation events (Abbott et al. 2013) and approximately 25% of plant species and 10% of animals exchange genes with relatives based on the observed frequency of hybrid individuals (Arnold 1997; Mallet 2005; Taylor and Larson 2019). Thus hybridization is a fundamental evolutionary process and it is critical to understand its role in adaptation (by introducing novel genetic variation through introgression), evolution, and in the establishment of species barriers.

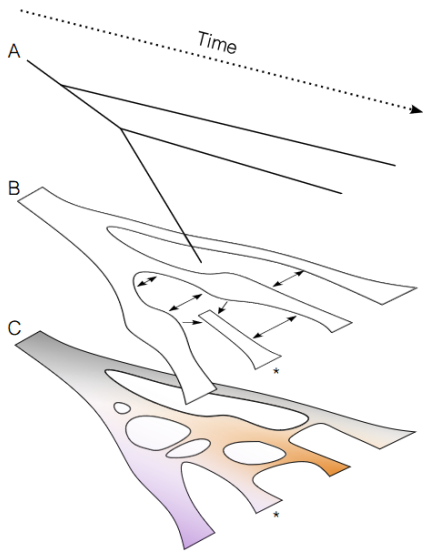


Figure 1. For simplicity, the tree of life is often depicted as single branches, with the only connections between them being divergent bifurcations **(A)**. In nature, hybridization results in gene flow between lineages during divergence **(B)**. This means the relationship between species is often better represented as the web of life **(C)**. Arrows indicate hybridization and gene flow, colors indicate divergence. * Hybrid species. Adapted and modified from Abbot et al. (2013).

When parents from two genetically distinct populations hybridize, their offspring can range from viable and fertile, to viable and infertile, to completely unviable, depending on the compatibility of the hybridizing genomes. An important factor in genomic compatibility is the level of genomic divergence between the populations from which the parent individuals come. Divergence is the accumulation of differences in DNA sequences between the parental populations caused by a lack of gene flow, which will increase the strength of reproductive isolation (RI). Classical RI barriers can include geography, ecology, behavior, physiology, and genetics (Coyne and Orr 2004; pp. 28-29). These are usually presented through their temporal order, from mating opportunities to offspring viability and fertility. The stronger these barriers are and the longer they persist, the greater the divergence between the populations. RI usually evolves progressively, and its strength may change over time and space, so hybridization occurs at varying levels of divergence (Abbott et al. 2013). Then, although the fitness of hybrid offspring is determined by the divergence between progenitors, hybridization can in turn reduce divergence through gene flow and recombination. In extreme cases this homogenizing effect can lead to speciation reversal and species collapse (Seehausen 2006). In other cases, the production of low fitness offspring may actually lead to further divergence through reinforcement, accelerating the speciation process. This sequence highlights the feedback relationship between hybridization and divergence. In all cases, the consequence of hybridization depends on whether the two parents are able to produce a hybrid offspring, and whether that offspring is viable and fertile. This means that understanding the factors that contribute to incompatibilities is an essential part of understanding hybridization and evolution.

1.1.2 What are transposable Elements

Transposable elements are DNA sequences of hundreds to thousands of bases that are capable of both changing positions within genomes and replicating independently of host DNA (Bourque et al. 2018; Wells and Feschotte 2020). Originally described by Barbara McClintock (1950), transposable elements (TEs) were noted for their ability to both move throughout the genome and affect the expression of genes proximate to their location (they were originally termed ‘controlling elements’). Since their discovery in maize, TEs have been found to be ubiquitous throughout the tree of life, with examples in nearly all known eukaryotic species (Wells and

Feschotte 2020) as well as prokaryotes (Kleckner 1981). In addition to their presence in nearly all organisms, TEs have had, and continue to have, a large number of evolutionary effects on their hosts at different timescales (Cavaller-Smith 1985; Pagel and Johnstone 1992).

TEs can take up shocking proportions of host genomes, up to 80% in some plant genomes (Lee and Kim 2014) and 45% in humans (Baltimore 2001). Such high TE abundances can inflate genome sizes, resulting in the complexity of organisms no longer correlating with their genome size. This observation is known as the C-value paradox (see Eddy 2012 for an overview of the C-value paradox). Initially ignored as "junk DNA", their adaptive and evolutionary contributions have been gaining recognition. TE activity has been found to play roles in gene creation and regulation (reviewed in Feschotte 2008; González and Petrov 2009), disease (Hancks and Kazazian 2016; Payer and Burns 2019), immune function (Broecker and Moelling 2019; Huang et al. 2016; Kapitonov and Koonin 2015), telomere extension (Pardue and DeBaryshe 2011), and hybridization (Bingham et al. 1982; Moyle and Nakazato 2010). The possibility of benefits resulting from TE activity have led to proposals of TEs as sources of genetic diversity and mutations that aid in adaptation to novel environments (Schrader and Schmitz 2019).

The mobility of TEs is their defining feature. There are two primary modes of transposition, which are used to classify TE sequences into classes, subclasses, superfamilies, and families (Finnegan 1989; Wells and Feschotte 2020). The highest-level division is into two classes. Class I is characterized by replication through an RNA intermediate and subsequent reverse-transcription back into the genome (Boeke et al. 1985); this is usually described as a "copy and paste" mechanism. Class II follows a "cut and paste" method whereby the transposon sequences are excised and reinserted in a new location in the genome (Bourque et al. 2018). Both of these mechanisms can result in sequence duplication and subsequent copy number increases. Families are further divided by transposition mechanisms as well as sequence homology and transposase specificity.

The transposition of TEs in a genome can have a range of effects, though they are predominantly negative. The most famous (though possibly neither the most important nor most common) effect occurs when transposing TEs insert themselves into functional DNA regions resulting in regulatory effects, gene mutation, and/or loss-of-function (Bennetzen and Wang 2014). Such

insertions are directly responsible for a range of diseases in humans, including some cancers (Payer and Burns 2019).

Given the downsides of TE activity within genomes, hosts have evolved mechanisms to silence (i.e. stop) TE activity. There are two types of mechanisms to silence and remove TEs: negative

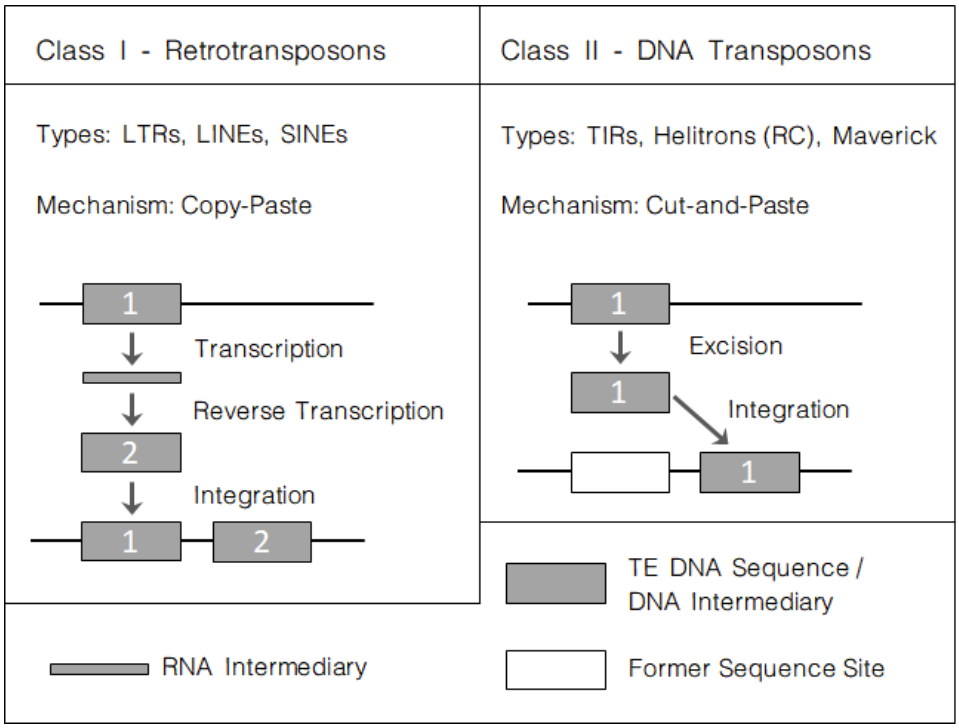


Figure 2. TE types and their transposition mechanisms. Duplication in Class I is a consequence of the transposition mechanism. Class II Transposons can duplicate if the excision occurs during DNA replication resulting in both the new and former sites being replicated. Adapted from Serrato-Capuchina and Matute (2018).

selective pressures and molecular defense mechanisms. There was previously some debate about the relative importance of each, but we now know that the former, which limits the accumulation of TE copy numbers (Brookfield 1996; Charlesworth and Charlesworth 1983) is likely the less important of the two (Kelleher et al. 2020). The second mechanism, molecular defenses, is still mostly unknown outside of model species. These defenses include DNA methylation and chromatin modification (Wood et al. 2016), zinc-finger protein silencing (Jacobs et al. 2014; Yang et al. 2017) , and post-transcriptional silencing by small-RNA pathways (Fu et al. 2014; Slotkin and Martienssen 2007). Together, these mechanisms work through a combination of transcriptional and post-transcriptional silencing to prevent TE transposition.

There are two important features of molecular TE silencing: 1) TE specificity, and 2) silencing delay. Sequence specificity is the requirement of a template for TE defenses to recognize TE sequences and their transcripts. Silencing delay is a consequence of defense mechanisms taking time, often generations, to fully recognize and silence TE activity. Together, these features underpin the dynamics of novel TE invasions in naive genomes, and result in reactivation of TEs and copy number expansion.

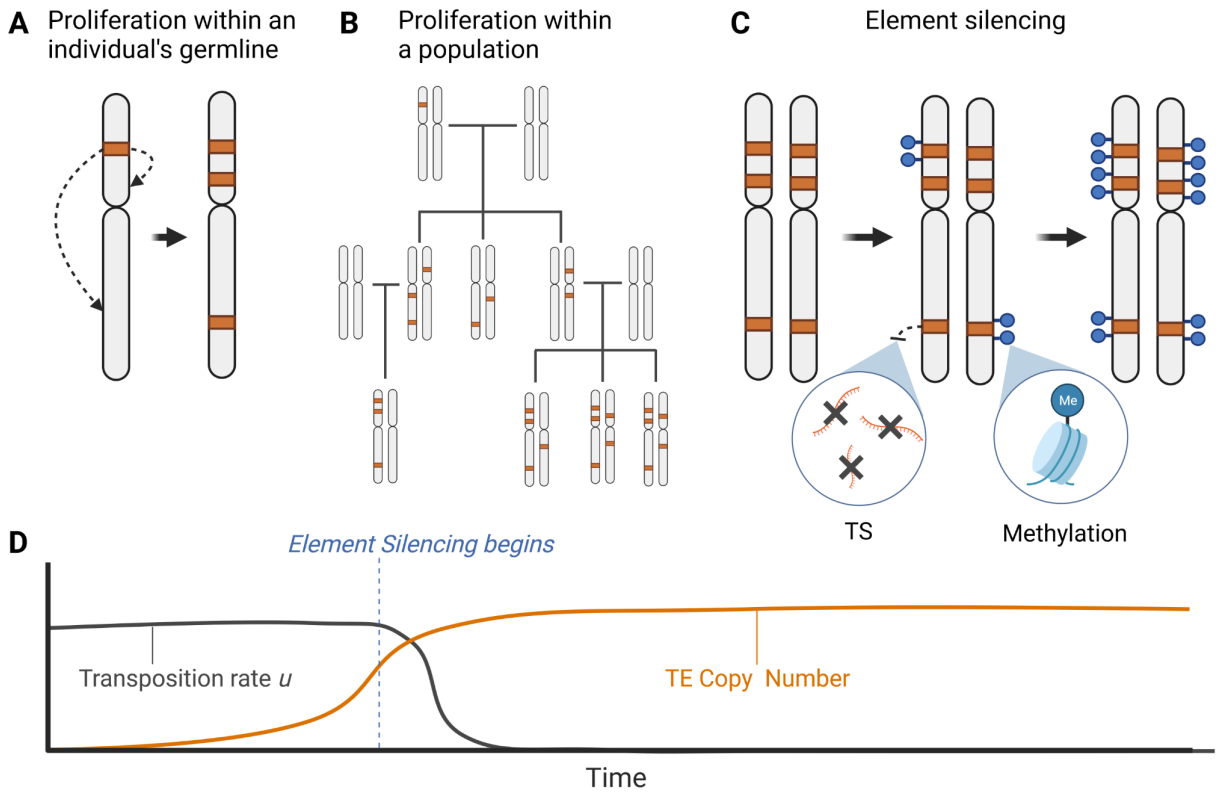


Figure 3. Conceptual process of a TE invasion in a population (or in hybrids after a hybridization event). Novel TEs both replicate within individual genomes **(A)**, likely within germline cells, and spread in the population through vertical transmission **(B)**, increasing copy numbers **(D)**. TE defense mechanisms, transcript silencing (TS) and heterochromatin formation (DNA methylation) **(C)**, begin to function. This leads to a decrease in transposition rate *u* **(D)**. Created with BioRender.com.

What does a TE invasion look like? The dynamics of TE spread in populations are a function of the rate of transposition (and subsequent copy number increase) and the rate of copy removal (Charlesworth and Charlesworth 1983). In most organisms, these rates are balanced, through transposition-selection equilibrium, with both being around zero. However, this picture is the

final stage of the TE invasion process. The process starts with the genome of a species that is naïve to a novel TE sequence. Such a TE sequence may enter either through introgression from a closely related species or through horizontal gene transfer (Silva et al. 2004). Kofler and colleagues (2015, 2018) describe this process in a natural and experimental invasion of the *P* element (a class II transposon, see above) in *Drosophila simulans*. After introduction, the TE can spread through the population, increasing a few copy numbers at a time each generation. As copy numbers increase, TE silencing mechanisms start to be recruited. This is still a poorly understood process, but the tailoring of molecular mechanisms, such as the pi-RNA pathway and heterochromatin expansion, begin to silence TE copies until the copy number increase is under control. The final result and signature of TE invasions is a "burst" of TE copy numbers in the genome. This leaves the picture we began with, large portions of genomes being made up of inactive and selectively neutral TE sequences.

1.1.3 Transposable Elements and Hybridization

In order to fully understand hybridization we must understand the factors that contribute to genomic incompatibilities, including TE activity. The primary relationship between TEs and hybridization is that TEs can act as a genetic incompatibility resulting in low-fitness hybrid offspring, which can have evolutionary effects in speciation or adaptation. The link was initially made with the discovery of hybrid dysgenesis (distinct germline abnormalities directly caused by TE activity) in *Drosophila* crosses being caused by a TE, the *P* element (Bingham et al. 1982; M. G. Kidwell et al. 1977; Margaret G. Kidwell 1983). Barbara McClintock (1984) formalized the relationship between TEs and hybridization with the genomic shock hypothesis, which predicts that stress on the genome, including from hybridization, can result in the deregulation of previously silenced TEs and their subsequent proliferation. With a testable hypothesis in genomic shock, research expanded to new taxa. This has yielded mixed results, with TE reactivation appearing in some species crosses, but not others (Table 1). Here I present the proposed mechanism of TE reactivation from hybridization, summarize the literature on the genomic shock hypothesis and present some broad conclusions, highlighting areas where we are still lacking information.

The mechanisms behind the genomic shock are conceptually quite simple. Two distinct species have functioning TE silencing mechanisms. However, as these mechanisms are each specific to TEs found in their respective lineages, when combined in a hybrid, there may end up being a mismatch of TEs and TE silencing mechanisms (Romero-Soriano et al. 2017). This mismatch can allow previously silenced TEs to become active and proliferate in hybrid genomes.

