Transposon Ecology: The Power of Community Ecology Methods on Genomic Transposable Element Data

by

Brent J. Saylor

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Doctor of Philosophy
in
Integrative Biology

Guelph, Ontario, Canada
© Brent Saylor, September, 2019
Transposon Ecology: Initial Exploration of Community Ecology Methods on Genomic Transposable Element Data

Brent Saylor
University of Guelph, 2019

Advisors:
Professor Karl Cottenie
Professor T. Ryan Gregory

More than half of the human genome is made up of self-replicating mobile sequences of DNA called transposable elements (TEs). Much of the research on TEs has focused on the molecular mechanisms by which they transpose, their mutagenic effects, and the influences of their activity on the evolution of the host genomes in which they reside. In the 1980s, the “selfish DNA” papers popularized the idea of TEs as parasites, an analogy that grew into the idea of the genome as an ecosystem in which TEs were equivalent to species. In this thesis, I put this idea to work by testing the effectiveness of community ecology methods at identifying community-level TE patterns and exploring the potential insights that this approach may provide. I used redundancy analysis, community phylogenetics, and the 4th corner analysis to analyze the relationship between the TE community, TE traits, and the genomic environment. Each of these methods was adaptable to genomic TE data. The redundancy analysis found that, on average, 50% of the differences between TE communities of 11 genomes were explained by where that community was positioned on the chromosome (spatial patterns). In the human genome, 60% of the differences between TE communities were explained by spatial patterns, of which 33% is explained by the chromatin state and the remaining 66% remained unexplained. I also showed
that TE traits affected the differences in TE communities, with older TE families having more structure within genomes. Phylogenetic community ecology methods indicated that 17-30% of the TE communities in the Drosophila melanogaster genome were phylogenetically non-random. This pattern could indicate important community structuring assembly processes. Finally, the 4th corner analysis showed that the TE community spatial patterns are associated with TE-specific traits. This association with the traits of individual TEs, and not TE families, makes it possible to predict where new TEs may be found and which TEs might occur in new environments. Overall, this thesis represents the first systematic application of ecological methods per se to genomic TE data and assessment of their utility, and lays the groundwork for many future studies in transposon ecology.
ACKNOWLEDGEMENTS

To start, I would like to thank my advisors Ryan Gregory and Karl Cottenie. I am incredibly lucky to have not one but two brilliant, supportive advisors willing to guide me through the process of completing grad school. You have both been an inspiration to me, not only in how to be a great scientist, but to do so while fostering an environment that is fun, productive, and accepting. I would also like to thank my committee member and former MSc advisor Stefan Kremer, for being a constant source of great advice, and for keeping me on track despite my tendency to wander down rabbit holes.

Along with my committee, I would also like to thank the other members of the transposon ecology working group, Dr. Tyler Elliott, for endless discussions and insights on TEs, and for company on the long journey to get our PhDs, and Dr. Stefan Linquist for the exposure to a whole new way of thinking about science, and for a much appreciated outside perspective that made me a better scientist. This group, consisting of members from three different colleges at the University of Guelph, was pivotal in building the groundwork on which I could construct this thesis.

Throughout working on this thesis, I have learned the importance of having a great support system both at work and at home, and I would like to thank the past and present members of the Gregory and Cottenie labs, for keeping me connected during my hardest times. I am forever grateful. Katherine Drotos, thanks for all you do, you are responsible for me being more engaged with what was going on in the department during the latter half of my thesis. Carolyn Trombley, thank you for the constant reassurance -- it has always been heartening to know that someone else is going through so many of the same issues I have. Darren Kelly, thanks for all the laughs, and the willingness to get outside and throw a frisbee around at lunch.

To my family, I am so lucky to have been born into such a loving family, and to have endless, unjudgmental support in whatever I do. I love you all, and don't say it nearly enough.
To my friends, Brendan, Anna, Matt, Jen, Holly, Kathleen, Mike, and Chanelle. You have kept me sane through this whole process. You are all amazing, and I can’t imagine what my life would be like without you.

This work would also not have been possible without grants from OGSST, NSERC funding through Karl Cottenie and Ryan Gregory, the resources at the University of Guelph Library and at counselling services, and the amazing office staff in the Department of Integrative Biology.

Finally, Tucker, my dog, you are a constant source of joy in my life. Now that this is done, it’s time to go outside.
# TABLE OF CONTENTS

Abstract ................................................................................................................................. ii

Acknowledgements .............................................................................................................. iv

Table of Contents ................................................................................................................ vi

List of Tables ........................................................................................................................ xii

List of Figures ....................................................................................................................... xiii

List of Abbreviations ............................................................................................................ xv

1 Chapter 1: Introduction to Transposable Elements and Transposon Ecology ............. 1

1.1 Introduction .................................................................................................................... 1

1.2 TE Background .............................................................................................................. 2

1.2.1 Discovery .................................................................................................................. 2

1.2.2 Current Known TE Diversity and Classifications .................................................... 2

1.2.3 Why Do We Care About TEs? .............................................................................. 4

1.3 Transposon Ecology ...................................................................................................... 6

1.4 What Makes Up a Transposon Community in Practice? ............................................. 12

1.5 Overview of Thesis ....................................................................................................... 14
# Chapter 2: Genomic Environments and Their Influence on Transposable Element Communities

## 2.1 Introduction

- 2.1.1 Transposable Elements in the Genome
- 2.1.2 Types of Transposable Elements
- 2.1.3 Evolution, Ecology, and the TE Perspective
- 2.1.4 TEs in a 3D Environment

## 2.2 Methods

- 2.2.1 Study Species and Genome Data
- 2.2.2 Transposable Element Categorization
- 2.2.3 Quantifying Within-Chromosome Spatial Community Structure
- 2.2.4 Window Size Analysis
- 2.2.5 Genomic and Chromosomal Properties
- 2.2.6 TEs in a 3D Environment
- 2.2.7 Indicator Species Analysis

## 2.3 Results

- 2.3.1 Window Size
- 2.3.2 Genomic and Chromosomal Properties
- 2.3.3 Spatial Importance of 3D Spatial Structure
2.3.4 Indicator Species Analysis ............................................................... 36

2.4 Discussion ......................................................................................... 39

2.4.1 Consistency of Results Across Genomes and Window Sizes ............... 40

2.4.2 Patterns of Transposable Element Distribution .................................. 43

2.4.3 3D Analysis and Indicator Species Analysis ..................................... 46

2.4.4 Conclusions and Future Directions ................................................ 48

2.5 Supplementary Tables ....................................................................... 50

3 Chapter 3: Transposable Element Traits and Their Influence on Transposable Element Communities ................................................................. 53

3.1 Introduction .......................................................................................... 53

3.1.1 TE Classification Systems............................................................... 53

3.1.2 Why TE Traits? .............................................................................. 54

3.1.3 Phylogenetic Community Ecology .................................................. 55

3.1.4 TE Age .......................................................................................... 60

3.2 Materials and Methods ........................................................................ 61

3.2.1 Genome Sequence Data ................................................................... 61

3.2.2 TE Location Data ............................................................................ 62

3.2.3 TE Communities ............................................................................. 62

3.2.4 TE Classification ........................................................................... 63
3.2.5 TE Ecophylogenetics ................................................................. 65
3.2.6 TE Age ................................................................................. 66
3.3 Results ..................................................................................... 67
  3.3.1 TE Locations and Classification ........................................... 67
  3.3.2 Community Phylogenetics ...................................................... 68
  3.3.3 Comparison of Hierarchical and Tripartite Classification Systems .......... 73
  3.3.4 TE Age ................................................................................. 73
3.4 Discussion ............................................................................... 74
  3.4.1 Community Phylogenetics and TE traits ................................. 75
  3.4.2 TE Age ................................................................................. 79
  3.4.3 Hierarchal Classifying vs Non-hierarchal Classification .......... 81
  3.4.4 Conclusions ........................................................................ 82
4 Chapter 4: Transposable Element Traits and Genomic Environments: The 4th Corner in Ecology ................................................................. 84
  4.1 Introduction ........................................................................... 84
    4.1.1 Traits and Environments: The 4th corner in Ecology ............. 84
    4.1.2 Building the 4th Corner in TE Ecology ................................. 85
  4.2 Materials and Methods .............................................................. 88
    4.2.1 Data .................................................................................. 88
4.2.2 4th Corner Analysis

4.3 Results

4.3.1 4th Corner Analysis Results

4.4 Discussion

4.4.1 Does the 4th Corner Analysis Work on Genomic TE Data?

4.4.2 How Future Data on the Genomic Environment Will Affect the 4th Corner Analysis

4.4.3 How Future Data on TE Traits Will Affect the 4th Corner Analysis

4.4.4 Conclusions

5 General Conclusions

5.1 Does Transposon Ecology Work?

5.1.1 Can Concepts and Methods from Community Ecology be Successfully Applied to Genomes?

5.1.2 What Sort of Variables are Important in a Transposon Ecology Context?

5.1.3 Are These Ecological Analyses Likely to Provide General Insights into Transposon Biology and Generate New Hypotheses in Transposon Biology?

5.2 What Makes Transposon Ecology Different from Classical Ecology?

5.3 What is the Future of Transposon Ecology?

5.3.1 The Future for the Methods in this Thesis
5.3.2 The Future for Transposon Ecology as a Field ........................................ 114

5.4 Concluding Remarks .................................................................................. 114

References ........................................................................................................ 115
**LIST OF TABLES**

Table 2.1 Summary of the 11 species included in the present analysis .............................................. 26

Table 3.1: Breakdown of all the TE families from *Dmel*, classified into groups according to
(Arkhipova, 2017) ........................................................................................................................................ 68

Table 3.2: Number of phylogenetically overdispersed and clustered windows identified along
*Dmel* chromosomes using MNTD and MPD .................................................................................................. 71

Table 3.3: Redundancy analysis results comparing community phylogenetics and non-hierarchal
trait results to community composition in *Dmel* .................................................................................................. 73
LIST OF FIGURES

Figure 1.1: Updates to the Finnegan classification system................................................................. 5

Figure 1.2: The Tripartite classification system. .................................................................................. 6

Figure 1.3: Thesis outline. .................................................................................................................. 15

Figure 2.1: $R^2_{adj}$ of the amount of variation within the TE community of each chromosome in 11 genomes at multiple window sizes. .......................................................................................... 32

Figure 2.2: Genome properties versus the $R^2_{adj}$ across 11 genomes.............................................. 34

Figure 2.3: Amount of Variation in TE community composition explained by each environmental factor in the human genome............................................................................................................. 35

Figure 2.4: Results of indicator analysis produced by multipatt function from R package vegan. .................................................................................................................................................. 38

Figure 2.5: Adjusted degrees of freedom versus window size in RDAs of the Bos taurus genome .................................................................................................................................................. 42

Figure 2.6: TE presence and abundance in 11 genomes........................................................................ 44

Figure 2.7: Subcompartment diagram.................................................................................................. 47

Figure 3.1: Calculating mean pairwise distanced and mean nearest taxon distance for two communities........................................................................................................................................... 58

Figure 3.2: Wicker classification tree in Dmel. ..................................................................................... 64

Figure 3.3: Mean nearest taxon and mean pairwise distance along each of the chromosomes of Dmel........................................................................................................................................... 70
Figure 3.4: Counts of clustered and overdispersed TE communities in *Dmel*..............................72

Figure 3.5: Average age of each TE family vs $R^2_{adj}$ of the amount of spatial variation explained by RDA ..........................................................74

Figure 4.1: Transposon community ecology workflow..........................................................87

Figure 4.2: General Results of the 4th corner analysis of *Dmel* chromosome 2 by trait type. ......95

Figure 4.3: Results of the 4th corner analysis of *Dmel* chromosomes 2 (a), 3 (b) and X (c) with TE traits decomposed into factor levels.................................................................99
LIST OF ABBREVIATIONS

A1-A2: euchromatic compartments

B1-B4: heterochromatic compartments

bp: base pair

CT: chromosome territory

CG: cytosine+ guanine base pairs

CTFC: insulator protein with a role in mediating DNA loops

*Dmel*: *Drosophila melanogaster*

FDR: false discovery rate

GEO: gene expression omnibus

Hp1: HP1 protein/ centromere chromatin state

IndVal: indicator value

IS: insertion sequence

LINE: long interspersed nuclear element

LTR: long terminal repeat

MITE: miniature inverted repeat transposable element

MNTD: mean nearest taxon distance

MPD: mean pairwise distance
PcG: Polycomb group chromatin state

PCNM: principal components of neighbouring matrices

Phi: coefficient of association

PIC: phylogenetically independent contrast

RDA: redundancy analysis

SINE: short interspersed nuclear element

srpRNA: signal recognition particle RNA

TE: transposable element

tRNA: transfer RNA

TSD: target site duplication
1 Chapter 1: Introduction to Transposable Elements and Transposon Ecology

1.1 Introduction

Most DNA in the biosphere – at least the portion of it packaged within the nuclei of eukaryotic cells – is not genes. The great majority of DNA in animals, plants, and fungi is repetitive, and much of that repetitive DNA is made up of highly diverse, self-replicating mobile pieces of DNA called transposable elements (Elliott & Gregory, 2015a). The nature of this part of the genome and its function, if any, has been a long-standing debate. Even large-scale sequencing projects, and the ever-growing number of fully sequenced genomes, have been unable to answer these fundamental questions. This thesis will adapt tools from ecology to explore the diversity, distribution, and abundance of the elements that make up the dominant component of most eukaryotic genomes.

An in-depth review of TE biology and community ecology methods is outside the scope of this introduction, and has been provided in more detail by various other authors (see e.g. Brookfield, 2005; Le Rouzic & Capy, 2005; Mauricio, 2005; Le Rouzic et al., 2007; Venner et al., 2009; Linquist et al., 2013, 2015; Saylor et al., 2013). Instead, the remainder of this introductory chapter will be divided into three sections. The first section will describe the general background of TEs including how they were discovered, what we currently know about them and why TEs are interesting. This will be followed by an overview of some of the important transposon ecology literature, how the work of the transposon ecology working group built on this literature, and finally an overview of how transposon ecology will be applied in this thesis.
1.2 TE Background

1.2.1 Discovery

Barbara McClintock discovered “controlling elements” in the 1940s while studying the genetic basis of colour variation in corn (McClintock, 1950, 1961). The discovery of “controlling elements” and transposition, a new form of mutation, led to McClintock being awarded the Nobel Prize in 1983. McClintock believed that these elements were responsible for regulating gene expression during development. Some of McClintock’s contemporaries also ascribed regulatory and evolutionary functions to these elements (Britten & Davidson, 1971; Cohen, 1976; Shapiro, 1977); however, this idea was controversial (Comfort, 1995, 2001). Others suggested that these repetitive sequences had no regulatory function (Peterson, 1970; Cavalier-Smith, 1978), leading controlling elements to be renamed to the less controversial name “transposable elements” in 1956 (Wood & Brink, 1956).

1.2.2 Current Known TE Diversity and Classifications

The discovery of TEs in a wide variety of eukaryotes, as well as their discovery in bacteria and yeast in the 1970s (Saedler & Heiß, 1973; Hu et al., 1975a,b; Ananiev et al., 1978; Ilyin et al., 1978; Cameron et al., 1979; Strobel et al., 1979; Young, 1979) eventually led to the need for a TE classification system. The first TE classification system, proposed by Finnegan (1989), divides TEs into two groups, those that transpose using an RNA intermediate, and those that transpose directly as the DNA element itself (Finnegan, 1989). The increase in genome sequencing projects, and the resulting increase in new TE discoveries, has only made the need for a comprehensive classification system for TEs more important than ever before. However, classifying TEs has proven to be very complex, as demonstrated by the amount of debate around the current classification systems (Kapitonov & Jurka, 2008; Seberg & Petersen, 2009; Kapitonov et al., 2009; Piégu et al., 2015; Hoen et al., 2015; Arensburger et al., 2016; Arkhipova, 2017).
Two updates have been proposed to the Finnegan classification system (Piégu et al., 2015). The first, and more widely used, is the “Repbase” classification system, which is used by Repbase, currently the largest TE database (Jurka et al., 2005; Kapitonov & Jurka, 2008), and the second is known as the Wicker classification system (Wicker et al., 2007). These classification systems are very similar (Figure 1.1). Both systems first classify TEs into groups based on whether they use a DNA or an RNA intermediate during transposition (Type by Repbase and Class by Wicker). TEs are then divided into Class (Repbase) versus Order (Wicker) according to mechanistic features, organization of proteins, and phylogeny when applicable. Within each class/order there are superfamilies – which are groups of TEs defined by consistent phylogenetic topology, structural features, and the size and consistency of target site duplications – and families – which are defined by sequence similarity and the 80-80-80 rule. This heuristic rule defines a TE family as a TE sequence with 80% sequence similarity to the consensus sequence, over 80% of its length and is at least 80bp long (Wicker et al., 2007).

The second style of classification system was proposed by Arkhipova (2017), who advocated abandoning the hierarchal classification of TEs in favour of a three-part trait-based classification (which I will call the “Tripartite” classification system in the remainder of this thesis). The three traits used to classify TE in this system are transposition mechanism, insertion mechanism, and additional structural protein (Figure 1.2).

TEs are complex and neither style of classification fully captures all of this complexity. For example, the hierarchal classification systems are able to take relatedness and shared history into account. The Wicker system is explicitly designed to be analogous to the phylogenetic classification of species. This aids in the interpretability of the classifications; however, some TE families/superfamilies do not share any proteins, and thus are presumed to be evolutionarily unrelated. Nevertheless, these phylogenetically unrelated TEs are grouped together under both the Wicker and the Repbase classification systems (see Piégu et al., 2015; Hoen et al., 2015; Arkhipova, 2017 for a more in-depth critique). The Tripartite classification system addresses this issue by removing the hierarchal structure completely and concentrating on functional TE traits.
However, this approach is unable to capture the relationship between TEs that possess a shared history and may have similar traits/distributions because of that shared history.

1.2.3 Why Do We Care About TEs?

Beyond their complexity, TEs have a fascinating evolutionary history that has captured the imagination of those who study them. TEs are found in every form of life, from bacteria to eukaryotes, and vary widely in both the diversity of TEs found in a genome and the amount of that genome they comprise (Elliott & Gregory, 2015b, a). TEs are a common source of (reversible) mutation, which can be seen in the many variegated colour patterns they cause in both plants and animals (Luo et al., 1991; Clark et al., 2006; Zhang et al., 2019). These mutations are likely generally deleterious (Hickey, 1982; Boissinot et al., 2001; Belancio et al., 2008; Nelläker et al., 2012), and the host organism must suppress them or mitigate their destructive effects (Law & Jacobsen, 2010). Some researchers have gone so far as to propose that the need for suppression of TEs was important in the origin of multicellularity (Britten & Davidson, 1971; McDonald, 1998; Gregory, 2005). TEs have also been co-opted to perform host functions in many organisms. In drosophilids, the silkworm moth Bombyx mori, and some green algae, TEs replace the enzyme telomerase, selectively inserting into the ends of the chromosome, and buffering them from degradation (Biessmann et al., 1990a; Takahashi et al., 1997; Miller et al., 1999; Casacuberta & Pardue, 2003; Pardue & DeBaryshe, 2003). Another example is the mammalian immune system that relies on a TE-derived mechanism to generate variable antigens to meet new cellular threats (for review see Jones & Gellert, 2004).

TEs make up between half and two-thirds of the human genome (de Koning et al., 2011), and we have over 1 million copies of one SINE element called Alu (Lander et al., 2001; Dewannieux et al., 2003). This large portion of TEs has, of course, had an impact on our biology, we too use TEs in our immune system, but TEs are also the cause of many diseases in humans. A recent study found that Alu is the cause of no less than 37 neurological disorders (Larsen et al., 2018).
Figure 1.1: Updates to the Finnegan classification system. This figure is reprinted from Piégu et al. 2015 with permission from Elsevier.
Figure 1.2: The Tripartite classification system. This figure was reproduced unaltered from Arkhipova, (2017) in compliance Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/) and shows how various TE families are broken down by Replicative integrative and structural components. The letters along the axis represent the single-letter codes for the replicative (y-axis) and integrative (x-axis) components of the TE, with the 3rd letter code for structural elements highlighted in blue boxes. The two-letter code for each combination of replicative/integrative components, an example of the most common polymerase or endonuclease and examples of which TEs fall in each category are shown at each intersection.

1.3 Transposon Ecology

From the earliest work on TEs, there has been a tendency to think of them in terms of their effect on the host. This thinking persists and can be see exemplified by claims such as those of the ENCODE project (Dunham et al., 2012), which concluded that 80% of the human genome is functional because it is transcribed, is nearby to a DNA binding site, or has some form of chromatic modifications. This claim is made despite criticisms that this view is simplistic, in part because much of this transcription involves TEs (Eddy, 2012; Doolittle, 2013; Graur et al., 2013;
Elliott et al., 2014). Indeed, the phrase “long dismissed as junk…” has been used so often to start papers discussing the impact TEs have had on the biology or evolutionary history of an organism that it has become somewhat cliché (see e.g. Biémont & Vieira, 2006; Volff, 2006; Muotri et al., 2007; Biémont & Voytas, 2010; Lisch & Slotkin, 2011; Arkhipova, 2017; Klein & O’Neill, 2018; Larsen, 2018; Larsen et al., 2018; Ricci et al., 2018). Although this thinking is still pervasive, there is a growing number of TE biologists who have embraced a multilevel framework for evolutionary biology (e.g., Brunet & Doolittle, 2015). This multilevel framework started a slow trend toward thinking of TEs at the level of the individual copy within the host genome.

