Rutgers University

Honors Thesis

Understanding The Evolutionary Dynamics Of Transposable Elements in *Drosophila* Via *de novo* Identification and Classification

Author:
Chinmay P. Rele
chinmay.rele@rutgers.edu

Supervisor:
Dr. Christopher E. Ellison
chris.ellison@rutgers.edu

Thesis Committee:
Dr. Premal Shah
Dr. Jinchuan Xing

A thesis submitted in fulfillment of the requirements for the degree of Honors in Genetics

in the

Ellison Lab
Department of Genetics

May 6, 2019

Declaration of Authorship

I, Chinmay P. Rele, declare that this thesis titled, "Understanding The Evolutionary Dynamics Of Transposable Elements in *Drosophila* Via *de novo* Identification and Classification" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: 

Date: Monday; May 06, 2019

"*Problems worthy of attack prove their worth by fighting back.*"

Piet Hein

"*The problem of automated repeat sequence family classification is inherently messy and ill-defined and does not appear to be amenable to a clean algorithmic attack.*"

Bao and Eddy, 2002

"*It is not so very important for a person to learn facts. For that he does not really need a college. He can learn them from books. The value of an education in a liberal arts college is not the learning of many facts, but the training of the mind to think something that cannot be learned from textbooks.*"

Albert Einstein

"*The world has more problems than it deserves and has more solutions that it is using.*"

Anon

RUTGERS UNIVERSITY

Abstract

Chinmay Rele
Department of Genetics

Honors in Genetics

Understanding The Evolutionary Dynamics Of Transposable Elements in *Drosophila* Via *de novo* Identification and Classification

by Chinmay P. Rele

Repeat Elements are some of the most misunderstood sequences in our genomes. They have a bad reputation of being harmful. However, there is clear evidence that though some repetitive sequences might be harmful, they can also be putatively adaptive as they might increase expression of genes that might, in very broad terms, increase fitness of an organism.

The reason for their absence in most genomic studies is in the difficulty in classifying their location and copy number due to the use of less than perfect sequencing techniques. In this thesis, my mentor and I hope to alleviate this disparity in the community by proposing a novel method of identification of repeats using a *de novo* approach.

We use available programs as well as custom pipelines to be able to identify the correct copy number and locus of these repeats to better understand genome architecture of the *Drosophila* genus and to gauge the evolutionary dynamics of transposons within it.

Drosophila, transposons, TE, evolution, *de novo*, phylogeny, computational, Nanopore, sequencing, novel identification

Acknowledgements

This project has been a labour of love for the better part of a year and has required the attention and advise of many people, too many to acknowledge individually.

However, some do stand out. I would like to thank Dr. Ellison for his continual input and advise in running the script and also Weihuan (Lucy) Cao, who has faithfully kept stocks of flies for our work alongside her own projects. Furthermore, I would like to thank Dr. Shah and Dr. Xing for being on my thesis committee and providing feedback and advise.

I would also like to thank Galen Collier and other `Amarel` admins who relentlessly keep the `Amarel` systems used up to date. I would also like to mention Dr. Meenakshi Kagda for her involvement in streamlining and optimizing the pipeline.

And finally, I would like to thank my colleagues and peers who have given me valuable input on this project.

Contents

List of Figures

List of Tables

Listings

List of Abbreviations

3C	**C**hromosome **C**apture on **C**hip
Alu	**A***rthrobacter* **lu***teus*
ARE	**A**daptively **R**elevant **E**nviromnent
bp / bps	**b**ase **p**air(s) | unit
BLAST	**B**asic **L**ocal **A**lignment **S**earch **T**ool
CNVs	**C**opy **N**umber **V**ariant(s)
CRAN	**C**omprehensive **R** **A**rchive **N**etwork
CRISPR	**C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats
dNTP	**d**eoxy-**N**ucleotide **T**ri-**P**hosphate
DNA	**D**eoxy Ribo**n**ucleic **A**cid
EEA	**E**nvironment of **E**volutionary **A**daptation
IDE	**I**ntegrated **D**evelopment **E**nvironment
iToL	**i**nteractive **T**ree **o**f **L**ife
gag	**G**roup **A**nti**g**en
GFP	**G**reen **F**luoroscent **P**rotein
GWAS	**G**enome **W**ide **A**ssociation **S**tudies
kb	**k**ilo **b**ase-pairs
LINE	**L**ong **I**ntersperced **N**uclear **E**lement
LTR	**L**ong **T**erminal **R**epeats
NCBI	**N**ational **C**enter for **B**iotechnology **I**nformation
NTP	**N**ucleotide **T**ri-**P**hosphate
plNTP	**p**hospho-**l**inked **NTP**
pol	**Pol**ymerase
RC	**R**olling **C**ircle | Helitron family
RNA	**R**ibo-**N**ucleic **A**cid
rRNA	**r**ibosomal **R**ibo-**N**ucleic **A**cid
RT	**R**everse **T**ranscriptase
SINE	**S**hort **I**ntersperced **N**uclear **E**lement
SMRT	**S**ingle **M**olecule **R**eal **T**ime
ssDNA	**s**ingle-**s**tranded **D**eoxy Ribo**n**ucleic **A**cid
SVA	**S**INE **V**NTR **A**lu
TPRT	**T**arget **P**rimed **R**everse **T**ranscription
TE	**T**ransposable **E**lement
TRF	**T**andem **R**epeat **F**inder
VNTR	**V**ariable **N**umber **T**andem **R**epeat

To my parents . . .

Chapter 1

Introduction

Most studies that bother with the matter at all usually tend to only worry about coding genes [1] – genes that have a say in what drives our cellular activity; however there is something more potent lurking in the background of our genomes, something that defines us more than what we can physically see. The effect of this source is both, alluring and elusive to us; we have not had the resources to study them as we did coding genes – knocking them out and using GFP tagging. These sequences are interesting, and also very repetitive.

1.1 Repeating Elements

Repeated sequences are patterns of repetitive nucleotides (dNTPs) that occur in multiple copies throughout genomes [2]. The most abundant sequence in DNA comprises of repeat elements. Repeat elements are found in all sorts of domains of life from *Animalia* [3] to *Plantae* [4] to even *Archaea* [5]. A significant fraction of the genome (about 40% to 70% of eukaryotic genomes) is composed of repetitive sequence, which influences genome organization, gene expression, and genome evolution [6]. The major categories of repeat elements include terminal repeats, tandem repeats and interspersed repeats.

1.1.1 Terminal Repeats

Telomeres are terminal repeats that prevent premature chromosomal degradation [7], [8]. Telomeres are the ends of chromosomes and contain a G-rich series of repeats. Telomerase recognizes the end of a repeat sequence, and using an internal RNA template, it extends the parent strand and adds additional repeats as it moves down the parental strand. The lagging strand is then completed by DNA Polymerase-α [9]. This is important for cell immortalization and fidelity of future generations [10].

1.1.2 Tandem Repeats

Tandem repeats are shorter bursts of repeats that occur right next to each other. They are thought to arise through DNA strand slippage during replication and are of unknown function [14]. They can have simple dNTP repeats such as $(AUG)^n$ [15] or can have much longer repeating dNTPs. Tandem repeats describe patterns that are useful when determining an individual's inherited traits, genetic profiling, and for forensic testing [16], [17].

1.1.3 Interspersed Repeats

Interspersed repeats or interspersed nuclear elements are like tandem repeats in that they include specific dNTP sequences that are repeated throughout the genome. But unlike tandem repeats, which occur right next to each other, interspersed repeats are non-adjacent and are spread throughout the genome of the organism. The most important of these elements are called Transposable Elements (TEs) or transposons, and they act pseudo-independently of other pathways [3], [20], [21].

1.2 Transposons

First discovered by McClintock in maize when trying to explain disparate nature of the colors found on kernels [22], transposons are the most abundant mutagenic locus [23]–[25]. McCLintock attributed transposons or "jumping genes" [26] to chromosome-breaking loci due to their recombination [22]. TEs are some of the most abundant elements that occur in eukaryotes [24].

The Selfish DNA hypothesis [15] implies that these are sequences that parasitically spread across the genome of a host by forming new copies of themselves. They are similar to viruses, which act as parasites and hijack host cell machinery to forcibly replicate themselves at the cost of the host's viability [27]. Activity and expression of TEs in the human brain reveal that neurons are susceptible to somatic genomic alterations [29].

TEs can be classified based on their independence into autonomous and non-autonomous TEs.

1.2.1 Autonomous TEs

Autonomous TEs can move by themselves [26], and they can either be retrotransposons or DNA transposons. They are defined by their ability to encode their own RTs and ligases in the case of retrotransposons [4], or their own transposases in the case of DNA transposons [25].

1.2.2 Non-autonomous TEs

Non-autonomous TEs require external machinery in order to transpose [30], usually from another autonomous TE. They differ from autonomous elements in their inability to code for their own RTs and integrases, increasing their dependency on mother elements. It is critical to note that retrotransposons still need host cell machinery in order to transcribe [31]–[33], but the transcribed RNAs are independent after that point.

Transposons are can also be split into two classes based on their replication mechanism [30] – Retrotransposons and DNA transposons.

1.2.3 Retrotransposons

Also called RNA transposons, retrotransposons increase their copy number within the genome via an RNA intermediate [36]. They have a copy-paste mechanism [26], by

which they first transcribe themselves into an RNA intermediate by hijacking host cell transcription machinery by either pretending to be genes or inserting themselves near genes [37]. Then, via a reverse transcriptase (RT) and integrase, they reverse transcribe that RNA back into DNA and insert it within the genome of the host. Most, but not all, retrotransposons encode their own RTs and integrases as eukaryotic cells often do not encode RTs or integrases [38], [39]; however, this is not always needed.

Their regulation can also be affected by so called "Mother elements", which can control their excision, transcription, reverse transcription and integration back into the genome [40]. Through this mechanism of "copying" themselves and leaving the original intact, at every cycle, the copy number of the particular element that has been transcribed and integrated increases [41].

They do this in a myriad of ways such as (1) by either disrupting gene pathways or genes, (2) plugging back in promoters inactivating them and halting the gene pathway, or (3) within insulators, inactivating the insulators and allowing that gene pathway to be over-expressed [26], [28], [42], [43].

1.2.4 LTR Retrotransposons

Long Terminal Repeat retrotransposons are classified due to their long terminal repeats flanking the internal region of the TE; this internal region can either code for RTs and integrases, transposases, or just be junk DNA [Figure 1.1]. Their size ranges from anywhere between 25kbp, such as *Pisum*'s Ogre TE [47], to the 100 bp range. All autonomous LTR retrotransposons encode two genes, a group specific antigen (*gag*) and a polymerase (*pol*) [48], [49]. A *gag* codes for the core structural proteins of retroviruses; whereas a *pol* is a DNA polymerase. They were first discovered because of their relative abundance in the maize genome [22].