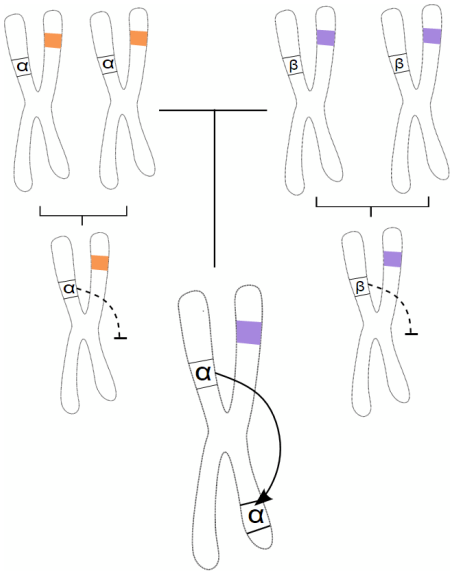


Figure 4. Within species the match of TEs (α and β) and the complete set of TE silencers (colored regions) is able to function to silence TE activity (dashed line). In hybrids, the mis-match of the alpha TEs and purple TE silencers results in the release of alpha TEs to actively transpose (solid arrow).

The initial quest for genomic shock took the form of systematic crosses between *Drosophila* species, looking for symptoms of hybrid dysgenesis. The outcomes from this quest are presented at the top of Table 1 and show mixed results, with TE activity detected in some crosses, but not in others. This pattern continues with crosses in other taxa. Such results are exemplified by the often cited *Helianthus annuus* and *H. petiolaris* sunflower crosses. In this case, initial results showed a signature of a past burst of TE activity in three independently formed hybrid species when compared to the parental species (Staton et al. 2009; Ungerer et al. 2006). This indicated a historic burst of TEs in hybrids in line with the genomic shock hypothesis. However, subsequent studies by Kawakami *et al.* (2011) revealed that when hybrids are made experimentally from the

Table 1. Report of previous results looking for evidence of TE bursts, TE activity, or TE copy number increases after hybridization. My literature search is not exhaustive, but rather intended to both show the diversity of results in the literature, and identify key characteristics and expectations of TE bursts.

Results		Species Cross	Relevant results	TE(s) studied	Reference
In *Drosophila*	no reactivation	*D. simulans* × *D. mauritiana*	No hybrid dysgenesis symptoms (e.g. high male recombination)	*	Coyne 1985, 1989
		D. pseudoobscura ×. *D. pseudoobscura bogotana*	No hybrid dysgenesis symptoms	*	Coyne 1986
		D. simulans × *D. sechellia*	No hybrid dysgenesis symptoms	*	Coyne 1986
		D. virilis × *D. lummei*	No hybrid dysgenesis symptoms	*	Coyne 1986
		Drosophila algonquin and *D. athabasca* lineages	No hybrid dysgenesis symptoms	*	Hey 1988
	TE reactivation	*Drosophila melanogaster* × *D. simulans*	*P* element hybrid dysgenesis	*	Kidwell *et al*. 1977
		D. koepferae × *D. buzzatii*	Hybrids saw an increase in TE mobilization relative to parent species in at least 28 TEs.	LTR-RT: *Osvaldo* LINE-like: *Helena* Class II TIR: *Galileo*	Labrador et al. 1999; Vela et al. 2014
Other species	no reactivation	*Saccharomyces cerevisiae* × *S. uvarum*	No increase in transposition rate in F1 hybrids compared to parentals. Mitochondrial inheritance influences transposition rates between hybrid strains.	LTR-RT: *Ty1*	Smukowski Heil et al. 2021
		Arabidopsis thaliana × *A. lyrata*	2% of ~40,000 elements differentially expressed between parent and F1 hybrid backgrounds	All superfamilies	Göbel et al. 2018
		Review of allopolyploid speciation	Mixed results, some cases show increase, some do not.	-	Parisod et al. 2010
		Helianthus anomalus, H. deserticola, and *H. paradoxus* (hybrid species of *H. annuus* × *H. petiolaris*)	TE transcription in F1 hybrids intermediate to parental levels.	LTR-RT: *Copia* LTR-RT: *Gypsy* Other elements	Renaut et al. 2014
		Helianthus annuus × *H. petiolaris*	LTRs remain transcriptionally active after hybridization, but no indication of copy number increases.	*Ty3*/*gypsy* *Ty1*/*copia* LTR-RTs	Kawakami et al. 2011

Other species	TE reactivation	*Yucca gloriosa* (hybrid species of *Y. aloifolia* × *Y. filamentosa*)	No current TE transcription in hybrids. Lower or intermediate abundances in hybrids relative to parent species.	All present (class I and II)	Heyduk et al. 2021
		Helianthus anomalus, H. deserticola, and *H. paradoxus* (hybrid species of *H. annuus* × *H. petiolaris*)	"Stunning" TE copy number increase (4-6x) in all hybrid taxa. TE insertions center around the pericentromeric regions of the hybrid genomes	*Ty3*/*gypsy*-like LTR-RT	Staton et al. 2009; Ungerer et al. 2006
		Macropus eugenii × *Wallabia bicolor*	TE copy number increase. Attributed to reduced methylation of the genome. Chromosomal changes also noted.	*KERV-1*	O'Neill et al. 1998
		Cottus rhenanus × *Cottus perifretum*	TE copy number increases in 20% of TEs present. TE increases biased by parental genome contribution (higher proportions from the parent with the larger genome size). Found outside of genetic regions, within a few hundred generations of admixture	All present (class I and II)	Dennenmoser et al. 2017, 2019
		Arabidopsis thaliana × *A. arenosa*	Parental effects of TE expression. Paternal TE expression was present in incompatible crosses.	*ATHILA*	Josefsson et al. 2006
		Allopolyploids *Nicotiana arentsii, N. rustica* and *N. tabacum*	TE mobilization after allopolyploid speciation	SINEs: *Au* and *TS* MITE: *Ns1* C*opia* LRT-RTs: *TNt1*, *Tnt2* LTR-RT: *TRIM*	Mhiri et al. 2019
		Rice RILs *RZ1*, *RZ2*, and *RZ35* (Hybrids from cultivar *Matsumae* × *Z. latifolia Griseb*.)	Concomitant TE mobilization via "cut-and-paste". Followed by efficient element repression.	MITEs: *mPing* and *Pong*	Shan et al. 2005
		A. geniculata and *A. triuncialis*; *Ae. cylindrica, Ae. geniculata* and *Ae. Triuncialis*	Novel TE insertions for all LTR-RTs. Symptoms of hybrid inviability or low fertility. Possible parental (maternal) effects.	LTR-RT families: *BARE1, Claudia, Egug, Fatima, Romani, Sabin*; 9 LTR-RTs	Senerchia et al. 2015, 2016

***** Hybrid dysgenesis is a syndrome of distinct germline abnormalities that are a direct consequence of TE mobilization (Kidwell *et al*. 1977). Thus, the presence or absence of dysgenesis symptoms can be used to infer TE activation in intraspecific crosses without the need for DNA sequence data.

TE: Transposable element. LTR-RT: Long Terminal Repeat retrotransposons. LINE: Long Interspersed Nuclear Element. SINE: Short interspersed nuclear element. MITE: Miniature inverted-repeat transposable element. TIR: Two inverted tandem repeats. For descriptions of Class I and Class II elements see section 1.1.2

parental species, there is no significant increase in TE copy numbers, though TEs did exhibit transcriptional activity.

Some broad conclusions can be drawn from the results of the literature search presented in Table 1. First, TE reactivation after hybridization is not systematic (e.g., Coyne 1985, 1986, 1989; Smukowski Heil et al. 2021). Second, phylogenetic distance between the species crossed increases the likelihood of TE reactivation, though it is not a perfect predictor. Third, the observed bursts are usually limited to one or a small number of TE sequences (e.g., Mihiri et al. 2019, Senerchia et al. 2016). Though some examples do show more (e.g., Dennenmoser et al. 2017, 2018 found increases in 20% of TE sequences), fitness costs caused by many TEs randomly inserting in genomes (potentially disrupting many functional regions) mean that we can likely generalize bursts as occurring in a small proportion of TE sequences. Fourth, copy number accumulation is generally slow, usually increasing less than one copy per haploid genome per generation. Fifth, the total copy numbers gained in a burst are likewise relatively small, with often less than 20 copies (Bonnivard and Higuet 1999; Kofler et al. 2015, 2018). These copy numbers remain stable after the burst, with little reduction in the subsequent generations. Sixth, some crosses repeatedly show reactivation in certain elements (e.g. the *p* element in *Drosophila*). Lastly, TE bursts are technically accompanied by an increase in genome size due to TE sequence duplications. These conclusions are general and derived from a limited number of study systems. Variation within and between species genomes, environments, and TE sequence dynamics mean we require much more data to begin to fully understand the characteristics of TE invasions in nature.

Despite extensive research in the laboratory and some field examples (Table 1), much remains unknown about TEs and hybridization beyond the generalizations above. This is especially true in non-model species and in natural hybridization events. The broad conclusions I identified may also be impacted by publication bias towards positive results (i.e., studies showing TE reactivation), model species and laboratory crosses. The repeatability of TE reactivation has only been shown for certain species crosses, meaning the relevance of these findings for natural populations is still unclear. As such, more examples of hybridization, such as the hybrid wood ant system in this thesis, need to be found and characterized in the wild.

1.2 The model system: *Formica* wood ants

In this thesis I use *Formica* red wood ants to investigate the genomic shock hypothesis in several hybrid populations. The two species in this project, *Formica aquilonia* and *F. polyctena*, are polygynous mound building wood ants, having up to hundreds of queens across multiple nests, forming a supercolony. The two species inhabit boreal forests across Europe and Asia, where they have contrasted latitudinal distributions (Martin-Roy et al. 2021) (Fig. 5A). The northern species, *F. aquilonia*, lives across Fennoscandia and the Baltics into northern and central Asia, with isolated populations in the Alps and Scotland. *F. polyctena* has a more southern distribution, ranging from central western Europe to central Asia, and extending north into central Sweden and southern Finland (Stockan and Robinson 2016). The overlapping ranges in southern Finland provide an area of sympatry where the two species hybridize (Beresford et al. 2017; Kulmuni et al. 2010).

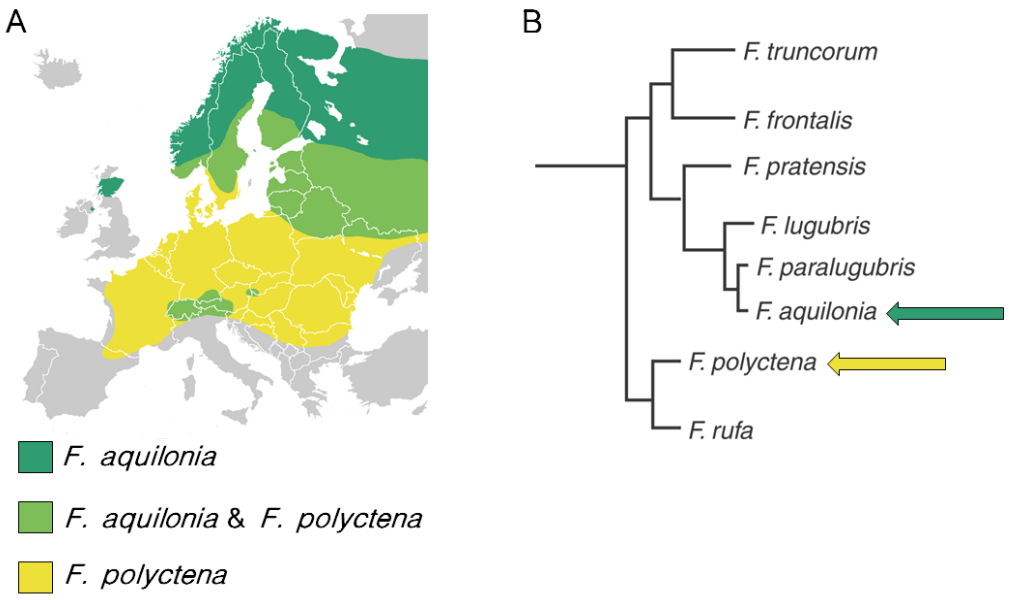


Figure 5. Map of species distributions across Europe (based on Stockan et al. 2016) **(A)**. Cladogram of the *Formica rufa* group with the species in this thesis highlighted (modified from Goropashnaya et al. 2004) **(B)**.

These two species are part of the *Formica rufa* species group (Fig. 5B) which likely diverged within the last 500,000 years (Portinha et al. 2021). Previous work has characterized several hybrid populations between the two species in Southern Finland (Kulmuni et al. 2010, Beresford

et al. 2017) (Fig. 6). In this project, I studied three of those populations, named Pikkala, Bunkkeri, and Långholmen (Fig. 6A).

The reconstruction of the evolutionary history of these populations suggests they arose recently (< 50 generations ago), likely from distinct hybridization events (Nouhaud et al. 2022). The Långholmen population contains two genetically distinct hybrid lineages (W and R) (Kulmuni et al. 2020) which share a unique mitochondrial haplotype and likely originated from the same hybridization event (Nouhaud et al. 2022).

The independent origins of the hybrid populations, along with the knowledge of their species history, and the availability of genomic resources for current populations make this an ideal system for testing the extent and repeatability of TE reactivation in hybrids in nature.

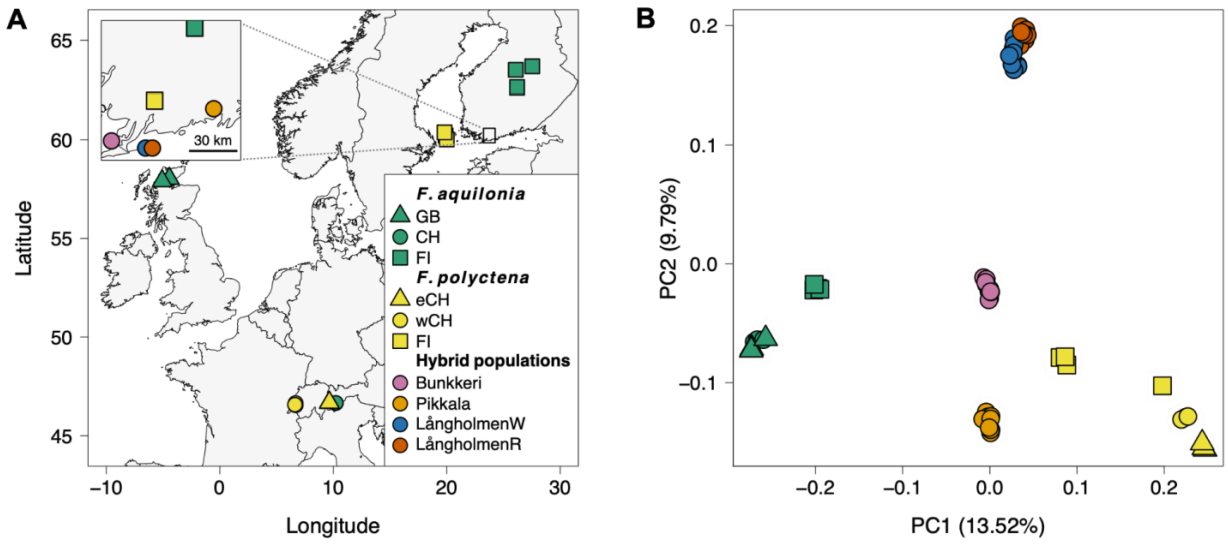


Figure 6. Map of the sample populations in Europe and Southern Finland **(A)**. PCA of female samples using 50,000 SNPs genome-wide, PC axis one shows differentiation between species, with hybrids intermediate to *F. aquilonia* and *F. polyctena* **(B).** These figures are from Nouhaud et al. (2022) and were not created by the author of this thesis.