As mentioned above, upon discovering TEs Barbara McClintock named them controlling elements, and attributed to them a role in controlling gene expression during development (McClintock, 1961). Controlling elements were then renamed Transposable Elements (Wood & Brink, 1956), which committed less to the impact of TEs on the regulation of host genes. The idea of TEs having a functional role persisted. For example, Britten and Davison (1971) proposed a model of eukaryotic gene regulation in which repetitive sequences influenced nearby genes. The concept of junk DNA, DNA without selection on its specific sequence at the host level (Ohno, 1972), was the beginning of a view that not all DNA must be beneficial for the host organism. The selfish DNA papers (Doolittle & Sapienza, 1980; Orgel & Crick, 1980) formalized the idea that TEs did not need to have a functional benefit for the host to exist, and that repetitive DNA could, in fact, be parasitic relative to the host genome. This resulted in the shift that TEs themselves could be under selection, and that the selection pressures on TEs do not need to act in the same direction as section pressures on the host organism (or genome) (Doolittle & Sapienza, 1980; Orgel & Crick, 1980). These papers popularized the view that that TEs were parasitic on, rather than adaptive for, the host, and remain highly sited (Elliott, 2016). The idea of TEs as genomic parasites eventually inspired the concept of the genome as an ecological system (Kidwell & Lisch, 2001). Kidwell & Lisch (2001), who coined the term genome ecology, popularized the view of TEs existing on a continuum, with parasitic TEs at one end, through TEs that were neutral with respect to host fitness, to mutualist TEs, which were selectively beneficial
for the host (Kidwell & Lisch, 2001). Comparing genomes to ecosystems grew in popularity and eventually, there were several attempts to formalize the analogy by systematically drawing parallels between organismal ecology and “genome ecology” (Brookfield, 2005; Le Rouzic et al., 2007; Venner et al., 2009).

This growing body of “genome ecology” literature was what inspired the formation of the transposon ecology working group that served as the foundation for both my MSc and this PhD thesis. The efforts of this working group began with a review of all the existing genome ecology and genome ecosystem literature. What we discovered was that there was a great deal of confusion about what the ecology of the genome was, and when or how the analogy might be useful. Our first paper (Linquist et al., 2013) addressed some of this confusion. We found that most of the questions that were asked in the genome ecology papers were TE level questions; however, many of these questions concerned the evolution of TEs rather than their ecology. To avoid this confusion, we proposed a formal definition of ecology and evolution:

“A strictly evolutionary approach investigates change (or the lack thereof) in some focal entity over successive generations. The focal entities can range from genes to traits or from populations to higher taxonomic units. A strictly ecological approach assumes no change in the focal entities themselves, but focuses instead on the relationships between these entities and their environment. Here we use ‘environment’ in a broad sense potentially to include any of the factors with which an entity interacts. We also suggest that focusing on TEs as species within the genome should more properly be called transposon biology, to distinguish it from host ecologic factors that may affect the distribution and abundance of TEs within the genome, which we call genome ecology.” (Linquist et al., 2013)

For the remainder of this thesis I will use the term transposon ecology to more properly refer to what is normally called genome ecology. We also identified the following questions as similar to those traditionally asked in ecology, and currently unexplained in transposon biology: What are the differences in the abundance of TEs among eukaryotes? What are the differences in the types of TEs that are most common in different organisms? How do the distributions of TEs
within a given genome differ? Finally, for the first time, we quantified the relative importance of proxies for transposon ecology and proxies for transposon evolution in explaining the abundance of TE superfamilies within a genome. We found that ecological factors explained more of the variation among the individual TE copies in closely related genomes (within drosophilids), but evolutionary properties explained more of the variation when the genomes were more distantly related (within mammals) (Linquist et al., 2013).

With a conceptual framework for what constitutes transposon ecology in hand, I was the first researcher to apply ecological methods to genomic TE data to learn about TEs (Saylor et al., 2013). I did this by obtaining the number and location of each known TE in the genome of the cow (Bos taurus) and by analyzing the TEs along each chromosome the same way an ecologist would analyze differences in species compositions along a transect over a mountain range. By employing this method, I found that spatial patterns explained ~60% of the variation in TE community of each chromosome, as well as the TEs that were most important in determining this structure (see Chapter 2 for an expansion of this idea).

The transposon ecology working group continued this work with a review on the differences between ecological and molecular neutral theory, and the implications of those differences for transposon ecology. This review was inspired by Serra et al., (2013) who identified TE patterns similar to those found in ecologically neutral communities. This was interesting as the application of ecological neutral theory to genomic data represented a molecular theory that inspired an ecological theory that was in turn applied to genomic data. We first found that ecological neutral theory has been slowly weakened from a good approximation of a community, to a null model, to an occasional alternative hypothesis (Linquist et al., 2015). The reason for this slow weakening is complex, but much of it is because of the differences in the basis of molecular and ecological neutral theory. Molecular neutral theory concerns the amount of expected allelic variation within a population. Non-neutral alleles, such as those influenced by selection or drift, are identified as those alleles with less within-population
sequence variation than would be expected under neutral processes. Ecological neutral theory concerns the lack of differential effects of competition between organisms within a community.

As with the issue of transposon ecology and transposon evolution identified in our previous paper (Linquist et al., 2013), ecological neutral and molecular neutral ideas are often compared incorrectly because of the conflation of processes happening at different levels. A molecular neutral process is neutral with respect to the fitness of the host organism, while an ecologically neutral process is only concerned with the competitive interactions of the entities within the community. With that background established, we assessed the ecological-neutral-theory-inspired analysis of Serra et al., (2013) to see how well it fit the assumptions of ecological neutral theory models. We identified four assumptions common to ecological neutral theory. The first assumption is an equiprobable success among species. This assumption seems unlikely to be met among TEs as genetic species, since DNA and RNA elements propagate through substantially different mechanisms, with substantially different rates of replication and mobility. The second assumption is that communities are structured into a larger metacommunity. In TEs this requires a structure of smaller communities within the genome. Some TE families may form smaller communities, by localizing to specific areas of the genome; however, others are able to disperse to anywhere in the genome, forming one large community and violating this assumption. The third assumption is zero-sum dynamics and a fixed carrying capacity. This means that there is only a fixed quantity of resources in the system, which limits total community size. While some species seem to have selective pressures that limit their genome size, and by extension restrict space for TEs, this limitation is unlikely to be universal (Kremer et al. in prep). This can be seen in the huge diversity of genome sizes among eukaryotes. The fourth assumption is that of limited dispersal. As previously mentioned, there are some TEs families that seem to have limits on their ability to disperse within the genome, however there are also examples of TE families that can disperse to anywhere in the genome. This serves as another example that one must be very careful about level of focus when applying transposon ecology, and that it is very important to be familiar with the literature when adapting the ideas within another field for application in a novel context.
Our analysis of the strengths and weaknesses of the (at that time) state-of-the-art literature on “genome ecology” clearly established that applying methods from other fields must be done carefully. There was still the question of what type of ecology would be the best for answering the questions about TEs posed in Linquist et al., (2013). The questions we raised in that paper concerned the differences in the abundance of TEs among eukaryotes, the differences in which specific TEs are most common among types of organisms, and the distribution of TEs within a given genome. Each of these concerns the diversity and abundance of TEs within and among genomes, and the interactions of those TEs with both each other and with the rest of the genomic environment. Within ecology, questions about the effects of interactions of groups of species with both their environment and with each other would fall under the purview of community ecology.

In fact, our first application of transposon ecology in Saylor et al., (2013) used a community ecology approach to identify spatial patterns in TE community organization in the Bos taurus genome. The application of concepts and methods from community ecology to a genomic context requires an assumption that there are broad generalities to which those concepts and methods are relevant. Within community ecology itself, there is an oft-cited claim that communities are too strongly influenced by historical contingencies to generate meaningful generalizations, and that those generalizations are instead found at levels below (populations) and above (ecosystems) communities (Lawton, 1999). We tested this claim using a survey of 240 meta-analyses in ecology. Each meta-analysis was categorized as testing one or more ecological hypotheses at different scales (population, community, ecosystem), as well as by habitat type and taxonomy (Linquist et al., 2016). We found that community ecology studies found generalities at the same rate as ecological studies conducted at other scales. This result not only demonstrated that community ecology as a field is able to produce meaningful generalizations but in doing so it also demonstrated that the methods used to find these patterns are potentially useful to find the type of pattern we are looking for in TEs.
With the conceptual groundwork and methodological suitability of community ecology for answering interesting TE questions established, this thesis will build transposon ecology methods to answer questions about the distribution and abundance of TEs within and between genomes.

1.4 What Makes Up a Transposon Community in Practice?

In this thesis I will apply community ecology methods to genomic TE data in order to take advantage of these well-developed methods to learn about TEs. I will do this by looking at three sets of variables commonly used in community ecology: 1) the communities themselves, 2) the environment in which those communities are found, and 3) the traits of the members of those communities.

In the context of transposon ecology, the community is a group of TEs located in a specific area, in our case this will be represented by the number of individual TE copies belonging to each TE family or superfamily along a contiguous linear section of DNA. Within this thesis, communities of TEs will be specified as “TE communities”, whereas the communities of organisms studied by classical community ecology will be called “classical communities” or simply “communities”. Within a TE community, individual TE copies will be counted at the family/superfamily level. This is done because family/superfamily is the level at which most variation in TE traits is seen. Individual TE copies will have a very similar sequence and will likely differ little in functional traits.

In transposon ecology, the environment refers to properties of the genome in which TEs are found. Just as the environment of an organism consists of both biotic and abiotic factors, the genomic environment consists of both the other TEs present in the genome as well as the non-TE parts of the genome. This includes genes, regulatory regions, structural regions such as centromeres and telomeres, other repetitive elements such as satellite and microsatellite DNA, and the proteins within the nucleus that interact with TEs, such as histones, DNA replication machinery, and the proteins involved in the host defence system.
Finally, traits in the context of transposon ecology consist of the functional properties of the TE families themselves. As with counting the members of the transposon community, TE traits will be measured at the level of the TE family/superfamily since TEs all share the same traits within families/superfamilies. In this thesis, I consider several TE traits, including the particular mechanisms by which TEs copy themselves, the mechanisms by which they insert themselves into the genome, other structural proteins that may be possessed by the TEs, and the ages of the TE families.

The specific TE community, genomic environment, and TE trait data variables in this thesis are in no way exhaustive; there are almost certainly undiscovered or unsequenced TEs within the genomes used in these chapters. Likewise, the genomic environment is as complex and varied as the environment in which we as organisms find ourselves, and as outlined above TE biology and classification are complex. Consequently, the properties of the genomic environment and TE trait variables in this thesis are only a small subset of the potential variables that could be important to TE biology. Since our TE community ecology approach relies on incorporating information on a whole suite of TE families/superfamilies, I could only include those environmental and trait data available for all of them.

Is the community ecology approach a useful way to study TEs, and what kinds of things might we learn from using this framework and associated methods? I will answer this question by asking the following three applied sub-questions about transposon ecology: 1) Can methods from community ecology be successfully applied to genomes? 2) If they can, what sort of variables are important in a transposon ecology context? 3) Are these ecological analyses likely to provide general insights into transposon biology and to generate new hypotheses about TE diversity and distribution? Success in answering these questions will be measured by assessing the feasibility of applying ecological analyses to genomic TE data, the ability of these analyses to produce patterns, and asking whether any patterns identified make biological sense.
1.5 Overview of Thesis

Each chapter of this thesis will use a different method or set of methods to examine the relationship between at least two of the TE community properties outlined above. Chapter 2 will investigate the relationship between the TE community and the properties of the genomic environment. Chapter 3 will explore the relationship between the TE community and the traits of the TEs in that community. Finally, Chapter 4 will examine the potential for associations between the properties of the genomic environment and TE traits (Figure 1.3).
Figure 1.3: Thesis outline. This figure shows the three community ecology variables examined in this thesis (top row and bottom left boxes). The fourth box is the result of the “4th corner” analysis which combines data from each of 3 variables to identify TE trait – genomic environment relationships. The arrows in the figure which relationships are examined in each chapter of the thesis.

In Chapter 2, I will build on the methods developed during my MSc and published in Saylor et al., 2013. This begins by addressing some methodological issue with selecting the ideal window (and therefore community) size. Once this has been addressed, I will apply this method to a set of 10 diverse genomes to determine whether these genomes also show evidence for
spatial structure in their TE communities. Finally, to associate these spatial patterns with a true genomic environment, I will extend this analysis beyond the linear chromosome to a 3D interaction map and the associated chromosome state of compartments and subcompartments in the human genome. These compartments are made up of pieces of DNA that are physically close together in 3D space when the genome is uncondensed and have been shown to be associated with different epigenetic regulation, making them truly different types of genomic environment. This will allow me to show an association between TE distribution of the TE community and a property of the genomic environment that is known to be importation to TE biology. I will also use an indicator species analysis to determine if any of the TE superfamilies in the human genome are predictive of a specific chromosome subcompartment.

In Chapter 3, the functional traits of each of the TE families in the *Drosophila melanogaster* (*Dmel*) genome will be measured by two trait-based TE classification systems, the Wicker classification system and the Tripartite classification system. One of the ways organismal community-trait relationships are analyzed in community ecology is by community phylogenetics. At the organism level, this analysis will use the phylogeny as a proxy for traits, and patterns of phylogenetic overdispersion or clustering can indicate assembly processes such as environmental filtering or competition. I will use the Wicker classification system to generate a trait “phylogenetic tree”, which can then be employed to detect patterns of phylogenetic overdispersion/clustering in the *Dmel* genome. In a second analysis, I will calculate the TE age trait for each TE family by taking the mean of the differences between the consensus sequence for that TE family and each copy of that TE in the *Dmel* genome. This trait will then be compared to the amount of variation in abundance between TE communities that is spatially structured for each TE family. This will allow me to compare the effect of time on spatial patterns in the *Dmel* genome.

In Chapter 4, I test the usefulness of the 4th corner analysis at detecting general patterns between TE traits and properties of the genomic environment. The 4th corner method will be used to detect patterns between the chromosome state and the distribution of TEs on each
chromosome (Chapter 2), and the TE trait data from the Wicker and Tripartite TE classification systems (Chapter 3). This approach will allow me to draw conclusions about the nature of the associations between a TE property and where the TE is likely to be found in the genome. Identifying these direct relationships between trait and environment is useful as it allows researchers to predict how species will react to changing environments, or how species will react in a new environment. In identifying these TE trait-genomic environment associations I will have used community ecology methods to look for patterns between each of the three transposon community ecology variables discussed above: TE community, genomic environment, and TE traits.
2 Chapter 2: Genomic Environments and Their Influence on Transposable Element Communities

2.1 Introduction

2.1.1 Transposable Elements in the Genome

Transposable elements (TEs) are mobile genetic elements that comprise a large portion of most eukaryotic genomes. The human genome, for example, contains more than 3 million copies of various types of TEs, making up between half and two-thirds of the total quantity of DNA (de Koning et al., 2011). The diversity and abundance of TEs in the genome is influenced by coevolution with the host and the interaction between properties of the genome and properties of the TEs. For example, some TEs persist because they have been co-opted for important regulatory or structural functions (Levis et al., 1993; Agrawal et al., 1998; Feschotte, 2008), whereas others are known as disease-causing mutagens (Han & Boeke, 2005; Medstrand et al., 2005; Belancio et al., 2008) that remain abundant as a result of their ability to make copies of themselves, despite their detrimental effects on the host genome (Orgel & Crick, 1980; Orgel et al., 1980). In this regard, the relationship between TEs and their host genomes may be considered along an ecological continuum from mutualism at one extreme, through commensalism, to strict parasitism at the other end of the spectrum (Kidwell & Lisch, 2001).

Beneficial TE insertions can be preserved by natural selection acting at the host level (Britten, 1996; Miller et al., 1999), and others may accumulate via mutation pressure (e.g., if net insertions outweigh TE deletions) or genetic drift (especially in small populations) if they simply do not exert a significant negative impact on host fitness (Lynch & Conery, 2004; Lynch, 2007; Lynch et al., 2011). However, TEs that impose fitness costs on the host, either as deleterious mutagens or simply as extra genetic baggage, can persist in the long term only if they are able to evade inactivation by the host. This can be done by a combination of replicating more rapidly than they are silenced by host defences (Chandler & Walbot, 1986; Chomet et al., 1987;
Bucheton, 1995) and/or spreading more quickly than they are removed from the population via host-level purifying selection (Charlesworth et al., 1994; Le Rouzie & Deceliere, 2005).

These processes result in substantial variability in the diversity and abundance of TEs within and among genomes. Within individual genomes, TE diversity can be seen in the large number of TEs and TE superfamilies (1355 and 37 respectively in the Human reference genome; Bao et al., 2015). The variation in TEs between genomes is even more evident, with the number of TE superfamilies ranging from 1 in Dirofilaria immitis (dog heartworm) to 39 in Branchiostoma floridae (lancelet), Bombyx mori (silkworm moth), and Hydra magnipapillata (freshwater hydra) (Elliott & Gregory, 2015a), and abundance of individual TEs varying widely, even among individuals of the same species (Zhang et al., 2000; Lepetie et al., 2002).

### 2.1.2 Types of Transposable Elements

In addition to varying in abundance and distribution, TEs are mobile within the genome and are grouped into two broad classes based on how they move/transpose. By the Wicker classification system Class I TEs, or retrotransposons, move within the genome using a copy-and-paste mechanism. Copy-and-paste transposition involves transcription of TE DNA into an RNA intermediate, with the element itself remaining in its original location and serving as a template. The RNA intermediate then exits the nucleus where it is reverse transcribed into DNA, which then re-enters the nucleus and inserts into a new location (Boeke et al., 1985; Garfinkel et al., 1985; Eichinger & Boeke, 1988; Luan et al., 1993). Most Class II elements transpose using a cut-and-paste mechanism without an RNA intermediate. Cut-and-paste transposition involves excising the TE itself from its location in the genome for reinsertion into a new location without leaving the nucleus. Increased copy number in cut-and-paste transposons is accomplished by the repair mechanisms of the host genome responding to the breaks left behind by the TEs excisions, which fills in the missing DNA from the complementary strand. Helitron and Maverick elements are Class II elements that use a form of cut-and-paste transposition with no RNA intermediate (Kleckner, 1990).
2.1.3 Evolution, Ecology, and the TE Perspective

These factors – TE effects on host fitness, suppression by the host genome, TE accumulation and dispersal within (and between, see for instance horizontal migration; Schaack et al., 2010) genomes, and the evolutionary relationships between the different TE families (Wicker et al., 2007) – are all important in shaping TE abundances in different genomes. Some of these factors, such as the phylogenetic relationships between the TE families and the coevolution between TEs and host mechanisms for suppressing TE replication, are explicitly evolutionary from the perspective of the TE. Other processes, such as the dispersal of TEs to other parts of the genome, are explicitly ecological from the perspective of the TE. According to Linquist et al. (2013), an explanation is ecological if it focuses on changes in composition of the community of TEs and how TEs interact with the host genome and other elements in it, whereas evolutionary explanations relate to changes in the TE sequences themselves over generations, including co-evolution with the host genome. This distinction becomes important when the mechanisms responsible for shaping TE distribution or abundance can be either evolutionary or ecological. For instance, an ecological explanation for the accumulation of TEs in a specific area of the genome could be that the area is available and the nearby TEs are able to disperse there. However, if a specific group of TEs changed in a way that let them exploit that same area, that would be an evolutionary explanation for that same observation.

The notion of “genome ecology” has been invoked numerous times in the TE literature, however, many of the purported examples actually relate to TE evolution, and conflating TE ecology and TE evolution can result in asking the wrong questions and using the wrong tools to answer them (Linquist et al., 2013). In a recent study, we applied an explicitly ecological approach to the analysis of patterns of TE distribution and abundance within the genome of the cow, Bos taurus (Saylor et al., 2013). Specifically, TE distribution was assessed using a well-established method derived from community ecology, akin to assessing community composition along an environmental gradient (e.g., how communities might vary along a mountain range; see
e.g., Whittaker 1960). Our genomic version of this analysis examined how communities of TE superfamilies varied along each chromosome.

To implement this community ecology approach in the study of TE distribution, each chromosome in the cow genome was divided evenly into discrete windows. The abundance and diversity of the TE superfamilies in each window were then assessed. Combining the superfamily counts in each window resulted in a TE community for each window. Various properties of the genome could then be examined as potential correlates of variation in local TE community composition. In Saylor et al. (2013), we considered the location of the window along the chromosome and local gene density as predictors, and these were used to test if the TE communities of each chromosome changed in predictable ways from one window to the next. The results in Bos taurus found that 50% of the within-chromosome variation in TE community composition was explained by examining physical position along that chromosome (Saylor et al., 2013). Our analysis demonstrated the power of this ecological approach, but it was largely a proof-of-concept and examined only one genome. Moreover, we implemented the most straightforward way to measure the location of any genetic element: the location of their sequence along the chromosome.

2.1.4 TEs in a 3D Environment

The above approach is the most intuitive way to represent loci on a chromosome. Most chromosome maps (physical, genetic, and karyotype) are linear in nature; however, this is an oversimplification of the actual distance between two loci on a linear chromosome. These idealized maps are representative of the phases of the cell cycle associated with cell division, which make up a relatively short part of the cell’s life cycle (Russell et al., 2010). During interphase, which makes up the majority of the cell’s life cycle, chromosomes are found uncondensed within the nucleus, where they are arranged into chromosome territories (CT) (Lichter et al., 1988; Pinkel et al., 1988; Lieberman-aiden et al., 2009; Zhang et al., 2012). Each CT contains one chromosome, with the interaction between chromosomes occurring at the borders between territories. CTs can be further divided into genomic compartments and
subcompartments (Rao et al., 2014). Genomic compartments are correlated to heterochromatic state and are categorized into two types: Compartment A is made up of loosely bound euchromatic DNA, which is accessible to proteins, is gene dense, and contains highly expressed genes. Compartment B is made up of tightly-bound heterochromatic DNA and is less assessable to proteins (Lieberman-aiden et al., 2009). Genomic subcompartments are areas within a genomic compartment that physically interact more often with each other than one would expect by chance. Although the specific reasons these subcompartments form is not clear, each of the six subcompartments identified by Rao et al. (2014) has a distinct interaction profile and similar histone modifications, indicating that they are regulated in similar ways. Furthermore, changes to the histone modification pattern of a specific area are associated with the changes in long-range interaction patterns, meaning that changes to the regulation of a locus can cause that locus to change position in the nucleus. Subcompartments A1 and A2 are both found within compartment A and are distinguished from each other by their histone modification pattern, and by when they finish replicating. Both compartments A1 and A2 finish replicating early in s phase, with A1 finishing in early S phase and A2 continuing replication into the middle of S phase (Rao et al., 2014). The B subcompartments, B1-B4, have more varied properties. Subcompartment B1 is associated with facultative heterochromatin. Subcompartments B2 and B3 both replicate late in S phase, have little histone modification and are associated with the nuclear lamina, a structural component of the nucleus located inside the nuclear membrane. Subcompartment B2 is largely made up of pericentric heterochromatin and is associated with the nucleolus, while subcompartment B3 is not. Subcompartment B4 is highly enriched for KRAB-ZNF superfamily genes and is only found in a small region on chromosome 19 (Rao et al., 2014).