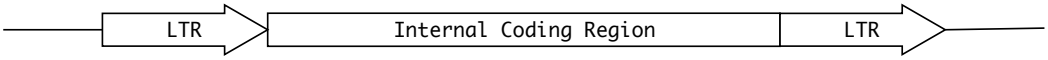


Figure 1.1: LTR Transposon Schematic

LTR retrotransposons can be further divided into the following classes based on their sequence homologies:

1.2.4.1 Endogenous retroviruses (ERVs)

These are the most important LTR transposons that comprise about 10% of the mouse and human genomes [50]. Endogenous retroviruses can play an active role in shaping genomes [51].

1.2.4.2 Ty1-*copia* retrotransposons

These are abundant TEs mainly in plants and algae. They code for the following domains: protease, integrase, RT and a ribonuclease in that order [24], [52] . In *Drosophila*, there are between 20 and 60 copies of a copia element within every genome [28].

1.2.4.3 Ty3-*gypsy* retrotransposons

Ty3-*gypsy* TEs and the Ty1-*copia* are similar in that they code for particular protein domains in a particular order. However, *gypsy* elements code for a protease, an RT, a ribonuclease and an integrase in that order. Based on specific protein domains and sequence motifs, they can be subdivided into categories such as *Chromoviruses*, *Errantiviruses* and OGRE elements, which are some of the largest TEs to infect plant life [47].

1.2.5 non-LTR Retrotransposons

Non-LTR retrotransposons are present in most eukaryotic genomes including that of humans Autonomous elements contain an RT domain [53]. Their general structure is presented in Figure [Figure 1.2]. Within the non-LTR family, there are subfamilies of TEs called SHORT INTERSPERSED NUCLEAR ELEMENTS (SINEs), LONG INTERSPERSED NUCLEAR ELEMENTS (LINEs), and SVA/*Alu* elements, the last of which are abundant in human populations.

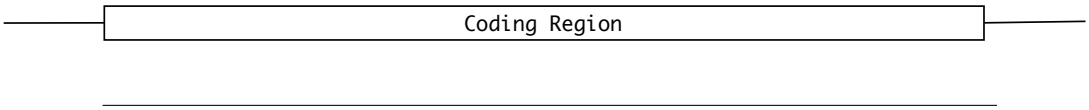


FIGURE 1.2: non-LTR Transposon Schematic

1.2.5.1 LINEs

LINEs are a very abundant group in eukaryotic cells and higher organisms. About 17% of the human genome is made up of LINE elements, but almost all have lost the ability to transpose [56] (except for LINE-1). They have evolved to be autonomous in nature by coding for their own *gag*, *pol*, RTs and integrases [57].

1.2.5.2 SINEs

SINES are the only TEs that have evolved to be non-autonomous by nature. They did not evolve from autonomous elements that lost their ability to self-transpose and use host-cell or intra-transposon proteins to excise. They have relied on LINEs to transpose from the start of their evolutionary lineage [56]. We know this as most of the sequence within this family is junk and thus cannot be annotated to protein domains.

1.2.5.3 SVA

SVAs are composite elements made from fragments of 3 others elements, namely SINEs, VNTRs and *Alu* elements. SVAs are the youngest retrotransposon family in the human genome and a number of diseases are known to be caused by SVA insertions [58]. SVA It is a family of repetitive sequences in the human genome and is classified as a SINE.

1.2.5.4 Alu

Alu elements are non-autonomous retrotransposons characterized by *Arthrobacter luteus* (Alu) restriction endonuclease activity [59]. A significant portion, over 10%, of the human genome, consists of *Alu* elements [2], [60], [61].

1.2.6 DNA Transposons

DNA transposons migrate through the genome via a cut-paste mechanism [53]. This system of motion needs an enzyme called a transposase [26], that caters to both the excision and integration of the DNA segment. Transposases (a subset of DNA transposons), the enzymes that cut DNA, are never perfect in their activity [62]; they leave irregularities at the locations, creating insertions of random base pairs, or delete preexisting base-pairs (indels) in DNA. These extra bases, or lack thereof, can interfere with gene and protein activity and thus are crucial to long-term viability [63]. Unlike retrotransposons, DNA transposons have an effect on the loci from which they are excised and also the loci to which they move [63]. They are found in both prokaryotic and eukaryotic genomes, and comprise a large part of the latter's genomes within non-coding regions [15], [64]. About 3% of the human genome is made up of DNA transposons, but they are "fossils", or remnants, of a past where our lineage had mobile TE activity [27], [65].

There is also a subsection of DNA transposons called Rolling Circle (RC) transposons a.k.a. Helitrons, which transpose via a rolling circle replication mechanism using an ssDNA intermediate [66].

Due to their repetitive structure [54] as well as their adaptive or debilitative effect on the genome, TEs have often been excluded from genomic studies [67].

It is important to know the major classes of TEs, as they affect gene expression differently from each other, which was most blatantly seen in McClintock's 1956 study of maize [22]. This occurred due to specific and stochastic modification of kernel color. This could only have occurred if there were random activation and repression of genes; something that could be easily explained by a highly mobile element/group of elements that inserted into or out of a locus, activating or deactivating it [22]. TEs affect gene expression by modifying gene expression near their sites of activity [25]. They can do this by moderating expression via promoters (initiates transcription of certain genes), enhancers (increase likelihood of transcription of gene) and repressors (sequence, if bound to by specific protein prevents transcription of the gene) of gene pathways [70]–[72]. They can modify methylation at sites [73] and their mobilization near promoters and/or enhancers can activate them [74]. TEs damage the genome by upregulation of mutagenic and oncogenic sites [75], but can also help us identify these sites [76], [77]. Transposons can permanently affect gene expression in an individual [78] and their progeny [7], [8], [21], [79].

1.3 Genome Evolution

TE evolution is different from host evolution in the following ways: (1) their sequence can evolve independently of host genome sequences as they do not code for anything

that their host cell needs; and (2) they are mobile, and their transposition affects host cell viability, making them adaptive if their insertion/excision makes the host cell more competitive, or debilitative if their insertion/excision makes the host cell less competitive.

1.4 Context of Study and Research

TE insertions can harm the genome, or insertions can be coopted by the genome to be beneficial [89]. Our research is interested in how these are copied and coopted by the genome to become adaptive.

We need unbiased annotations of the TEs to know how they have been coopted. The pipeline described in this paper is the result of a search in the literature and resources that have been developed by other groups. The set of annotations gained as a result of this pipeline will serve as a valuable resource for future studies in the lab as well as the field.

1.5 Sequencing and *de novo* Classification

It is easy to identify TEs based on sequence similarities with a database or other individuals of the same genus/species [95], [96], but what of recent stochastic TE genome insertions? These new TE insertions, which might have come from another species, cannot be classified as TEs if they are not in an existing database, and would be classified as host cell sequence by default [30].

We need to be able to classify sequences *de novo* so that there is no interference from other repeats and no sequencing bias. This bias arises from incomplete assemblies due to using short reads. Each method has its own advantages and drawbacks, which need to be compensated for. Whole genome sequencing has given us the opportunity to study organisms and systems in a way that would not have been previously possible. However, all sequencing technologies are not alike and choosing a method is dependent on the type of data required for the study.

1.5.1 Illumina Sequencing

The most common and cheapest form of DNA sequencing, Illumina benefits from high accuracy due to high coverage reads. It does this by sequencing all parts of the genome multiple times and then aligning those contiguous sequences [97] to form a whole genome assembly. Most of the read lengths range from 75-150 bp. This is not very effective in sequencing and analyzing the copy number of TEs as most of them are much larger than this length [54] and occur many times in the genome, thus making assembly nearly impossible.

1.5.2 PacBIO SMRT

PacBIO's Single Molecule Real Time (SMRT) sequencing is a real time sequencing method that eavesdrops on and harnesses the power of the DNA Pol as a sequencing engine as it works to replicate DNA. Instead of normal dNTPs, SMRT uses phospholinked nucleotides (plNTPs) attached to a different colored florescent label, which is attached to each of the 4 dNTPs, which is removed and emits a light when DNA Pol

base-pairs it to the parent strand. This emission is recorded as a base-pair in DNA, a phenomenon which is compounded as more bases are added, creating a strand a sequence of letters [99]. It allows for much longer read lengths, but is a very expensive method, and cannot be used by most labs that require both high-throughput and regular sequencing strategies.

1.5.3 Nanopore MinION

Oxford Nanopore's MinION is much cheaper than Illumina and PacBIO for a single sequencing run. Unlike Illumina, it produces long reads [100], but those reads are inconsistent and not very accurate in their sequence [100], [101]. So, this is a method better suited for determining the genomic location of repeat elements, but not the particular sequence of those repeat elements (which can be sequenced properly using Illumina).

Most species that are regularly studied have already been sequenced using Sangar sequencing, and then later improved upon by a mixture of Sangar and Illumina sequencing protocols [98], [102], [103]. This implies that their sequence is well known, but at long repeat loci, their arrangement is up to debate due to the inability to properly align contiguous sequences (contigs) to a particular locus [28]. Most of these get collapsed to the same genetic locus over and over, not showing their spread throughout the genome. They are missing many of their longer TEs from the current version of their genome assembly. However, 16 different species of *Drosophila*, spanning millions of years of evolution, have high quality genome assemblies made from long-read sequencing technologies [104].

These assemblies across species differ in their reliance on short vs. long read data, the programs used in their analysis, and most importantly, in the parameters chosen to analyze the data. In order to compare TE evolutionary dynamics between species, we need a comprehensive annotation of TEs in these genomes that use the same approaches for each of the species such that they are comparable. We also hope to estimate the age of each of the present TE families based on sequence divergence within Drosophila as well as sequence divergence between other species. This information can be obtained from running the pipeline on all Drosophila species and another model organism separately, and then comparing their similarities. This would tell us a few things: (1) the time the TE arrived within the Drosophila genus, (2) its activity across the genus, (3) its mutagenic capabilities, and (4) its effect on the host genome. We will also be looking at sequence similarities between and across these species to estimate the activity of these TEs.

1.6 Repeat Identification

As seen above, sequencing strategies are not all alike; they have their own advantages and downfalls. In the same way, methods of identification of TEs are also very different. They are divided into Homology-based identification and de novo identification.

1.6.1 Homology based

This is a much faster approach to identifying TEs, which is accomplished by comparing the sequence of the individual to a database of known TEs. Its accuracy depends on

the completeness of its library. This is very effective and accurate if all the known TEs are in the database, and it will miss TEs in the genome if their sequence is absent from the database. It is unlikely that every single TE is present in the database as new TE families and TEs are regularly discovered.