1.3 Study questions and hypotheses

Following the predictions of the genomic shock hypothesis I outlined two research questions. I use abundance to refer to the quantitative amount of TEs within a genome (i.e. TE copy numbers

or counts), and content to refer to the qualitative aspect of TE identity (e.g. which orders of TEs). I have the following study questions:

Q1. Is there a difference in TE abundance between hybrids and both *F. aquilonia* and *F. polyctena* genomes that would be consistent with a TE burst following hybridization?

This question has three possible outcomes for the data (Fig. 7). The null hypothesis is that there is no effect of hybridization on TE abundance, shown by TE abundance in hybrids equal or intermediate to the TE abundances of *F. aquilonia* and *F. polyctena* (Fig. 7B). Alternative hypothesis one is that TE abundance in hybrids is higher than the average TE abundance of either species. This result would be in line with the predictions of the genomic shock hypothesis (Fig. 7A). Alternative hypothesis two is that TE abundance is below that of either species, indicating strong selection against TE sequences in the hybrid genomes (Fig. 7C). This result is included as a complement to options A and B, but I found no previous studies suggesting if it is likely, or even possible, Table 1.

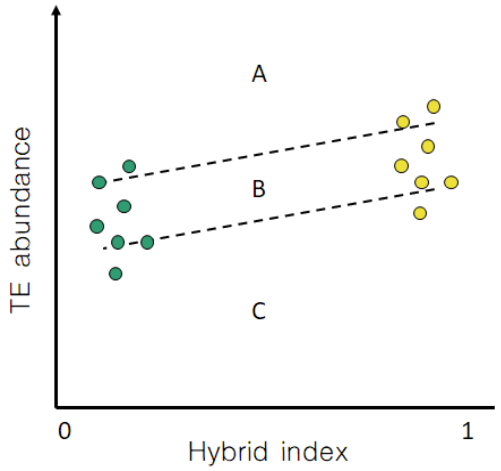


Figure 7. Possible outcomes of question 1. Hybrid abundances in area **(A)** would be consistent with a TE burst. The null expectation is copy numbers intermediate to abundances in the parent species in area **(B)**. Abundances in area **(C)** would result from TE copies being removed from hybrids, this is theoretically and empirically unlikely.

Q2. If there is a TE burst, is it repeated in each hybridization event, and if so, are the same TEs active in each event?

To test repeatability of a TE burst, each hybrid population can be compared to the parent species. If TE reactivation is not repeatable, I expect that a TE burst does not occur in all populations (Fig. 8A). Alternatively, repeatable TE reactivation will result in a burst in all hybrid populations (Fig. 8B).

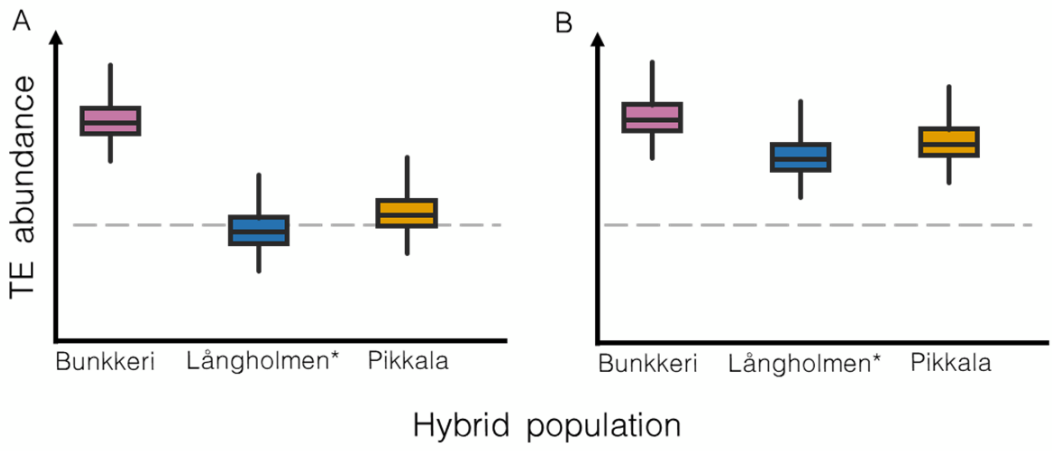


Figure 8. Possible results of the repeatability of TE bursts. The null hypothesis of no relationship between abundances in each population **(A)** A significant association between abundances in each population **(B)**. The dashed line represents the mean abundances of the parent species. * The Långholmen population includes two distinct lineages, W and R, that are represented together for illustrative purposes.

Furthermore, repeatability may be seen not only in TE bursts, but in TE content as well. Do the same orders or sequences increase in every population? The null hypothesis for this question is that there is no relationship between TEs that burst in each population (i.e. random sequences become reactivated). Alternatively, only the same TE sequences may become active in each population, which would indicate some TEs are more prone to reactivation. Finally, do a small or large portion of sequences show evidence of increase? Previous research indicates that one or a few sequences become active, which would be my expectation in cases of reactivation.

2. Materials and Methods

2.1 Data

2.1.1 Individual sampling and whole-genome sequencing

To be able to compare TE abundance and content between the parental species and hybrids I used samples collected from *F. aquilonia*, *F. polyctena*, and three *F. aquilonia* × *F. polyctena* hybrid populations (Table 2). The Långholmen hybrid population has two distinct lineages designated W and R (Kulmuni et al. 2020), both of which were sampled. Overall, the data consist of 99 whole genomes (Table 2).

DNA extractions were done from whole-bodies following a SDS (sodium dodecyl sulfate) protocol and DNA libraries were constructed using NEBNext DNA Library Prep Kits (New England Biolabs). Sequencing and processing was performed by Novogene (Hong Kong) on an Illumina Novaseq 6000 (150 base pairs paired-end reads), targeting 15× per female (diploid), and 8× per male (haploid). Trimming of raw reads and adapter sequence removal were performed using Trimmomatic (v0.38; parameters LEADING:3, TRAILING:3, MINLEN:36; Bolger et al. 2014). The female whole-genome resequencing data was published in (Nouhaud et al. 2022).

Table 2. Species, location, and sex of the samples used in this thesis. All hybrid sample locations are in southern Finland. $N_{total} = 99$.

Species	Location	n (female)	n (male)	Total (f/m)
F. aquilonia	Switzerland	3	-	10 / -
	Scotland	3	-	
	Finland	4	-	
Hybrid	Bunkkeri	10	10	39 / 40
	Långholmen (W)	10	13	
	Långholmen (R)	9	7	
	Pikkala	10	10	
F. Polyctena	Switzerland	6	-	10 / -
	Finland	1	-	
	Åland Islands (Finland)	3	-	

2.1.2 Manual curation of the consensus TE library

The consensus TE sequences used in this thesis were identified *de novo* from the genome assembly of a single hybrid male (Nouhaud et al. 2021) using RepeatModeler2 (Flynn et al. 2020), implemented in the Dfam TE Tools Container (v1.1, https://github.com/Dfam-consortium/TETools). The identified TE sequences were classified (into TE orders and families) and collected into a consensus library for use as a reference to find TEs in other genomes. These steps were performed by Pierre Nouhaud.

As it contains all repeats found in the genome, I curated the TE consensus library to remove non-TE repetitive sequences which were annotated by RepeatModeler2. Of the 1415 TE consensus sequences initially found, ~66% remain unclassified into any group, which is common for new sequencing projects in non-model organisms. Preliminary analyses revealed a large increase of some repeats in hybrids relative to parents (see results). After examining the specific sequences responsible for this increase, I found, through an homology search on NCBI (with BLASTn, Altschul et al. 1990), that the sequences were not TEs but rather DNA satellite sequences. Satellite DNA is a type of repetitive sequence made up of tandem repeats and often associated with DNA structure and centromere function (Garrido-Ramos 2017), but which do not belong to TEs (no self-replication). This prompted a systematic search through all sequences in the TE reference library to find and remove other non-TE sequences. I ran a BLASTn search (Altschul et al. 1990) on all TE sequences in the reference library and found nine which had hits to DNA satellite sequences (one e-value 0.22, all others $\leq 3\times10^{-15}$). I then ran a discontinuous-megablast on those nine to find more dissimilar sequences and make sure that there were in fact satellite DNA sequences. Six out of nine returned 100% satellite DNA hits (ranging from 46 to 289 hits). From this I am confident that those six sequences are satellite DNAs and they were removed from the TE data set (but later analyzed on their own). The other three sequence hits had either 1 or 0 hits to satellite DNA sequences from the dc-megablast. These three were further investigated with TE-Aid (Goubert et al. 2022; https://github.com/clemgoub/TE-Aid.git) and were confirmed to likely be satellite DNA, or at least non-TE, sequences. One sequence identified by RepeatModeler2 as a TE sequence (ltr-*copia*) is likely the internal portion of a longer LTR and confirmed through submitting the sequence to Giri CENSOR, a repeat-sequence specific repository (Bao et al. 2015). Thus, of the

9 potential satellite sequences initially identified, 8 were removed from the analyses and one sequence (ltr-*copia*) was kept.

Further curation was done to remove non-TE sequences that had been classified by RepeatModeler2. These included 2 more satellite DNA sequences, 5 rRNA, and 5 snRNA consensus sequences. In total, my curation removed 19 non-TE sequences from the reference library.

2.2 *de novo* TE annotation with dnaPipeTE

The initial software I used for the project was dnaPipeTE (Goubert et al. 2015), a *de novo* TE search tool that identifies repeated sequences from raw sequence reads without the need for a reference library. This involves assembling sequences from a low coverage sample of the sequence data (<1×). Only repetitive elements will have enough coverage to be fully assembled because of their high abundance compared to other sequences (Goerner-Potvin and Bourque 2018; Clément Goubert et al. 2015). For example, if 0.1× coverage is used, only sequences that appear at least ten times in the data can be assembled. After repetitive elements have been discovered and assembled, they can be identified and annotated with a repository-based homology search. *De novo* TE discovery allows for finding novel TEs as the initial search is not limited by database coverage, but characterization may be limited for unknown or novel repeat sequences.

Since the program cannot handle paired end sequence data, I used only the forward (F) read data from each sample in this step. Moreover, to avoid assembling organelle genomes, reads mapping on mitochondrial DNA and bacterial genomes (such as *Wolbachia*) were removed from the data by Pierre Nouhaud. To do this, reads were mapped against the reference genome and the *Wolbachia* genome (Nouhaud *et al.*, 2021) with BWA MEM using the default parameters (v0.7.17; Le and Durbin 2011) and Picard Tools with default parameters (v2.21.4; http://broadinstitute.github.io/picard) was used to remove duplicates. Only reads mapping against the nuclear ant genome were then extracted, converted back to FASTQ files and used for subsequent analyses with DnaPipeTE.

DnaPipeTE takes two parameters as input, genome coverage and sample number. Genome coverage sets the sample coverage used, e.g. 0.1 for 0.1× coverage per sample, and sample number is the number of times forward reads were chosen to be resampled from the total read pool (genome coverage = 0.5× and sample number = 2 means two samples of 0.5× are taken). I tested different parameter combinations in order to identify an optimal combination that balanced efficiency of computation speed and resources, and reliability. One ant genome was used to test a combination of parameters (coverage ranging from 0.01× to 0.4× and samples ranging from 1-4). The results of this testing found that with the same sample, dnaPipeTE identified highly variable numbers of TE sequences (Fig. 9). The cause of this variability was that there were high levels of ant satellite sequences present, which are not well annotated in the databases used. Because of this, the assignment of the low-coverage samples to specific TE families or satellite types was inconsistent between runs of the program. This resulted in inter-run variance which was roughly equal to inter-sample variance, making interpretation of the dnaPipeTE results unreliable (Fig. 9). As I could not find a set of parameter values where results would consistently converge, the dnaPipeTE analysis was abandoned.

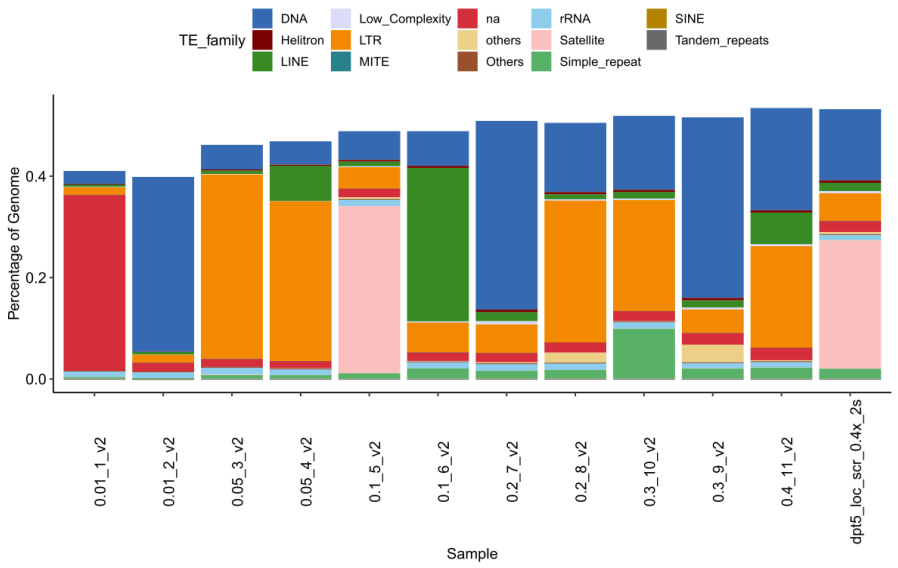


Figure 9. Parameter testing runs of dnaPipeTE on a single genome (the same individual was used in all runs). Genome coverage ran from 0.01x to 0.4x. The number of samples for all runs was 2. While the total percentage of the genome identified as repeats (40-50%) is similar between runs, the assignment of the repeats varied considerably between runs. This was determined to be due to high numbers of ant satellite sequences present in the data that are not in databases used for annotation.

2.3 Guided TE annotation with deviaTE

After dnaPipeTE, analysis moved forward with another TE quantification tool: deviaTE (v0.3.8; Weilguny and Kofler 2019), which requires a database of consensus sequences for the homology search. Here I used my curated hybrid TE library. Because only TE sequences from the library are used, DeviaTE can handle raw read data as input. After finding matching consensus sequences in the read data, absolute TE copy numbers are quantified using read depth across the sequence (Fig. 10) normalized to single copy genes (for which the number of copies is known, i.e. two in a diploid genome). For single copy genes I used 10 BUSCO genes (Benchmarking Universal Single-Copy Orthologs) located on different ant chromosomes. BUSCO genes are highly conserved sequences that are expected to be found in all genomes of a given group (here, Hymenoptera), and in single copy (Simão et al. 2015; Waterhouse et al. 2013). DeviaTE gives the copy number per haploid genome for each TE sequence, hence it controls for ploidy variations (an important consideration as male ants are haploid and females are diploid).

The minimum read length and minimum alignment length were both set to 30 to filter out small and low quality match results. Normalization method was set to --single_copy_genes, using BUSCO genes. All sequences in the reference library were included in each genome search by using the --families ALL switch.