This physical 3D structure of a chromosome within a cell has important implications for how genes are expressed, how they interact with regulatory elements, and how accessible the DNA is to proteins (Cremer & Cremer, 2001; Sexton et al., 2007; Bickmore, 2013). Like genes, TEs also need to be accessed by protein machinery such as transposases, and polymerases to function. It is thus likely that it will also influence the TE distribution along the chromosome. This, however, has never been studied. If the physical structure does have an influence on TE
community dynamics, I would expect sub-compartments that are physically close to each other, and thus more likely to be regulated in similar ways, will be more similar than predicted based on chromosome location alone. This is consistent with findings by Sanyal et al. (2012), who found only 7% of looping interactions are with the closest gene and a strong correlation between long-range promoter-enhancer interaction and gene expression. If genomic subcompartments are acting as different genomic environments, I would also expect heterochromatic regions (subcompartments B1-B4) and euchromatic regions (subcompartments A1-A2) to have different TE community compositions. If these chromosome structural properties are important determinants of TE location, I would expect strong relationships between these properties and TE locations along the chromosome, similarly to the analysis of the B.taurus genome (Saylor et al., 2013).

While this relationship between functional chromosome structure and TE chromosomal distribution is the primary focus of this study, I will also study the generality of these TE spatial patterns. Chromosome subcompartment data were only available for the human genome, as it is the only chromosome interaction analysis with a sufficiently high resolution to detect subcompartments (Rao et al., 2014). To confirm our results in other genomes, I will also explore the generality of spatial findings across a suite of species with genomes with high sequencing depth, scaffolds assembled into full chromosomes, and well-annotated TEs. There are 11 species that fit these criteria available from Genbank.

Chromosome structure within genomes appears to be a universal genome property, from chromosome territories at the coarsest level of organization (Cavalli & Misteli, 2013), to chromosome looping which has been found in a wide variety of prokaryotes and eukaryotes (Hofmann & Heermann, 2015). If the TE spatial distribution is (partially) determined by these universal chromosome structural properties, then I predict that the spatial patterns in TE distributions would be detected across a wide variety of genomes.

To replicate the spatial analysis across multiple species, one methodological issue must be solved first. In the primary analysis of Saylor et al. 2013, the window sizes were determined
independently for each chromosome so that each window contained an average of 100 TEs. This ensured that a TE community would be examined within each window, no matter the size of the chromosome. However, it had the less-desirable effect of normalizing TE density as a chromosome property, possibly obscuring TE density itself as an explanatory factor of the importance of spatial location. Using a systematic approach with evenly distributed fixed window sizes should make it possible to identify the effect of window size on this analysis. The Saylor et al. (2013) study used a fixed window size across the entire genome. The fixed window method produced very similar results to the dynamic window approach when the fixed window size was similar to the average size of the dynamic windows, with the added benefit that windows on any chromosome, in any genome, were directly comparable. However, since the window size affects the number of communities on each chromosome, the average size of each individual community of TEs, and in turn the computational resources required to conduct the analysis, were very different. How to choose an appropriate window size for the analysis of a given genome was not fully explored. Additionally, it remains to be determined whether similar window sizes can be used across widely different genomes in such a way that the results can be compared.

The present study investigates the importance of a chromosome’s 3D spatial structure (the “genomic environment”) on the distribution of the TEs on each of the chromosomes in the human genome and assesses the usefulness of using TEs as indicators of specific genomic environments. Additionally, I investigate if the spatial patterns observed in the Bos Taurus TE communities are also found in a set of 11 genomes from diverse eukaryotic. To accomplish this, I also assess the impact of window size on detecting the spatial structure of the TE community within each of our 11 study genomes.
2.2 Methods

2.2.1 Study Species and Genome Data

Of the available whole genome sequences, only those that were assembled into chromosomes were considered. Eleven eukaryote genomes – including representative vertebrates, invertebrates, plants, and fungi – were included in the present study, on the basis of genome size, chromosome number, and phylogenetic diversity. These are summarized in Table 2.1. The reference genome was used where available. Reference genomes are the highest quality genomic dataset in the NCBI database and are highly curated and annotated. If no reference genome was available a representative genome was used. Representative sequences are also of high assembly quality; however, they have less annotation and are deemed representative of a species by clustering of at least similar 10 variant genomes. If neither one of those was available, the most recent assembly was used. The output of RepeatMasker searches of each genome were downloaded from the Genbank FTP site. These files contain the names and locations of any region in each genome that matched TEs in the RepBase TE database.
Table 2.1 Summary of the 11 species included in the present analysis, including information on genome size, chromosome number, and source sequence accession.

<table>
<thead>
<tr>
<th>Species</th>
<th>Common name</th>
<th>Genome size(bp)</th>
<th>Chromosome number (n)</th>
<th>Assembly accession</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Homo sapiens</em></td>
<td>Human</td>
<td>3088269832</td>
<td>23</td>
<td>GCF_00000140.5.28</td>
</tr>
<tr>
<td><em>Bos taurus</em></td>
<td>Cow</td>
<td>2670123310</td>
<td>30</td>
<td>GCF_00000305.5.6</td>
</tr>
<tr>
<td><em>Mus musculus</em></td>
<td>Mouse</td>
<td>2730855475</td>
<td>21</td>
<td>GCF_00000163.5.23</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>Vinegar fly</td>
<td>143706478</td>
<td>7</td>
<td>GCF_00000121.5.4</td>
</tr>
<tr>
<td><em>Takifugu rubripes</em></td>
<td>Pufferfish</td>
<td>391484725</td>
<td>22</td>
<td>GCF_00018061.5.1</td>
</tr>
<tr>
<td><em>Danio rerio</em></td>
<td>Zebra fish</td>
<td>1371702787</td>
<td>25</td>
<td>GCF_00000203.5.5</td>
</tr>
<tr>
<td><em>Anopheles gambiae</em></td>
<td>Mosquito</td>
<td>265011681</td>
<td>5</td>
<td>GCF_00000557.5.2</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td>Roundworm</td>
<td>100272607</td>
<td>6</td>
<td>GCF_00000298.5.6</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>Mustard weed</td>
<td>11914634</td>
<td>6</td>
<td>GCF_00000173.5.3</td>
</tr>
<tr>
<td><em>Oryza sativa</em></td>
<td>Rice</td>
<td>382150945</td>
<td>12</td>
<td>GCF_00000542.5.2</td>
</tr>
<tr>
<td><em>Zea mays</em></td>
<td>Maize</td>
<td>2059701728</td>
<td>20</td>
<td>GCF_00000500.5.1</td>
</tr>
</tbody>
</table>

2.2.2 Transposable Element Categorization

The bins used to categorize the TEs within each genome are based on the RepeatMasker output downloaded from the RefSeq database on GenBank. Our analysis used the superfamily level classification of TEs as it is the most well-defined classification below the more general TE Class. It is also the most commonly reported, which increases the degree to which data can be compared across these different genomes. Modifications to the RepBase classification were
carried out to reflect updates to TE superfamilies subsequent to the record submission to Repbase (Kapitonov et al., 2009; Yuan & Wessler, 2011; Vassetzky & Kramerov, 2013). Several groups of LINEs, SINEs, LTR retrotransposons, and DNA transposons could still not be identified to the superfamily level in their original publications. These will be referred to as superfamilies for simplicity, however, they reflect less well-defined groupings.

2.2.3 Quantifying Within-Chromosome Spatial Community Structure

To detect the relative impact of within-chromosome community structure for all of the spatial analyses, I used redundancy analysis (RDA) as implemented in the vegan package in R (Legendre & Legendre, 1998). This performs a multivariate multiple regression with the counts for the number of TEs in each window as the dependent variable, and the properties of the genomic/chromosomal environment (i.e. spatial pattern or chromatin subcompartment) as the independent variable. The spatial environment was modelled with the principal components of neighbouring matrices (PCNM) (Dray et al., 2006; Borcard et al., 2012) procedure. The input for the PCNM for the analyses that use linear spatial structure was a dissimilarity matrix representing the distances between each window, and the input for the 3D analysis was a dissimilarity matrix based on the interaction frequencies from the HiC data of each chromosome. For more details, see Saylor et al. (2013). This results in an $R^2_{adj}$ value for each chromosome which represents how well the TE community in each window can be predicted based on the environmental variable used. In other words, the $R^2_{adj}$ value is an indication of how well the patterns generated by the PCNM procedure can explain the variation in the TE communities of each window.

2.2.4 Window Size Analysis

The window size analysis used RDAs of TE abundances across windows as a function of the spatial location of the window to explain TE communities within each chromosome as above and in Saylor et al. 2013. This was done for each chromosome in each of the 11 genomes at each of 20 different window sizes ranging between 10x the size of the largest Bos taurus TE
(14,753bp) at minimum to the size of the smallest *Bos taurus* chromosome at a maximum (4,404,134bp). This resulted in windows ranging from 14,753bp to 4,404,134bp by increments of 219,469bp.

### 2.2.5 Genomic and Chromosomal Properties

In addition to assessing the impact of window size on the detection of spatial patterns in TE community composition along the chromosome, I also assessed the impact of changing the window size on chromosomes with different environmental properties. I selected one “large” (790,718bp) and one “small” (144,808bp) window size because they represented extremes of window sizes while avoiding sizes small enough to cause statistical issues (see Discussion). The genomic properties investigated with these two window sizes were: 1) The total length of the available genome sequence; 2) The C-value, an independent genome size estimate of physical size for that species taken from the Animal Genome Size Database (Gregory, 2016) or Plant DNA C-values Database (Benett & Leitch, 2012); 3) the difference between the genome size estimate and the available sequence length, which serves as a measure of how complete the sequence is; and 4) the number of chromosomes. The chromosomal properties, downloaded from the GenBank entries for each genome (Supplementary Table 1), were: 1) Genome, which consisted of which genome the chromosome was from and was used in the phylogenetic independent contrast to account for non-independence of the data; 2) Chromosome length, which consisted of the length of the sequence for that chromosome and was used to measure the amount of space for the TEs to insert; 3) GC%, the percentage of the sequence made of guanine-cytosine base pairs, which is highly variable, easily calculated from the sequence, and has been correlated with the presence of some TE families and other genomic features (see Eyre-walker & Hurst, 2001 for review); 4) Number of genes, which directly measures the number of genes on the chromosome; and 5) Number of proteins, which measures the number of those genes with known or putative protein products. Each of the genomic properties was compared to average $R^2_{adj}$ for all of the chromosomes in that genome and each of the chromosomal parameters were compared to the $R^2_{adj}$ for each chromosome across genomes.
The phylogenetic distances based on the protein architecture of 11 host species were downloaded from the sequenced tree of life webpage (Fang et al., 2013). The resulting phylogeny was used to run a phylogenetically independent contrast (PIC) analysis on the $R^2_{adj}$ from the RDA for each chromosome, and on the genomic and chromosomal properties. In the chromosome property analysis, polytomies were added to the tips with each chromosome in each genome being equally related to each other. The contrasts of the average $R^2_{adj}$ values were then compared to the contrasts of the genomic properties as above.

### 2.2.6 TE in a 3D Environment

10kb resolution interaction frequency data with MAPQ scores above 30 generated by Rao et al. (2014) were downloaded from the Gene Expression Omnibus (GEO) database (GEO accession GSE63525). These frequency data were normalized using the method from Knight & Ruiz, (2013) to adjust for methodological artifacts according to the instructions downloaded with the data. The genomic compartment locations were calculated by taking the first principal component of the normalized interaction matrix (Lajoie et al., 2015), using the cmdscale function in the stats package of R.

The genomic subcompartment data were also downloaded from the Rao et al. dataset hosted in the GEO database. The subcompartment data consisted of start and end positions for each subcompartment along the sequence of each human chromosome, and which of the seven subcompartments types (A1, … NA) that section is classified as. The boundaries between each compartment were identified by Rao et al. (2014) using an arrow matrix transformation to generate “corner score”. I then associated that structural information to our TE distribution data. For each window in our chromosome spatial analysis, I calculated the proportion of the window made up of each subcompartment. This resulted in seven variables consisting of the proportion for each window made up of that subcompartment.

Finally, I repeated the spatial RDA for the human chromosomes (see Quantifying within-chromosome spatial community structure section above), but this time in addition to the spatial
patterns obtained with PCNMs, I used spatial patterns generated by PCNMs of the frequency data, the seven additional explanatory variables from the subcompartments, the first principal component of the interaction frequency matrix, and the number of genes in each window.

2.2.7 Indicator Species Analysis

The usefulness of each TE and each pair of TEs as an indicator of genomic subcompartment in the human genome was determined using the multipatt function found in the indicspecies package for R (Dufrêne & Legendre, 1997; De Cáceres & Legendre, 2009; De Cáceres et al., 2010). This analysis computes two types of an indicator entity: IndVal, which evaluates the strength of using each TE as an indicator of a specific environment; and Phi, which is a measure of the correlation between the species presence/absence matrix, and the genomic subcompartment. Each of these statistics was measured for each TE in each genome.

IndVal scores range from 0 to 1, with 1 indicating a TE always occurs in a given environment, and never occurs in other environments, and 0 indicating a given TE never occurs in a given environment and is always found in other environments. Within each chromosome, IndVal scores were generated for each TE for each subcompartment/pair of subcompartments. The significance of each score was assessed using a permutation test, and significant scores, where $p < 0.05$ were reported.

Phi scores were also produced for each of the 22 human chromosomes. Phi scores also range from 0 to 1, with 1 being a perfect correlation between two binary vectors and 0 being no correlation between binary vectors. The significance of each score was assessed using a permutation test, and scores where $p < 0.05$ were reported.

2.3 Results

In this study I used tools from community ecology to look for spatial structure in the TE communities of 11 genomes. I found that ~ 60% of the variation in TE communities can be explained by spatial patterns. Furthermore, in the human genome 40% of the variation in the TE
community was explained by the 3D structure of the genome, and half of that was explained by the chromosomal environment (genomic subcompartment).

### 2.3.1 Window Size

The results of the window size analysis are shown in Figure 2.1. Spatial patterns were significant predictors of TE community composition in 131 of the 149 chromosomes analyzed at all window sizes ($p < 0.05$). Of the 18 other chromosomes, the 16 *Saccharomyces cerevisiae* chromosomes were only significant at the smallest windows size, 14,753bp at the $p < 0.05$ level. The other two chromosomes that were not significant at all window sizes were the X and Y chromosomes in the *D. melanogaster* genome. Both of those chromosomes were significant at the $p < 0.05$ for window sizes below 100,917pb. The X chromosome was not significant at any larger window size, and the Y chromosome was also significant at the 144,808bp window size but not at any window sizes above this size.

Among the TE communities of the chromosomes that had a significant spatial component an average of ~50% of the variation can be explained by spatial patterns alone, with the highest mean $R^2_{adj}$ of 60% found in *Danio rerio* and the lowest mean $R^2$ of 37% found in *Mus musculus* (Figure 2.1).
Figure 2.1: $R^2_{adj}$ of the amount of variation within the TE community of each chromosome in 11 genomes at multiple window sizes. This figure shows the $R^2_{adj}$ of for each of the 11 analyzed genomes at each of 20 different window sizes. Each coloured line represents one chromosome and each pane is a different genome. The genomes lacking results for some of the larger window sizes do so because at those sizes the smaller chromosomes in that genome would have been made up of less than three windows.

2.3.2 Genomic and Chromosomal Properties

The relationship between the average amount of variation in TE communities explained for each genome ($R^2_{adj}$) and whole-genome properties are shown in Figure 2.2. None of the whole genome properties, Chromosome number, C-value, Sequenced length, or the difference
between C-value and sequence length, showed a significant relationship with average $R^2_{adj}$ (all $p$ values $> 0.05$).

This remained true after accounting for differences based on phylogeny using phylogenetic independent contrast. After correcting phylogeny with PIC, $R^2_{adj}$ and GC% were positively correlated ($p < 0.05$), while the other chromosome properties showed no significant correlations with $R^2_{adj}$.

### 2.3.3 Spatial Importance of 3D Spatial Structure

The 3D spatial structure measured by the HiC interaction frequency explained on average 43$\pm$12% of the TE community distribution within each human chromosome. This was always a subset of the variation explained by our initial analysis in which distances were generated using the PCNM procedure ($R^2_{adj}$ 69$\pm$7%).

Of the variation explained by HiC data, nearly half of that variation (a total of 22$\pm$12%) is explained by chromosome subcompartments. An additional 7$\pm$4% is explained by the type of subcompartment, but not by HiC data. Gene location data was also analyzed, however, it only explained a total 2$\pm$3% of the variation in the TE community (Figure 2.3).
Figure 2.2: Genome properties versus the $R^2_{adj}$ across 11 genomes. a) shows the correlation between $R^2_{adj}$ and 4 properties of the genome: Chromosome number, C-value, sequenced length, and the difference between C-value and sequence length. A) Show the raw results, while b) shows the results after PIC analysis. In both cases none of the correlations were significant.
Figure 2.3: Amount of Variation in TE community composition explained by each environmental factor in the human genome. This figure shows a breakdown of the $R^2_{adj}$ for the TE communities of each chromosome in the human genome by what environmental factor explains that variation. In this case, $R^2_{adj}$ indicates the amount of variation in the TE community explained by each variable. The $R^2_{adj}$ for the TE communities of each chromosome are partitioned into those explained by 3 explanatory variables. 1) Number of Genes in each window. 2) Which subcompartments the window was made up of and 3) How close the windows were, as measured by HiC interaction frequency. Each boxplot in the bottom panel represents one of the sections in the Venn diagram above the above. The Gene, HiC and SubC boxplots represent the whole circle in the Venn diagram, while the remaining boxplots represent the 7 subsections.
2.3.4 Indicator Species Analysis

Indicator value (IndVal) scores were produced for each TE within each of 22 human chromosomes. The mean Indval score across all chromosomes was 0.55, with the scores ranging from the highest DNA/hAT-Tip100 and LINE/RTE (IndVal= 0.95 and 0.89 as indicators for subcompartment B2 or NA on chromosome 22), to the lowest, satellite/centromere (IndVal = 0.35 for subcompartment NA on chromosome 16) (Figure 2.4a and Table S1).

Phi scores were also produced for each of the 22 human chromosomes. This resulted in 93 significant potential indicator TEs among the 22 chromosomes. The mean Phi score across all chromosomes was 0.38, with the scores ranging from the highest DNA/hAT-blackjack (IndVal= 0.82 for subcompartment A2 on chromosome 21) to the lowest, rRNA (Phi = 0.23 for subcompartment B1 on chromosome 14)(Figure 2.4 b and Table S2).

Overall, the consistency of indicator scores between chromosomes was low, as shown in the lower panels of Figure 2.4 a and b. The majority of TEs were not significant indicators on more than one or two chromosomes, and those that were indicators on multiple chromosomes were rarely indicators of the same environment type (Figure 2.4a and b, lower panel). The exceptions to this were found in the Phi scores of Alu and scores for satellite DNA. The Phi scores of Alu showed a significant correlation with an environment on 8 chromosomes, with 4 of those correlations associated with the A1 subcompartment and a fifth being with the A1+NA subcompartment pair. The various categories of satellite DNA showed a more consistent pattern. When the Phi/IndVal score was significant, Satellite DNA was always an indicator of the NA subcompartment. For centromeric satellite DNA, this relationship was detected in 5 chromosomes by the Phi score and 3 chromosomes by the IndVal score.
**Figure 2.4:** Results of indicator analysis produced by `multipatt` function from R package `vegan`. a) Shows the IndVal scores resulting from the analysis. b) Shows the Phi scores resulting from the analysis. For both a) and b) the top boxplot shows the distribution of IndVal/Phi scores, while the lower stacked bar plot shows how often TEs are indicators of a given environment. Blue colours are scores for single environments while reds are environment pairs.
2.4 Discussion

The underlying spatial structure that is present in the TE communities in these 11 genomes, like the underlying spatial structure of communities consisting of organisms, can only be explained by a complex mix of both evolutionary and ecological factors. In TE communities, these patterns are further complicated by selection pressures occurring at both the level of the host and the level of the TE, which can work to either reinforce or counteract each other. This complexity necessitates careful consideration of both evolutionary and ecological processes, and the scale at which they are acting, before making conclusions about TE communities. At the host level, ecology focuses on interactions between different species, which rarely if ever involves TEs. Evolutionary processes at the host level can involve TEs, but mainly as sources of mutation, as they cause changes in the focal entity, the host, or by host-level processes affecting TEs, such as drift fixing TE insertions in small populations. TE evolutionary processes are those in which the TEs themselves are changing. This is the focus of most TE research. This analysis focused on the ecology at the level of the TE, by examining the relationships between various types of TEs and their environment. Although TE ecology is often discussed, the boundaries between these levels and processes are often not considered before designing experiments or making conclusions, violating many of the assumptions for those processes (Linquist et al., 2013, 2015).

The analysis presented by Saylor et al. (2013) focused on explicitly ecological methods adapted from community ecology. That study demonstrated the utility of such an approach in principle and in practice. In this Chapter, I continued that analysis by extending this proof-of-concept in four major ways: 1) By examining the impact of window size selection in the implementation of the method; 2) by applying the approach to 11 genomes of varying sizes and compositions; 3) by considering additional genomic and chromosomal factors that may influence TE abundance and distribution; and 4) by examining the role of how physically close areas of the genome are in predicting TE community.
2.4.1 Consistency of Results Across Genomes and Window Sizes

Importantly, the results obtained using various window sizes and multiple genomes were remarkably consistent, suggesting that this approach will be broadly applicable in analyzing TE abundances and distributions across a wide range of taxa. In particular, the present analyses demonstrated that a large amount of the spatial variation in the TE community of each genome was explained by accounting for the spatial location of those TEs along the chromosome. In other words, using RDAs a large proportion of the TEs likely to be found in a TE community (window) of any chromosome can be predicted based on the relative location of that window along the chromosome.