1.6.2 *de novo* based

de novo identification is much slower than homology-based identification but is much more complete as it searches for new repeats without the dependence on a database. It is more complete than homology-based identification as it can identify potential TEs that are not present in the homology database. It is able to do this based on the fact that TEs have a very similar sequence and are usually present in multiple copies in the genome.

We are in the process of creating a set of comparable TE annotations across *Drosophila* via a repeat analyzing pipeline. We classified these repeats *de novo* using RepeatModeler and RepeatMasker, and created custom `Python` scripts to filter out new repeats based on previously classified ones and assign them to a family based on sequence similarity. We hope to have a single-command pipeline that can be used for *de novo* identification of repeat sequences from any species; however, we have been working solely with *Drosophila* ourselves so that we could compare our results and postulate whether they were accurate for the pipeline to be used by other species. Having a database or a pipeline to create such a database is crucial to understanding TE activity across species and their evolutionary constraints and effects on the host genome.

We hope to be able to answer the following three questions after the completion of the pipeline, and after we have run the pipeline on the sequences from the *Drosophila* genus attained through different sequencing technologies:

1. Is the abundance of simple repeats positively correlated with the abundance of transposable elements?
2. Does sequencing technology (Nanopore vs. Illumina vs. Sangar) affect identification of TEs?
3. *D. melanogaster* has many young, active LTR retroelements. Are there evolutionary transitions within *Drosophila* where some species become dominated by other families of TEs, or do all species of *Drosophila* have abundant LTR retroelements?

1.7 Advantages of Our Study

Studying TEs across species of the *Drosophila* genus is confounded by:

1. Effort in identifying TEs within each species; and
2. the Quality of the assembly.

Studies such as the *Evolution of genes and genomes on the Drosophila phylogeny* [1] and *Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies* [105] have a few very potent disadvantages, when accounting for the repeat content within *Drosophila*. These specific studies were confounded by:

1. having used short read sequencing technologies such as ILLUMINA in order to estimate TE content.
    - This method is inherently faulty as it does not adequately assemble repeating elements longer than a certain number of basepairs (maximum as 100 for ILLUMINA).
    - Regions that actually have repeats within them will get collapsed onto themselves, increasing the coverage values for those regions.
2. having variable quality of assemblies that do not allow for proper annotation to sequences.
    - In the *12 Genomes* paper [1], some species were sequenced very deeply, and others were sequences with low coverage.
3. differences in the amount of effort used to generate each of these assemblies.
    - This might also affect in identification of TEs.
    - A bias in the amount of TEs annotated per species is created when not having the same strategy and parameters when sequencing and assembling the genome of each of these species, respectively.

Since our study was aimed at identifying the variability of TE content within the *Drosophila* genus, we needed to account for all the aforementioned scenarios. Thus, our study has the following advantages:

1. standard approach to identify TEs across all species.
    (a) Differential protocols to elucidate TEs within a genome lead to differential annotations of TEs and thus, an imperfect correlation.
2. Using assemblies generated using the same approach.
    (a) We used the genomes provided in the *Drosophila 15 Genomes Project* by Miller et. al [106].
    (b) We used these as they had been assembled using the same protocols, and using NANOPORE (long read) data, which also allowed us to properly annotate the frequency of each of these TEs.

Chapter 2

Methods

All work was done on Amarel, a "condominium" style computing environment developed to serve the university's wide-ranging research needs. We had to download Anaconda, a Python package management software in order to manage python packages and for easy downloading of particular software. Each program was called using a shell script specific to Amarel's guidelines.

2.1 Installing Programs and Packages

We would have omitted this section, as most programs and packages can be installed using Anaconda; but the two programs used most, RepeatMasker and RepeatModeler are not contained within the Anaconda environment, and have dependencies of their own.

What follows is a brief description of how we installed each program we used to create this *de novo* pipeline along with the command(s) and/or methods required to install programs not contained within Anaconda. A brief description of programs installed using the websites of their respective developers is listed in [Table 2.1]; and, programs installed using Anaconda is provided in [Table 2.2].

2.1.1 Main Programs

The major programs we used were downloaded from the developper site, and main programs we used downlaoded from developer site. mentioned in table x

2.1.2 Other programs

All other programs could be installed using Anaconda, a Python package manager. All programs could be installed using `conda install <name of program>` or an equivalent command on the Bash command line. Bash is a simple, yet powerful programing language that runs on the Linux shell.

Most Anaconda install commands could be found on the Anaconda website, or with a quick Google search.

Further information of all programs is contained within Section 6.

TABLE 2.1: Programs installed directly from developer site

Program Name	Version	Function
RepeatModeler	Open-1.0.11	*de novo* repeat family identification and modelling.
RepeatScout-1	1.0.5	Discover repetitive substrings in DNA.
RepeatMasker	Open-4.0.7	Align REPEATMODELER sequences to proper name if possible.
REPEATMASKER Libraries	Vol.23; Issue 10	Includes TRF libraries; library of all annotated repeats in multiple species.
RECON	1.08	REPEATMODELER dependency; automatic *de novo* identification.
nseg	○	RepeatScout dependency; low complexity sequence identification

2.2 Creating the *de novo* pipeline

The programs mentioned above were used in order to create the pipeline for *de novo* identification of repeat elements.

Many challenges needed to be overcome and optimizations applied in order to faithfully represent the presence of repeats, specifically TEs, in the genomes of *Drosophila*. The overview of the whole pipeline is shown below in [Figure 2.1] with a more simplified version in [Figure 2.2], and reasoning and particular information for each stop follows it.

2.2.1 Obtaining Sequence Data

2.2.1.1 Nanopore Sequences

NANOPORE sequences were also obtained from the GITHUB repository by DANNY MILLER titled Drosophila15GenomesProject, which contained assembled NANOPORE sequences of the following species that were created using their corresponding stock numbers [Table 2.3].

We then had to unzip the files, which was done by the single command:

```
gunzip $FASTA_name
```

LISTING 2.1: Unzipping Fasta

2.2.2 Running RepeatModeler

REPEATMODELER is a *de novo* repeat family identification and modeling package that is a pipeline that runs many other programs such as RepeatScout, RECON, TRF, all of whose information is given in the section "Programs Used"[6].

A REPEATMODELER Database was needed so that REPEATMODELER could be run.

TABLE 2.2: Programs installed using Anaconda

Program Name	Version	Function
`Python3`	`3.6.5 [GCC 7.2.0]`	General purpose file management and calculation.
`BedTools`	`2.27.0`	Genome arithmetic and format conversions.
`Perl5`	`5.8.8`	RepeatModeler dependency; programming language
`TRF`	`4.0.4`	RepeatModeler dependency; public database of tandem repeats
`NCBI BLASTX`	`2.5.0+`	Find possible gene alignments from all sequences.

RepeatModeler simply identifies repeats in the genome assembly and wherever possible, renames them to repeats that have already been identified. This is beneficial, as it saves time for identifying repeats *de novo*, for those sequences already present in the literature. This segment of RepeatModeler does a homology-based search to minimize on time the algorithm takes to run.

Despite RepeatModeler identifying repeat elements within the genome assembly, it does not know what to annotate novel repeats as; furthermore, it also identifies repeats that are not transposons, such as: repeating genes from multi-copy gene families.

2.2.3 Running UCLUST

RepeatModeler runs itself multiple times in order to better annotate repeats. It then creates a directory in which it stores all the information of these runs within subdirectories. It also creates a few summary files [Table 2.5].

Since `consensi.fa.classified` is in a very easily parsable format and also has more information, this file was used for further analysis. This file had the whole sequence of the repeat identified, and wherever possible, a possible annotation to a class.

UCLUST, a clustering algorithm that "merges" similar sequences to clusters, was run. This process is described in [Figure 2.4] and [Figure 2.5]. Sequences that might have been caught by RepeatModeler multiple times and annotated as something different were eliminated in this step. Those sequences were clustered and reported only a single sequence that encapsulated all the information from other sequences.

2.2.4 BLAST

Basic Local Alignment Search Tool (BLAST), created by the National Center for Biological Information (NCBI) is a tool that finds regions of similarity between biological sequences of DNA, RNA and peptides. There are many flavors of BLAST [Table 2.4] that finds similarities across and between these biological sequences.

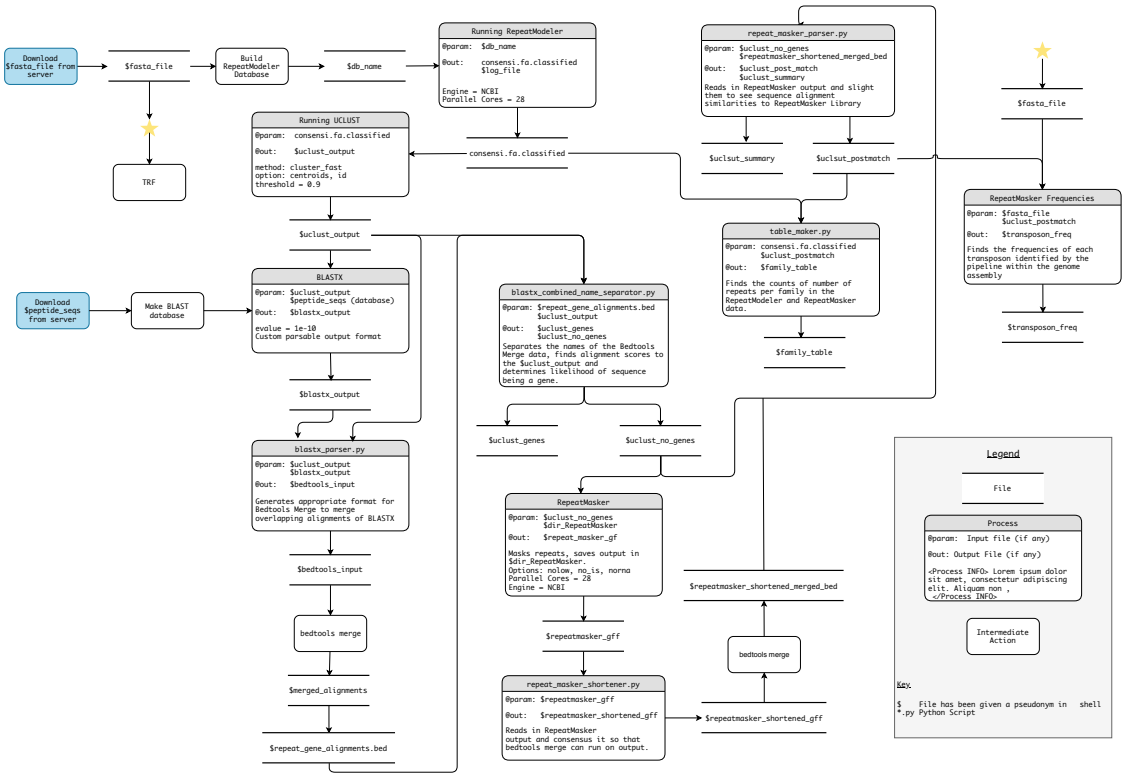


FIGURE 2.1: Pipeline Overview

It compares the query that we give it against a database, which we might provide, or that is already provided in the program suite.