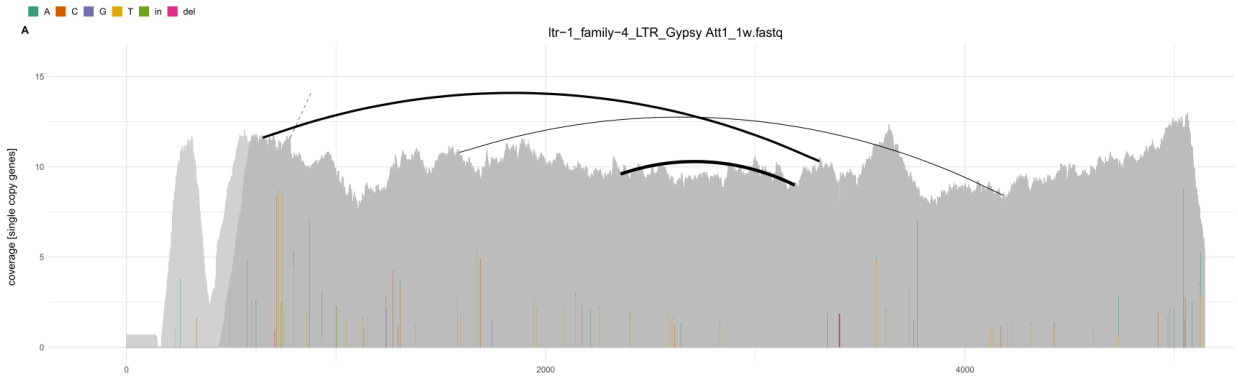


Figure 10. Example sashimi plot from the deviaTE output. The x-axis is the base-pair position in the sequence. The y-axis is the number of times each base appeared in the read data, normalized by the number of BUSCO genes (single copy genes) to show number per haploid. Copy number per haploid is calculated as the average across all bases, in this case around 10.

Copy numbers are calculated by deviaTE using the base and physical coverage (sequences spanned by split reads) for each TE consensus sequence in the individual's raw reads from the FASTQ file. This value is then normalized using the average coverage of the 10 BUSCO sequences. For example, if a sequence has a read depth of 100× and the average depth of the 10 BUSCO sequences is 10× then the sequence has 100/10 = 10 copies in the genome and would appear similar to Figure 10.

2.4 K-mer genome size estimates

TE bursts can drive genome size expansion (Romero-Soriano et al. 2016; Talla et al. 2017; Ungerer et al. 2006), but there may be other causes including heterochromatin (which contains many TEs) and satellite DNA expansion (Craddock et al. 2016). This means that an increase in TE copy numbers may not be due to TE reactivation. The expected genome size in hybrids is the mean size of the parental genomes weighted by their genomic contribution to the hybrid population (Romero-Soriano et al. 2016; Tiersch and Goudie 1993). Following this, if TE copy numbers are only random contributions from each parent, then the expected copy number would be the average of the parental copy numbers weighted for genomic contribution. However, this expectation can be violated by large increases in TEs or satellite DNA. The implication for this thesis is that if there is a genome size increase in the hybrid samples, then it can cause a signal in TE copy numbers that is similar to that expected under the genomic shock hypothesis, but without TE reactivation. I test for these alternative explanations in two ways, 1) by correcting TE abundances by genome size for each sample, and 2) by looking for increases in known satellite DNA sequences in hybrids and their copy numbers relative to TE copy numbers.

Genome size estimates were computed for all female samples (n = 58) except Lai_2w. This sample was excluded due to a data error where the reverse sequence reads were lost. Males were excluded due to model convergence errors in findGSE (Sun et al. 2018), possibly due to fewer amounts of data used for *k*-mer histogram estimation (because of haploidy in males). K-mer estimation uses substrings of length *k* to find the length of a sequence length L. The number of kmers k that fit into L is given by (L - k) + 1 = n. As L becomes large ($> 10^5$), n becomes an accurate estimate of L (Li and Waterman 2003). Bioinformatic tools can use raw sequence reads

to count the number of times kmers of size k appear to compute genome size. I generated k-mer histograms for each female sample with jellyfish 2.2.10 (Marçais and Kingsford 2011). For input to Jellyfish, both forward and reverse reads for each sample were combined. Counts up to 50,000 were kept, the rest of commands were left at the defaults described in https://github.com/gmarcais/Jellyfish/tree/master/doc. K-mer length was set to 15 to help model convergence with high levels of genomic repeats. Inputs for estimated genome sizes were based on assembly sizes of 254, 312, and 280Mb for *F. polyctena*, *F. aquilonia*, and hybrids respectively (based on unpublished genome assembly results). The k-mer histogram outputs were then run through findGSE (Sun et al. 2018; https://github.com/schneebergerlab/findGSE) to calculate genome sizes.

Ten satellite DNA sequences were identified during curation of the TE repeat library. I computed the number of copies in all samples and plotted total satellite copies versus total TE copies per individual for comparison. To find the contribution of satellite DNA to individuals genome size I multiplied the satellite consensus sequence length by the number of copies. This calculation gives the total contribution in base pairs of the identified satellite DNA per individual.

2.5 Statistical analyses

2.5.1 Question 1 data summarization and analysis

To investigate if there is a difference in TE abundance between hybrids and both species (Q1), I calculated the total TE copy number per TE sequence per individual with deviaTE. To quantify total TE abundance per individual, the TE sequence abundances within each individual were summed.

Statistical testing was performed with R 4.1.2 (R Core Team 2021). A linear mixed effects (LMM) model was run using the package nlme (Pinheiro et al. 2022) to compare the total TE copy number T in species x (species i.e. *F. aquilonia*, hybrid, *F. polyctena*). To account for non-independence between individuals samples from the same population, population was included as a random effect z in the model, Eq. 1.

$$T_{ij} = \beta x_{ij} + b_i z_{ij} + \epsilon_{ij} \quad (1)$$

Visual inspection of the residual qq plot confirmed the residuals to be normally distributed. No outliers were identified.

To correct for changes in genome size, I divided the total TE copy number by the genome size for each female sample to get a normalized TE copy number (TE copy number per Mb) per individual. I ran the analysis using corrected TE copy numbers for the female samples using the same LME model as question above for the normalized data. Visual inspection of the residual qq plot confirmed the residuals to be normally distributed and again no outliers were identified.

2.5.2 Question 2 data summarization and analysis

To look for repeatability of TE abundance between hybrid populations I tested for differences in total TE copy numbers between each hybrid population and the *F. aquilonia* individuals. I hypothesized that if TE bursts are repeatable, then all hybrid populations should have higher TE abundances than the mean of the parent species. I used *F. aquilonia* samples as a reference for significance testing because they have a higher mean than *F. polyctena*. Therefore significant results between hybrids and *F. aquilonia* would also be significant for the mean of the parent species.

To see the effect on TE abundance *T* I ran a linear mixed effects model using the nlme package with the parent species and hybrid populations as groups *x*. I included population as a random effect *z* to account for non-independence of samples from the same *F. aquilonia* populations, Eq. 2a.

$$T_{ij} = \beta x_{ij} + b_i z_{ij} + \epsilon_{ij} \quad (2a)$$

To understand the effect of dependence within groups I ran the same linear model without population as a random variable, Eq. 2b.

$$T_{ij} = \beta x_{ij} + \epsilon_{ij} \quad (2b)$$

To understand consistency in TE content I visualized TE abundances 1) by TE order, and 2) by individual TE sequences. This analysis was only qualitative, looking for trends in the data from graphical analysis. For individual TE sequences I generated a heatmap based on TE counts to visualize patterns of consistency within TE sequences (i.e., do the same TE sequences burst in each hybrid population?). For each TE *t* I computed the difference between the abundance *T* of each individual *i* and the median abundance of the parent species *P* (i.e., the median of all *F. aquilonia* and *F. polyctena* samples, which represents the expected abundance in hybrids), Eq. 3a. I then took the median difference for each population *j*, Eq. 3b. Finally I filtered only TEs with differences of biological interest. I defined an arbitrary cutoff of TE sequences with less than or greater than four copies in any group. This cutoff is arbitrary, but was informed by my literature search (see Table 1) where the smallest observed increase in copy numbers during TE invasions was two copies.

$$\Delta T_{ti} = T_{ti} - Med\left[P_t\right] \tag{3a}$$

$$\Delta T_{tj} = Med\left[\Delta T_{ti}\right] \tag{3b}$$

To identify candidate TEs that may have burst after hybridization, I generated a second heatmap. The idea behind this heatmap to remove TEs where copies in a hybrid population could be explained by the TE being present in at least one of the parent groups. Therefore, I filtered out TEs where the median difference in all parent species groups was <2 copies and at least one hybrid population was >4 copies. This guaranteed a minimum difference of 2 copies between hybrid populations and the parent species. The scale of the heatmap is binned from -10 or fewer (blue) to 10 or more copies (red), with TEs in a population that are within 4 copies of the parents are shown in shades of gray. This allows for visualizing whether there are TEs that are consistently higher in all hybrid populations, and this difference is not explained by TE copies contributed by ancestry.

3. Results

3.1 Hybrid genome sizes are intermediate to parental genome sizes, potential expansion likely driven by satellite DNA

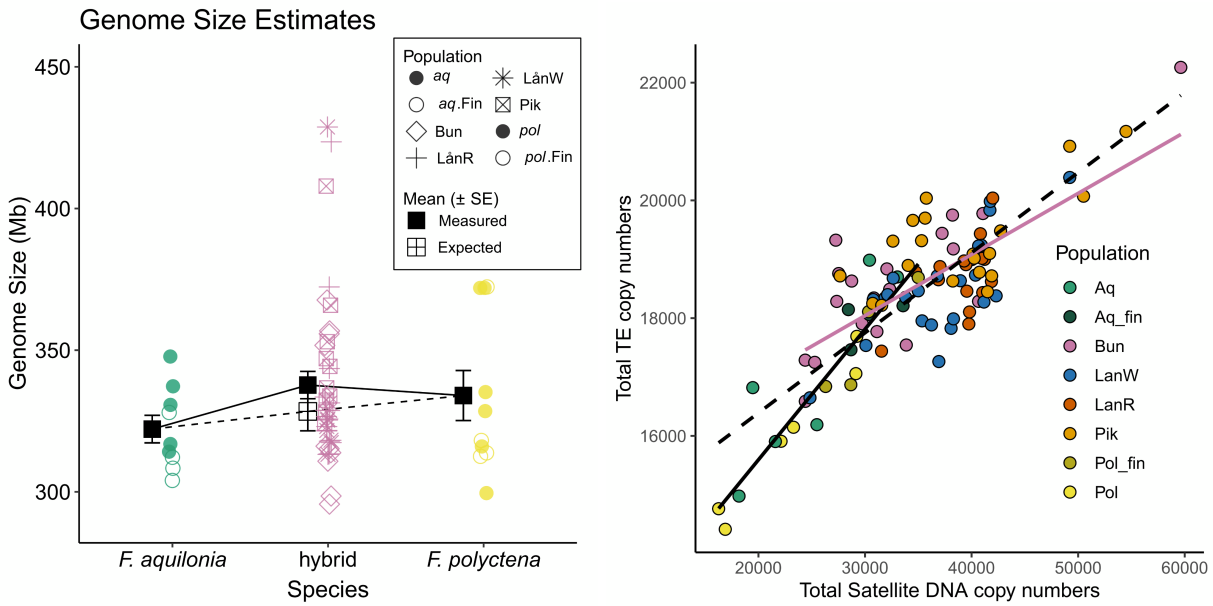


Figure 10. Genome size estimates of female samples (n =58) from k-mers. Median hybrid genome size (328.7 Mb) is equal to the expected average genome size of the two species (328.3 Mb). There are no significant differences between species (F = 1.129, df = 2, p = 0.331) **(A)**. Total copy numbers of satellite DNA and TEs in each sample **(B)**. The black trend line is for parent species samples only, the purple trendline is hybrid samples only, and the dashed trendline is all samples.

The k-mer genome size estimates were run to account for TE copy number increases due to factors other than TE activity. The null expectation for genome size is the average of the two parental genome sizes weighted by genome contribution (Romero-Soriano et al. 2016; Tiersch and Goudie 1993), here 328Mb, Fig 10. There are no significant differences in genomes sizes between hybrids (mean 337Mb) and *F. aquilonia* and *F. polyctena* (F = 1.129, df = 2, p = 0.331). However, satellite DNA does show significantly increased copy numbers in hybrids relative to both parent species (*F. aq* - Hyb, t = -3.71, p = 0.002; *F. pol* - Hyb, t = -4.02, p = .001), and on average contributes an extra 13Mb (S.D 8.8Mb) to hybrids over parents. Additionally, comparing

satellite DNA and TE copy numbers in individual samples does not show a significant bias towards TEs (Fig. 10B) which would be expected if there was reactivation.

3.2 Q1. Hybrids have significantly higher TE abundances compared to the parental species

When comparing hybrid populations to the two parent species, species status had a significant effect on TE abundance (ChiSq = 9.22, df = 2, p = 0.009) while controlling for individuals population of origin. Hybrids showed significantly higher TE abundances compared to both *F. aquilonia* (t = -2.177, df = 16, p = 0.0448) and *F. polyctena* (t = -2.985, df = 16, p = 0.0087) (Fig. 11). This result is consistent with the expected signal from TE activation after hybridization.

However there are two potential causes for an increase in TE abundance, 1) TE activity can cause an increase in copy numbers by sequence duplication, or 2) TE sequence may duplicate as a byproduct of genome size increases, without becoming active themselves. Correction for

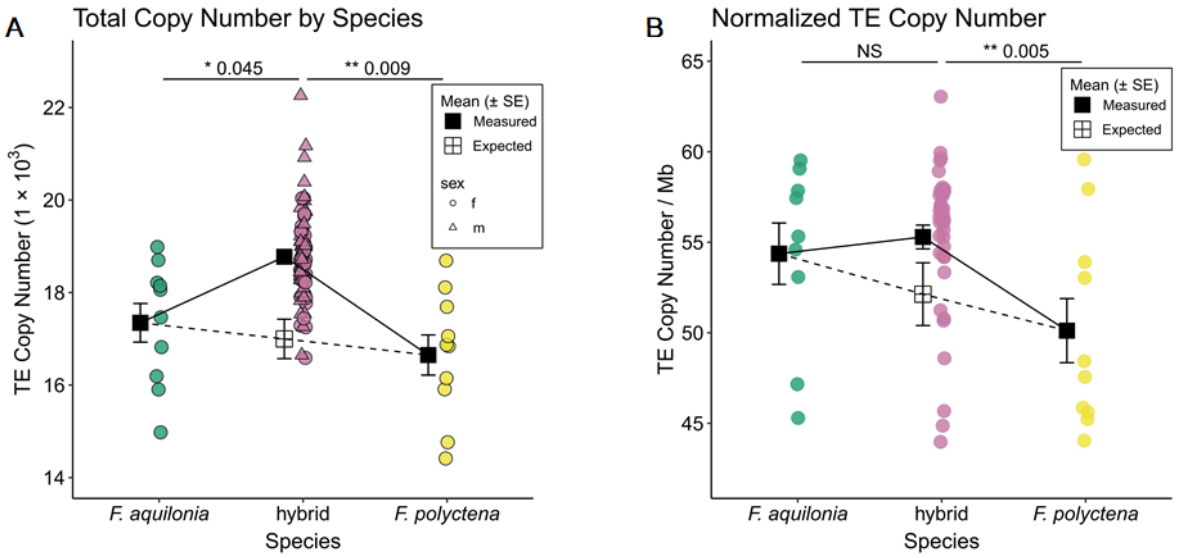


Figure 11. TE abundances by species (n = 99). Hybrid samples have significantly higher TE abundances than both *F. aquilonia* (t = -2.18, df = 16, p = 0.045) and *F. polyctena* (t = -2.99, df = 16, p = 0.009) **(A)**. TE abundances by species corrected for estimated genome sizes (n = 58). Hybrid samples have significantly higher TE abundances than *F. polyctena* (t = -3.22, df = 16, p = 0.005), but not *F. aquilonia* (t = -0.55, df = 16, p = 0.59) **(B)**.

genome size in the female genome samples (males were discarded because their sequencing depth was too low for proper genome size estimation) showed that TE copy numbers in hybrids are not significantly different from *F. aquilonia* (t = -0.551, df = 16, p = 0.589), but remain significantly different from *F. polyctena* (t = -3.221, df = 16, p = 0.005) (Fig. 11 B).