Care must be taken when adapting any set of tools for a new use. The analysis of TE communities using methods from community ecology appears promising, as the spatial location of TEs was correlated to the composition of the TE community in all 193 chromosomes analyzed across all 11 genomes. Although the spatial distribution of TEs was significant on each chromosome, the amount of the TE community in each window that could be explained in this way was not, and the TE superfamilies that were spatially structured were not consistent across chromosomes. One would expect that decreasing the window size would increase the explanatory power of spatial patterns as it would allow finer-scale spatial patterns to be detected. However, reducing the window size too much results in a steep drop-off in explanatory power, which is most evident in the smaller genomes (Figure 2.1). This fact highlights one notable limitation of the method when applied to genomes vs. ecological communities. Ecological communities typically have less complete sampling, and the statistical methods used on this ecology data are designed with this in mind. In the context of analyzing whole-genome sequences, there is a risk of creating statistical overpowers as the degrees of freedom are extremely high, which can cause the model to falsely identify patterns as significant.

This was a known issue in the original ecological application of the PCNM method, and the solution was to use $R^2_{\text{adj}}$ instead of the raw $R^2$ value (Equation 1). This $R^2_{\text{adj}}$ value reduces the $R^2$ based on the inflated statistical power; however, the hundreds of thousands of
observations typical of a whole-genome analysis are too extreme for even the $R^2_{\text{adj}}$ calculation, and $R^2_{\text{adj}}$ plummets as the difference between the number of windows and the number of potential spatial patterns generated from the PCNM procedure increases (Figure 2.5). This brings up an importation point raised by Linquist et al. (2013), namely that the assumptions and limits of any model need to be carefully considered before applying them to a different type of data. In this case, without considering the assumptions and function of the model, one might assume that all TE interactions happen at a very large scale, as $R^2_{\text{adj}}$ is lower in analyses with small window sizes. This may be true in some cases, fine-scale patterns in TE distribution may also be obscured by lack of statistical power.

Equation 1:

$$R^2_{(Y|X)\text{adj}} = 1 - \frac{n - 1}{n - p - 1}(1 - R^2_{(Y|X)})$$

Where $n$ is # of observations (windows) and $p$ is # parameters (Potential PCNMs graphs)
Despite this limitation, the analysis of 11 genomes differing in size and chromosome number revealed some interesting patterns. Notably, similar TE families were implicated in accounting for spatial variation of the TE community in each individual genome, regardless of the window size used. Moreover, the significant TE families at each window size were consistent, with those with the weakest support becoming non-significant at smaller window sizes. This weakened support for at smaller window sizes was due to the increasing adjustment to the $R^2_{\text{adj}}$ value. Thus, results of the spatial analysis are relatively robust to window size even across very different genome sizes or numbers of chromosomes. By the time window size becomes sufficiently small to engender computational limitations, the majority of the TE families identified as spatially relevant continue to be identified as such on each chromosome.

**Figure 2.5:** Adjusted degrees of freedom versus window size in RDAs of the *Bos taurus* genome. The higher adjusted degrees of freedom (# windows - # PCNMs) the greater the adjustment used to calculate the $R^2_{\text{adj}}$. Each coloured line represents one of the 30 chromosomes in the *Bos taurus* genome.
By contrast, those TEs that are not considered significant in terms of spatial structure at certain window sizes are typically the ones that had the lowest explanatory power initially. This robustness notwithstanding, it would still be advisable to implement the analysis multiple times with different window sizes when dealing with previously unstudied genomes as a matter of best practice.

2.4.2 Patterns of Transposable Element Distribution

The results of the present analyses indicated that spatial patterns explain ~50% of the variation between TE communities in each of 11 distantly related genomes and that larger chromosomes exhibit more spatially structured TE communities. This consistency suggests that there are some common factors influencing the locations of TEs within a given genome. The cause(s) of the observed spatial patterns is still not completely clear, however our evidence suggests that the genomic environment itself may play some role. This is shown by TEs with similar abundances in different genomes being spatially structured in one genome, but not in another (Figure 2.6a). These differences in the amount of spatial structure in different genomes may indicate that the same TEs in different genomes may not be showing the same patterns because they are responding to their environment – in this case, the genome. Although a genome’s properties were not related to the amount of spatial organization of that genome’s TE community, it was related to the composition of the TE community. This is shown in the difference between the number and identity of the TE superfamilies organized by a spatial pattern in different genomes. For example, organisms that are closely related phylogenetically have similar groups of TEs that are spatially structured. The TE superfamilies found in plants were almost all spatially structured, while in mammals only about half of the superfamilies were shown to have some degree of spatial structure (Figure 2.6).
Figure 2.6: TE presence and abundance in 11 genomes. This figure shows the evolutionary distance between each of the genomes, alongside a table showing which TE superfamilies are present in each genome. The size of each square represents the log of the abundance of each TE family. Black squares represent superfamilies that have at least some of their between-community variation explained by spatial patterns and white squares show superfamilies where no spatial pattern was found. In a) the squares are scaled so that the largest square with each superfamily is the same size. This allows for the comparisons of superfamilies that have different higher or lower average numbers across genomes. In b) the squares are normalized so that the largest square in each genome are the same size to allow for comparisons across genomes with vastly different numbers of TEs.
This large-scale difference among taxa could be explained in several ways: 1) a shared history of TE insertions among similar closely related genomes, 2) interactions between TEs and their genomic environment, which are more similar in more closely related species; or 3) some properties of the TEs themselves, with the types of elements differing among taxa. It seems likely that options 1 and 2 are connected, as more closely related species are more likely to have more similar genomic environments than more distantly related species. For example, plants and animals have similar but distinct systems to suppress genome activity (Law & Jacobsen, 2010). As a result of these differences, plants and animals have unique patterns of methylation. In plants, methylation is preferentially found in repetitive areas of the genomes (Feng et al., 2010), including TEs, and the methylation occurs on cytosines irrespective of the sequence surrounding them (Henderson & Jacobsen, 2007). In mammals, methylation is found primarily on CG dinucleotides, and rarely in any other context. This methylation is found throughout the genome and is estimated to be found on ~70-80% of mammalian CG sites (Ehrlich et al., 1982). These differences could affect how and when individual TEs are suppressed, potentially contributing to particular spatial distributions. TE distribution can also be caused by TE-specific properties, such as insertion site preference. These preferences range from TEs that are enriched in specific regions, to those with very specific target sites. TEs that display insertion preference for specific regions include MITEs in genic regions (Wessler et al., 1995; Zhang et al., 2000), Ty5 elements in heterochromatic regions of Saccharomyces (Zou et al., 1996; Zou & Voytas, 1997), Ty3/gypsy elements in the centromeres of plants (Neumann et al., 2011), and the non-LTR elements that maintain chromosome ends in Drosophila (Levis et al., 1993; Pardue & DeBaryshe, 2011a; DeBaryshe & Pardue, 2011). TEs with very specific target sites are often found in various RNA genes, such as Pokey and R2 elements in 28s RNA genes, and the Dada DNA element family, some of which target U6 and U1 snRNA, and various tRNA genes (Kojima & Jurka, 2013). In that regard, spatial structure in TE distributions could reflect the spatial patterns of different insertion sites in different genomes.
2.4.3 3D Analysis and Indicator Species Analysis

By incorporating interaction frequency data from high-resolution HiC data I was able to explain $2/3$ of the variation in the TE community that was explained by our more complete PCNM model. The variation explained by the HiC data is a subset of the PCNM, which generates artificial community abundances in such a way that any spatial pattern between the windows along a chromosome is accounted for. The HiC analysis is a specific subset of these based on the frequencies at which the windows along the chromosome interact. The HiC dataset explaining $2/3$ of the variation means that a large part of the variation in the TE community is explained by the physical closeness of the sections of the chromosome when they are uncondensed in the nucleus. Additionally, half of the community data explained by the HiC data is also explained by genomic subcompartments. Areas of the chromosome that have been classified as the same compartments have been shown to have consistent epigenetic marks, which play a role in how accessible these areas are to specific TEs (Zou & Voytas, 1997; Lippman et al., 2004; Brunmeir et al., 2010). Based on the banding patterns of the HiC analysis, they also form clusters of chromosome loops, the borders of which are physically bound together with CTCF anchor protein (Figure 2.7).
Figure 2.7: Subcompartment diagram. The left panel of this diagram shows the patterns you would expect to see on an interaction map generated on a HiC analysis. The right panel of the diagram shows how subcompartments are made up of chromosome loops, the boundaries of which are bound together with CTCF anchor proteins. Reprinted from (Rao et al., 2014) Figure 1d. with permission from Elsevier
Knowing that genomic subcompartments were able to explain a large amount of the variation in TE communities, I examined the predictive power of this environmental variable on the TE community. Figure 2.4 shows far fewer significant IndVal scores than Phi scores, and that neither of these scores was as high as one would want to see for use as a traditional indicator used in something like biomonitoring. For example, in biomonitoring, a high score could indicate that the presence of a particular species is strongly indicative of a particular pollutant. The differences in the number of significant scores are likely that the TEs tend not to be found exclusively in one environment. The IndVal score weights this specificity more heavily than the Phi score, which is a measure of correlation (Dufrêne & Legendre, 1997). Thus, our results show TE superfamilies that are found in greater abundance in some subcompartments than others. For example, Figure 2.4b shows that Alu elements are found most often in A1 or A1+NA subcompartments on five different chromosomes.

Overall, with the exception of Alu, and the NA compartment, which seems to be associated with satellite DNA, the TE superfamilies identified by the IndVal and Phi analysis were not consistent across chromosomes (Figure 2.4). This inconsistency indicates that the ecological patterns structuring TE communities do not extend to the chromosome level. This indicates that those involved in transposon ecology may need to think of TE communities at a more local scale than that of the genome, which is currently the standard (e.g. see Brookfield, 2005; Mauricio, 2005; Le Rouzic et al., 2007; Venner et al., 2009). Instead, the TE communities may be structured at a smaller scale, with the TEs of a whole chromosome, or a whole genome, being more analogous to a metacommunity, with dispersal occurring between more local communities, or from the metacommunity (the TE population of the chromosome/genome) at large (Holyoak et al., 2005).

2.4.4 Conclusions and Future Directions

By taking an ecological approach to TEs and drawing inspiration from existing ecological methods, I have found that the distribution of ~60% of the TEs, within a diverse set of genomes, is distributed along the chromosome in a detectable pattern. Across these genomes, the TE
superfamilies in which these patterns are detectable are correlated with the phylogeny of the host taxa. In a more focused analysis of the impact of 3D spatial relationships on TE community, I found that a large part of TE community composition was structured by the physical distance between the communities, and the genomic subcompartment the community was found in. From those results, I suggest that, along with producing and examining more high resolution genomic HiC data, in order to more explicitly define the scale of TE communities, those interested in the ecology of the genome should continue to look at community ecology, and perhaps more specifically metacommunity theory, to better understand the distribution of TEs within and between genomes.
### 2.5 Supplementary Tables

**Table S1**: Significant IndVal scores for each TE on each produced by the indicator species analysis

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>IndVal</th>
<th>p-value</th>
<th>TE</th>
<th>SubCompartment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.421272</td>
<td>0.047</td>
<td>DNA/TcMar</td>
<td>NA</td>
</tr>
<tr>
<td>1</td>
<td>0.538299</td>
<td>0.023</td>
<td>SINE?</td>
<td>NA</td>
</tr>
<tr>
<td>2</td>
<td>0.544331</td>
<td>0.001</td>
<td>LTR?</td>
<td>B1</td>
</tr>
<tr>
<td>3</td>
<td>0.412272</td>
<td>0.024</td>
<td>LTR/ERV</td>
<td>A1</td>
</tr>
<tr>
<td>3</td>
<td>0.444923</td>
<td>0.042</td>
<td>SINE?</td>
<td>A1+B2</td>
</tr>
<tr>
<td>3</td>
<td>0.660182</td>
<td>0.004</td>
<td>srpRNA</td>
<td>A1</td>
</tr>
<tr>
<td>4</td>
<td>0.59728</td>
<td>0.002</td>
<td>DNA/hAT?</td>
<td>NA</td>
</tr>
<tr>
<td>6</td>
<td>0.41502</td>
<td>0.005</td>
<td>Satellite</td>
<td>NA</td>
</tr>
<tr>
<td>7</td>
<td>0.396417</td>
<td>0.015</td>
<td>Satellite</td>
<td>NA</td>
</tr>
<tr>
<td>7</td>
<td>0.628799</td>
<td>0.001</td>
<td>Satellite/centr</td>
<td>NA</td>
</tr>
<tr>
<td>8</td>
<td>0.638628</td>
<td>0.005</td>
<td>SINE/tRNA</td>
<td>A1+B2</td>
</tr>
<tr>
<td>11</td>
<td>0.607026</td>
<td>0.002</td>
<td>Satellite/centr</td>
<td>NA</td>
</tr>
<tr>
<td>12</td>
<td>0.615626</td>
<td>0.026</td>
<td>Satellite/centr</td>
<td>NA</td>
</tr>
<tr>
<td>13</td>
<td>0.482719</td>
<td>0.041</td>
<td>scRNA</td>
<td>A1</td>
</tr>
<tr>
<td>14</td>
<td>0.433492</td>
<td>0.004</td>
<td>DNA/TcMar</td>
<td>B1</td>
</tr>
<tr>
<td>16</td>
<td>0.350686</td>
<td>0.026</td>
<td>Satellite/centr</td>
<td>NA</td>
</tr>
<tr>
<td>18</td>
<td>0.675529</td>
<td>0.007</td>
<td>srpRNA</td>
<td>NA</td>
</tr>
<tr>
<td>19</td>
<td>0.423761</td>
<td>0.004</td>
<td>Satellite/centr</td>
<td>NA</td>
</tr>
<tr>
<td>19</td>
<td>0.426401</td>
<td>0.006</td>
<td>Satellite/telo</td>
<td>NA</td>
</tr>
<tr>
<td>20</td>
<td>0.667816</td>
<td>0.026</td>
<td>SINE</td>
<td>A2+B1</td>
</tr>
<tr>
<td>20</td>
<td>0.515964</td>
<td>0.034</td>
<td>srpRNA</td>
<td>B1</td>
</tr>
<tr>
<td>21</td>
<td>0.951528</td>
<td>0.013</td>
<td>DNA/hAT-Tip100</td>
<td>B2+NA</td>
</tr>
<tr>
<td>21</td>
<td>0.892915</td>
<td>0.037</td>
<td>LINE/RTE</td>
<td>B2+NA</td>
</tr>
<tr>
<td>22</td>
<td>0.495763</td>
<td>0.039</td>
<td>DNA/hAT?</td>
<td>B1</td>
</tr>
</tbody>
</table>