There are certain genes that are present in multiple copies called gene families [107]. Therefore, after clustering the REPEATMODELER data using UCLUST, removal of those repeat sequences which were genes was done in order to identify probable TEs.

Only the peptide sequence for *D. melanogaster* was downloaded from Fly Base Genome Releases FTP Client, as it has been extensively studied.

Though this was a peptide sequence, BLASTX requires a custom database format for lookups, which had to be generated; after which, the output from UCLUST could be run through BLAST and we could get a BLASTX output.

Contained within the BLAST output were a list of matches with the name of the UCLUST query along with its aligned gene in the *D. melanogaster* peptide sequence. Also contained within it was alignment information, encapsulating the start and end of the query, start and end of the sequence (database entry), and the length of the alignment.

This file could be easily parsed in order to see which repeat sequences aligned to genes.

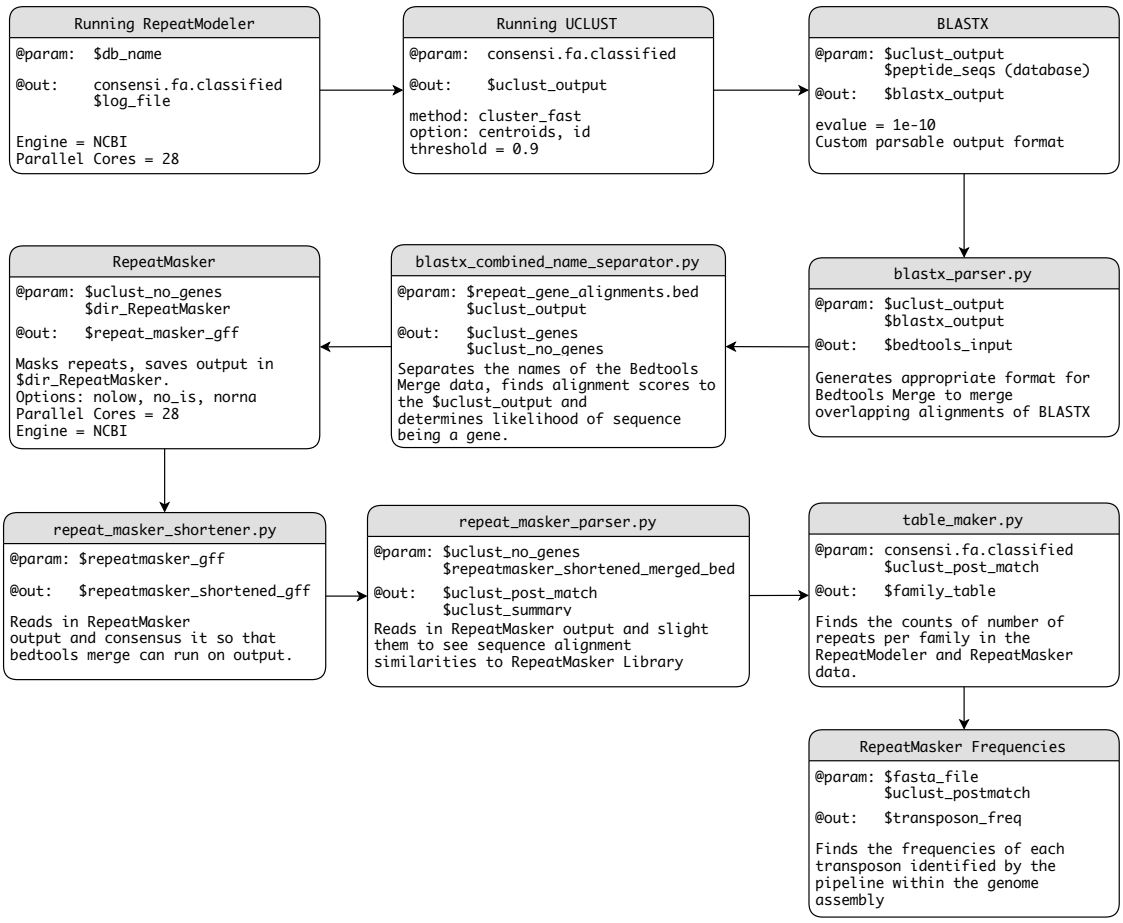


FIGURE 2.2: Pipeline Overview [Simplified]

2.2.5 BedTools Implementation and Custom Pipeline 1

After finding particular repeats that were aligned to genes, we needed to parse through them and find out which ones were actually genes. No programs existed to perform the tasks we required, so we had to generate our own pipeline using PYTHON and BEDTOOLS.

Alignments from the BLAST output were grabbed and formatted it in a way that was readable by BEDTOOLS.

IDs and their alignment information had to be given to BEDTOOLS MERGE (`mergeBed`). `mergeBed` merges overlapping alignments and outputs a single longer alignment [Figure 2.8]. Alignments to the reverse strand were identified by having the start alignment locus to be after the end alignment locus, and it was switched before running the data through mergeBed.

TABLE 2.3: NANOPORE Assemblies from GITHUB

Species	Stocks and Stock Numbers
D. ananassae	14024-0371.13
D. biarmipes	14023-0361.02
D. bipectinata	14024-0381.07
D. erecta	14021-0224.01
D. eugracilis	14026-0451.02
D. mauritiana	14021-0241.01
D. mojavensis	15081-1352.22
D. persimilis	14011-0111.01
D. pseudoobscura	14011-0121.94
D. sechellia	14021-0248.01
D. simulans	14021-0251.006
D. triauraria	14028-0691.9
D. virilis	15010-1051.87
D. willistoni	14030-0811.00
D. yakuba	14021-0261.01

MERGEBED would identify the length of the alignment, but not isolate those alignments based on highest alignment length or score. This had to be done using another custom piece of PYTHON titled `blastx_combined_name_separator.py`. This would merge single gene-repeat alignments and report the maximum alignment to the gene.

Figure 2.8 shows the procedure of `blastx_combined_name_separator.py`. It first merges the two alignments for `Repeat-ProteinA alignment`; after which it compares the lengths of those alignments.

$$Alignment_A + Alignment_B < Alignment_C$$

As the length of `Repeat-ProteinB` is greater than `Repeat-ProteinA`, it is more likely that the repeat aligns to `ProteinB`.

Despite having a higher alignment to the `Repeat` than `ProteinA`, the repeat element might still not be a gene. If the repeat aligned to the gene very little, it is unlikely that the repeat is a gene. To account for this, we had to check if the following condition was met:

$$\frac{Length\ of\ Alignment}{Length\ of\ Repeat} \geq 0.5$$

If the above condition was met, then it is likely a gene as it aligns very well to the gene with some small gaps at the end. The threshold of 0.5 was chosen as it is a standard, but can be easily altered if need be. It was chosen as it is a good estimate of whether a

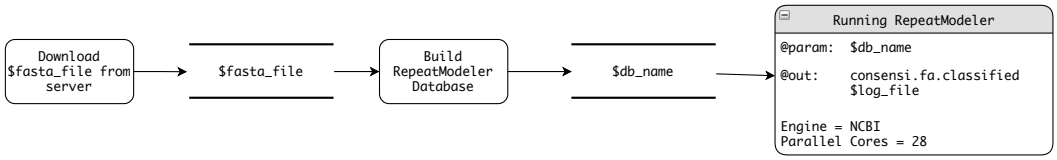


FIGURE 2.3: Running RepeatModeler

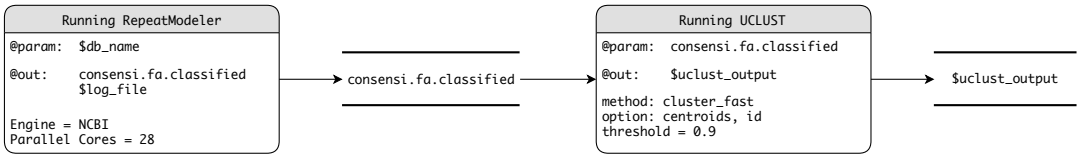


FIGURE 2.4: UCLUST

sequence is a gene or not. Too high more genes occur in the output file and thus won't be efficiently filtered.

If the condition passed, then those repeats were added to the file `$uclust_genes`, and if the condition did not pass for that alignment, they were added to the file `$uclust_no_genes` for further filtering and analysis.

2.2.6 RepeatMasker and Custom Pipeline 2

Two files were returned by our implementation above, one with repeats likely to be genes, and another not including those sequences.

We used REPEATMASKER to annotate the `$uclust_no_genes` if they matched previously described TEs. The REPEATMASKER parameters we used are present in Table 2.5.

We also had to account for Chimeric sequences, which are sequences that are made from two or more TEs. We used the same threshold as we did for Unknown sequences. More information can be found out from `repeat_masker_parser.py` on GITHUB.

REPEATMASKER returned sequences that had aligned to repeats that have already been identified to be of particular families and classes. Our custom pipeline then accounted for those repeats and included them in the `$uclust_post_match` file.

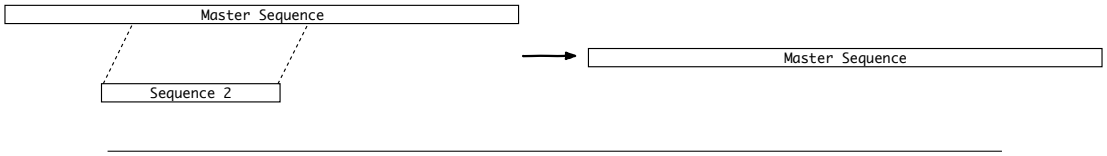


FIGURE 2.5: UCLUST Algorithm

TABLE 2.4: Types of BLAST algorithm

BLAST Name	**Query**	**Database**
BLASTN	Nucleotide	Nucleotide
BLASTX	Translated Nucleotide	Peptide
tBLASTN	Peptide	Translated Nucleotide
BLASTP	Peptide	Peptide

The `$uclust_summary` file encompasses all repeats that have been classified and those that have not.

For the newly unidentified sequences, we needed to summarize if they had been identified as new repeat sequences, or if they had been identified by REPEATMASKER.

2.2.7 Transposon Frequencies

Transposon counts within the genome were calculated by adapting REPEATMASKER to output the alignments of the identified transposon sequences, and then counting those alignments with code.