3.3 Distinct TEs may be reactivated in different hybrid populations.

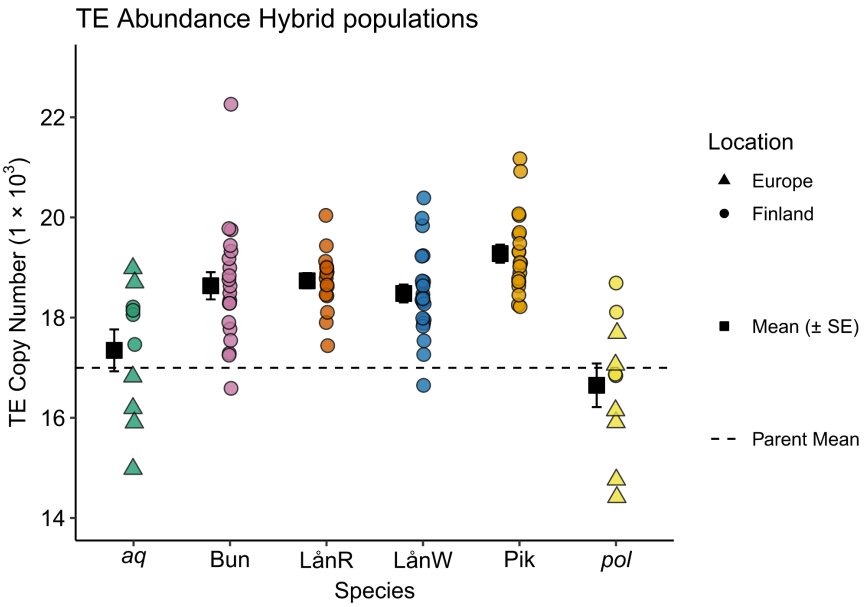


Figure 13. TE abundances in hybrid populations. Square points with bars are mean ± SE. For reference, the dashed line is the mean of all parent samples. All contrasts between each hybrid population and *F. aquilonia* are not significant (t-test, t <= 1.838, p >= 0.109).

To find consistency in the occurrence of TE bursts after hybridization events, I tested for differences in TE abundances between each hybrid population and *F. aquilonia* (a conservative approximation of the mean of the parent species (Fig. 13). Differences between each hybrid population's TE abundance and *F. aquilonia* is expected if all populations experienced a TE burst resulting in increased TE copies. Parent species samples are from populations distributed across Europe, Table 2. This may lead to dependence within parent species groups, i.e. *F. aquilonia* samples from Scotland are more closely related to each other than to *F. aquilonia* samples from Switzerland. Pairwise contrasts with corrections for dependence between populations revealed

Table 3. Contrasts between each hybrid population and the parent species mean (represented by *F. aquilonia*) testing for differences in TE abundance.

Test	Contrast	df	t	p
t-test from lmm with population as random variable	Aq-Bun	7	1.195	0.271
	Aq-LånW	7	1.045	0.331
	Aq-Pik	7	1.838	0.109

no significant differences in TE abundance (Table 3). A lack of statistical significance is likely due to dependence within populations (Supplementary Table 1) and low sample sizes (n = 10 in *F. aquilonia* and n <= 23 in hybrid populations).

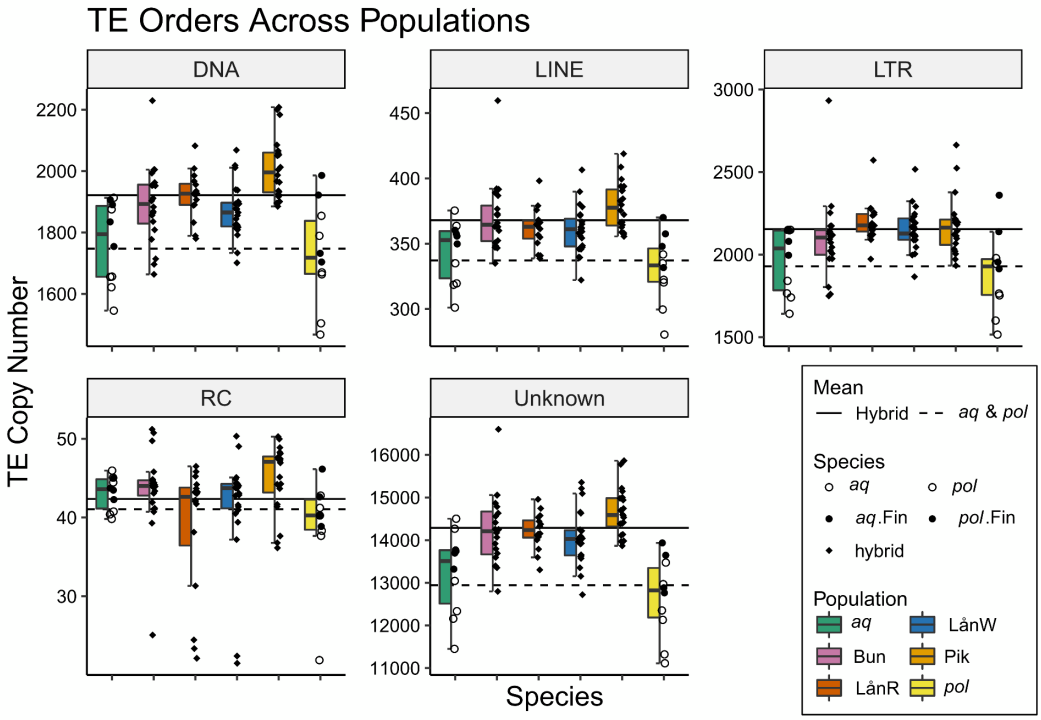


Figure 14. TE abundances in hybrid populations partitioned by TE order. Solid lines in each panel are the hybrid mean and the dashed lines are the mean of *F. aquilonia* and *F. polyctena*.

To find consistency in TE content I qualitatively assessed patterns of TE abundance across TE orders and individual sequences. If there is evidence of a TE burst, is the TE content the same or

different in each hybrid population? Results show consistency in copy number increases across TE orders between hybrid populations: DNA and LINE sequences increase in all hybrid populations, and LTR sequences increase in both Långholmen (W & R) and Pikkala (Fig. 14). There were also increases across hybrid populations in the unknown sequences. These results show that if there was TE reactivation then it was not limited to a single TE order.

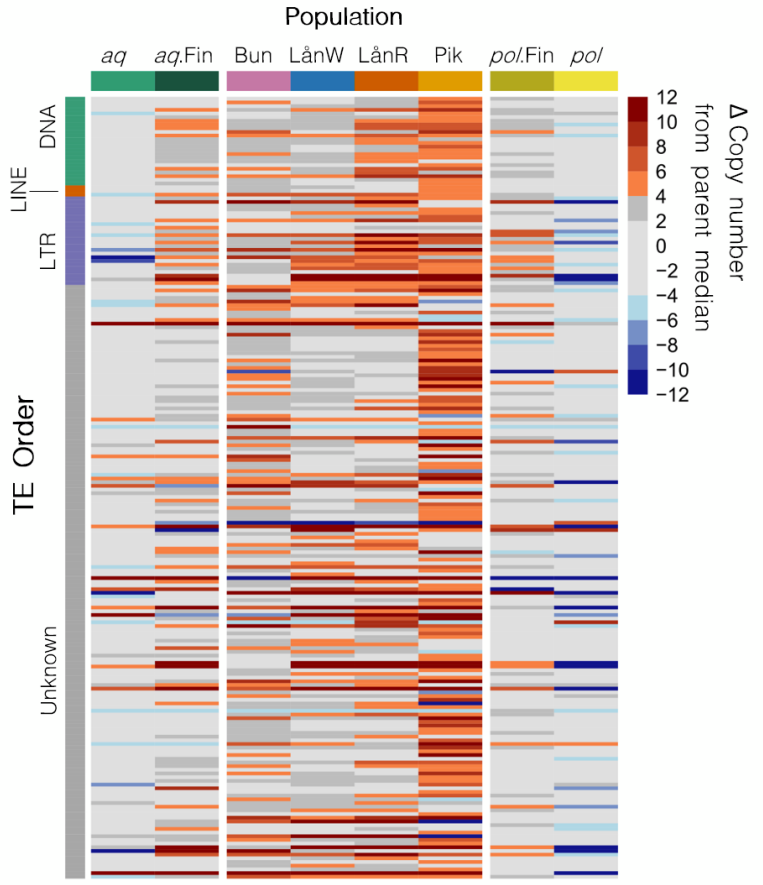


Figure 15. Heatmap of the most significant TE sequences. Each cell the median of the difference from the parent species median for each TE in each population, Eq. 3a,b. Significance is defined as TEs with at least a 4 copy number difference from the median TE abundance of *F. aquilonia* and *F. polyctena* for each hybrid population. The expected pattern in the heatmap if the same TEs burst in each hybrid population is red horizontal bands, with either gray or blue in the parent species. Hybrid populations TE abundances are generally above that of the parent species. Samples of *F. aquilonia* and *F. polyctena* from inside Finland are generally higher in abundances than the European samples for most TEs.

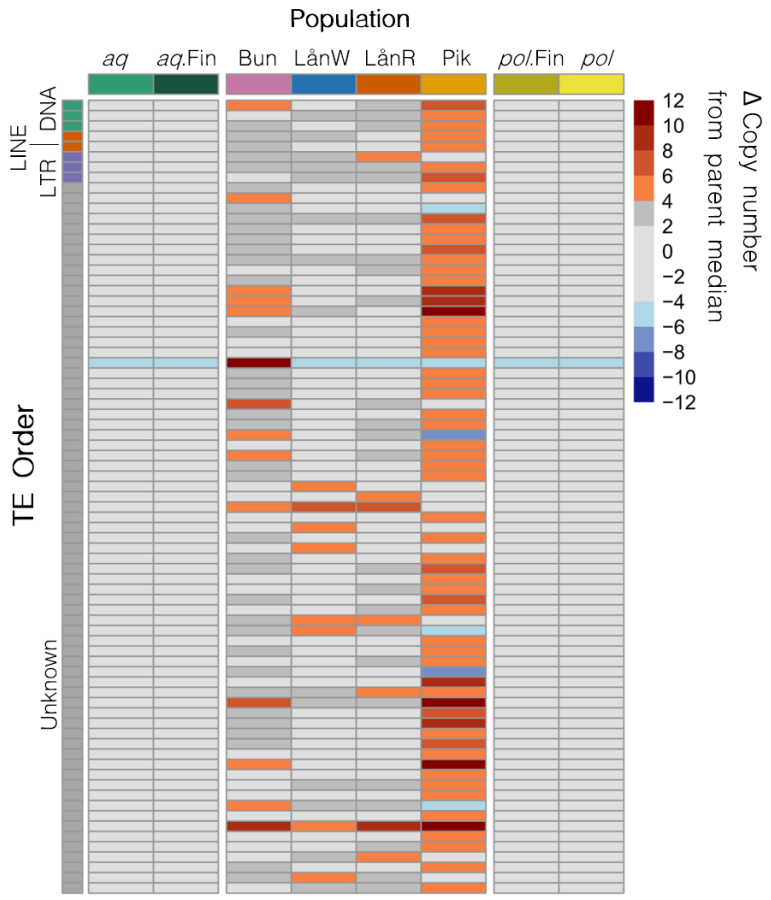


Figure 16. Heatmap with the 77 candidate TEs that may have burst after hybridization. Note that TE profiles are population-specific, with more TEs showing signs of reactivation in Pikkala. The sequence with all parent populations below the parent species mean is due to zero copies present in most parental samples.

To qualitatively address the repeatability of TE reactivation across hybridization events I generated two heatmaps. These allowed me to calculate the number of TEs that potentially reactivated, and visualize whether the same TEs reactivated in each population. The first heatmap reveals that a large number of sequences (212 out of 1386) may have differences of biological interest (± 4 copies) in at least one population (Fig. 15). Of these 212, only two sequences have consistently fewer copies in hybrids than one of the parent groups (one is abundant in Finnish *F. polyctena*, one is abundant in European *F. polyctena*). All other TEs show equal or greater differences in hybrids. This heatmap also shows differences in TE abundances are usually higher in samples from within Finland than from samples from across Europe. To further clarify these results, I filtered copy number differences to find TEs with copy number

differences that are not explained by ancestry, i.e., TEs which are above the expected value (parent species median) only in hybrids, but not the parent species. (Fig. 16). This led to 77 candidate reactivated TEs which have higher copy numbers in hybrid populations than all parent species populations. From these sets it can be seen that 1) there are consistently higher TE abundances in Finnish than in European *F. aquilonia* and *F. polyctana* populations 2) there is little consistency in the same sequences between hybrid populations, and 3) there are consistently high copy numbers in the Pikkala hybrid population.

4. Discussion

Hybridization is a widespread and fundamental evolutionary process that plays a role in adaptation and the establishment of species barriers. Hybridization can lead to previously silenced TEs becoming reactivated, which potentially results in hybrid inviability or sterility. It remains an open question whether or not this mechanism is a frequent part of hybridization, as a majority of research has focused on model organisms.

In this thesis I contribute to filling this knowledge gap by searching for TE activity after natural hybridization events between two wood ant species, *F. aquilonia* and *F. polyctena*. The hybridization events that gave rise to the three populations studied here have limited shared ancestry, and possibly independent origins (Nouhaud et al. 2022). This allowed for the opportunity to not only look for TE activation after hybridization, but to see if it is repeatable.

To test for TE reactivation, I used data on TE copy numbers from 99 whole genome sequences. I tested for the expected signal of TE copy number increases in hybrid populations by comparing TE abundances between hybrid samples and samples from the two parent species (Ungerer et al. 2006). My analysis showed that hybrids have higher TE abundances than either species, consistent with the expected signal from TE reactivation. However, correction for genome size increases showed that hybrid TE abundances were only significantly different from one species, *F. polyctena*, though the mean was still above the expected average of the parent species. I discovered that satellite DNA (non-TE repetitive sequences) also increased in abundance, which could explain increases in hybrid genome sizes. To assess the repeatability of TE reactivation I

performed quantitative analysis of TE abundances between hybrid populations and mean parent species TE abundance. This result was inconclusive, though hybrid populations have higher abundances, the difference is not statistically significant. Finally, qualitative analysis of repeatability revealed that TE reactivation could have occurred after each hybridization event, but candidate reactivated TEs show little or no consistency between hybrid populations, suggesting the TE dynamics are unique to each hybridization event. Below, I will discuss whether my results demonstrate TE reactivation and possible alternative explanations, as well what my results show about the repeatability of TE reactivation after hybridization.