**Table S2**: Significant Phi scores for each TE on each produced by the indicator species analysis

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Phi</th>
<th>p-value</th>
<th>TE</th>
<th>SubCompartment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.390022</td>
<td>0.042</td>
<td>DNA/TcMar</td>
<td>NA</td>
</tr>
<tr>
<td>1</td>
<td>0.468183</td>
<td>0.01</td>
<td>SINE</td>
<td>NA</td>
</tr>
<tr>
<td>1</td>
<td>0.438437</td>
<td>0.018</td>
<td>SINE?</td>
<td>NA</td>
</tr>
<tr>
<td>2</td>
<td>0.386531</td>
<td>0.006</td>
<td>DNA</td>
<td>B2</td>
</tr>
<tr>
<td>2</td>
<td>0.390029</td>
<td>0.005</td>
<td>DNA/hAT-Charlie</td>
<td>A1+A2</td>
</tr>
<tr>
<td>2</td>
<td>0.301511</td>
<td>0.05</td>
<td>DNA/TcMar-Pogo</td>
<td>B1</td>
</tr>
<tr>
<td>2</td>
<td>0.31944</td>
<td>0.043</td>
<td>LINE/CR1</td>
<td>A1</td>
</tr>
<tr>
<td>2</td>
<td>0.318877</td>
<td>0.046</td>
<td>LINE/L1</td>
<td>B1</td>
</tr>
<tr>
<td>2</td>
<td>0.490398</td>
<td>0.001</td>
<td>LTR?</td>
<td>B1</td>
</tr>
<tr>
<td>2</td>
<td>0.446805</td>
<td>0.003</td>
<td>Other</td>
<td>A1+A2</td>
</tr>
<tr>
<td>2</td>
<td>0.531889</td>
<td>0.001</td>
<td>SINE/Alu</td>
<td>A1+A2</td>
</tr>
<tr>
<td>3</td>
<td>0.289144</td>
<td>0.008</td>
<td>DNA/hAT-Charlie</td>
<td>A2+B3</td>
</tr>
<tr>
<td>3</td>
<td>0.240699</td>
<td>0.033</td>
<td>DNA/TcMar-Tigger</td>
<td>A1+A2</td>
</tr>
<tr>
<td></td>
<td>Value</td>
<td>p-value</td>
<td>Description</td>
<td>Strain</td>
</tr>
<tr>
<td>---</td>
<td>---------</td>
<td>---------</td>
<td>-----------------------</td>
<td>---------</td>
</tr>
<tr>
<td></td>
<td>0.374202</td>
<td>0.001</td>
<td>LINE/CR1</td>
<td>A1+B1</td>
</tr>
<tr>
<td></td>
<td>0.256213</td>
<td>0.026</td>
<td>LTR/ERV</td>
<td>A1</td>
</tr>
<tr>
<td></td>
<td>0.391048</td>
<td>0.001</td>
<td>LTR/ERVL-MaLR</td>
<td>A2+B3</td>
</tr>
<tr>
<td></td>
<td>0.307281</td>
<td>0.003</td>
<td>Other</td>
<td>A1+A2</td>
</tr>
<tr>
<td></td>
<td>0.521076</td>
<td>0.001</td>
<td>SINE/Alu</td>
<td>A1</td>
</tr>
<tr>
<td></td>
<td>0.270362</td>
<td>0.009</td>
<td>SINE/Deu</td>
<td>A1</td>
</tr>
<tr>
<td></td>
<td>0.295915</td>
<td>0.007</td>
<td>SINE/MIR</td>
<td>B1</td>
</tr>
<tr>
<td></td>
<td>0.45817</td>
<td>0.001</td>
<td>srpRNA</td>
<td>A1</td>
</tr>
<tr>
<td></td>
<td>0.47326</td>
<td>0.001</td>
<td>DNA/hAT</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>0.487134</td>
<td>0.001</td>
<td>DNA/hAT-Charlie</td>
<td>A2+B2</td>
</tr>
<tr>
<td></td>
<td>0.328546</td>
<td>0.026</td>
<td>DNA/TcMar-Mariner</td>
<td>A2</td>
</tr>
<tr>
<td></td>
<td>0.415051</td>
<td>0.001</td>
<td>LINE/CR1</td>
<td>A2+B2</td>
</tr>
<tr>
<td></td>
<td>0.314412</td>
<td>0.033</td>
<td>LINE/L1</td>
<td>A1+NA</td>
</tr>
<tr>
<td></td>
<td>0.379171</td>
<td>0.004</td>
<td>LTR/ERV</td>
<td>B1</td>
</tr>
<tr>
<td></td>
<td>0.44484</td>
<td>0.001</td>
<td>SINE/Alu</td>
<td>A1</td>
</tr>
<tr>
<td></td>
<td>0.411097</td>
<td>0.022</td>
<td>DNA/TcMar</td>
<td>A1</td>
</tr>
<tr>
<td></td>
<td>0.421954</td>
<td>0.007</td>
<td>LTR</td>
<td>A1</td>
</tr>
<tr>
<td></td>
<td>0.475392</td>
<td>0.005</td>
<td>LTR/Gypsy</td>
<td>A1</td>
</tr>
<tr>
<td></td>
<td>0.272772</td>
<td>0.036</td>
<td>DNA/hAT-Blackjack</td>
<td>B1</td>
</tr>
<tr>
<td></td>
<td>0.376929</td>
<td>0.006</td>
<td>Satellite</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>0.281919</td>
<td>0.019</td>
<td>scRNA</td>
<td>A1</td>
</tr>
<tr>
<td></td>
<td>0.405552</td>
<td>0.002</td>
<td>SINE/Alu</td>
<td>A1+A2</td>
</tr>
<tr>
<td></td>
<td>0.327294</td>
<td>0.014</td>
<td>SINE/MIR</td>
<td>A1+B1</td>
</tr>
<tr>
<td></td>
<td>0.277798</td>
<td>0.026</td>
<td>DNA/hAT-Tip100</td>
<td>A2+B3</td>
</tr>
<tr>
<td></td>
<td>0.274149</td>
<td>0.029</td>
<td>DNA/TcMar-Tigger</td>
<td>A2+B1</td>
</tr>
<tr>
<td></td>
<td>0.285805</td>
<td>0.015</td>
<td>LTR/ERV1</td>
<td>B2</td>
</tr>
<tr>
<td></td>
<td>0.280151</td>
<td>0.02</td>
<td>LTR/Gypsy</td>
<td>A2+B3</td>
</tr>
<tr>
<td></td>
<td>0.334762</td>
<td>0.008</td>
<td>Satellite</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>0.470311</td>
<td>0.001</td>
<td>Satellite/centr</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>0.257107</td>
<td>0.046</td>
<td>SINE/tRNA</td>
<td>A2</td>
</tr>
<tr>
<td></td>
<td>0.251641</td>
<td>0.033</td>
<td>tRNA</td>
<td>B2</td>
</tr>
<tr>
<td></td>
<td>0.38914</td>
<td>0.02</td>
<td>DNA/hAT-Charlie</td>
<td>A1</td>
</tr>
<tr>
<td></td>
<td>0.360321</td>
<td>0.049</td>
<td>LINE/CR1</td>
<td>B1</td>
</tr>
<tr>
<td></td>
<td>0.344584</td>
<td>0.05</td>
<td>LTR/ERVL</td>
<td>A1+B3</td>
</tr>
<tr>
<td></td>
<td>0.441672</td>
<td>0.017</td>
<td>LTR/Gypsy</td>
<td>A1</td>
</tr>
<tr>
<td></td>
<td>0.389586</td>
<td>0.028</td>
<td>SINE/Alu</td>
<td>A1</td>
</tr>
<tr>
<td></td>
<td>0.460133</td>
<td>0.008</td>
<td>SINE/tRNA</td>
<td>A1</td>
</tr>
<tr>
<td></td>
<td>0.364473</td>
<td>0.037</td>
<td>DNA/hAT</td>
<td>A1+B1</td>
</tr>
<tr>
<td></td>
<td>0.560383</td>
<td>0.001</td>
<td>LINE/CR1</td>
<td>A1+B2</td>
</tr>
<tr>
<td></td>
<td>0.567816</td>
<td>0.001</td>
<td>LINE/L2</td>
<td>A1</td>
</tr>
<tr>
<td></td>
<td>0.393198</td>
<td>0.022</td>
<td>LTR/ERVL</td>
<td>A1+B2</td>
</tr>
<tr>
<td></td>
<td>0.438822</td>
<td>0.007</td>
<td>LTR/Gypsy</td>
<td>A2+B1</td>
</tr>
<tr>
<td></td>
<td>0.38801</td>
<td>0.027</td>
<td>scRNA</td>
<td>A1+NA</td>
</tr>
<tr>
<td></td>
<td>0.414296</td>
<td>0.018</td>
<td>SINE</td>
<td>B2</td>
</tr>
<tr>
<td></td>
<td>0.506475</td>
<td>0.001</td>
<td>SINE/Alu</td>
<td>A1+NA</td>
</tr>
<tr>
<td></td>
<td>0.489273</td>
<td>0.001</td>
<td>SINE/MIR</td>
<td>A1</td>
</tr>
<tr>
<td>#</td>
<td>Value</td>
<td>Score</td>
<td>Description</td>
<td>Layer</td>
</tr>
<tr>
<td>---</td>
<td>--------</td>
<td>-------</td>
<td>--------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>9</td>
<td>0.432523</td>
<td>0.006</td>
<td>snRNA</td>
<td>B1</td>
</tr>
<tr>
<td>10</td>
<td>0.593752</td>
<td>0.05</td>
<td>DNA/TcMar-Tc2</td>
<td>A1</td>
</tr>
<tr>
<td>10</td>
<td>0.758431</td>
<td>0.017</td>
<td>LINE/L1</td>
<td>A1</td>
</tr>
<tr>
<td>11</td>
<td>0.275413</td>
<td>0.042</td>
<td>LTR?</td>
<td>NA</td>
</tr>
<tr>
<td>11</td>
<td>0.429624</td>
<td>0.002</td>
<td>LTR/ERVL-MaLR</td>
<td>B3</td>
</tr>
<tr>
<td>11</td>
<td>0.360884</td>
<td>0.009</td>
<td>Satellite/centr</td>
<td>NA</td>
</tr>
<tr>
<td>11</td>
<td>0.585723</td>
<td>0.001</td>
<td>SINE/Alu</td>
<td>A1</td>
</tr>
<tr>
<td>11</td>
<td>0.30542</td>
<td>0.018</td>
<td>srpRNA</td>
<td>A1+B2</td>
</tr>
<tr>
<td>13</td>
<td>0.360202</td>
<td>0.04</td>
<td>LTR/ERVK</td>
<td>A1</td>
</tr>
<tr>
<td>13</td>
<td>0.388481</td>
<td>0.032</td>
<td>scRNA</td>
<td>A1</td>
</tr>
<tr>
<td>14</td>
<td>0.256536</td>
<td>0.034</td>
<td>DNA?</td>
<td>B1</td>
</tr>
<tr>
<td>14</td>
<td>0.34768</td>
<td>0.004</td>
<td>DNA/TcMar</td>
<td>B1</td>
</tr>
<tr>
<td>14</td>
<td>0.338939</td>
<td>0.005</td>
<td>LINE/CR1</td>
<td>A2+B2</td>
</tr>
<tr>
<td>14</td>
<td>0.26431</td>
<td>0.034</td>
<td>LINE/RTE</td>
<td>A1+B2</td>
</tr>
<tr>
<td>14</td>
<td>0.248342</td>
<td>0.047</td>
<td>LTR/ERVL</td>
<td>A1+B1</td>
</tr>
<tr>
<td>14</td>
<td>0.25564</td>
<td>0.027</td>
<td>LTR/ERVL?</td>
<td>B1</td>
</tr>
<tr>
<td>14</td>
<td>0.280897</td>
<td>0.024</td>
<td>LTR/ERVL-MaLR</td>
<td>A1</td>
</tr>
<tr>
<td>14</td>
<td>0.230319</td>
<td>0.05</td>
<td>rRNA</td>
<td>B1</td>
</tr>
<tr>
<td>14</td>
<td>0.265871</td>
<td>0.021</td>
<td>Unknown</td>
<td>B1</td>
</tr>
<tr>
<td>15</td>
<td>0.308691</td>
<td>0.018</td>
<td>DNA/PiggyBac</td>
<td>A2</td>
</tr>
<tr>
<td>15</td>
<td>0.26474</td>
<td>0.039</td>
<td>LTR/ERVK</td>
<td>A2+NA</td>
</tr>
<tr>
<td>15</td>
<td>0.376674</td>
<td>0.005</td>
<td>SINE/Alu</td>
<td>A2+NA</td>
</tr>
<tr>
<td>16</td>
<td>0.262042</td>
<td>0.027</td>
<td>DNA/hAT-Blackjack</td>
<td>A2</td>
</tr>
<tr>
<td>16</td>
<td>0.23774</td>
<td>0.022</td>
<td>Satellite</td>
<td>NA</td>
</tr>
<tr>
<td>18</td>
<td>0.535679</td>
<td>0.005</td>
<td>srpRNA</td>
<td>NA</td>
</tr>
<tr>
<td>19</td>
<td>0.375373</td>
<td>0.007</td>
<td>Satellite/centr</td>
<td>NA</td>
</tr>
<tr>
<td>19</td>
<td>0.346194</td>
<td>0.009</td>
<td>Satellite/telo</td>
<td>NA</td>
</tr>
<tr>
<td>20</td>
<td>0.534522</td>
<td>0.049</td>
<td>SINE?</td>
<td>A2</td>
</tr>
<tr>
<td>21</td>
<td>0.825999</td>
<td>0.048</td>
<td>DNA/hAT-Blackjack</td>
<td>A2</td>
</tr>
<tr>
<td>22</td>
<td>0.376163</td>
<td>0.019</td>
<td>DNA?</td>
<td>A1+B1</td>
</tr>
<tr>
<td>22</td>
<td>0.349662</td>
<td>0.027</td>
<td>DNA/hAT?</td>
<td>B1</td>
</tr>
<tr>
<td>22</td>
<td>0.383177</td>
<td>0.038</td>
<td>DNA/hAT-Blackjack</td>
<td>B1</td>
</tr>
<tr>
<td>22</td>
<td>0.378181</td>
<td>0.017</td>
<td>LTR/ERVK</td>
<td>A1+NA</td>
</tr>
<tr>
<td>22</td>
<td>0.388074</td>
<td>0.02</td>
<td>snRNA</td>
<td>B2</td>
</tr>
</tbody>
</table>
Chapter 3: Transposable Element Traits and Their Influence on Transposable Element Communities

3.1 Introduction

As much as 85% of all eukaryotic genomes are made up of mobile DNA segments (Schnable et al., 2009) which increase their copy numbers through transposition. Although these transposable elements (TEs) were discovered as a source of mutation in corn (McClintock, 1950), and TE insertions tend to have a negative impact on fitness (Belancio et al., 2008; Nelläker et al., 2012), current evidence suggests that individual elements range on a continuum from parasitic (Doolittle & Sapienza, 1980; Orgel & Crick, 1980) to mutualistic (Kidwell & Lisch, 2001), with the majority being neutral or nearly neutral (Lynch & Conery, 2003, 2004; Pasyukova et al., 2004; Lynch, 2007, 2011). In addition to this range in life history strategies, there is also a large variation in abundance and diversity of TEs in different genomes, ranging from 1 to 39 superfamilies within the same genome (Elliott & Gregory, 2015a).

3.1.1 TE Classification Systems

Understanding the large and diverse TEs communities found in different genomes requires a comprehensive TE classification system. Two of the most current classification systems are the one proposed Wicker et al., (2007) and the tripartite classification system proposed by Arkhipova, (2017). The Wicker system, which builds on the most prevalent system used by RepBase by splitting several families based on more up to date phylogenetic information, organizes TEs in a hierarchal system analogous to a phylogeny based on TE traits such as the proteins the TEs use, the overall genetic structure of the TE, and the target site duplication (TSD), which are duplications of host DNA sequence leftover by imperfect transposition (Wicker et al., 2007). In this system, TEs are first organized into Class I and Class II elements based on their mode of transposition. Class I elements or copy-and-paste elements transpose via retrotransposons, which use an RNA intermediate, and is distantly related to the
mechanism viruses use to transpose. Class II or cut-and-paste-elements transpose via transposase, which has a DNA intermediate and often relies on host repair of the gap left by their transposition to increase in copy number. The second hierarchical level, order, is determined by major differences in transposition mechanism within each class. Within orders, superfamilies group TEs based on consistent topology, structural features, and the size and consistency of TSDs. Within superfamilies, individual TE insertions can be grouped into families based on sequence similarity. Finally, within family, subfamilies are often used to distinguish autonomous elements, which are able to transpose themselves, from non-autonomous elements, in which some of the proteins needed for that element to copy itself are degraded or missing, so they need proteins from autonomous elements to transpose (Wicker et al., 2007) (Figure 1.1).

The second system used to classify TEs is the Tripartite classification system, which does away with the evolutionary hierarchy, and uses three-letter codes to represent each TE family. Each of the letters represents a trait of a TE family with the first being transposition mechanism, the second being the proteins used for integration (integrase/endonuclease), and the third being any additional structural proteins (Arkhipova, 2017) (Figure 1.2). In both systems, there can be many individual copies of each TE family, often determined using the 80-80-80 rule, which states that copies within a family are at least 80% similar over 80% of their length and at least 80bp long.

3.1.2 Why TE Traits?

In an effort to begin explaining the observed variation in TE communities, previous work (Linquist et al., 2013, 2015; Saylor et al., 2013) built a framework for differentiating between ecology and evolution of the genome, and demonstrated that variation in the distribution and abundance of TEs could be explained using an explicitly ecological framework. Evolution will also play a role but is not always required to explain observed patterns. This is different from mainstream TE thinking, which tends to think of TEs in terms of co-evolution with the host. In this thesis I distinguish classical communities of organisms from our genomic TE communities, which consist of each individual copy of a TE within a window of bases on a chromosome. I also
distinguish the classical organismal environment from the genomic environment where TEs are found, which consists of the TEs themselves, the rest of the DNA/RNA in the genome, and any proteins in the genome.

Chapter 2 builds on this notion of studying TEs within the framework of the ecology of the genome, and I showed that ~50% of the variation in the communities of TE superfamilies could be explained by where on the chromosome a TE community was located. In the human genome, 20% of the variation in TE composition between TE communities was explained by the 3D structure of the chromosome, and what type of genomic environment the community was in. This demonstrated that the genomic environment of the TE community was important in structuring its community composition.

Community ecology also influenced the next important step in the analysis of spatial TE patterns. As Legendre et al. pointed out in 1997, community ecologists try to explain the variation in communities in two ways: by comparing community composition to properties of the environment, and by comparing community composition to the traits of the species within the communities themselves. Our previous work has established the former, and the effect of TE traits on the structure of TE communities is the focus of this chapter.

3.1.3 Phylogenetic Community Ecology

In community ecology, the effect of organism traits on community composition is examined using a plethora of methods. One popular method is phylogenetic community ecology. This approach relies on the assumption that species that are more closely related phylogenetically tend to also have more similar traits (Webb et al., 2002). Assuming this is true, the phylogeny can then be used as a proxy for the general similarity of all traits. In turn, phylogenetic patterns in each community can be evidence for specific community assembly processes, which are based on traits. This is done by obtaining a phylogeny for all the species that could potentially be interacting (often done by including all measured species across all
communities that are part of the study) and comparing the phylogenetic patterns of all co-occurring species with each community.

The rationale for linking these phylogenetic patterns to co-occurrence is through the concept of the niche. Hutchinson (1957) described the niche of a species as a multidimensional trait space, in which species can persist. The traits in this space can be anything from the properties of the environment to the availability of resources. The extent of niche overlap within a community can be used to help predict which species will co-occur within that community (Webb et al., 2002; Mouillot et al., 2005; Banta et al., 2012). The measurement of both traits and phylogeny are used to represent the extent to which the niche of the species within a community overlap, and from there to predict patterns in species co-occurrence.

Gerhold et al. (2015) recently reviewed the logic and use of this area of research in ecology: ecophylogenetics. They identified several types of possible inferences. Firstly, researchers have used a pattern-as-proxy view (Cavender-Bares et al., 2004; Cavender-Bares et al., 2009; Cadotte & Tucker, 2017; Tucker et al., 2017). In this approach, phylogenetic patterns can be used to detect the underlying assembly processes responsible for structuring the community (Webb et al., 2002), with two main types of patterns. Phylogenetic clustering is seen as evidence of environmental filtering, which occurs when some property of the environment excludes species without a specific trait or set of traits. This causes clustering because closely-related species are more likely to have similar traits, which results in the environment excluding those species which are more distantly related and therefore less similar to one another. Co-occurring species are thus more closely related than would be expected by chance (see Figure 3.1).

The opposite pattern, phylogenetic overdispersion, is seen as evidence of competitive exclusion. Competitive exclusion causes the more closely related species to exclude each other within a community. This exclusion happens because the more similar species will share more of their resources/niche and therefore competition will be stronger among more closely related species, and one species will outcompete the other, resulting in its exclusion from the local
community. This process will cause the species of a community to be less related to each other than would be expected by chance.

These assembly patterns are detected by comparing the phylogeny of each community to the larger phylogeny of all potentially interacting species. Two measures are often used to detect phylogenetic clustering or overdispersion of a community: mean nearest taxon distance (MNTD) and mean pairwise distance (MPD) (Webb, 2000; Webb et al., 2002). MNTD is measured by determining the distance between each species and its nearest phylogenetic neighbour on a phylogeny of all species within each community in the analysis and then taking the mean across all communities. This method biases the detection of patterns towards the tips of the phylogenetic tree, because MNTD may not pass through any deeper nodes at all (Figure 3.1). MPD is calculated by measuring the phylogenetic distance between each pair of co-occurring species in a local community and then calculating the mean. This measure is less biased towards patterns at the tip of the tree, so it can detect patterns deeper in the tree (Miller et al., 2017). Low values of MNTD and MPD indicate a phylogenetically clustered community, meaning the species in the community are more related to each other than would be expected in a random community of the same size. High values of MNTD and MPD indicate phylogenetically overdispersed communities, meaning the species in a community are less related to each other than would be expected in a random community of the same size. If the p-values of MNTD and MPD are neither above 0.95 or below 0.05, then the community is said to be randomly distributed phylogenetically.
Figure 3.1: Calculating mean pairwise distance and mean nearest taxon distance for two communities. Communities 1 and 2 are drawn from a common phylogeny (top). Community 1 is phylogenetically clustered while community 2 is overdispersed. Distance matrices show the distance between each pair of species, measured here as the number of nodes between species pairs (middle). MPD was calculated by taking the average of distance between each pair of species; MNTD was calculated by taking the average distance to the closest neighbor (bottom).
Gerhold et al. (2015) identified seven assumptions needed for this pattern-as-proxy view of phylogenetic community ecology: 1) The phylogeny accurately reflects the species trait differences, 2) A specific trait is needed for a specific ecological function, 3) Trait similarity causes competition, 4) Competition causes exclusion, 5) The system is at rest, so phylogenetic patterns reflect currently ongoing assembly processes, 6) Habitat filtering and competition are alternative processes, and 7) Community phylogenetic patterns depend on local processes. They showed that violation of each of these assumptions has a range of impacts, from affecting the interpretation of the results to invalidating the method.

Of the above assumptions, 1 and 2 are the most important as they are required for the detected patterns to represent meaningful ecological patterns. The remainder of the assumptions have to do with how to interpret the data to get the appropriate assembly process. For instance, phylogenetic clustering and overdispersion have alternate explanations. In addition to environmental filtering, phylogenetic clustering can also be caused by competition. For example, if height is limiting in a plant system, then only the tallest plants, which should be most closely related as they are the most similar, will survive (Mayfield & Levine, 2010). In addition to competitive exclusion, overdispersion can also be caused by convergent evolution of traits for similar niches (Cavender-Bares et al., 2004; Cavender-Bares et al., 2004), and by organisms that are prone to repeated explosive adaptive radiations (Cavender-Bares et al., 2009). These exceptions have resulted in a patterns-as-result view, in which phylogenetic dispersal patterns are more useful in understanding the effect of coexistence on habitat-lineage pools and how it affects present-day species coexistence; this is opposed to the pattern-as-proxy view, in which the results are interpreted as proxies for assembly processes (Gerhold et al., 2015).

While Gerhold et al. (2015) convincingly argued that the ecophylogenetically pattern-as-proxy view is fraught with problems, this view could actually form a very useful framework to study TE traits and community assembly. TE “species” are much simpler compared to ecological species, being made up of between 2-6 (Wicker et al., 2007) genes while even the minimal bacterial genome contains ~473 (Hutchison et al., 2016). This simplicity increases the strength of
phylogeny as a proxy for traits because the phylogeny is based directly on these few traits, and thus the usefulness of the ecophylogenetic analysis is higher in transposon ecology than in traditional ecology. Thus, assumptions 1 and 2 are easily met by TEs and their two trait-based classification systems.

In this chapter, I will use the Drosophila melanogaster (Dmel) genome and its TE communities to study the importance of TE traits on the structure of TE communities. The Dmel genome is well suited to this type of analysis because it is small, has a highly diverse TE community that has a greater proportion of active elements than most other host species (Eickbush & Furano, 2002; Barrón et al., 2014) each present in low copy number (Bartolomé et al., 2002). The diversity of the Dmel TE community means that there are more TE types that could be interacting, making the detection of ecological patterns more likely and the high proportion of active TEs means that they are still potentially interacting, which is required for detecting ecological processes.

### 3.1.4 TE Age

In addition to the classification of TEs, I included one other trait that may have some influence on the TE community in the Dmel genome: the average age of the TE family. This is based on a recent ecological model of TEs propagating in a new genome (Kremer et al. in prep), that over time an increasing amount of the variation in TE communities was explained by spatial location on the chromosome (Wei, 2018).

TE age thus captures some of the history of co-evolution between that TE family and the host genome. This could be explained by older TEs having more opportunities to transpose, and if transposition is influenced by the genomic environment, then more transpositions should cause a stronger the association will be between genomic environment and the distribution of that TE superfamily.
The results of Wei, (2018) and Kremer et al. (in prep), combined with the large number of active individual TE copies and the high removal rate of inactive TEs, leads to the prediction that, on average, older TE families will have more spatial structure in their locations. I predict this because newer active elements would not have had the time to accumulate differences in their sequence, so a high turnover of inactive elements maintains a high level of sequence similarity within a TE family. One might then expect that if a family is older (has more sequence divergence within the family), then the individual copies have inserted into niches where they are less detrimental to the host.

In this study, I used community ecology techniques to explore the relationship between TE traits and the TE communities in which they are found. I assessed the impact of community structure on the distribution of TEs within their communities using the two dominant TE classification systems as stand-ins for TE traits and analyzed those communities with tools developed for phylogenetic community ecology. Because TEs adhere to the assumptions of phylogenetic community ecology, I can adopt a pattern-as-proxy view, despite its controversy when used in classical community analysis. I predicted that if TE traits influence the distribution of TEs within the genome, I will detect phylogenetic clustering and/or overdispersions. If co-occurrence of TE copies is influenced by genomic environmental filtering, then I predict TE communities will be phylogenetically clustered. If co-occurrence of TE copies is influenced by competition, then I expect to detect phylogenetic overdispersion. I also predicted that if TEs persisted longer in the Dmel genome because they have found a niche they are well suited to, then older TE families will be more spatially structured.

3.2 Materials and Methods

3.2.1 Genome Sequence Data

The sequences of each chromosome from the 6th release of the Dmel ISO-1 genome were used for this analysis. Dmel is a common model organism in genetics. Consequently, an above average amount of time and money has been invested into understanding Dmel genetics, and into
the sequencing of its genome. This effort has resulted in a genome sequence with well
c charact erized genes and TEs, that has been assembled into full chromosome scaffolds. These full
chromosome scaffolds, which are not often generated in smaller sequencing projects, are
important for the spatial portion of the analysis. Finally, the *Dmel* TE community itself is well
suited for the age analysis. Unlike the TE community in many larger genomes, the TE
community in drosophilids is characterized by a high diversity of TE families each present at low
copy numbers. A genome with high TE copy number and low diversity would be much more
time and resource-intensive to analyze using our method, as the amount of time required to run a
multiple sequence alignment increases non-linearly with increasing numbers of sequences. The
human genome, for example, contains millions of copies on element named Alu and aligning
these millions of copies would be very time-consuming.

### 3.2.2 TE Location Data

The identities and locations of TEs in the *Dmel* genome were identified using
RepeatMasker (Smit *et al.*, 2004) with the default settings and using the *Dmel* repeat library. The
RepeatMasker Family classifications were updated to bring them in line with more modern TE
classification (Kapitonov *et al.*, 2009; Yuan & Wessler, 2011; Vassetzky & Kramerov, 2013).
The “one-code-to-find-them-all” perl script (Bailly-Bechet *et al.*, 2014) was used to reassemble
nested TEs, filter entries for quality, and extract FASTA sequences of each individual TE.
Finally, the FASTA files for the *Dmel* chromosomes 2 and 3 are split into L and R. While these
parts are actually one chromosome, the entries for each TE on 2R and 3R had the length of
chromosomes 2L and 3L respectively added to them. This resulted in one set of TE locations for
each chromosome.