The REPEATMASKER algorithm is described graphically in [Figure 2.11]. This shows how REPEATMASKER collects the data from the 5 predefined TEs (for illustrative purposes), finds the sequence within the genome assembly, and then outputs their frequencies, as well as their locations within the assembly. [1]

2.2.8 Summarizing output

After completing the run of the pipeline, we needed to confirm that it has worked as to our expectations. We also needed to summarize all identified repeats. We ran a

[1] Please note that the numbers that appear within [Figure 2.11] are pseudo-random and do not mean anything. They are simply present to help explain the methods with which we adapted REPEATMASKER to attain our results.

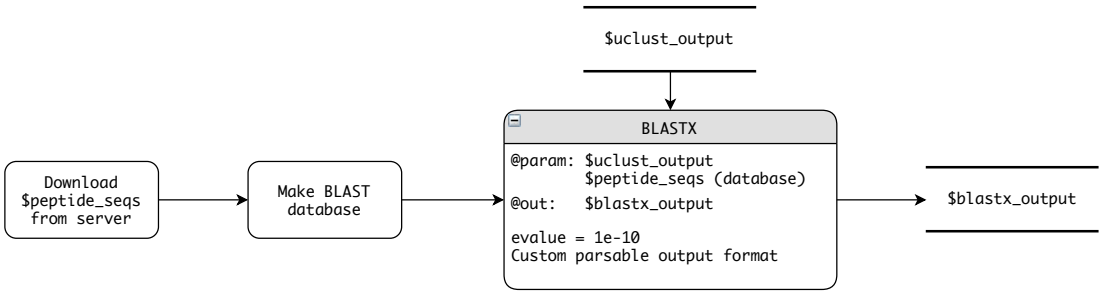


FIGURE 2.6: BLASTX

custom PYTHON script to parse through the output files and grab TE families that were identified.

We created a summary file that has the following information about the run:

1. Species of the run.
2. Recent modification date/time.
3. Date/Time of run.
4. Total number of sequences identified by REPEATMODELER.
5. Total number of sequences clustered by UCLUST.
6. Total number of base pairs of TEs within the genome assembly.
7. Total number of base pairs of simple repeats within the genome assembly.
8. Size of the genome assembly.
9. Number of new sequences identified that are not genes.
10. Names of aforementioned sequences.
11. Number of new sequences identified that are genes.
12. Names of aforementioned genes.
13. Working directory information.
14. File summaries of:
    (a) Python scripts used.
    (b) Input/Output files.
    (c) Intermediate files used.

NOTE: Summaries of unused intermediate files and temporary files not included.

All of this information was useful in creating and analyzing the data, within and outside the pipeline. It also assisted us in easily parsing the data without requiring to open each intermediate file and reading its data/metadata.

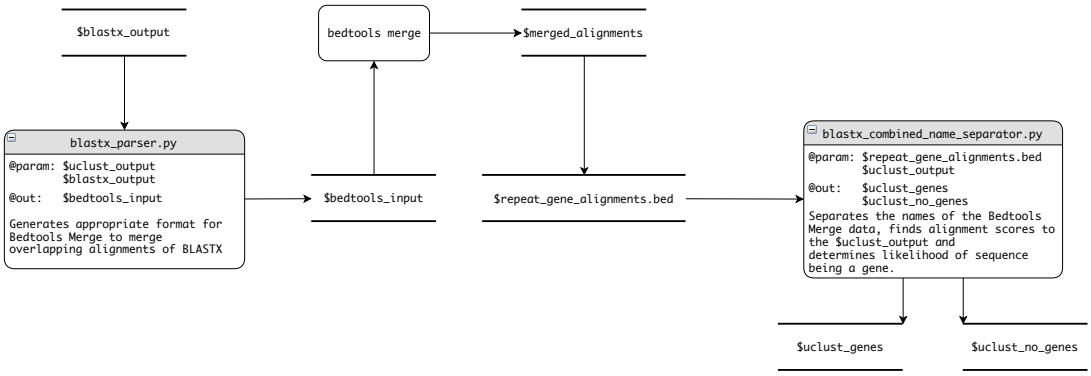


FIGURE 2.7: Custom Pipeline 1

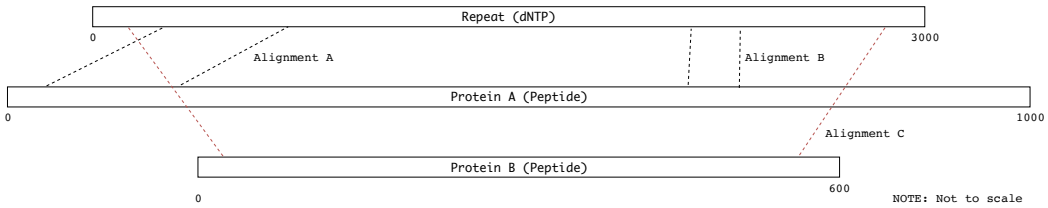


FIGURE 2.8: BEDTOOLS Merge Algorithm

2.3 Analyzing Data

2.3.1 Quality of Assemblies

Assessment of assembly qualities was required in order to ascertain whether we had good data to begin with. We estimated the assembly quality for each species using N_50, a common metric defined as:

> The N_{50} is defined as the minimum contig length needed to cover 50% of the genome.

This means that at least half (50%) of the assembly is contained within the contigs that are the N_{50} or larger [2].

[2] Keith Bradnam gives a good explanation of the N_{50} statistic on his blog acgt.me.

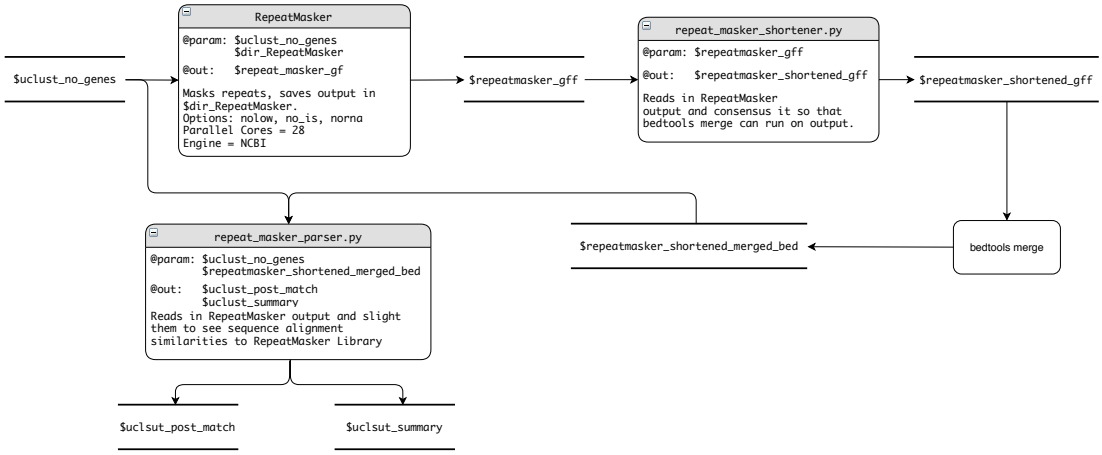


FIGURE 2.9: REPEATMASKER and Custom Pipeline 2

2.3.2 Runs

The run of each species was contained within a single directory with the name structure:

`<species>_<sequencing method>_<date of run>`

This included the output files, the summaries as well as the data files which we needed to analyze collectively.

In order to account for this scattering of data, we copied all relevant files into another directory that was out of the directories which contained the runs for each species called `identified_TEs`.

2.3.3 Spearman Correlation

A lot of correlation plots were created (emphasized in §2), for which we needed a correlation coefficient.

The most common form of correlation, the Pearson's correlation test, also includes outliers, which would skew the data quite a bit. In order to account for this, we used Spearman's correlation.

r_s is the Spearman Correlation, as defined by Equation 2.1:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{2.1}$$

where:

- $d_i = rg(X_i) - rg(Y_i)$, the difference between the two ranks of each observation; and

TABLE 2.5: REPEATMASKER Parameters

Parameter	**Description**
`-e ncbi`	Using NCBI as a search engine.
`-pa 28`	Using max number of cores assigned by the system.
`-norna`	Eliminates/excludes RNA sequences.
`-no_is`	Eliminates/excludes bacterial sequences.
`-gff`	Parsable output file format.
`-species drosophila`	Specifies species database for identified repeats; though stating drosophila, it only encapsulates *D. melanogaster*.
`-dir $dir_RepeatMasker`	Specifying the directory to be made and data copied into.

- n is the number of items.

Spearman's correlation was used as it proves/disproves correlation appropriately by insulating the effects of outliers within the data [109].

Since our dataset is unlikely to contain ties, the above formula would report the correct values. When the data contains ties, Equation 2.2 would have had to be used.

$$r_s = \rho_{rg_x, rg_y} = \frac{cov(rg_x, rg_y)}{\sigma_{rg_x} \sigma_{rg_y}} \tag{2.2}$$

where:

- ρ is the standard Pearson's correlation;
- $cov(rg_x, rg_y)$ is the covariance; and
- $\sigma_{rg_x}, \sigma_{rg_y}$ are the standard deviations of the ranked variables.

Formula selection was done within R by default. We did not have to specify the formula used.

2.3.4 Repeat Frequency

In order to find the frequency of certain classes of TEs, we needed a way to isolate the classes. This was accomplished by a few scripts of PYTHON code, namely:

1. `trf_repmask_condensor.py`
2. `merged_Sat_to_out.py`
3. `class_condensor.py`; and
4. `class_name_condensor.py`

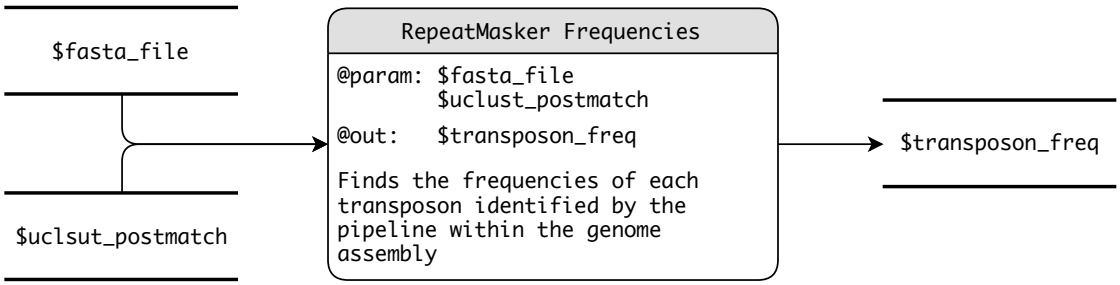


FIGURE 2.10: Using REPEATMASKER to identify TE copy number within the genome assembly

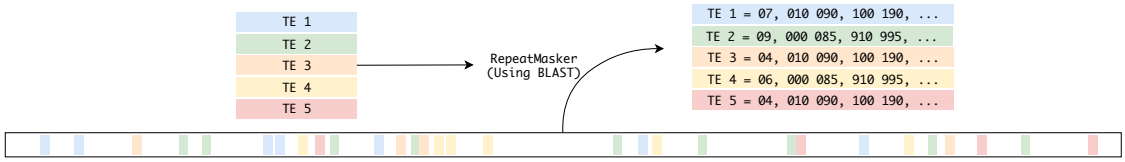


FIGURE 2.11: REPEATMASKER Algorithm

The first three scripts were performed on each individual run of a single species, but the last scrip (`class_name_condensor.py`), was carried out on the sum of the output of all runs, so that we could get a single file that contained all the data we wanted.