4.1 Increases in TE and satellite DNA copies point to heterochromatin expansion

The pattern of increased total abundances I have detected in hybrids (Figs. 11, 13) is not explained by very high copy numbers in a low number of sequences, but rather small copy number increases across many sequences. Genome size expansion would, one, explain increases in many TE sequences without TE reactivation, two, not have major fitness consequences, and three, show the same pattern that is seen in the data. Such expansion has been previously observed in hybrids in other species (Craddock et al. 2016; Romero-Soriano et al. 2016; Ungerer et al. 2006). I attempted to test this hypothesis by normalizing TE copy numbers by genome size to find TE copies per million bases (copies/Mb). Assuming random distribution of TEs across the genome, genome size increases with no TE activity should result in a proportional increase in TE abundance, so that copies/Mb remains constant. If there is only TE reactivation, or simultaneous TE reactivation and genome size increase then copies/Mb should increase. However, the assumption of random TE distribution does not hold (Supplementary Fig. 1). In this case, increases in heterochromatic regions, which are high in both satellite DNA and TEs, could lead to my observed patterns of TE abundance without TE reactivation. My results showed increased abundances of satellite DNA in hybrids, and no clear bias towards TE copies (Fig. 10B), which supports genome size expansion as a parsimonious explanation. Despite this, my analysis still revealed candidate reactivated TEs (Fig. 16). This suggests that both genome size expansion and TE reactivation may have occurred after hybridization. An additional bioinformatic analysis would be able to disentangle genome expansion with no TE reactivation

from TE reactivation. Characterizing TE insertion polymorphisms would be able to provide detailed information on TE insertion sites. If hybrids would have more novel TE insertion sites compared to the two species, that would provide direct evidence of TE activity in hybrids. The 77 TEs I have identified (Fig. 16) are good candidates for this additional analysis.

If genome size does not fully explain hybrid TE abundances, are there other alternative explanations to TE activity? One possible cause of the observed patterns is that a TE burst occurred in the hybrid populations after hybridization, but is from causes other than hybridization, such as horizontal gene transfer (Peccoud et al. 2017; Silva et al. 2004). I consider this unlikely for multiple reasons. From previous studies, Table 1, we know that bursts are often only tens of generations in duration and difficult to catch while they are happening. The populations in this analysis are likely <50 generations old, meaning that, while not impossible, it's very unlikely that a TE burst unconnected to hybridization has occurred, and in all populations. Another possibility is that gene flow between hybrid populations transferred TEs between them. However, as there is no consistency in the candidate reactivated TEs this would not fit the data. Finally the geographic relationship of the populations, especially considering the very limited dispersal ability of these species (Vitikainen et al. 2015), does not provide an easy route for simultaneous TE invasion in each hybrid population, which would be required for the signal shown in the data.

Genome size increase does not fully explain increased TE abundance in hybrids on its own. Also, it is compatible with TE reactivation occurring as a parallel process. I thus conclude that the signal of higher satellite DNA and TE abundances in hybrids than in the parent species is indicative of genome expansion in heterochromatic regions. This expansion does not exclude TE reactivation, which could have happened concurrently after hybridization.

4.2 Reactivation potentially still occurred in a limited number of TEs

While there is some variation, most previous studies that identified TE reactivation found it happened in a limited number of sequences, Table 1. To answer this question in these hybrid

populations I identified candidate TE sequences (Fig. 15). From this, within the Bunkkeri and Langholmen hybrid populations there are 7-14 potential TEs. This number which potentially burst is thus in-line with increases seen in previous studies (e.g. Vela et al. which found 28 reactivated TEs, Table 1). As stated above, TE insertion polymorphism analysis would provide a definitive answer to the number of TEs that burst in each hybrid population.

Pikkala stands from the other hybrid populations in terms of TE abundances. There, 61 sequences have large differences from the parent species. Given what is known about the fitness effects of TE transposition it seems incredibly unlikely that all of these sequences underwent reactivation. One possible explanation might come from the presence of other members of the *Formica rufa* group in Finland. Specifically *F. polyctena* and *F. rufa* are sister species (Fig. 5B) and difficult to distinguish by genomic analyses such as ADMIXTURE (I. Satokangas, personal communication). This leaves the possibility that the Pikkala population (*F. aquilonia* × *F. polyctena*) have also hybridized with *F. rufa*, resulting in the population actually being [*F. aquilonia* × *F. polyctena*] × *F. rufa* hybrids. While this scenario is not fully supported by the analysis of genomic data (Nouhaud et al. 2022), it could lead to different TE dynamics than in the other two hybrid populations. To examine this I included a single sample, Fis2-1w, which outside analysis has shown to have some *F. rufa* ancestry, as a comparison (supplementary Fig. 2). This sample shows little or no similarity to the Pikkala population. Though this is a limited analysis, it implies that an additional cross is not an explanation for the pattern in Pikkala.

There is a difference in castes sampled (workers or queens) for the hybrid populations. The Bunkkeri population has female workers while both Langholmen and Pikkala females are queens. This means that a difference in caste samples does follow the pattern of higher TE abundances in Pikkala.

If the caste sampled and an alternative species cross does not explain TE abundances in Pikkala, what if there were high levels of TE reactivation? Could there be aspects of the ecology and environment that serve to reduce the impact of TE transposition and thus explain the higher number of potential sequences in Pikkala? There is, however, a high level of similarity between the Pikkala population and the other hybrid populations both ecologically and environmentally,

so further data and analysis would be required to identify any consistent microclimatic differences. Here I briefly speculate about three mechanisms that could in theory reduce the fitness costs of TE transposition and result in higher TE abundances: temperature-dependent TE dynamics, insertion site bias, and hybrid fitness benefits (heterosis). As seen in Kofler et al. (Kofler et al. 2015, 2018), TE transposition rate is highly dependent on temperature. Average temperatures in Salo (~60km from the populations) during the Spring months, when eggs are produced, range from 10-15 °C (Finnish Meteorological Institute n.d.). In theory, this is enough to impact and slow down TE transposition rates, leading to fewer fitness effects per generation. Second, insertion site bias can vary considerably between sequences. An example of this is differences in insertion site preference between some TEs in *Drosophila* leading to either severe (Spradling et al. 1995), or less deleterious effects (Metaxakis et al. 2005). The active TEs may simply have lower fitness costs. Finally, alternative processes outside of TE activity may select for hybrid formation. Benefits to climate adaptation or other causes could be a strong selective force that promotes hybridization or offsets the fitness costs associated with it (Martin-Roy et al. 2021). Whether these mechanisms are possible or relevant in these populations is difficult to consider, and beyond the data available for this thesis.

4.3 Unique TEs may be reactivated across hybrid populations

A second major objective of the thesis was to look at repeatability of evolution. Do we see the same patterns emerge many times? That would suggest predictable, underlying processes are at work. I looked at repeatability in two ways, quantitative and qualitative. Quantitatively my results are inconclusive, likely due to corrections for dependence within parent species samples and low sample sizes, supplementary Table 1. I found no statistically significant difference between each hybrid population and *F. aquilonia* (used as a conservative proxy for the mean of parent species abundance). Therefore, I focus on the qualitative analysis. There, I expected TE sequences from the same TE order to increase in each population, and this was observed. As shown by my literature review, when looking at different species, TE bursts are not the rule. My results broadly suggest that intraspecific bursts may be more repeatable, that is, if there is TE activation in one hybridization event, that it will happen with each successive hybridization

event. However, in regards to repeatability of which TEs burst, there is little or no consistency between hybrid populations.

The clearest answers to which TEs burst in each hybrid population will be provided by TE insertion polymorphism analyses. Should those results confirm that different TEs burst in each population, then what does that imply about the underlying TE reactivation process? Reactivation of TEs occurs when TE suppression mechanisms fail in hybrids through a mismatch of TEs and TE suppressors (Romero-Soriano et al. 2017). Different TEs reactivation in each hybrid population implies then that TE defences fail in a unique way in each hybridization event. This may come about through unique TE and TE defense mis-matches in each hybrid population due to differences in the location of defense sequences along the genome. Alternatively (or additionally), heterochromatin formation around TE insertions correlates with TE activity in hybrids (O'Neill et al. 1998). Alternative de-methylation of heterochromatin resulting from hybridization events could release different TEs in each population and also account for the observed pattern. Investigation of either of these explanations would require additional data beyond the genomic data available for this thesis.

It is currently difficult to make conclusive statements about the repeatability of TE reactivation after hybridization. Despite this, my analyses do suggest that TE bursts likely occur after each hybridization event, but they differ in which TE sequences become active. This limitation could be overcome with future work and new analyses that are more direct observations of TE activity, including TE insertion polymorphisms or even new RNA seq data.

4.4 TE reactivation and past hybridization in Finland

Past demographic and hybridization history of *F. aquilonia* and *F. polyctena* in Finland is an important context for the thesis results. The species split an estimated 500,000 years ago, but in Finland there is evidence of past limited and asymmetric gene flow between the species up to the present, ranging from 0.2 - 3.2 migrants per generation (Portinha et al. 2021). How might this history help to explain my result of Finnish *F. aquilonia* and *F. polyctena* having higher

differences of TE abundances more often than European *F. aquilonia* and *F. polyctena* (Fig. 15)? There is potential for the parental species genomes in Finland to have undergone historical TE bursts. If this is true, and if there are TE bursts after every hybridization event, then I would expect European samples to have the lowest abundances, hybrids the highest, and Finnish samples intermediate. Total TE abundance data does not support this, as Finnish samples are within the range of the European samples, though larger sample sizes would give a much clearer picture (Fig. 13). However, this stepped pattern does seem to appear when visualizing the data as difference from the parent medians (Fig. 16, supplementary Fig. 1). Alternatively, historical heterochromatin expansion after hybridization, which increased TEs without reactivation, would also result in higher TE abundances in parent species from Finland. As in my above analysis, this is a parsimonious explanation. However, as seen in Figure 10A, genome sizes of Finnish parents are not consistently higher than those of European parents, though the variance in estimates is high. A possible explanation for this is that as historical hybrids with higher genome sizes and TE abundances backcrossed with *F. aquilonia* and *F. polyctena*, later generations had decreasing genome sizes on average, yet kept the increased TE copy numbers. In light of my main results, the second option seems more likely, though further analysis of the data is warranted. How representative the ant system is of typical species divergence and secondary contacts is still an open question, and without an answer the generalization of this result is difficult. The wood ant system is still emerging with new data and analyses there are surely even more insights to reveal.

5. Acknowledgments

I would like to thank Pierre Nouhaud for his mentoring in bioinformatics, scientific and life advice, and patience with explanations and review, without all of which I would not have been able to complete this thesis. I would like to thank Jonna Kulmuni for her advice, feedback and unwavering support and encouragement throughout the project. I thank Clement Goubert for his advice on Transposable Elements and support while I presented results. And to the whole SpecIAnt group, thank you for the many discussions and positive research environment that kept me going throughout the project. I would like to thank the CSC for the use of computational

resources and data storage. Finally, I would like to thank the Societas pro Fauna et Flora Fennica for funding a master's thesis grant for this project.

6. References

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., et al. (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, *26*(2), 229–246. https://doi.org/10.1111/j.1420-9101.2012.02599.x

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Arnold, M. L. (1997). *Natural Hybridization and Evolution*. Oxford University Press, USA.

Baltimore, D. (2001). Our genome unveiled. *Nature*, *409*(6822), 815–816. https://doi.org/10.1038/35057267

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*(1), 11. https://doi.org/10.1186/s13100-015-0041-9

Barton, N. H., & Hewitt, G. M. (1985). Analysis of Hybrid Zones. *Annual Review of Ecology and Systematics*, *16*(1), 113–148. https://doi.org/10.1146/annurev.es.16.110185.000553

Bennetzen, J. L., & Wang, H. (2014). The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. *Annual Review of Plant Biology*, *65*(1), 505–530. https://doi.org/10.1146/annurev-arplant-050213-035811

Beresford, J., Elias, M., Pluckrose, L., Sundström, L., Butlin, R. K., Pamilo, P., & Kulmuni, J. (2017). Widespread hybridization within mound-building wood ants in Southern Finland results in cytonuclear mismatches and potential for sex-specific hybrid breakdown. *Molecular Ecology*, *26*(15), 4013–4026. https://doi.org/10.1111/mec.14183

Bingham, P. M., Kidwell, M. G., & Rubin, G. M. (1982). The molecular basis of P-M hybrid dysgenesis: The role of the P element, a P-strain-specific transposon family. *Cell*, *29*(3), 995–1004. https://doi.org/10.1016/0092-8674(82)90463-9

Boeke, J. D., Garfinkel, D. J., Styles, C. A., & Fink, G. R. (1985). Ty elements transpose through an RNA intermediate. *Cell*, *40*(3), 491–500. https://doi.org/10.1016/0092-8674(85)90197-7

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bonnivard & Higuet. (1999). Stability of European natural populations of Drosophila melanogaster with regard to the P–M system: a buffer zone made up of Q populations. *Journal of Evolutionary Biology*, *12*(4), 633–647. https://doi.org/10.1046/j.1420-9101.1999.00063.x

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., et al. (2018). Ten things you should know about transposable elements. *Genome Biology*, *19*(1), 1–12. https://doi.org/10.1186/s13059-018-1577-z

Broecker, F., & Moelling, K. (2019). Evolution of Immune Systems From Viruses and Transposable Elements. *Frontiers in Microbiology*, *10*. https://doi.org/10.3389/fmicb.2019.00051

Brookfield, J. F. Y. (1996). Models of the spread of non-autonomous selfish transposable elements when transposition and fitness are coupled. *Genetics Research*, *67*(3), 199–209. https://doi.org/10.1017/S0016672300033681
Cavaller-Smith, T. (1985). The evolution of genome size. https://www.osti.gov/biblio/6076917. Accessed 1 May 2021
Charlesworth, B., & Charlesworth, D. (1983). The population dynamics of transposable elements. *Genetics Research*, *42*(1), 1–27. https://doi.org/10.1017/S0016672300021455
Coyne, J. A. (1985). Genetic studies of three sibling species of Drosophila with relationship to theories of speciation. *Genetics Research*, *46*(2), 169–192. https://doi.org/10.1017/S0016672300022643
Coyne, J. A. (1986). Meiotic Segregation and Male Recombination in Interspecific Hybrids of Drosophila. *Genetics*, *114*(2), 485–494.
Coyne, J. A. (1989). Mutation rates in hybrids between sibling species of Drosophila. *Heredity*, *63*(2), 155–162. https://doi.org/10.1038/hdy.1989.87
Coyne, J. A., & Orr, H. A. (2004). *Speciation*. *Speciation*. Sunderland: Sinauer Associates, Inc.
Craddock, E. M., Gall, J. G., & Jonas, M. (2016). Hawaiian Drosophila genomes: size variation and evolutionary expansions. *Genetica*, *144*(1), 107–124. https://doi.org/10.1007/s10709-016-9882-5
Dennenmoser, S., Sedlazeck, F. J., Iwaszkiewicz, E., Li, X.-Y., Altmüller, J., & Nolte, A. W. (2017). Copy number increases of transposable elements and protein-coding genes in an invasive fish of hybrid origin. *Molecular Ecology*, *26*(18), 4712–4724. https://doi.org/10.1111/mec.14134
Dennenmoser, S., Sedlazeck, F. J., Schatz, M. C., Altmüller, J., Zytnicki, M., & Nolte, A. W. (2019). Genome-wide patterns of transposon proliferation in an evolutionary young hybrid fish. *Molecular Ecology*, *28*(6), 1491–1505. https://doi.org/10.1111/mec.14969
Eddy, S. R. (2012). The C-value paradox, junk DNA and ENCODE. *Current biology: CB*, *22*(21), R898-899. https://doi.org/10.1016/j.cub.2012.10.002
Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, *9*(5), 397–405. https://doi.org/10.1038/nrg2337
Finnegan, D. J. (1989). Eukaryotic transposable elements and genome evolution. *Trends in Genetics*, *5*, 103–107. https://doi.org/10.1016/0168-9525(89)90039-5
Finnish Meteorological Institute. (n.d.). Climate Statistics from 1961 onwards. *Finnish Meteorological Institute*. https://en.ilmatieteenlaitos.fi. Accessed 19 April 2022
Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, *117*(17), 9451–9457. https://doi.org/10.1073/pnas.1921046117
Fu, A., Jacobs, D. I., & Zhu, Y. (2014). Epigenome-wide analysis of piRNAs in gene-specific DNA methylation. *RNA Biology*, *11*(10), 1301–1312. https://doi.org/10.1080/15476286.2014.996091
Garrido-Ramos, M. A. (2017). Satellite DNA: An Evolving Topic. *Genes*, *8*(9), 230. https://doi.org/10.3390/genes8090230
Göbel, U., Arce, A. L., He, F., Rico, A., Schmitz, G., & de Meaux, J. (2018). Robustness of Transposable Element Regulation but No Genomic Shock Observed in Interspecific Arabidopsis Hybrids. *Genome Biology and Evolution*, *10*(6), 1403–1415. https://doi.org/10.1093/gbe/evy095