### 3.2.3 TE Communities

Once the TEs were located, each chromosome was split into 144,500bp windows
following Chapter 2. This resulted in a total of 947 windows across 5 chromosomes.
3.2.4 TE Classification

Each TE in the *Dmel* genome was classified by both the Wicker and Tripartite TE classification systems.

3.2.4.1 Hierarchal

A phylogeny of based on the Wicker classification was created manually by adding each TE family found in the *Dmel* genome to a tree constructed as a newick tree file then visualized in R (Figure 3.2). The length of each branch was set to 1.
**Figure 3.2:** Wicker classification tree in *Dmel*. This tree shows the hierarchical breakdown of all the TEs in the drosophila genome classified by the Wicker method.
3.2.4.2 Tripartite

The second classification system is based on the Tripartite trait classification system (Arkhipova, 2017). This system makes no assumptions about the evolutionary history (i.e. phylogeny) of the TEs. Instead, it groups individual TEs based on 3 traits: 1) Method of transposition, 2) Method of insertion, and 3) additional structural proteins (Figure 1.2). The Trait grouping data consists of the 3-letter code associated with each of these traits.

3.2.5 TE Ecophylogenetics

3.2.5.1 Hierarchal Classification System Analysis

The Wicker tree from the hierarchal TE classification section was used to look for evidence of phylogenetic clustering and overdispersion in the TE community. The TE abundances of the TE families that make up each TE community was determined for each TE community by dividing each chromosome into 144kb windows and counting each TE copy in each window. Mean pairwise distance (MPD) and Mean nearest pairwise distance (MNPD) described in (Webb et al., 2002) were calculated using the Picante package in R (Kembel et al., 2010) and using the number of nodes passed through as a measure of distance. This analysis will report p-value instead of effect size because the effect size in this analysis is dependent on the phylogeny of the specific community in each window. p-values above 0.95 indicate phylogenetic overdispersion, and p-values below 0.05 indicate phylogenetic clustering.

3.2.5.2 Classification System Comparison

Redundancy analyses (RDAs) for each chromosome were also run to assess how well each of the classification systems described the TE community of that chromosome (see chapter 2 for details on RDAs). For the Wicker classification system, the independent variables were the effect sizes (z) from calculating MPD and MNTD for classification of each window as either phylogenetically clustered, overdispersed, or no pattern. The z score represents the difference from the expected phylogenetic relatedness of co-occurring species for that specific window. It
was used in this analysis instead of the p-value because the intent is not to say if the TEs in that window are phylogenetically overdispersed or clustered, but how big the effect in that window is. This was done for both MNTD, which can detect patterns towards the tips of the tree and MPD which is more sensitive to patterns in deeper nodes of the tree. For the Tripartite classification system, the dependent variables were counts of the numbers of each trait label in each window.

3.2.6 TE Age

3.2.6.1 Sequence Alignment

The consensus sequences for each TE in *D. melanogaster* were extracted from the Repbase RepeatMasker library using RepeatMasker’s extract taxon tool. LTR elements with separate entries were combined so that the terminal inverted repeats flanked the internal sequence.

For each TE the consensus sequence was aligned to each TE copy of the same element using the pairwise Alignment function in R’s Bioconductor package (Ramos et al., 2017). The distance between each copy and the consensus sequence for the family that copy was a member of was then calculated using the dna.dist function in the R’s ape package (Paradis, 2012). The distance was calculated using the Kimura 80 distance measure ignoring sites that were not in both sequences (Kimura, 1980). This was done to provide the most conservative information for TE fragments, which would have lower scores depending on their size if missing sites were included. This resulted in a distance score between each copy of each TE family.

3.2.6.2 Measures of TE Age

The pairwise distances produced from the alignments described above were used to calculate the age of each TE. The distances were summed, and the total was divided by the number of TE sequences in that alignment. This resulted in an average distance per TE, or distance per leaf measurement. Our measure is similar to Faith’s measure of phylogenetic
diversity (Faith, 1992), calculated without making a tree. This method was used over simpler methods, such as maximum distance, as those methods can obscure some of the variation caused by back mutations, and sequences that mutate in different ways (Felsenstein, 2004; Christin et al., 2012). This method also allowed us to compare across TEs with different numbers of sequences. The spatial structure within each TE family was measured by running a redundancy analysis (RDA) for each TE family and taking the average $R^2_{\text{adj}}$ across all chromosomes. Pearson correlation coefficients were then calculated between $R^2_{\text{adj}}$ and TE age for each TE family.

3.3 Results

3.3.1 TE Locations and Classification

A total of 50,270 individual TE copies were identified in the Dmel RepeatMasker files. After running the program, one code to find them all (Bailly-Bechet et al., 2014) there were 45,634 TE copies, of which 25,889 were full length. After removing TE sequences not assigned to a chromosome scaffold there were 17,523 elements covering 26% of the Dmel genome. After correcting LTR entries with one code to find them all there was a total of 143 TEs families distributed among 14 TE superfamilies. The results of classifying each of the TE families in Dmel by the Wicker classification can be found in Figure 3.2.

The Dmel TE families were also classified according to the Tripartite method (Figure 1.2) This resulted in the 143 families being divided into five character-trait groups, the DD- class had 29, the DH- class had 2, the RAN class had 38, the RDN class had 73 and the RPN class had 1 (Table 3.1).
Table 3.1: Breakdown of all the TE families from *Dmel*, classified into groups according to (Arkhipova, 2017)

<table>
<thead>
<tr>
<th>Count of Element</th>
<th>Integrate/endonuclease</th>
<th>D</th>
<th>H</th>
<th>P</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transposase</td>
<td>A</td>
<td>29</td>
<td>2</td>
<td></td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>38</td>
<td>73</td>
<td>1</td>
<td>112</td>
</tr>
<tr>
<td>Grand Total</td>
<td></td>
<td>38</td>
<td>102</td>
<td>2</td>
<td>143</td>
</tr>
</tbody>
</table>

3.3.2 Community Phylogenetics

Classification of the TE families in the *Dmel* genome by the hierarchal Wicker classification system (Figure 3.2) allowed the measurement of MNTD and MPD for the TE community of each window in the *Dmel* genome. Figure 3.3 shows the *p*-values for each window along each chromosome. *p*-values below 0.05, which indicate the TE community of that window is more phylogenetically clustered than would be expected by chance, are found mostly at the ends of chromosomes 2 and 3. *p*-values above 0.95 indicate communities that are more overdispersed than would be expected by random chance and are found across the whole chromosome, with a higher concentration toward the center.

MNTD detected a non-random phylogenetic species distribution pattern in 103 of *Dmel*’s 947 windows. Of those, 100 (97%) were overdispersed, and 3 (3%) were phylogenetically clustered. These phylogenetically non-random windows were predominantly found on chromosomes 2, 3, and X. Both of the windows with a non-random pattern on the Y chromosome were phylogenetically clustered, while the majority of the non-random windows on chromosome 2 and all of the non-random windows on chromosomes 3 and X were overdispersed (Figure 3.3 and Table 3.2).
MPD detected a non-random phylogenetic species distribution pattern in 87 of *Dmel*’s 947 windows. Of those, 62 (71%) were overdispersed, and 25 (29%) were phylogenetically clustered. These windows were predominantly found on chromosomes 2, 3, and X, however, non-random windows were found on each chromosome. All 6 of the non-random windows on the Y chromosome were phylogenetically clustered, while all the 4 non-random windows on chromosome 4 were overdispersed.

MNTD primarily detected overdispersed windows, while phylogenetically clustered windows were detected by both MPD and MNTD. Forty-one overdispersed windows were identified by both measures, with the remainder only identified by one analysis. Of the phylogenetically clustered windows only one was identified by both measures. Moreover, only one window was identified as overdispersed by one method and clustered by the other.

The distribution of TEs in non-random windows can be seen in Figure 3.5. TEs are present in overdispersed communities (Figure 3.4a) at a relatively equal rate across the phylogeny, while in phylogenetically overdispersed communities (Figure 3.4b), there is more variation in how often a given TE is present, and some TE families are present in more windows than others.
Figure 3.3: Mean nearest taxon and mean pairwise distance along each of the chromosomes of *Dmel*. The X axis in this figure is the index of each window numbered sequentially along the chromosome. The observed P is the result of the MNTD or MPD analysis implemented by picante in R, with orange points showing values for MNTD and blue points showing values for MPD. The dashed lines at the top and bottom of the figure are at .95 and .05 respectively. A $p$-value above .95 indicates that the TE community in that window is more phylogenetically overdispersed than would be expected in a randomly generated phylogeny with the same number of species. A $p$-value below .05 indicates that the TE community in that window is more phylogenetically clustered than would be expected in a randomly generated phylogeny with the same number of species.
Table 3.2: Number of phylogenetically overdispersed and clustered windows identified along *Dmel* chromosomes using MNTD and MPD.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>MNTD</th>
<th></th>
<th></th>
<th>MPD</th>
<th></th>
<th></th>
<th># of windows</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overdispersed</td>
<td>Clustered</td>
<td>Total</td>
<td>Overdispersed</td>
<td>Clustered</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>43</td>
<td>1</td>
<td>44</td>
<td>21</td>
<td>8</td>
<td>29</td>
<td>336</td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>0</td>
<td>40</td>
<td>22</td>
<td>9</td>
<td>31</td>
<td>415</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>X</td>
<td>17</td>
<td>0</td>
<td>17</td>
<td>15</td>
<td>2</td>
<td>17</td>
<td>162</td>
</tr>
<tr>
<td>Y</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>6</td>
<td>6</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>3</td>
<td>103</td>
<td>62</td>
<td>25</td>
<td>87</td>
<td>947</td>
</tr>
</tbody>
</table>
Figure 3.4: Counts of clustered and overdispersed TE communities in *Dmel*. This figure shows counts of how many times each TE was present in a phylogenetically overdispersed (a) or clustered (b) TE community in *Dmel*. The panel on the left of each figure is the phylogeny of TEs in *Dmel* genome classified by the Wicker classification system. The size of the bars in each column represents the counts of each TE family. The first column to the right of the phylogeny shows the pattern across the whole genome, while the subsequent columns show the results for each chromosome. Each histogram is scaled by relative abundance, so the maximum value in each histogram spans the whole column. Histogram bars are alternate colours by TE family.
3.3.3 Comparison of Hierarchical and Tripartite Classification Systems

After using RDAs to assess the relationship between our classification systems and the distribution of the TE community, I found the Tripartite trait-based classification system was able to explain much more of the TE community than the detection of community patterns based on the Wicker hierarchical classification system (Table 3.3). The Tripartite system was able to explain an average of \(~70\%\) of the variation in TE community (p < 0.001 on all chromosomes), while the effect size of MNTD and MPD, based on the Wicker classification system, was an average of .024 and .006 when the relationship was significant (Table 3.3).

Table 3.3: Redundancy analysis results comparing community phylogenetics and non-hierarchal trait results to community composition in Dmel. The traits column contains the results for the RDA of the TEs in Dmel classified by the Tripartite classification system, and the remaining column are the results of RDAs from various statistics of MPD and MNTD produced by the community phylogenetics analysis. * p<0.05 ** p<.01 *** p<.001

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Tripartite Traits</th>
<th>MPD Z Score</th>
<th>MNTD Z Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM_Chr2</td>
<td>0.740 ***</td>
<td>0.013 *</td>
<td>0.044 *</td>
</tr>
<tr>
<td>DM_Chr3</td>
<td>0.688 ***</td>
<td>0.034 ***</td>
<td>0.004</td>
</tr>
<tr>
<td>DM_Chr4</td>
<td>0.882 ***</td>
<td>-0.113</td>
<td>-0.014</td>
</tr>
<tr>
<td>DM_ChrX</td>
<td>0.757 ***</td>
<td>0.035 **</td>
<td>0.003</td>
</tr>
<tr>
<td>DM_ChrY</td>
<td>0.477 ***</td>
<td>0.060 **</td>
<td>0.082 **</td>
</tr>
<tr>
<td>Average</td>
<td>0.709</td>
<td>0.006(0.036)</td>
<td>0.0240.024(0.063)</td>
</tr>
</tbody>
</table>

3.3.4 TE Age

Average distance per leaf shows a gradual distribution of TE ages and has a stronger positive relationship between age and average $R^2_{adj}$ (Figure 3.5).
Figure 3.5: Average age of each TE family vs $R^2_{adj}$ of the amount of spatial variation explained by RDA. This figure shows the average distance between the sequence of each TE copy and each of the consensus sequence for that family versus the average of the $R^2_{adj}$ values for that family across each chromosome p-value = 2.037e-06, cor=0.4387538.

3.4 Discussion

In this study, I used community ecology methods to determine the effect of TE traits on the makeup of TE communities. I did this by examining two traits: how that family is classified, which serves as a proxy for their combination of traits, and the age of a TE family.
3.4.1 Community Phylogenetics and TE traits

Given the correlation between TE community and genomic environment (niche) seen in Chapter 2, one might expect that the properties of the TEs could have a role in determining the observed distribution of TEs within a genome. I found evidence for the role of TE traits in the structuring of TE communities. I observed that 15-17% (Table 3.2) of the TE communities within the Dmel genome are either phylogenetically clustered or overdispersed.

Overdispersion was detected by both MPD and MNTD across each of the Dmel chromosomes. As these measures are sensitive to deeper and shallower patterns in the metacommunity tree, respectively, this indicates that TEs are more different than would be expected by chance both within similar groups of TEs and between groups of TEs with very different traits.

Within TEs there is ample evidence for potential competition among similar elements. Evidence of competition among TEs includes population models, examples of lack of co-occurrence, and mechanisms of differential expression. Simulated population models of autonomous and non-autonomous element dynamics have shown that increasing the number of non-autonomous elements decreases the replication of autonomous elements (Le Rouzic & Capy, 2006).

There are numerous examples of lack of co-occurrence of similar elements: Early radiolabelling experiments in Dmel observed that mdg-1 and P elements rarely occurring near each other, while mdg-1 and Copia elements were often found together (Biemont & Gautier, 1989). In the South African frog, Xenopus laevis, the retroelements Tx1L and Tx2L are found within the DNA elements Tx1D and Tx2D respectively. Despite the fact that the endonuclease retroelements used to insert into the DNA elements will bind to both targets, Tx1L has never been found in T1x2D and Tx2L has never been found in T1x1D (Christensen et al., 2001). More recent evidence within Dmel showed that R1 and R2 retroelements, which share the same target site in ribosomal DNA, avoided inserting into sites which already contained the other element.
(Ye et al., 2005). Molecular evidence showed that this is due to the rearrangement of the nucleosome structure, limiting access to the machinery required for insertion (Ye et al., 2005).

Mechanisms of differential expression between TEs allow one TE to increase its own replication rate or decrease the replication rate of other TEs. Evidence for mechanisms of differential expression between TE families includes eukaryotic MITE elements, as well as the interaction of Hermes and hobo elements in insects. Mite elements have two mechanisms of boosting transcription rate. First, they have internal enhancers, which increase rates of expression. Second, they can use the transposase of distantly related TEs for their own expression (Sundararajan et al., 1999). Another example of TEs affecting the expression of their competitors is found between the human Alu and LINE L1 elements. The Alu element physically competes with its autonomous counterpart L1 for replication machinery (Ahl et al., 2015), a direct example of TEs competing over a shared resource.

Given the number of examples and mechanisms of competition within and between TE families, it seems reasonable that the observed overdispersion within the Dmel genome could be caused by what is known in classical ecology as competitive exclusion.

The second pattern detected in the communities of Dmel was phylogenetic clustering. Phylogenetic clustering in the Dmel genome is restricted to the center of chromosomes 2, 3 and the Y chromosome. In chromosomes 2 and 3 this pattern is detected by MPD but not by MNTD. This indicates that the clustering occurs deeper within the tree, i.e. closer to the root of the tree. This can be seen in Figure 3.4, which shows that Class I elements, which are grouped closer to the root of the tree, are more common in these phylogenetically clustered communities. There is a well-known relationship between similar TEs occurring in the same region, for example in the Dmel telomeres. In most organisms, telomeres act as buffers at the end of the DNA, which can be degraded over time as the cell divides. The enzyme telomerase is responsible for maintaining the telomeres by extending them, to stop them from degrading fully. However, Dmel has no telomerase, instead the telomere region is maintained by a set of Class I Jockey elements, namely TART, TAHRE, and HETA (Traverse & Pardue, 1988; Levis, 1989; Biessmann et al., 1990a,b,
1992; Levis et al., 1993; Biessmann & Mason, 1997; Pardue & DeBaryshe, 2003, 2011a; Villasante et al., 2007). These 10-50kb regions at the end of each chromosome are not normally part of the sequenced region, as they are variable in length over the lifetime of each cell and highly repetitive, which makes them difficult to sequence accurately. They do, however, show up on the smaller and easier to sequence Y chromosome, which is known to be highly repetitive. Jockey is a relatively large TE family, found throughout drosophilids. Unlike TART, TAHRE, and HETA, most Jockey elements are not involved in telomere maintenance, and their transcripts localize to the cytoplasm (Pardue & DeBaryshe, 2011a). Some of these Jockey elements are found to occur more often in the windows with evidence for clustering than in windows with evidence for overdispersion (Figure 3.4). The specificity of these few Jockey elements, and their cooption to replace telomerase suggests a likely co-evolutionary history between these elements and their host.

Despite the evidence for a co-evolutionary relationship, a closer look at the interactions between TART, TAHRE, and HETA reveals that this system could be maintained by an ecological mechanism, mutualism. All three of these TEs are predominantly found in telomeres, but HETA is the only one of these TEs that can target itself to the telomeric region (Pardue & DeBaryshe, 2011a,b). TAHRE and TART can only insert themselves into the telomere region by interacting with the HETA replication machinery, which recruits TAHRE and TART to the telomere region. This appears to be parasitic at first glance, however, HETA may use parts of the TAHRE and TART replication machinery in addition to its own, allowing it to replicate more frequently (Pardue & DeBaryshe, 2011a,b). As there is a known mechanism that causes these Class I elements to cluster at the ends of the chromosome, it seems reasonable that the regions proximal to the telomeres may be a preferred niche for other Class I elements, and that it is plausible that a combination of both evolutionary and ecological mechanisms could maintain this pattern.

As discussed above, the community phylogenetics methods used here have their limitations in classic ecology. The application of these methods to TE communities avoids some
of these issues, but the nature of the TEs in *Dmel* also poses some other limitations. The primary one is that the phylogeny generated using the Wicker classification system (Figure 3.2) contains many polytomies and is not evenly distributed (the Gypsy family makes up a disproportionate part of the tree). As a result of both the polytomies and the uneven distribution of TEs, an increased number of TE communities will be identified as phylogenetically clustered. This makes the number of communities identified as overdispersed conservative. These limitations likely had little effect on the communities identified as phylogenetically clustered, as the pattern appears consistent in Class I transposons outside the Gypsy family (Figure 3.4). An additional limitation of the number of polytomies in the analysis would be to limit the power of MNTD to detect patterns. This occurs because MNTD preferentially detects patterns occurring at the tips of the tree, and the lack of species separation caused by polytomies would make these patterns more difficult to detect.

Within ecology the use of phylogenetic community ecology is not without some criticism. As discussed in the introduction, there are very important assumptions required to make this type of inference (Mayfield & Levine, 2010; Gerhold *et al.*, 2015). I suggested that TE ecology fits assumptions 1 and 2, which have to do with the relationship between traits and phylogeny, better than classical ecology. The remainder of the assumptions do not violate the validity of the method; however, they do impact the interpretation of the results.

In light of the TE community’s history having been encoded in the genome, and the complete nature of the *Dmel* genome sequence, the scale-based criticisms also hold less weight in TE communities than in classical organismal communities. This record of the TEs history within the genome makes assumptions 5 and 7 more likely. The effect of local vs global patterns in communities are also less of a problem in the genome, where most of the genome has the potential to interact. The assumption that the system is at rest, and that communities may not have caught up to reflect the processes currently responsible for assembly is less important when the history of the community is still there, and the age of each TE can be determined by comparing its sequence to the sequences of other copies of those same TEs. This allows one to
look at patterns in the community throughout the history of the TEs, so those assembly processes should always be detectable.

3.4.2 TE Age

The TE community in the *Dmel* genome is characterized by a large number of active TEs, each of which is present in low copy number. This is thought to be caused by a high TE turnover rate (Petrov et al., 2003). Given recent evidence that there is a structural pattern in TE community distribution (Saylor et al., 2013), I predicted that the TE families that survive the longest are those that either dispersed to a place where they are less likely to be removed, or have a preference for a niche where they are less likely to be deleted. I found evidence to support this, with older TEs having more structure to their communities in each of the *Dmel* chromosomes.

I predicted that older TEs will have more spatially structured communities than younger TEs. As shown in our previous work (Chapter 2), ~54% of the distribution of the TE community the *Dmel* genome shows evidence of spatial structure, and much of this structure is correlated with the type of environment in which those TEs are found. Distance per leaf measures the average amount of divergence between two TEs. I found that distance per leaf was significantly correlated to spatial structure (Persons correlation coefficient is 0.43 vs 0.21) (Figure 3.5).

In *Dmel*, there are several elements that are preferentially located in different areas of the genome and thus could help explain this pattern. For instance, some elements are preferentially found in heterochromatic regions of the genome. TART, THARE, and HETA are exclusively found in the *Dmel* telomeres, which is the heterochromatic regions at either end of each chromosome (Traverse & Pardue, 1988; Biessmann & Mason, 1997; Pardue & DeBaryshe, 2011b). Other elements show the opposite distribution, preferentially being found in gene-rich euchromatic regions. For example, The P element shows such a strong preference for inserting into gene regions that it was used in the Drosophila gene disruption project to target mutations to Drosophila genes (Bellen et al., 2011).
The correlation between age and how structured these TEs are spatially in the genome may reflect either a co-evolutionary strategy in which niche specialization improves the longevity of that TE within the genome by avoiding important genic regions, or ecological habitat filtering, in which some TEs are able to better take advantage of some niches based on their traits. This is especially true in the Dmel genome, which has a high rate of deletion/turnover and a large variety of active TEs in low copy number. The combination of low copy number and high deletion means that TEs are in increased danger of being deleted if they do not either find a niche where they are safe from deletion or increase in copy number faster than they can be deleted.