The code for all these scripts is provided in the appendix in the order they were run.

2.3.5 Simple Repeats, Satellite Sequences and Unknown elements

REPEATMODELER identified all repeats in the genome and attempted to classify them into categories such as:

1. LTR,
2. LINE,
3. DNA,
4. RC...

and for those it could not classify with vigor, it identified them as UNKNOWN. For sequences it attempted to identify, but is not completely sure of their class, it assigns them the class, and appended the class name with a question mark (`?`), for example with `[ SINE? ]`.

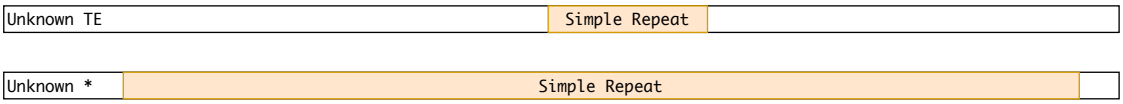
During analysis, UNKNOWN sequences got classified as either Satellite sequences and/or Simple Repeats by TRF. At the end of our analysis, most if not all UNKNOWN repeating sequences were identified as either Satellites or Simple Repeats, so in the following file:

`./identified_TEs/<species>_condense_classes.txt`

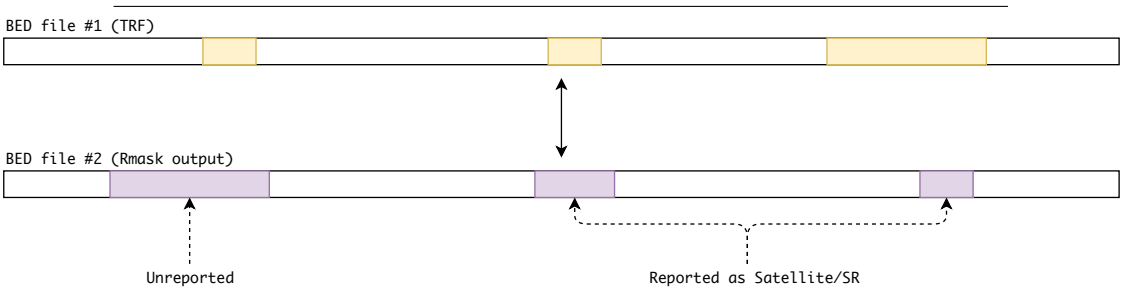
it is key to note that Simple Repeats and Satellites are the same as all UNKNOWN repeat elements, so the base-pairs of either set can be ignored for further analysis as they mean the same. UNKNOWN repeats were ignored as they have been classified as one of the other.

BEDTOOLS INTERSECT (`intersectBed`) was used in order to find the overlap between these sequences, and a diagram visually elaborating this process is shown in [Figure 2.12].

Figure 2.12a shows whether a certain repeat is likely a TE or wholly comprised of a Simple Repeat. 0.8 rational coverage was used by `intersectBed` to define this boundary. Figure 2.12b shows the input for `intersectBed`. It only outputs the two sequnces on the right of the second BED file, as the ratio of $rmask : TRF \geq 0.8$. The *rmask* signifies the raw REPEATMASKER ouput that encapsulates all identified repeats (including UNKNOWN repeat annotations). This was intersected with the TRF ouput (which contained Simple Repeats) and the resultant was shown. Reassignment of IDs could then be done to those reported sequences.



(A) UNKNOWN TE vs. UNKNOWN Simple Repeat



(B) `intersectBed` input

FIGURE 2.12: UNKNOWN accomodation

2.4 GitHub

All the code used to create the data we will soon talk about is provided in the GITHUB repository called `de_novo-identification` curated by me, `Dr-Drosophila`.

Due to the inability to store large files due to GitHub's size restrictions on free accounts, we are not able to provide the FASTA files for the sequences we used, but they have been linked to earlier in the document.

Chapter 3

Results

3.1 Genome Assemblies

We retrieved 16 assemblies from two studies [106], [110]. These assemblies were created using the same protocols, thus allowing us to compare our data across species. They were produced using NANOPORE sequencing, which produces long reads of the order $\sim 10\frac{kbp}{read}$. This gives us confidence in our ability to compare our data across each species.

A tabular and graphical representation of the quality of the assemblies is provided in Table 3.1 and Figure 3.1.

TABLE 3.1: Genome Assembly Qualities of *Drosophila*

Species	**Assembly size**	**Contigs**	**Average contig size**	N_{50}
D. ananassae	189221946	371	510 032.199 461	2612784
D. biarmipes	182453935	661	276 027.133 132	2791184
D. bipectinata	163165444	570	286 255.164 912	567431
D. erecta	130293209	58	2 246 434.637 931	16960765
D. eugracilis	159429531	546	291 995.478 022	1010701
D. mauritiana	134165749	266	504 382.515 038	4738483
D. melanogaster	131856353	208	633 924.774 038	3866686
D. mojavensis	168142858	122	1 378 220.147 541	5220960
D. persimilis	163933157	415	395 019.655 422	3429058
D. pseudoobscura	159031139	361	440 529.470 914	2983193
D. sechellia	138120607	109	1 267 161.532 110	7712364
D. simulans	133725236	76	1 759 542.578 947	7762389
D. triauraria	173623250	482	360 214.211 618	741655
D. virilis	169714588	141	1 203 649.560 284	4170062
D. willistoni	194955081	489	398 681.147 239	1515988
D. yakuba	143252825	111	1 290 565.990 991	5227393

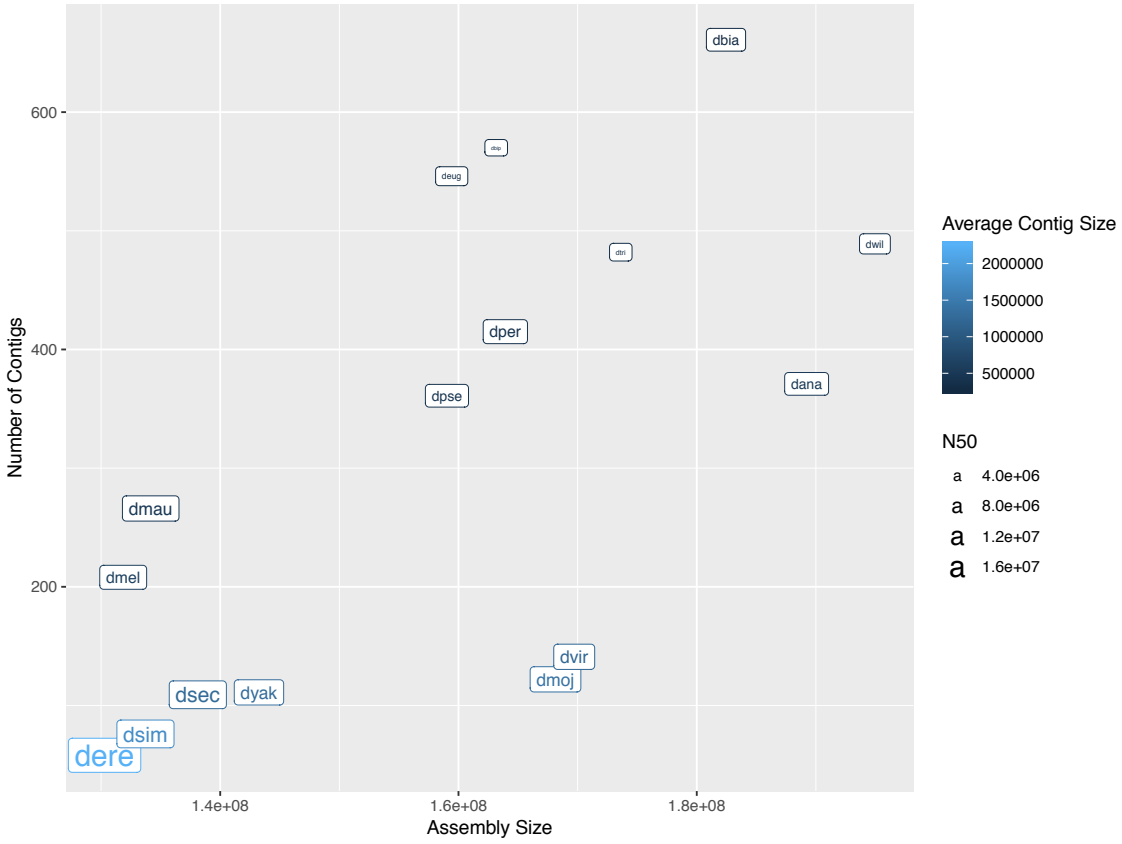


FIGURE 3.1: Summary of Genome Assembly Qualities

3.2 Pipeline Results

3.2.1 Pipeline Summary

The pipeline we ran consisted mainly of running RepeatModeler, and then cleaning up the data, which included assigning repeat elements identified by RepeatModeler to genes and annotating those not identified as genes to TE classes using RepeatMasker. A detailed version of the pipeline is shown in Figure 2.1 and a simplified version is shown in Figure 2.2.

We ran RepeatModeler to identify all repeating elements *de novo*. RepeatModeler takes anywhere between $15 - 27$ hours based on the size of the genome assembly; larger the assembly, the more time RepeatModeler takes to run. RepeatModeler identified between 446 sequences for *D. simulans* to 1459 sequences for *D. biarmipes*.

RepeatModeler outputs a FASTA of all sequences it identified to be repeats. Sometimes, those repeats recur within the output FASTA as RepeatModeler identifies sub-sequences as repeats themselves. In order to account for this, we ran UCLUST, a clustering algorithm that clusters sequences that are contained within larger sequences. This allows us to work with fewer sequences, and thus makes downstream annotation easier. In total, we clustered between 13.13% (for *D. melanogaster*) and 26.03% (for *D. ananassae*) of all sequences identified by RepeatModeler.