Goerner-Potvin, P., & Bourque, G. (2018). Computational tools to unmask transposable elements. *Nature Reviews Genetics*, *19*(11), 688–704. https://doi.org/10.1038/s41576-018-0050-x

González, J., & Petrov, D. A. (2009). The Adaptive Role of Transposable Elements in the Drosophila Genome. *Gene*, *448*(2), 124–133. https://doi.org/10.1016/j.gene.2009.06.008

Goropashnaya, A. V., Fedorov, V. B., & Pamilo, P. (2004). Recent speciation in the Formica rufa group ants (Hymenoptera, Formicidae): inference from mitochondrial DNA phylogeny. *Molecular Phylogenetics and Evolution*, *32*(1), 198–206. https://doi.org/10.1016/j.ympev.2003.11.016

Goubert, Clement, Craig, R. J., Bilat, A. F., Peona, V., Vogan, A. A., & Protasio, A. V. (2022). A beginner's guide to manual curation of transposable elements. *Mobile DNA*, *13*(1), 7. https://doi.org/10.1186/s13100-021-00259-7

Goubert, Clément, Modolo, L., Vieira, C., ValienteMoro, C., Mavingui, P., & Boulesteix, M. (2015). De Novo Assembly and Annotation of the Asian Tiger Mosquito (Aedes albopictus) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (Aedes aegypti). *Genome Biology and Evolution*, *7*(4), 1192–1205. https://doi.org/10.1093/gbe/evv050

Hancks, D. C., & Kazazian, H. H. (2016). Roles for retrotransposon insertions in human disease. *Mobile DNA*, *7*(1), 9. https://doi.org/10.1186/s13100-016-0065-9

Hey, J. (1988). Speciation via hybrid dysgenesis: negative evidence from the Drosophila affinis subgroup. *Genetica*, *78*(2), 97–103. https://doi.org/10.1007/BF00058840

Heyduk, K., McAssey, E. V., Grimwood, J., Shu, S., Schmutz, J., McKain, M. R., & Leebens-Mack, J. (2021). Hybridization History and Repetitive Element Content in the Genome of a Homoploid Hybrid, Yucca gloriosa (Asparagaceae). *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/fpls.2020.573767

Huang, S., Tao, X., Yuan, S., Zhang, Y., Li, P., Beilinson, H. A., et al. (2016). Discovery of an Active RAG Transposon Illuminates the Origins of V(D)J Recombination. *Cell*, *166*(1), 102–114. https://doi.org/10.1016/j.cell.2016.05.032

Jacobs, F. M., Greenberg, D., Nguyen, N., Haeussler, M., Ewing, A. D., Katzman, S., et al. (2014). An evolutionary arms race between KRAB zinc finger genes 91/93 and SVA/L1 retrotransposons. *Nature*, *516*(7530), 242–245. https://doi.org/10.1038/nature13760

Josefsson, C., Dilkes, B., & Comai, L. (2006). Parent-Dependent Loss of Gene Silencing during Interspecies Hybridization. *Current Biology*, *16*(13), 1322–1328. https://doi.org/10.1016/j.cub.2006.05.045

Kapitonov, V. V., & Koonin, E. V. (2015). Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. *Biology Direct*, *10*(1), 20. https://doi.org/10.1186/s13062-015-0055-8

Kawakami, T., Dhakal, P., Katterhenry, A. N., Heatherington, C. A., & Ungerer, M. C. (2011). Transposable Element Proliferation and Genome Expansion Are Rare in Contemporary Sunflower Hybrid Populations Despite Widespread Transcriptional Activity of LTR Retrotransposons. *Genome Biology and Evolution*, *3*, 156–167. https://doi.org/10.1093/gbe/evr005

Kelleher, E. S., Barbash, D. A., & Blumenstiel, J. P. (2020). Taming the Turmoil Within: New Insights on the Containment of Transposable Elements. *Trends in Genetics*, *36*(7), 474–489. https://doi.org/10.1016/j.tig.2020.04.007

Kidwell, M. G., Kidwell, J. F., & Sved, J. A. (1977). Hybrid Dysgenesis in DROSOPHILA

MELANOGASTER: A Syndrome of Aberrant Traits Including Mutation, Sterility and Male Recombination. *Genetics*, *86*(4), 813–833.
Kidwell, Margaret G. (1983). Evolution of hybrid dysgenesis determinants in Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, *80*(6), 1655–1659. https://doi.org/10.1073/pnas.80.6.1655
Kleckner, N. (1981). Transposable Elements in Prokaryotes. *Annual Review of Genetics*, *15*(1), 341–404. https://doi.org/10.1146/annurev.ge.15.120181.002013
Kofler, R., Hill, T., Nolte, V., Betancourt, A. J., & Schlötterer, C. (2015). The recent invasion of natural Drosophila simulans populations by the P-element. *Proceedings of the National Academy of Sciences*, *112*(21), 6659–6663. https://doi.org/10.1073/pnas.1500758112
Kofler, R., Senti, K.-A., Nolte, V., Tobler, R., & Schlötterer, C. (2018). Molecular dissection of a natural transposable element invasion. *Genome Research*, *28*(6), 824–835. https://doi.org/10.1101/gr.228627.117
Kulmuni, J., Nouhaud, P., Pluckrose, L., Satokangas, I., Dhaygude, K., & Butlin, R. K. (2020). Instability of natural selection at candidate barrier loci underlying speciation in wood ants. *Molecular Ecology*, *29*(20), 3988–3999. https://doi.org/10.1111/mec.15606
Kulmuni, J., Seifert, B., & Pamilo, P. (2010). Segregation distortion causes large-scale differences between male and female genomes in hybrid ants. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(16), 7371–7376. https://doi.org/10.1073/pnas.0912409107
Labrador, M., Farré, M., Utzet, F., & Fontdevila, A. (1999). Interspecific hybridization increases transposition rates of Osvaldo. *Molecular Biology and Evolution*, *16*(7), 931–937. https://doi.org/10.1093/oxfordjournals.molbev.a026182
Le, S. Q., & Durbin, R. (2011). SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Research*, *21*(6), 952–960. https://doi.org/10.1101/gr.113084.110
Lee, S.-I., & Kim, N.-S. (2014). Transposable Elements and Genome Size Variations in Plants. *Genomics & Informatics*, *12*(3), 87–97. https://doi.org/10.5808/GI.2014.12.3.87
Li, X., & Waterman, M. S. (2003). Estimating the Repeat Structure and Length of DNA Sequences Using ℓ-Tuples. *Genome Research*, *13*(8), 1916–1922. https://doi.org/10.1101/gr.1251803
Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, *20*(5), 229–237. https://doi.org/10.1016/j.tree.2005.02.010
Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770. https://doi.org/10.1093/bioinformatics/btr011
Martin-Roy, R., Nygård, E., Nouhaud, P., & Kulmuni, J. (2021). Differences in Thermal Tolerance between Parental Species Could Fuel Thermal Adaptation in Hybrid Wood Ants. *The American Naturalist*, *198*(2), 278–294. https://doi.org/10.1086/715012
McClintock, B. (1984). The significance of responses of the genome to challenge. *Science*, *226*(4676), 792–801. https://doi.org/10.1126/science.15739260
McClintock, Barbara. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, *36*(6), 344–355. https://doi.org/10.1073/pnas.36.6.344
Metaxakis, A., Oehler, S., Klinakis, A., & Savakis, C. (2005). Minos as a Genetic and Genomic Tool in Drosophila melanogaster. *Genetics*, *171*(2), 571–581. https://doi.org/10.1534/genetics.105.041848

Mhiri, C., Parisod, C., Daniel, J., Petit, M., Lim, K. Y., Borne, F. D. de, et al. (2019). Parental transposable element loads influence their dynamics in young Nicotiana hybrids and allotetraploids. *New Phytologist*, *221*(3), 1619–1633. https://doi.org/10.1111/nph.15484

Moyle, L. C., & Nakazato, T. (2010). Hybrid Incompatibility "Snowballs" Between Solanum Species. *Science*, *329*(5998), 1521–1523. https://doi.org/10.1126/science.1193063

Nouhaud, P., Beresford, J., & Kulmuni, J. (2021). Cost-effective long-read assembly of a hybrid Formica aquilonia × Formica polyctena wood ant genome from a single haploid individual. *bioRxiv*. https://doi.org/10.1101/2021.03.09.434597

Nouhaud, P., Martin, S. H., Portinha, B., Sousa, V. C., & Kulmuni, J. (2022, January 18). Rapid and repeatable genome evolution across three hybrid ant populations. bioRxiv. https://doi.org/10.1101/2022.01.16.476493

O'Neill, R. J., O'Neill, M. J., & Graves, J. A. (1998). Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature*, *393*(6680), 68–72.

Pagel, M., & Johnstone, R. A. (1992). Variation across species in the size of the nuclear genome supports the junk-DNA explanation for the C-value paradox. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *249*(1325), 119–124. https://doi.org/10.1098/rspb.1992.0093

Pardue, M.-L., & DeBaryshe, P. G. (2011). Retrotransposons that maintain chromosome ends. *Proceedings of the National Academy of Sciences*, *108*(51), 20317–20324. https://doi.org/10.1073/pnas.1100278108

Parisod, C., Alix, K., Just, J., Petit, M., Sarilar, V., Mhiri, C., et al. (2010). Impact of transposable elements on the organization and function of allopolyploid genomes. *New Phytologist*, *186*(1), 37–45. https://doi.org/10.1111/j.1469-8137.2009.03096.x

Payer, L. M., & Burns, K. H. (2019). Transposable elements in human genetic disease. *Nature Reviews Genetics*, *20*(12), 760–772. https://doi.org/10.1038/s41576-019-0165-8

Peccoud, J., Loiseau, V., Cordaux, R., & Gilbert, C. (2017). Massive horizontal transfer of transposable elements in insects. *Proceedings of the National Academy of Sciences*, *114*(18), 4721–4726. https://doi.org/10.1073/pnas.1621178114

Pinheiro, J., Bates, D., & R Core Team. (2022). *nlme: Linear and Nonlinear Mixed Effects Models*. https://CRAN.R-project.org/package=nlme

Portinha, B., Avril, A., Bernasconi, C., Helanterä, H., Monaghan, J., Seifert, B., et al. (2021). Whole-genome analysis of multiple wood ant population pairs supports similar speciation histories, but different degrees of gene flow, across their European range. *bioRxiv*. https://doi.org/10.1101/2021.03.10.434741

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Renaut, S., Rowe, H. C., Ungerer, M. C., & Rieseberg, L. H. (2014). Genomics of homoploid hybrid speciation: diversity and transcriptional activity of long terminal repeat retrotransposons in hybrid sunflowers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1648), 20130345. https://doi.org/10.1098/rstb.2013.0345

Romero-Soriano, V., Burlet, N., Vela, D., Fontdevila, A., Vieira, C., & García Guerreiro, M. P. (2016). Drosophila Females Undergo Genome Expansion after Interspecific Hybridization. *Genome Biology and Evolution*, *8*(3), 556–561. https://doi.org/10.1093/gbe/evw024

Romero-Soriano, V., Modolo, L., Lopez-Maestre, H., Mugat, B., Pessia, E., Chambeyron, S., et

al. (2017). Transposable Element Misregulation Is Linked to the Divergence between Parental piRNA Pathways in Drosophila Hybrids. *Genome Biology and Evolution*, *9*(6), 1450–1470. https://doi.org/10.1093/gbe/evx091

Schrader, L., & Schmitz, J. (2019). The impact of transposable elements in adaptive evolution. *Molecular Ecology*, *28*(6), 1537–1549. https://doi.org/10.1111/mec.14794

Seehausen, O. (2006). Conservation: Losing Biodiversity by Reverse Speciation. *Current Biology*, *16*(9), R334–R337. https://doi.org/10.1016/j.cub.2006.03.080

Senerchia, N., Felber, F., North, B., Sarr, A., Guadagnuolo, R., & Parisod, C. (2016). Differential introgression and reorganization of retrotransposons in hybrid zones between wild wheats. *Molecular Ecology*, *25*(11), 2518–2528. https://doi.org/10.1111/mec.13515

Senerchia, N., Felber, F., & Parisod, C. (2015). Genome reorganization in F1 hybrids uncovers the role of retrotransposons in reproductive isolation. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1804), 20142874. https://doi.org/10.1098/rspb.2014.2874

Serrato-Capuchina, A., & Matute, D. (2018). The Role of Transposable Elements in Speciation. *Genes*, *9*(5), 254. https://doi.org/10.3390/genes9050254

Shan, X., Liu, Z., Dong, Z., Wang, Y., Chen, Y., Lin, X., et al. (2005). Mobilization of the Active MITE Transposons mPing and Pong in Rice by Introgression from Wild Rice (Zizania latifolia Griseb.). *Molecular Biology and Evolution*, *22*(4), 976–990. https://doi.org/10.1093/molbev/msi082

Silva, J. C., Loreto, E. L., & Clark, J. B. (2004). Factors That Affect the Horizontal Transfer of Transposable Elements. *Current Issues in Molecular Biology*, *6*(1), 57–72. https://doi.org/10.21775/cimb.006.057

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Slotkin, R. K., & Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, *8*(4), 272–285. https://doi.org/10.1038/nrg2072

Smukowski Heil, C., Patterson, K., Hickey, A. S.-M., Alcantara, E., & Dunham, M. J. (2021). Transposable Element Mobilization in Interspecific Yeast Hybrids. *Genome Biology and Evolution*, *13*(3), evab033. https://doi.org/10.1093/gbe/evab033

Spradling, A. C., Stern, D. M., Kiss, I., Roote, J., Laverty, T., & Rubin, G. M. (1995). Gene disruptions using P transposable elements: an integral component of the Drosophila genome project. *Proceedings of the National Academy of Sciences*, *92*(24), 10824–10830. https://doi.org/10.1073/pnas.92.24.10824

Staton, S. E., Ungerer, M. C., & Moore, R. C. (2009). The genomic organization of Ty3/gypsy-like retrotransposons in Helianthus (Asteraceae) homoploid hybrid species. *American Journal of Botany*, *96*(9), 1646–1655. https://doi.org/10.3732/ajb.0800337

Stockan, J. A., & Robinson, E. J. H. (2016). *Wood Ant Ecology and Conservation*. Cambridge University Press.