Of the seven assumptions Gerhold et al., (2015) proposed to properly interpret phylogenetic community ecology patterns as assembly processes, I have addressed assumptions 1,2,5, and 7. Assumptions 3, 4 and 6 have to do with interpreting which assembly process can be linked to what observed phylogenetic patterns. In classical ecology this is often linked to scale (see Cavender-Bares et al., 2009 and Emerson & Gillespie, 2008 for review). As the scale of the study gets larger, the processes likely causing the observed patterns in community structure move from density-dependent interactions (such as competition, mutualism, parasitism) to environmental filtering, to biogeographical processes (speciation and extinction). This scale pattern holds for spatial, temporal and phylogenetic scales. Spatially density-dependent interactions such as competition may structure communities at small neighbourhood scales. As the scale increases more heterogeneous environments are included, thus communities might be expected to be structured by environmental filtering, with more closely related species being able to tolerate more similar environments (Cavender-Bares et al., 2009; Mouquet et al., 2012). Finally, as the scale gets large enough, biogeographical processes, such as the regional species pool, begin to structure the community (Cavender-Bares et al., 2009). Time scales function along a similar continuum, with the biogeographical mechanism being most important at longer time scales. Phylogenetically, competition will be most evident among closely related species, and as the phylogeny gets larger more diverse, inter-clade differences may overwhelm these interspecific processes (Emerson & Gillespie, 2008; Cavender-Bares et al., 2009).
3.4.3 Hierarchal Classifying vs Non-hierarchal Classification

When the two classification methods are compared, the Tripartite method explains each TE community better than our derivatives of Wicker’s hierarchal method (Table 3.3).

The Tripartite method’s average $R^2_{adj}$ was 0.71, meaning that abstracting from 143 conventional TE families to 6 different combinations of three functional TE traits still captured much of the same information. This classification captured an average of 71% of the variation in TE family abundances between different windows (TE communities) between chromosomes. This much simpler way of describing the community reduced the number of TE labels by 96% while only losing 29% of its descriptive power. This suggests that it likely performs well in describing TEs using traits relevant to TE function.

The derivatives of Wicker’s method were far less effective in describing the TE communities (Table 3.3). Neither MPD or MNTD were significant on all the chromosomes, and $R^2_{adj}$ was an order of magnitude lower than the Tripartite system (Table 3.2).

MPD was significant on Chromosomes 2, 3, X, and Y. This is to be expected as these are the chromosomes which had the most communities that were phylogenetically non-random (Table 3.2), and which had the most obvious patterns spatially (Figure 3.3), with phylogenetically clustered groups at the ends and overdispersed patterns in the middle. It is in fact surprising that community assembly patterns were correlated to spatial structure at all, as the ability of the Wicker classification to explain spatial patterns needed another step of analysis, the community phylogenetics analysis, to be correlated to the results of the spatial analysis. On the other hand, it was possible to compare the classification from the Tripartite system to the spatial patterns directly.

MNPD detected significant patterns on only two chromosomes and detected more phylogenetically non-random communities than MPD (Table 3.2). As mentioned above, MNPD is best at detecting phylogenetic patterns at the tips of the phylogeny. So, finding community
structure correlated to MNPD on fewer chromosomes further supports the idea that the patterns in community structure observed here are occurring deeper in the phylogeny, i.e. at the family/superfamily level.

Overall, our results suggest that the Tripartite method successfully classifies the diversity of the TE communities in Dmel, however, the hierarchical structure of the Wicker classification is more readily applicable in methods designed for phylogenetic community ecology, as the Tripartite method is not convertible into a phylogeny.

3.4.4 Conclusions

The aim of this study was to build on the previous work, done in Chapter 2 and in (Saylor et al., 2013), using methods from community ecology to learn about the role of TE traits in structuring TE communities. In doing this I had 2 predictions: 1) Older communities will have more spatial structure than younger ones, 2) There would be some phylogenetic patterns in some TE communities. I confirmed both as 1) Older TEs were more spatially structured than younger ones, and 2) 143 of the 947 TE communities in the Dmel genome had some non-random phylogenetic co-occurrence structure. In detecting these patterns, I found that overdispersion was more common than phylogenetic clustering. This could be evidence of convergent evolution, however, there is also ample evidence of competition among TEs, so these patterns could also plausibly be explained by purely ecological mechanisms. I also observed that phylogenetically clustered communities occurred almost exclusively at the ends of chromosomes. This is interesting as Dmel has an atypical telomere maintenance system, which could plausibly be maintained by ecological processes. These same processes or something about this area of the genome could also favour specific types of TE communities in these areas, causing the observed pattern in of phylogenetic clustering. Future analysis of this type will be needed to investigate if the same type of patterns is detected in other drosophilids with this atypical type of telomere maintenance and if phylogenetic clustering is also found at the ends of chromosomes with more typical telomerases.
TEs are predominantly studied from the perspective of how they evolve/coevolve with the host and even attempts to integrate ecological ideas to TEs often end up conflating TE evolution and TE Ecology (Linquist et al., 2013). Thus, in addition to Gerhold et al. (2015)’s suggestion that community phylogenetic patterns can be an interesting result in their own right, the identification of potentially ecological patterns in TE communities opens up the whole field of TE community ecology, which can in turn draw inspiration from the more well-established field of community ecology.
4 Chapter 4: Transposable Element Traits and Genomic Environments: The 4th Corner in Ecology

4.1 Introduction

4.1.1 Traits and Environments: The 4th corner in Ecology

In the previous chapters of this thesis, I successfully applied community ecology approaches to find patterns between TE traits / the genomic environment and the TE community. Chapter 2 explored the patterns between the TE community and spatial and environmental predictor variables, such as chromatin state; Chapter 3 explored patterns between the TE community and TE traits such as TE age and the proteins responsible for transposition and excision. In this chapter, I test the suitability of another approach derived from community ecology, the 4th corner analysis, for identifying TE trait / genomic environment relationships in genomic TE data.

In ecology, identifying broad trait-environment relationships involves examining large groups of comparable taxa within many varying environments. A common community ecology method for identifying these patterns is called the 4th corner analysis (Legendre et al., 1997; Dray & Legendre, 2008; Ter Braak et al., 2012; Braga et al., 2018). This method analyzes the relationship between species traits and properties of the environment using three data tables. The first is a community matrix which counts the abundance (or presence) of each species (columns) at each site (rows). The second is a matrix of the environmental properties (columns) of each site (rows). The third table consists of each species (columns) and their biological traits (rows). The 4th corner analysis then calculates the “4th corner”, which is a 4th matrix containing the relationships between the environmental properties (columns) and species traits (rows; Figure 1). After computing the correlations between the traits and the environmental properties, the 4th corner analysis tests for the significance of those relationships by permuting the rows and columns of the community matrix. This significance testing is done by permuting the rows and
columns of the community matrix to see the effect on the correlations in the 4th corner matrix. One of the main attractions of this method is, instead of trying to ask how each species in an environment will change when that environment changes, which is a large, difficult task, you can ask how specific traits tend to react in each environment.

The 4th corner analysis has been used to identify many trait-environment relationships in classical ecology. For example, traits associated with surviving without leaf litter are strongly associated with success in post-fire communities of collembola (Huebner et al., 2012). Traits associated with canopy-nesting, ground-nesting, bark- and foliage- insectivores, cavity-nesting and ground foraging birds determine their success in post-fire boreal forest stands (Azeria et al., 2011). Finally, the 4th corner analysis was able to identify species traits related to water pH and turbidity of the water as the best predictors of a post-flooding success in Mediterranean macroinvertebrate communities (Gallardo et al., 2009).

4.1.2 Building the 4th Corner in TE Ecology

In the previous chapters of this thesis, I have identified patterns between the composition of the TE community with both the properties of the genomic environment such as heterochromatin state and spatial location in the genome and TE traits, such as the mechanism of transposition and reinsertion into the genome. In this chapter, I will use the 4th corner analysis to examine the direct associations between the genome properties identified in Chapter 2 and the TE traits identified in Chapter 3, bypassing the need to make predictions about where each individual TE might be found.

In Chapter 2 of this thesis, I examined the relationship between the TE communities of 11 genomes and their genomic environment. I did this by examining how TE communities changed linearly along each chromosome. I discovered that ~50% of the variation in TE communities across the chromosome could be explained by large-scale distribution patterns along the chromosome. Further analysis revealed that 3D spatial structure accounted for ~40% of the variation in TE communities of the human genome, and that about half (~20%) of those spatial
location patterns (PCNMs) were explained by discrete types of subcompartments (which are a property of the genomic environment) created by the folding of the chromosome into areas that are regulated in a similar way. These subcompartments are correlated with many epigenetic modifications that help to regulate how tightly the DNA is bound, and therefore how accessible that DNA is to the machinery of the cell, as well as marking regions of DNA that serve various structural purposes. In this paper, I will use these large-scale linear spatial patterns, as well as a functional classification of chromosome state similar to that of subcompartments as environmental properties of the genome.

In Chapter 3, I examined the relationship between TE traits and the structure of the TE community in *Drosophila melanogaster* (*Dmel*). I used the two predominant classification methods for TEs, the Tripartite classification system proposed by (Arkhipova, 2017) and the phylogenetically inspired classification system proposed by (Wicker *et al.*, 2007). Both classification systems group TEs by their traits. The Tripartite classification systems groups each TE family using a 3-part classification consisting of the type of replicative machinery, the type of integrative machinery, and any structural proteins present. The type of replicative machinery consists of R for RNA polymerase, D for those with only integrate functions, and B for those that use a B DNA polymerase for replication. The integrative components are broken down into 10 categories based on the type of integrase/endonuclease present. The third structural trait consists of 3 types of a capsid protein that allow TEs to occasionally move between cells (for further details see Chapter 3).

The Wicker classification categorizes TEs hierarchically by sorting the TEs into Class, Subclass, Order, and Superfamily based on traits (i.e. different proteins and their organization), then further sorting TEs into family by sequence similarity (see Chapter 3). I found that the Tripartite classification system did a better job of describing the diversity of TEs within the *Dmel* genome, however, the Wicker classification system allowed for more complex analysis of TE community dynamics such as TE community assembly. These conclusions will be tested
Figure 4.1: Transposon community ecology workflow. This figure shows the three community ecology variables examined in this thesis (top row and bottom left boxes). The box is the results of the “4th corner” analysis which combines data from each of 3 variables to identify TE trait – genomic environment relationships. The arrows in the figure which relationships are examined in each chapter of the thesis. a) shows the workflow in community ecology and b) shows the workflow in transposon community ecology.
further in this chapter by investigating if one classification system shows stronger associations between TE traits and the properties of the genomic environment.

This chapter combines the data generated in previous chapters to examine the relationship between the TE traits of our classification systems, and the environmental patterns I found to be relevant in structuring the TE community. I will examine if the 4th corner analysis can be used to determine the relations between TE traits and properties of the genomic environment. To do this I will use the counts of the individual copies of each currently annotated TE in the Dmel genome as the community matrix, broken into windows, or sites, linearly along each chromosome. The properties of the genomic environment will consist of the properties of each window (Chapter 2), and the TE traits will be the classification properties of each TE family in the Dmel genome (Figure 4.2). I expect that, like in the previous two chapters, it will be possible to use an ecological tool, the 4th corner analysis, to detect associations between TE traits and properties of the genomic environment in the TE communities of the Dmel genome.

4.2 Materials and Methods

4.2.1 Data

4.2.1.1 TE Community Data

The TE community data were obtained by running RepeatMasker (Smit et al., 2004) to identify TEs within the JSAE01000000 distribution of the Dmel genome sequence. The results were then run through one code to find them all script (Bailly-Bechet et al., 2014) to recombine nested TEs and accurately count LTR elements (see Chapter 2). This resulted in a TE community matrix consisting of the counts of each TE family (columns) within each window/site (rows; upper left panel in Figure 4.1).
4.2.1.2 Genomic Properties Data

*Dmel* chromatin state data were obtained from Sexton *et al.* 2017 (Kharchenko *et al.,* 2011; Sexton *et al.,* 2012; Li *et al.,* 2017), Table S1. These data classify each chromosome into 4 physical domains based on epigenetic marks and HiC contact interaction frequency. In chapter 2, similar HiC derived chromatin state data was significantly correlated to differences in variation among TE communities. The active domain contains more genes and is more actively transcribed; the PcG contains more densely packed heterochromatic regions which are transcribed less often; the HP1 domain is structural heterochromatin found in the centromeres; and, the Null domain contains areas that could not be clearly classified. These chromatin state classifications are similar to those found by Rao *et al.,* (2014); however, the unlike the Rao *et al.* analysis, which was carried out at a lower resolution, and detected fewer types of chromatin states (called subcompartments by Rao). Rao identified two euchromatic regions (A1 and A2), analogous to the 1 active region found in *Dmel*; four heterochromatic subcompartments (B1-B4), analogous to the Hp1 and PcG chromatin states in *Dmel*; and one undetermined subcompartment, analogous to the Null chromatin state in *Dmel.*

Spatial patterns that were significantly correlated to community composition were also used as environmental data. All possible distributions were generated using the principal components of neighbouring matrices procedure, and those correlated to the TE community were identified by redundancy analysis (RDA; see Chapter 2).

The “Genomic Properties” data matrix in this analysis consists of ecological properties of the genome described above (columns) for each site/window(rows) in each chromosome (upper right panel in Figure 4.1).

4.2.1.3 TE Trait Data

Trait data for each TE family in the *Dmel* genome were determined using the Wicker and Tripartite classification systems (Wicker *et al.,* 2007; Arkhipova, 2017). The Wicker
classification system is phylogenetically inspired and classifies TE families hierarchically by trait. Each column in the 4th corner analysis thus consisted of one categorical variable that classifies each TE family by superfamily.

The Tripartite classification system uses a three-letter code to classify TEs by trait; one letter each for transposase, endonuclease, and structural proteins (see Chapter 3 for more details). The columns in this 4th corner analysis are categorical variables for type of polymerase and type of endonuclease. The structural code was excluded because the classification included NAs, which were not compatible with the analysis.

The TE traits data matrix in this analysis consists of TE families (columns) and TE traits (rows; lower left panel in Figure 4.1). The traits consisted of one row for each TE family in the Dmel genome according to the Wicker classification system, and one row for each character state in each of the 3 traits used by the Tripartite classification system.

4.2.2 4th Corner Analysis

The 4th corner analysis links TE traits to ecological properties of the genome through TE community data. This is done in one of two ways depending on the type of trait and environment data being compared. 1) If both variables are categorical, a contingency table of each site/species (window/TE family) combination is generated with each entry consisting of one combination of a trait and environment variable. The expected distribution of values for the contingency table is generated by permuting the values in the rows and columns of the TE community table and remaking the contingency table 99,999 times. The significance of each cell in the contingency table is then calculated using a Chi-squared test and the newly calculated expected distributions for each variable. 2) If a categorical variable is being compared to a continuous variable it is done by decomposing the categorical variable into multiple binary variables and calculating the Pearson correlation coefficient for each level of the categorical variable.
In both of these cases, the significance of each relationship is then calculated by permuting windows (rows) and TE families (columns) independently to generate an expected null distribution given the sites and species present in the community (Legendre et al., 1997; Dray & Legendre, 2008). A Holm correction for multiple hypothesis testing was then applied to each p-value (Holm, 1979).

The ‘fourthcorner’ function as implemented by the ade4 package in R (Chessel et al., 2004; Dray et al., 2007; Dray & Dufour, 2015; Bougeard & Dray, 2018) was used to run the 4th corner analysis (Legendre et al., 1997; Dray & Legendre, 2008; Ter Braak et al., 2012; Pontarp et al., 2019). The analysis was run using model 6, which permutes both sites and species, with 99,999 permutations and using Holm correction to account for multiple hypothesis testing (Holm, 1979). Chromosome Y was excluded due to lack of chromosome domain data, and chromosome 4 was excluded because of its small size (only 9 TE communities) and the low number of significant spatial distributions (Chapter 2).
4.3 Results

a)
c) TE trait and genome property on Dmel chromosome 3 without correction

Genome Property:
- HP1 centromeric
- Null
- PRC1
- PCNM1
- PCNM10
- PCNM11
- PCNM13
- PCNM15
- PCNM17
- PCNM18
- PCNM2
- PCNM22
- PCNM24
- PCNM26
- PCNM28
- PCNM3
- PCNM31
- PCNM35
- PCNM37
- PCNM39
- PCNM4
- PCNM41
- PCNM43
- PCNM45
- PCNM5
- PCNM9
- PCNM12
- PCNM20
- PCNM26
- PCNM28
- PCNM8
- PCNM83
- PCNM85
- PCNM88
- PCNM89

TE Trait
- Family
- Polymerase
- Endonuclease

Strength of association: 0, 1, 2, 3
Figure 4.2: General Results of the 4th corner analysis of Dmel chromosome 2 by trait type. White cells represent no significant relationship, and blue cells represent significant positive associations. For a concise representation, the rows and columns have been switched relative to Figure 1. The rows represent the chromosome domains (Active, Null, H3P1 and PcG) and the significant spatial distributions (PCNMs). The PCNMs are ordered from bottom to top by those that explained the most community variation in chapter 2. The columns represent the TE traits: The Wicker family classification, and type of DNA Polymerase and endonuclease used by the TE. a), b), and c) show the results of the 4th corner analysis with no correction for multiple hypothesis testing for on Dmel chromosomes X, 2, and 3 respectively. d) shows the 4th corner results for chromosome x with Holm correction for multiple hypothesis testing.

4.3.1 4th Corner Analysis Results

Of the three Dmel chromosomes large enough to conduct the 4th corner analysis, only chromosome 2 had significant associations after the Holm correction. Figure 4.2d shows the
associations between overall trait variables and each ecological variable in chromosome 2. Of these, the type of endonuclease a TE has was correlated to 4 spatial patterns, and the TE family type was correlated to 2 spatial patterns. None of the heterochromatin states were correlated to any of the TE traits.

Because our number of statistical tests and degrees of freedom is high relative to ecological studies (see also Chapter 2), Holm’s correction is very conservative in our analysis. Since the goal of this analysis was to identify the feasibility of this method in a genomic context, I also performed the analysis without the Holm correction (Figure 4.2 a-c), to evaluate what types of information would be provided by the method. The centromeric domains were associated with the type of endonuclease and TE family on chromosome 3 (Figure 4.2c), and Null domains were associated with TE family on the X chromosome (Figure 4.2a). A total of 78 associations between trait type and environment were detected, 3 of which were related to chromosomal domain type, and 75 were associated with spatial patterns. None of the associations were found across more than 2 chromosomes.
Figure 4.3: Results of the 4th corner analysis of Dmel chromosomes 2 (a), 3 (b) and X (c) with TE traits decomposed into factor levels. Decomposing the TE traits allowed me to detect pattern associations with individual levels within a TE trait. White cells represent no significant relationship, red cells show a significant negative association, and blue cells represent significant positive associations. The rows represent the chromosome domains (Active, Null, Pp1 and PcG) and the significant spatial distributions. The PCNMs are ordered from bottom to top by those that explained the most community variation in Chapter 2. The rows represent the TE traits: The Wicker family classification, and type of DNA Polymerase and endonuclease used by the TE.
The 4th corner analysis also allowed me to examine the associations between the TE trait and property of the genomic environment at a finer grain. This is done by decomposing the levels of each categorical variable (Figure 4.3). Once again, no significant results were found after the Holm correction, so the unadjusted significance values were used here to see if the 4th corner method was detecting any trait-environment associations at all. A total of 18 associations were found between chromosome states and TE traits (Figure 4.3 top four rows of each plot). Four of those were with a specific endonuclease, while the remaining 14 were with associations with a specific TE family. Of these 18 associations, the only association that occurred on more than one chromosome was that between PcG (Heterochromatic) regions and DNA/hAT elements. The remaining 332 associations across the three chromosomes were with significant spatial distributions (PCNMs that were significant in chapter 2).

4.4 Discussion

4.4.1 Does the 4th Corner Analysis Work on Genomic TE Data?

In this chapter, I investigated whether the 4th corner analysis can be used to examine transposon ecology data in the same way it is used in ecology. The 4th corner analysis was successfully applied to the TE communities of Dmel, as the analysis could be carried out on genomic TE data, and it was able to detect some realistic associations between TE traits and the genomic environment, however many were not significant after multiple hypothesis testing. Surprisingly, there was no strong relationship between any TE traits and properties of the genomic environment used in this paper. Most of the TE trait – genomic environment associations identified in this analysis were with the spatial patterns from Chapter 2. These PCNMs represent community TE composition along each chromosome. These patterns were generated separately for each chromosome, so the identifier of each PCNM is not comparable across chromosomes. Even so, the PCNMs used in this chapter all explained some variation in the individual TEs that structured the TE population in Chapter 2. The results of this 4th corner analysis are able to go a step further by showing that pattern holds when looking at TE traits.
rather than individual TE families. Although the current state of knowledge is lacking on what may cause these specific patterns in the Dmel genome, I was able to show that the TE community in this genome is spatially structured, and that spatial structure was associated with specific TE traits.

As in Chapter 2, the large amount of data involved in genome-scale community analysis can be problematic for analyses designed for ecological communities. In this chapter, the p-value adjustment to correct for multiple hypothesis testing is an issue when combined with the large and varied nature of the TEs in the Dmel genome. This complex community and environment, and the resulting high number of hypotheses that need to be tested resulted in a very stringent p-value adjustment. The consequence of this was that only associations with very strong statistical relationships remained significant after adjustment.