RepeatModeler identifies repeating sequences within the genome. However, not all repeating sequences are TEs; some might also be repeating genes. To account for these, we ran blastX, which reports significant alignments to genes. We used the `e-value` of 10^{-10}, as it would show proper alignments to genes within the *D. melanogaster* peptide sequence. We chose *D. melanogaster* as it has a very well annotated transcriptome. blastX further reduced the number of RepeatModeler sequences by 0.036% (for *D. melanogaster*) to 1.59% (for *D. ananassae*) of the dataset from UCLUST.

Though RepeatModeler attempts to annotate repeating elements as TEs, it is not perfect. We used RepeatMasker to properly annotate those repeats present in RepBase. RepBase has most if not all repeating elements that have been classified across the genus, and it was very effective in identifying most TEs.

More detailed information about the number of sequences identified at each step of the pipeline are given in Table 3.2.

3.2.2 Annotated TEs

RepeatModeler provides TE family annotations within its algorithm. However, from the RepeatModeler output, we are not able to identify what family the TE belongs to and whether it has been identified by RepBase.

RepeatModeler assigns each consensus sequences to a known class (LTR, DNA ...), as a Satellite or Simple Repeat; if unable to classify the element, it lists it as Unknown. RepeatModeler was able to assign a majority of the consensus sequences to a particular TE class. However, for the elements not assigned to a particular TE class between 1 (*D. pseudoobscura*) and 18 (*D. mojavensis*) were assigned to Sat/ST and between 27 (*D. melanogaster*) and 415 (*D. triarauria*) were assigned as Unknowns [Table 3.3]. It is apparent that *D. melanogaster* would have the least number of Unknown identifications as this species has been extensively annotated.

TABLE 3.2: Number of Putative TE Families Identified After Each Corresponding Step of Pipeline

Species	RepeatModeler	After UCLUST	After blastX
D. ananassae	1361	1006	990
D. biarmipes	1457	1151	1136
D. bipectinata	1161	937	920
D. erecta	482	386	372
D. eugracilis	1027	769	746
D. mauritiana	499	383	371
D. melanogaster	632	549	547
D. mojavensis	575	490	481
D. persimilis	1113	902	881
D. pseudoobscura	951	797	772
D. sechellia	619	479	465
D. simulans	446	369	357
D. triauraria	1019	902	886
D. virilis	651	514	506
D. willistoni	1380	1167	1140
D. yakuba	848	649	635

Some sequences were identified as rRNAs, but they have been omitted from the counts in Table 3.3.

3.2.3 Using RepeatMasker to Identify TE Classes

In order to classify each identified repeat as a family, we needed to run these sequences through REPEATMASKER which runs a homology-based search of a query, against the REPBASE sequences. REPEATMODELER annotated those sequences in Table 3.3.

Table 3.4 shows how many sequences moved from the Unknown category to being annotated as TEs and vice versa. More elements moved from being annotated as TEs to Unknowns as REPEATMODELER was rather lax in assigning classes to those elements. This is apparent by a decrease in number of TEs identified and a stark increase in Unknown sequences. We trusted REPBASE as it contains all the sequences we know so far, and is most widely used in homology-based searches.

Please note, that there is not a perfect correlation, as REPEATMASKER also assigns some Unknowns to Sat/SR, a category of repeats not shown here.

3.2.4 Unknowns Identified

We plotted the percent of Unknown sequences identified against the total amount of repeating sequences identified and the size of the assembly in Figure **??**. This was mainly done as a sanity check.

FIGURE 3.2: Number of Repeats Identified per Species

TABLE 3.3: REPEATMODELER Assignment of Families

Species	Classified as TEs	Sat/SR	Unknown
D. ananassae	1134	14	208
D. biarmipes	1121	11	326
D. bipectinata	950	10	201
D. erecta	414	9	59
D. eugracilis	883	16	128
D. mauritiana	448	11	40
D. melanogaster	595	10	27
D. mojavensis	439	18	118
D. persimilis	838	6	269
D. pseudoobscura	683	1	267
D. sechellia	574	8	37
D. simulans	396	10	40
D. triauraria	594	10	415
D. virilis	510	4	137
D. willistoni	1084	11	285
D. yakuba	729	9	110

D. melanogaster has a highly annotated genome, so we expected a low percent of Unknowns are accommodated within the genome assembly. This was correct.

We also noticed that *D. persimilis* and *D. bipectinata* have much lower percent of Unknowns, but this might be because:

1. There are no new TE insertions within these two species, or
2. All TEs within these two species have already been identified in other species.

3.2.5 Copy number for RepeatModeler Consensus Sequence

For each species, we needed to find the copy number of TEs of each class within the genome. In order to do this, we used REPEATMASKER again with different parameters and a custom library to find the frequency of each TE within the genome assembly. The custom library we used was the sequence information that had been clustered by UCLUST and the filtered for genes. This gave us the copy number of each sequence.

Figures 3.4 and 3.5 show the percent of repeats identified as each class and the base-pairs of each class across all the species.

Our study showed the same pattern of repeats as a previous study [105], in that we report the same percentages of annotations per total annotations. However, we were able to classify many more sequences as TEs and each of our categories has a higher number of basepairs as compared with the study. This was because their use of short read data, which reduces the frequency of TEs identified within the assembly. The genome coverage percent we identified is shown in Table 3.6.

TABLE 3.4: Number of Repeats Identified per Species

Species	TEs	Unknown
D. ananassae	679	311
D. biarmipes	656	480
D. bipectinata	646	247
D. erecta	286	86
D. eugracilis	547	199
D. mauritiana	307	64
D. melanogaster	496	51
D. mojavensis	220	261
D. persimilis	543	338
D. pseudoobscura	455	317
D. sechellia	375	90
D. simulans	296	61
D. triauraria	318	568
D. virilis	267	239
D. willistoni	674	493
D. yakuba	482	153

3.3 TE and Sat/SR Content vs. Genome Size

According to a previous study, the number of basepairs in *Drosophila* corresponding to a coding region are relatively constant [108]. This would also mean that abundance of TEs is directly related to the size of the genome assembly, and in turn, the genome.

It has been known for a long time that genome size is correlated with repeat content [111]; with larger genomes, come more repeats. We sought out to find which repeats correlate higher with increase in genome size. We also wanted to see how this correlation varies across different TE classes.

In Figure 3.7, a $r_s = 0.9088235$ between the repeats identified by REPEATMODELER (i.e. TEs) shows that there is a high correlation between TEs and the size of the genome. This correlation was highly significant as we noticed a $p < 2.2 \cdot 10^{-16}$.

Furthermore, some TE classes show a higher correlation with genome size than others [Figure 3.8]. The line shown for each facet in Figure 3.8 is a regression line, but it is clear to see that DNA, RCs and LTRs have a higher correlation to genome size as compared to LINEs, Sat/SRs and Unknowns.

We noticed a large range of correlations from $r_s = 0.761$ for the artificial class *Unknown* all the way to $r_s = 0.479$ for LTR elements. all of these correlations were significant with a $p < 0.05$ except for LTR, which and a $p = 0.062$.

With a rather high $r_s = 0.9088235$ for all repeating elements, and the relatively lower correlations of each of the other TE classes, we can say that there is not a single element, but rather all elements play a compounding role on the genome size of the species.