Sun, H., Ding, J., Piednoël, M., & Schneeberger, K. (2018). findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics*, *34*(4), 550–557. https://doi.org/10.1093/bioinformatics/btx637

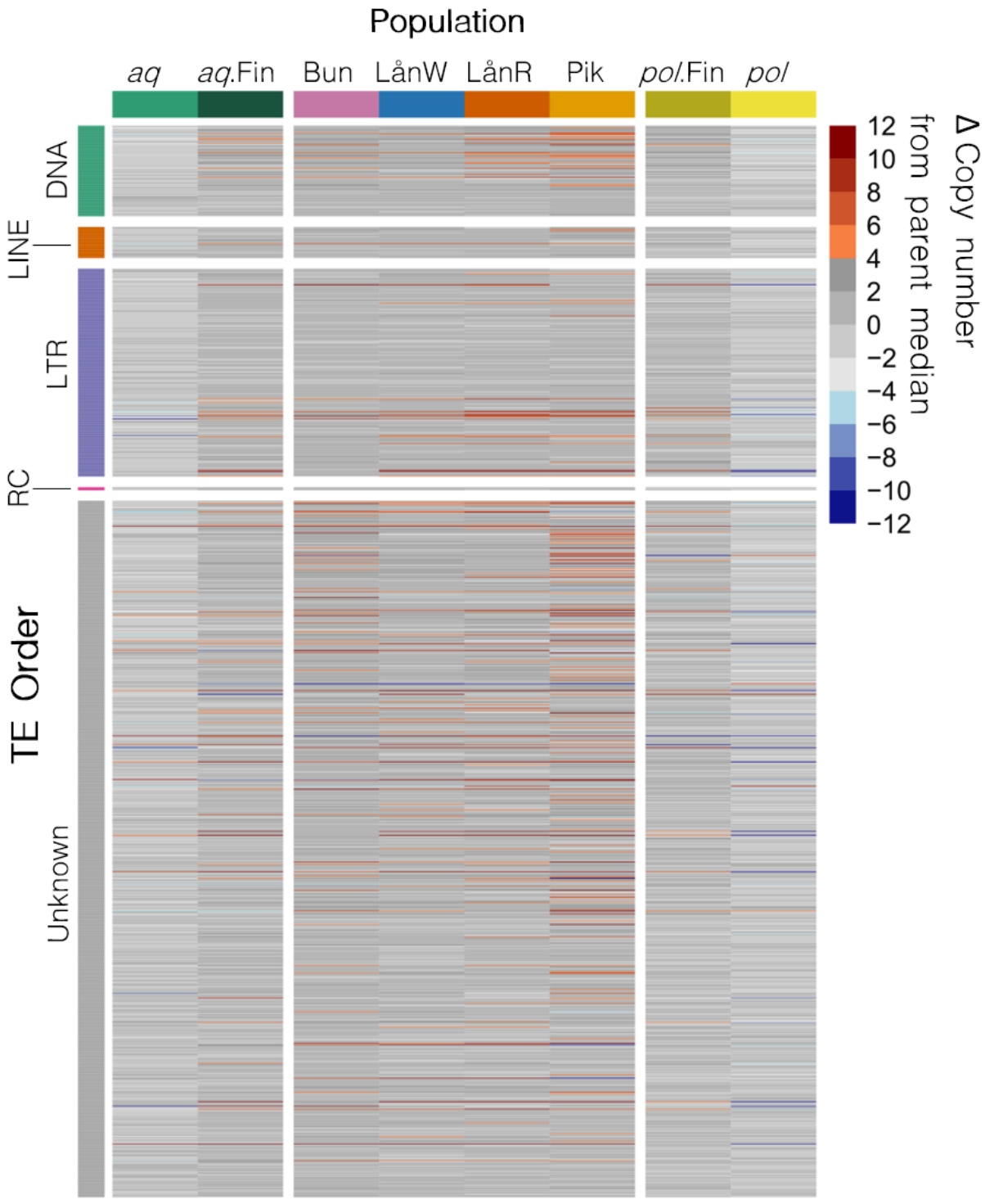
Talla, V., Suh, A., Kalsoom, F., Dinca, V., Vila, R., Friberg, M., et al. (2017). Rapid increase in genome size as a consequence of transposable element hyperactivity in wood-white (leptidea) butterflies. *Genome Biology and Evolution*, *9*(10), 2491–2505.

https://doi.org/10.1093/gbe/evx163
Taylor, S. A., & Larson, E. L. (2019). Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nature Ecology & Evolution*, *3*(2), 170–177. https://doi.org/10.1038/s41559-018-0777-y
Tiersch, T. R., & Goudie, C. A. (1993). Inheritance and Variation of Genome Size in Half-Sib Families of Hybrid Catfishes. *Journal of Heredity*, *84*(2), 122–125. https://doi.org/10.1093/oxfordjournals.jhered.a111292
Ungerer, M. C., Strakosh, S. C., & Zhen, Y. (2006). Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. *Current Biology*, *16*(20), R872–R873. https://doi.org/10.1016/j.cub.2006.09.020
Vela, D., Fontdevila, A., Vieira, C., & Pilar García Guerreiro, M. (2014). A genome-wide survey of genetic instability by transposition in Drosophila hybrids. *PLoS ONE*, *9*(2). https://doi.org/10.1371/journal.pone.0088992
Vitikainen, E. I. K., Haag-Liautard, C., & Sundström, L. (2015). Natal Dispersal, Mating Patterns, and Inbreeding in the Ant Formica exsecta. *The American Naturalist*, *186*(6), 716–727. https://doi.org/10.1086/683799
Waterhouse, R. M., Tegenfeldt, F., Li, J., Zdobnov, E. M., & Kriventseva, E. V. (2013). OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Research*, *41*(D1), D358–D365. https://doi.org/10.1093/nar/gks1116
Weilguny, L., & Kofler, R. (2019). DeviaTE: Assembly-free analysis and visualization of mobile genetic element composition. *Molecular Ecology Resources*, *19*(5), 1346–1354. https://doi.org/10.1111/1755-0998.13030
Wells, J. N., & Feschotte, C. (2020). A Field Guide to Eukaryotic Transposable Elements. *Annual Review of Genetics*, *54*, 539–561. https://doi.org/10.1146/annurev-genet-040620-022145
Wood, J. G., Jones, B. C., Jiang, N., Chang, C., Hosier, S., Wickremesinghe, P., et al. (2016). Chromatin-modifying genetic interventions suppress age-associated transposable element activation and extend life span in Drosophila. *Proceedings of the National Academy of Sciences*, *113*(40), 11277–11282. https://doi.org/10.1073/pnas.1604621113
Yang, P., Wang, Y., & Macfarlan, T. S. (2017). The role of KRAB-ZFPs in transposable element repression and mammalian evolution. *Trends in genetics : TIG*, *33*(11), 871–881. https://doi.org/10.1016/j.tig.2017.08.006

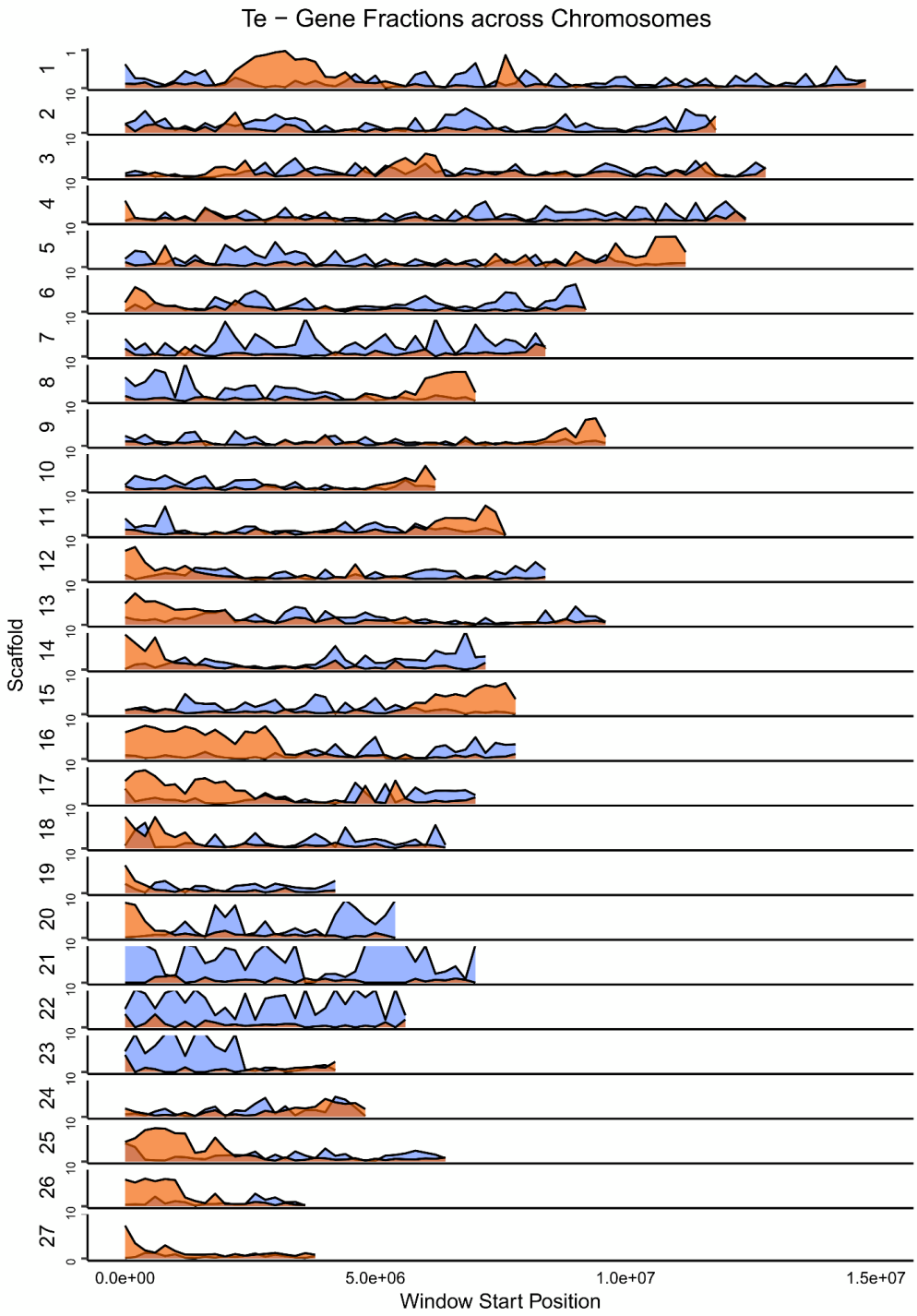
7. Appendix

Supplementary Table 1. Understanding the effect of population as a random variable in the linear mixed model testing for differences in TE abundance between the hybrid populations and the parent species mean (represented by *F. aquilonia*). The top of the table are test statistics based on the linear mixed model which included population as a random variable to control for dependence within the *F. aquilonia* samples. The bottom of the table is a linear model without population as a random variable.

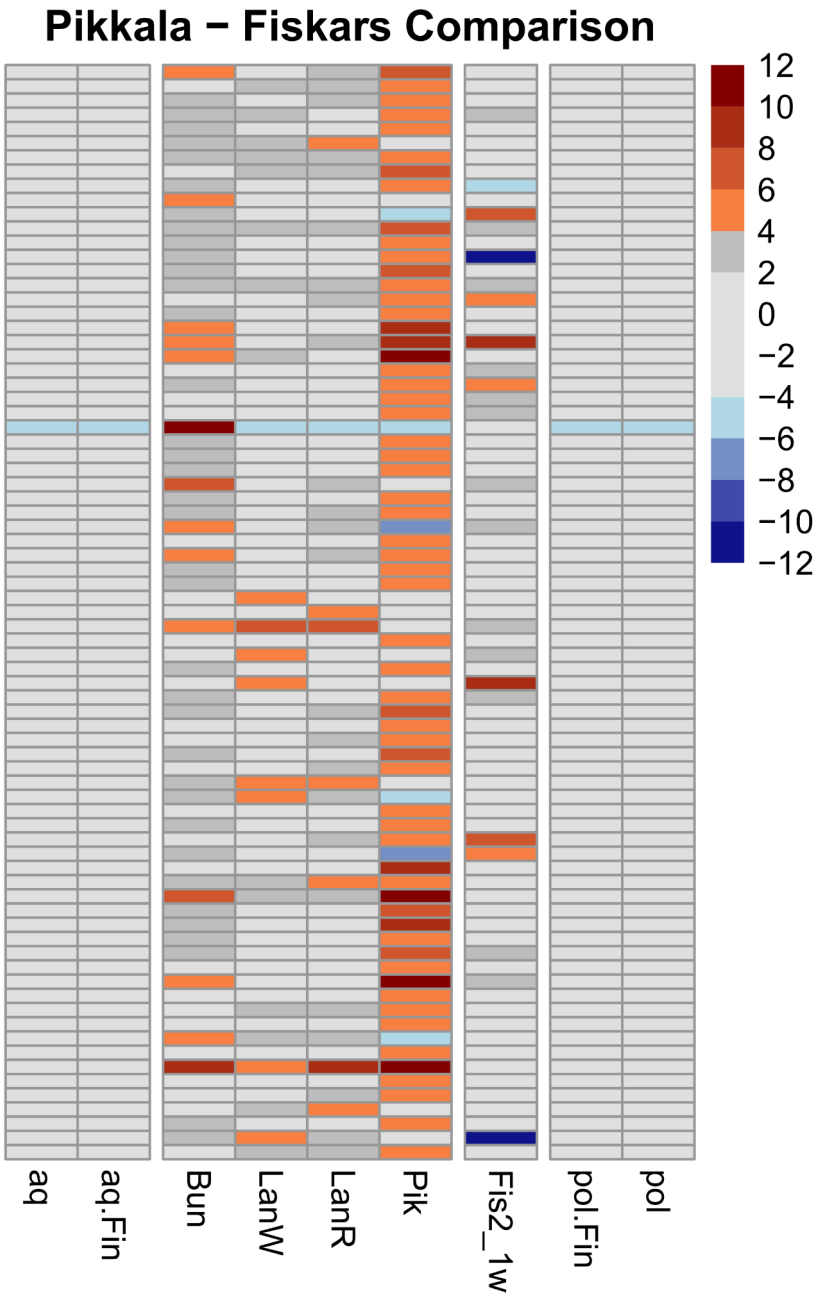
Test	Contrast	df	t	p
t-test from lmm with population as random variable	Aq-Bun	7	1.195	0.271
	Aq-LanW	7	1.045	0.331
	Aq-Pik	7	1.838	0.109
t-test from lm with no random variable	Aq-Bun	69	3.252	0.002**
	Aq-LanW	69	2.929	0.005**
	Aq-Pik	69	4.872	1.5e-06***



Supplementary Figure 1. Heatmap for all TE sequences. Demonstrates that the pattern of higher TE copies in hybrids and Finnish parent samples in heatmap in Figure 15 is generally true. Only one sequence has a median difference of <4 copies from the parent species in all populations.



Supplementary Figure 2. TE (orange) and gene (blue) abundances across the hybrid genome. Shaded regions show the proportion of repeats and genes in 10kb windows across each scaffold in the hybrid genome assembly. TEs are non randomly distributed, but tend to cluster in telomeric regions on certain chromosomes. Figure produced from data generated by Pierre Nouhaud.



Supplementary Figure 3. To understand potential *F. rufa* ancestry in the Pikkala population, a Fiskars individual (a likely *F. aquilonia* × *F.* rufa sample) is presented for comparison. There is little similarity to the Pikkala population suggesting misidentification is unsupported.