This issue is compounded by the likely range of p-values produced by this analysis. The 4th corner analysis generates p-values by permuting the community matrix to generate a likely null distribution for each environment/ trait variable within that community, then comparing the observed value to the null distribution. For categorical variables, this is done using a chi-squared test and for continuous and ordinal variables this is done using a Pearson correlation coefficient. In both these tests the p-value required to remain significant after correction would require an almost perfect correlation; correlations of this type seem unlikely in genomic and community data, which tend to be complex and noisy. Future investigation into the application of this type of ecological analysis may benefit from the use or development of less stringent multiple hypothesis corrections such as the false discovery rates (FDR) correction (Benjamini & Hochberg, 1995), as well as different test statistics. For example, the chi-squared test is known to perform less well on sparse contingency tables, which occurred in the TE communities in some windows. Tests like the Fisher’s exact test (Renter J J Higgins J M Sargeant et al., 2000) or versions of Ku’s correction (Ku, 1963; Finkler, 2010) are known approaches that may be able to account for this in the future.
The goal of each chapter of this thesis was not only to investigate the role of a small subset of the most obvious TE traits/genomic environmental properties but to generate one or more tools and additional hypotheses in this new field. With the suitability of TE community data for this type of analysis demonstrated, these methods are now open to examine a whole host of TE trait – genomic environment patterns.

4.4.2 How Future Data on the Genomic Environment Will Affect the 4th Corner Analysis

This exploratory 4th corner analysis indicated that the broad level classification of TEs and the genomic environment is not sufficient to detect meaningful mechanistic associations after correction for multiple hypothesis testing. Given the amount of spatial structure identified in each of the genomes in Chapter 2, the relationship between age and TE traits and the spatial patterns in the TE communities along each chromosome in Chapter 3, and the few traits associated with TEs, the lack of many significant associations from the 4th corner analysis result suggests that there are ecological patterns we are not capturing. A necessary next step in this analysis will be to examine both TE traits and the genomic environment at a finer resolution.

A finer resolution view of the genome environment could include the application of more detailed 3D structural data, epigenetic tag data, and the data on the flexibility of the DNA in a specific area. The HiC available for *Dmel* was at a lower resolution than the human genome data used in Chapter 2. As a result, the chromosome state data in *Dmel* had only 1 heterochromatin state, whereas the finer resolution of the human HiC map allowed them to identify 4 classes of heterochromatin, which was found to be the best predictor of TE community structure (Chapter 2). It may also be the case that the meaningful associations between TE Trait and genomic environment happen at a finer scale than our windows can capture, obscuring meaningful heterogeneity in the genomic environment. As I showed in Chapter 2, window size can influence the detection of patterns when dealing with large genomic datasets, and thus it may also be useful to vary the window size and use the epigenetic tags that are correlated with chromosome state to classify these smaller windows. Interactions at this finer resolution are consistent with
the finding that the flexibility of the DNA immediately surrounding piggyBac elements can result in up to a 60% difference in transcription (Esnault et al., 2011).

These finer-scale properties of the genomic environment, along with other nearby environmental properties such as access to promoters and suppressors, either internally or nearby, could be what allows specific element copies to become successful. It is possible that a TE’s success in spreading may depend on it finding a good genomic niche. Upon transposition, some TEs have the potential to disperse anywhere in the genome, allowing them access to all of the niches in the genome. If comparing this universal dispersal to an ecological species’ dispersal this would be equivalent to an organism that could disperse to anywhere on the planet. Thus, it might be expected that this relatively high level of dispersal would select for TEs to have strong associations between their traits and the optimal locations for those traits, as each potentially has access to the most ideal sites for it in the genome. Thus, I would expect that universal dispersal in TEs should logically result in TEs finding their ideal environment, and thus similar genomic environments should have similar TE communities. However, in the indicator value analysis in Chapter 2, I found that similar genomic environments on different chromosomes could not be predicted by which TEs were present. Knowing that TE traits are playing at least some role in structuring the TE community, this indicates that TE biologists may not yet be measuring the right properties of the genomic environment, and maybe the chromosomes are more different than currently realized.

4.4.3 How Future Data on TE Traits Will Affect the 4th Corner Analysis

Our lack of TE trait / genomic environment associations could also indicate a problem with how TE traits were classified. It is possible that the TE classifications methods used here are either too broad or not relevant to TE ecology. TE classification is complex and filled with heuristics such as the 80-80-80 rule (Wicker et al., 2007), and competing classification systems (Kapitonov & Jurka, 2008; Seberg & Petersen, 2009; Kapitonov et al., 2009; Piégu et al., 2015; Arensburger et al., 2016; Arkhipova, 2017), as outlined in the Introduction. This has resulted in numerous calls to modernize TE classification and make the classification both more systematic
and more transparent (Piégu *et al.*, 2015; Hoen *et al.*, 2015; Arensburger *et al.*, 2016; Arkhipova, 2017). I would suggest that these problems reflect a lack of understanding about the functional role of TE traits in relation to their environment, and as phylogeny and traits are related in organismal biology (Pavoine *et al.*, 2013), perhaps a better understanding of what TE traits are functionally important and how may help the effort of finding a better TE classification system.

In Chapter 3, I used a tree of TE families based on the Wicker classification; the tree was not used directly in this analysis, however, Wicker based family classifications were included. Not using the phylogeny directly does have some drawbacks. Phylogenetic relationships violate the assumption of non-independence and thus can cause inflated type I error (Li & Ives, 2017). One solution to this process can be found in (Braga *et al.*, 2018) which uses the phylogeny to constrain the permutations which are used to generate the expected distributions of each variable for the chi-squared test. This approach is promising; however, it has only been implemented for categorical variables, and thus could not currently be used in our analysis. A second option is to use the phylogeny directly as a trait variable (Pavoine *et al.*, 2011) however in this analysis it would have been highly correlated to the Wicker family classifications. Inclusion of a true phylogeny, which is available for some subsets of TE families that share the same proteins, may allow us to detect TE trait/genomic environment relationships that are obscured by categorical classification systems used here.

As TEs have relatively few traits compared to species ecology, it is surprising that more significant associations between these traits and the genomic environment were not forthcoming. In Chapter 3, the community phylogenetics analysis found that TE communities were more or less phylogenetically related to each other depending on what part of the genome they were in. If the 4th corner analysis were using meaningful measures, it would be expected that stronger TE trait - properties of the genomic environment associations would be found. The lack of strong associations in this analysis might indicate that some of our TE traits may either be functionally equivalent or that some may be artificially grouped. It is also possible that there are more relevant TE traits that we have yet to be examined.
4.4.4 Conclusions

In this chapter, I was able to show that it is possible to run a 4\textsuperscript{th} corner analysis on genomic TE data. In doing so, I was able to detect some significant associations between the spatial patterns in the TE community that make up part of the genomic environment, and TE traits. These associations were weaker than expected, given the strong associations between the TE community and the genomic environment found in Chapter 2, and between the TE community and TE traits found in Chapter 3. In the future, modification of the test statistic used in this analysis and inclusion of more phylogenetic data may improve the accuracy and detection of associations. I also discussed how the current state of knowledge in TE biology may limit and suggest that further work into which TE traits are functionally different, the scale at which TE traits are analyzed and what traits of the genomic environment are relevant for TE biology may improve our understanding of the relationship between TE traits and the properties of the genomic environment.
5 General Conclusions

5.1 Does Transposon Ecology Work?

Is the community ecology approach to TEs a useful way to study them, and what kinds of conclusions might be made from using this framework and methods? The overall success of this approach was measured by asking three sub-questions, each of which is revisited below.

5.1.1 Can Concepts and Methods from Community Ecology be Successfully Applied to Genomes?

The answer is a resounding Yes. In this thesis, I was able to successfully adapt four community ecology methods to TE data, each of which provided new insights into the relationship between two important community ecology variables at the genomic level. In each case, the input transposon ecology data easily converted into the same format that would be used for analysis of classical community ecology data. The one major difference was that the whole genome datasets used in transposon ecology analyses were much larger than classical community ecology datasets. This sometimes resulted in a tremendous amount of output data that must be parsed; for example the redundancy analysis used in Chapter 2 produced a tremendous amount of data for each chromosome of each of the 11 genomes analyzed, so much so that there was only time to summarize the total amount of TE community variation explained in this thesis, leaving data on which specific TEs are most or least often found together in which environments to be analyzed at a later date.

5.1.2 What Sort of Variables are Important in a Transposon Ecology Context?

I introduced three important types of community ecology variables at the beginning of this thesis. These were the TE community, the properties of the genomic environment, and the traits of the individual TEs within each TE family. Even using a subset of the potential properties of the genomic environment, and using a necessarily limited exploration of potentially important
TE traits I was able to demonstrate that each of these three variables was important in characterizing the distribution and abundance of TEs in various genomes.

5.1.3 Are These Ecological Analyses Likely to Provide General Insights into Transposon Biology and Generate New Hypotheses in Transposon Biology?

This question too, clearly, can be answered in the affirmative. Each of the three chapters was able to either demonstrate something new about TE biology or provide a more nuanced understanding of the factors that influence the distribution and abundances of TEs in different areas of the genome.

In Chapter 2, I used a redundancy analysis to show how much of the variation in the TE community was spatially structured in each of 11 diverse genomes. Within these genomes, ~50% of the variation within the TE community could be explained by large scale spatial patterns alone. This is important as it shows that TEs are distributed non-randomly in each of the genomes analyzed, and as these genomes are quite diverse it would be reasonable to expect that this is true in many genomes. With this spatial structure demonstrated, I sought to find the effect of more explicit properties of the genomic environment on TE communities. As chromosomes are not laid out in a straight line inside the nucleus, I used an interaction frequency map of the human genome to represent the 3D structure of each of the uncondensed chromosomes. This 3D interaction map was able to explain a total of ~40% of the variation in the TE community. This was closer to demonstrating a relationship between the TE community and a property of the genomic environment. By including data on chromatin state, which (Rao et al., 2014) showed bunches together in the nucleus and is predictive of how that DNA is regulated, I was able to explain ~20% of the total variation in the TE community. This indicates that there are truly different environment types in the human genome, and that certain TEs are preferentially found in some of these areas. Finally, the TEs found in these subcompartments were not the same across all chromosomes, indicating that the chromosomes themselves may represent different environments to the TE.
In Chapter 3 I showed that TE older families are more spatially structured than younger TE families in *Dmel*. *Dmel* has a small genome with mainly active TEs, and TEs that become inactive are quickly deleted (Petrov *et al.*, 2003). Thus, it would be expected that those TEs that persist long enough to become old are those that have been able to find a niche where they have been able to survive. This chapter also used a phylogenetic community ecology approach to analyze how phylogenetically clustered or overdispersed each TE community was. I found that 15-17% of the TE communities in the *Dmel* genome were non-random phylogenetically. Within these phylogenetically non-random communities, the majority of the communities were phylogenetically overdispersed, while the TEs in the center of chromosomes 2 and 3 were phylogenetically clustered. In classical organismal ecology, the interpretation of these patterns can be complicated, mainly because of the validity of using phylogeny as a proxy for traits. TEs, however, only have a very limited number of proteins and thus associated traits. Moreover, the phylogeny used in our analysis is based entirely on these limited traits. This means that the classical interpretations of phylogenetic overdispersion as competition, and phylogenetic clustering as environmental filtering, are more likely in transposon ecology.

In chapter 4, I used a 4th corner analysis to show associations between TE traits and properties of the genomic environment. Using this analysis, I was able to show not only TE families, but the traits of the TE’s themselves are associated with the spatial patterns that structure the TE community. The weak nature of the associations in this analysis also shows that that either the relevant functional traits of TEs in our classification systems are not being captured, or that the properties of the genomic environment that are importation to TEs have not yet identified, or both.

### 5.2 What Makes Transposon Ecology Different from Classical Ecology?

The above analyses show that while at the functional level transposon ecology and classical ecology are largely similar, there are some important differences that must be kept in mind. The first is that of sampling. In classical community ecology work, an important part of
collecting community data is deciding how to plan research sites and collect your data. It is rarely feasible to measure an entire community, and statistics must therefore be used to ensure that the sample is representative and to confirm that the observed patterns are not artifacts of sampling. By contrast, when using complete assembled genome sequences, transposon ecology is far less limited by sampling challenges. The whole TE community is sampled because every known TE that is sequenced can be counted. Of course, the sampling is only quasi-complete because even high-quality genome assemblies such as the human genome are missing a small fraction, and indeed that fraction tends to be made of TEs and other repetitive sequences (Treangen & Salzberg, 2012; Elliott & Gregory, 2015a). Nonetheless, the differences between subsampling in classical ecology and nearly complete sampling in transposon ecology are striking.

This nearly complete sampling provides transposon ecology with some advantages over classical ecology. Nearly every TE in the whole community and what environment they are found in is known. Any correlations that are found are not potentially biased, so no statistics are needed to confirm what the relationship is, the pattern is what it is, and there is no need to show that it could be the product of how the data was sampled. On the other hand, nearly complete sampling can pose its own challenges. Notably, the amount of data in a genome often makes analysis more time consuming, and computationally expensive, so it may not be feasible for the largest genomes or for researchers who lack sufficient computational resources. Another drawback is that the methods in Chapter 2 produce so much output that it can be difficult to narrow in on interesting patterns.

In my thesis, I have focused primarily on how classical ecology can help transposon ecology. Transposon ecology, however, also has the potential to aid classical ecology. One way it can do this is by providing completely sampled communities that can then be used to test the effect of sampling on a real-world whole community. Currently, this is only possible in classical ecology using a simulated community approach. Another way transposon ecology could influence classical ecology is that genomes provide replicates of a whole ecosystem. There are
many copies of the genome of each species, and many copies of the genome of similar species. This allows transposon ecology greater statistical power to evaluate whether observed patterns arise by chance or because of some mechanism of interest. Classical ecology often struggles with the ability to test mechanisms or predictions because there are no planetary replicates, or because experimental manipulation of ecosystems would be unethical, or because ecosystem-level processes operate on too large a scale or over too long a time frame. For the local scale, organismal ecologists often use lab experiments with small and fast-growing organisms (bacteria, protists, etc.), often combined with theoretical work, but at larger scales, this quickly becomes less feasible or informative. Since genomes provide an entire ecosystem that can be replicated and manipulated, these theories can be tested at a much larger scale on real data, and with replicates involving many distinct genomes.

5.3 What is the Future of Transposon Ecology?

5.3.1 The Future for the Methods in this Thesis

Three follow-up projects have already been generated by this work. One project used the spatial methods in Chapter 2 on Insertion Sequences (IS) in bacterial genomes. Even though bacterial genomes are much smaller and have many fewer TEs than most eukaryotes, identifiable spatial structure was observed in ~30% of the genomes analyzed (Lanteigne, 2019). In addition, the transposon ecology working group is working on a simulation that models one TE family that spreads using only ecological means (i.e. the TEs do not change in any way through the simulation) within a population of host genomes. We use this simulation to show purely transposon ecology methods of genome size change (Kremer et al, in prep). This also represents a good start at generating simulated TE communities to test the methods in this thesis. The data generated by this model have also been used in a thesis project, which also used the spatial analysis from Chapter 2 to look at the effect of TE traits and properties of the host environment on these simulated bacterial TE communities (Wei 2018).
In this thesis, I have only scratched the surface of what these methods have to offer transposon ecology. With this work as a starting point, future work can go in two equally useful directions. The first direction is increasing the level of detail in describing the different variables. For properties of the genomic environment, this would involve exploring the vast amount of genetic and cellular data (e.g., methylation patterns and histone modifications) that has already been generated to see which are able to explain more of the remainder of the ~50% of the spatial patterns I identified in Chapter 2. For example, aspects of epigenetic regulation are known to be associated with TE presence (Lippman et al., 2004), and could be included more directly and at a finer scale of resolution than chromatin state. The epigenetic modifications of each TE could also be taken into account and may provide a way to further classify TEs functionally. Likewise, the flexibility of the DNA surrounding TEs can affect the transposition rate of some TEs (Esnault et al., 2011), and should be investigated at a wider scale.

For a finer scale exploration of TE traits, the history and the age of the TE should be investigated further. In large genomes, which are less likely to delete inactive TEs than Dmel, there are more inactive TEs. Classification of TEs as active or inactive is currently difficult and time-consuming. However, inactive TEs act as a type of fossil record encoded in the genome, and thus a TE community’s spread through the genome over time. By more specifically ageing each individual TE copy, it might be possible to estimate when individual elements spread. In fact, this method will be far more detailed than tracking the spread of species in ecosystems, for which inferring historical population properties can be much more difficult.

The ageing, and establishing more and other TE properties at the scale of the genome, can be greatly influenced by machine learning. These approaches have increased in popularity in ecology, including with reference to genetic data. One recent example used DNA barcodes to identify differences between native and non-native species at the DNA level (Andersen et al., 2019). It may be possible to use a similar method to determine a TE’s age, or what type of environment a TE might be found in at a much larger scale than possible for a DNA barcoding analysis. Finer-grained exploration of the TE traits that are important for their ecology is clearly
needed, as none of the theoretically comprehensive traits, from the tripartite classification system, included in the 4th corner analysis explained the spatial variation seen in Chapter 2. Further work on which traits of TEs can explain TE distribution and abundance and how to classify them could go a long way in explaining this unexplained spatial variation. It may also be useful to consider TE traits at a finer scale than the proteins they use to function. Perhaps the specific protein domains that make up the protein are more important for the ecological function of a TE family.

The alternative to gathering more detailed data is to increase the breadth of analysis by exploring whether patterns observed in some species are found in a wider range of genomes, or in different tissue types with different patterns of expression. One of the most exciting aspects of TE ecology from an ecology perspective is that each genome could represent a vastly different environment with vastly different patterns and communities. The genome of each species will have a different history of evolutionary pressures (e.g., different restrictions on genome size), and different chance events (e.g., population bottlenecks and fixation of neutral alleles). There also may be a different history of coevolution between TEs and the genome (e.g., different silencing mechanisms and different TE families becoming dominant). It is thus possible that the strength of the TE trait – genomic environment varies from genome to genome. In this case, I would expect that this same analysis in other genomes may give different results.

Though this thesis was successful in demonstrating that transposon ecology methods work and that they can provide new insights, in some cases these methods would benefit from ongoing efforts at optimization. As discussed in Chapter 2 and Chapter 4, the large amount of data involved in transposon ecology approaches can cause problems with p-value adjustments becoming so strict that significant patterns become near impossible to identify, so other corrections should be investigated. This is not a new problem, although less stringent corrections for multiple hypothesis testing exist, such as the false discovery rate (FDR) correction (Benjamini & Hochberg, 1995). An alternative approach that should be investigated is to adopt a predictive modelling approach. This approach would allow cross-validation or subsampling.
methods to be used for statistical validation. This would be advantageous as these methods test the accuracy of predictions directly and do not need to be corrected for multiple comparisons. Further development of the methods could also include biological positive an negative controls. For instance, non-TE repetitive DNA that results from tandem duplication or replication slippage could be expected to have a strong spatial pattern, and thus, instead of being excluded, these elements could serve as a positive control in future analyses. In genomes where the active and inactive TEs are known, if there are consistent patterns between the strength of the spatial structure of active and inactive TEs in the same family, then these patterns may serve as useful controls as well. Likewise in genomes with TEs that are known to be active in specific tissues, if high-quality data of those tissues exits those TEs may be able to be used as positive controls.

Now that the utility of ecological methods to find meaningful patterns in genomic TE data has been demonstrated in principle, it may be useful to more systematically test the methods. This could be done by using simulated data, for example, with relationships of differing strength between specific TE traits and properties of the genomic environment, or by spiking the data with a known pattern. By rerunning the methods in this thesis on these simulated datasets it would be possible to directly show the power of each method to detect these known patterns in the data. An additional benefit of this approach is that by using simulated datasets with known patterns, the alternate corrections for statistical significance could be assessed. Using this approach both rates of type I and type II error could be assessed before further use on biological data. Biological controls, such as other repetitive parts of the genome, which show a spatial pattern, could be included in future analyses.

The reliance on windows for defining and counting TE communities is another feature of these methods that should be investigated. I showed in Chapter 2 that the size of the window used in transposon ecology analysis can affect the results. One solution that has not yet been investigated is to abandon using windows altogether. This could be done by individually tracking all TEs and features of interest in the genome. This change would necessitate either careful
consideration of what constitutes a TE community or moving away from community transposon ecology into population transposon ecology methods.

5.3.2 The Future for Transposon Ecology as a Field

Expanding to broader levels of study from transposon community ecology, it would be possible to apply concepts and methods from classical biogeography to TEs. The field of Biogeography studies space at a wide geographic scale and time at a geological scale. Through studying distributions of species at these large scales, biogeography is often concerned with patterns of migration, speciation, and extinction. Studying these processes necessitates including evolutionary processes and may be helpful at further clearing up the impact of transposon ecology and transposon evolutionary processes. Although (Linquist et al., 2013) demonstrated that it is easy to conflate evolutionary and ecological processes, this type of question is common in TE biology, and biogeographical methods may be a good approach for dealing with those types of questions.

5.4 Concluding Remarks

In this thesis, I have carried out the first systematic transposon ecology analysis at the genome scale. In doing so, I have demonstrated the utility of these methods on genomic TE data. I have also generated a large amount of specific spatial interaction data in 11 diverse genomes and identified new hypotheses relating to the ways in which properties of the genomic environment structure the TE population, the effect of TE properties on how the TE community is distributed, and how TE traits are related to the distribution of TEs within the genomic environment. My work represents the first foray into the field of transposon ecology and judging by its results the plethora of methods available to ecologists of all fields promise to be a great aid to the transposon ecologists of the future.
REFERENCES


Benett MD, Leitch IJ. 2012. Plant DNA C-values Database.


Clark LA, Wahl JM, Rees CA, Murphy KE. 2006. Retrotransposon insertion in SILV is responsible for merle patterning of the domestic dog. Proceedings of the National Academy of
Sciences 103: 1376–1381.


Gregory TR. 2016. Animal Genome Size Database.


Hofmann A, Heermann DW. 2015. The role of loops on the order of eukaryotes and


**Seberg O, Petersen G. 2009.** A unified classification system for eukaryotic transposable


Wei X. 2018. TE distribution patterns comparison between simulated models and prokaryotic genomes.