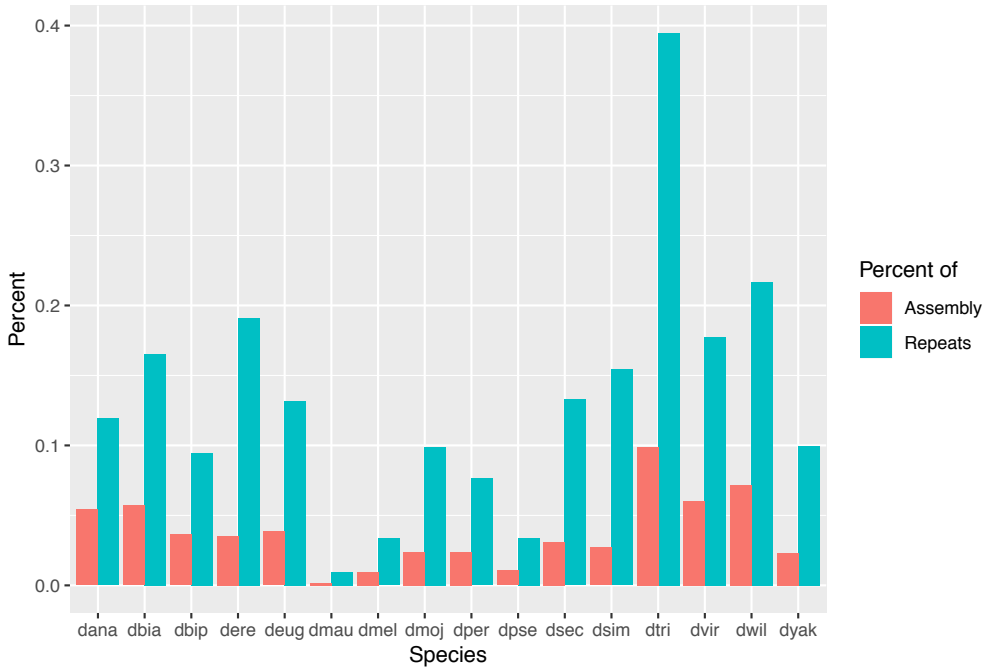


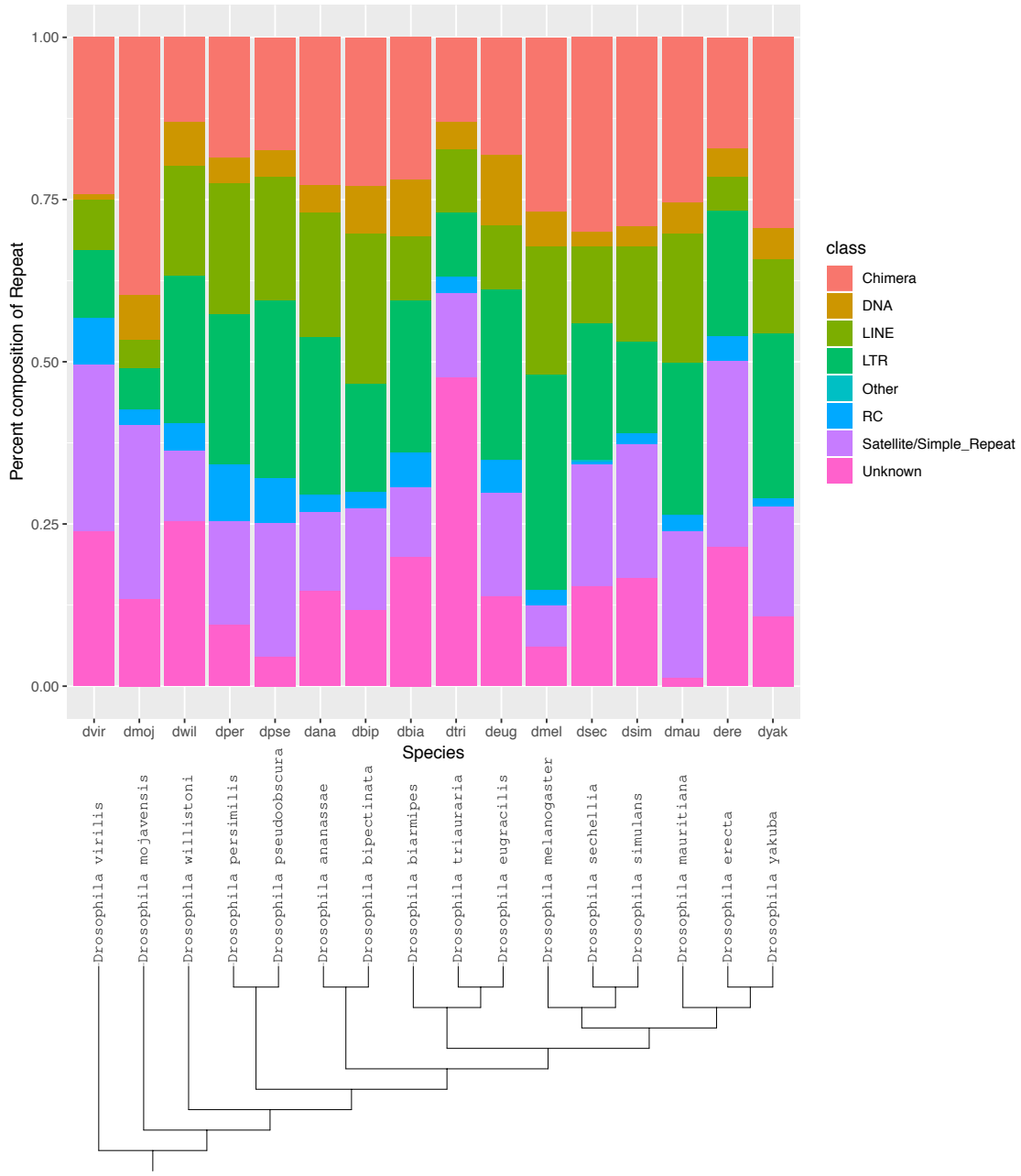
FIGURE 3.3: Unknowns identified as a percent of total Assembly Size and All Repeats

It is key to note that *Chimera* and *Unknown* are artificial categories created by us in order to simplify analysis. *Unknown* contains all elements that might have been identified by REPEATMODELER, but did not align sufficiently with any sequence in REPEATMASKER. *Chimera* contains all elements that match with something in REPEATMASKER, but match very highly when two elements are combined, and are a product of two classes of elements being present in a single repeating element identified by REPEATMASKER.

3.4 Genome Size contraction in *melanogaster* subgroup

As is evident in 3.5, the *melanogaster* subgroup, consisting of *D. melanogaster*, *D. simulans* and *D. sechellia* all seem to have a large decrease in TE content, especially DNA elements. This implies that there was a large decrease in DNA element content in the ancestor of the *melanogaster* subgroup.

This effect also seems to be expanded to *D. erecta*, *D. yakuba* and *D. mauritiana*. All of these species have a drastic reduction in DNA element content relative to other species in *Drosophila*.



Percent composition of each class or repeating elements as a ratio to the total amount of repeats within each species.

FIGURE 3.4: Abundance of TE Classes – Ratio

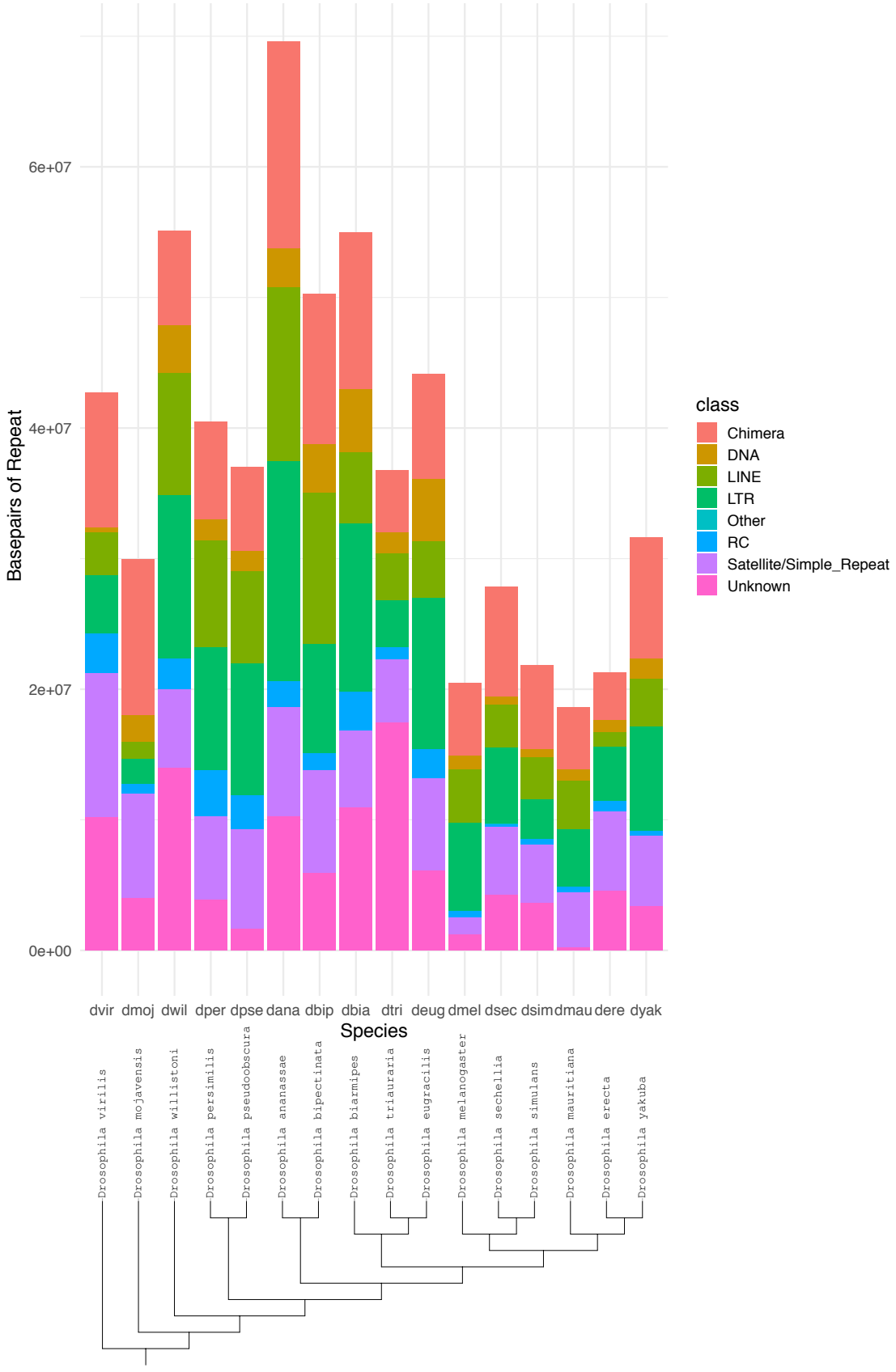


FIGURE 3.5: Abundance of TE Classes – Raw BPs

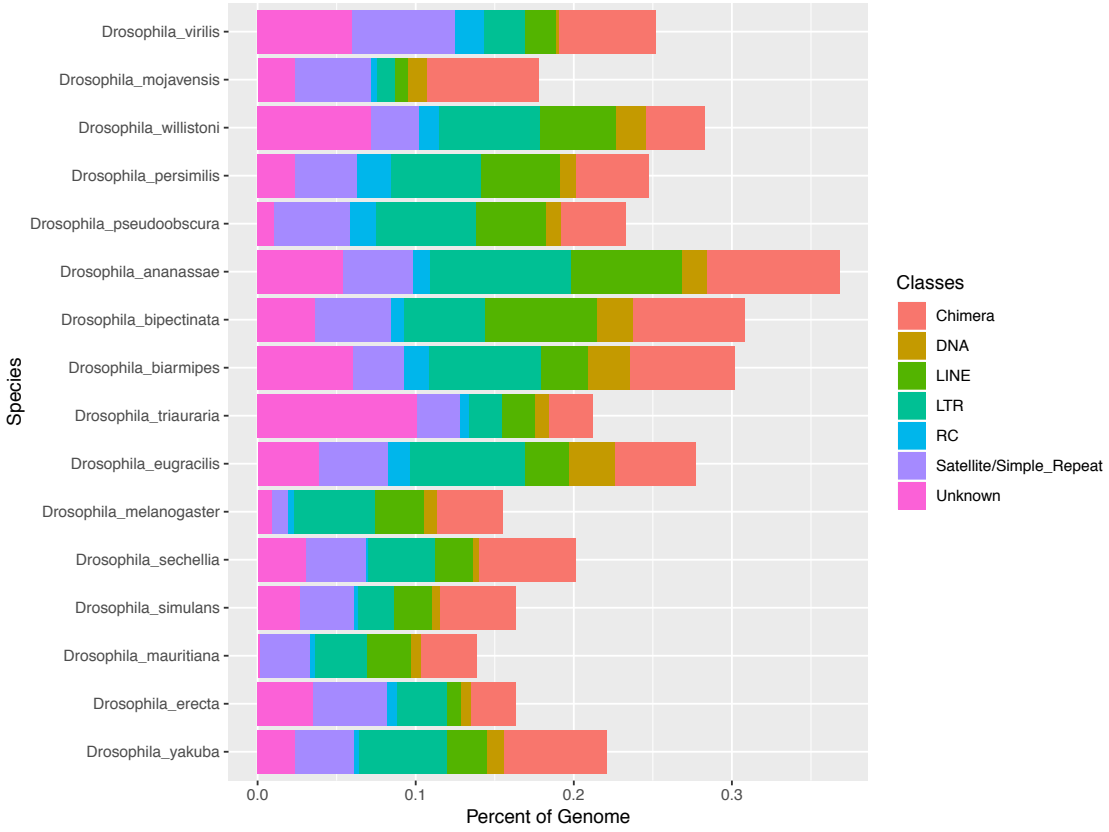


FIGURE 3.6: Percent of Genome covered by Repeating Elements

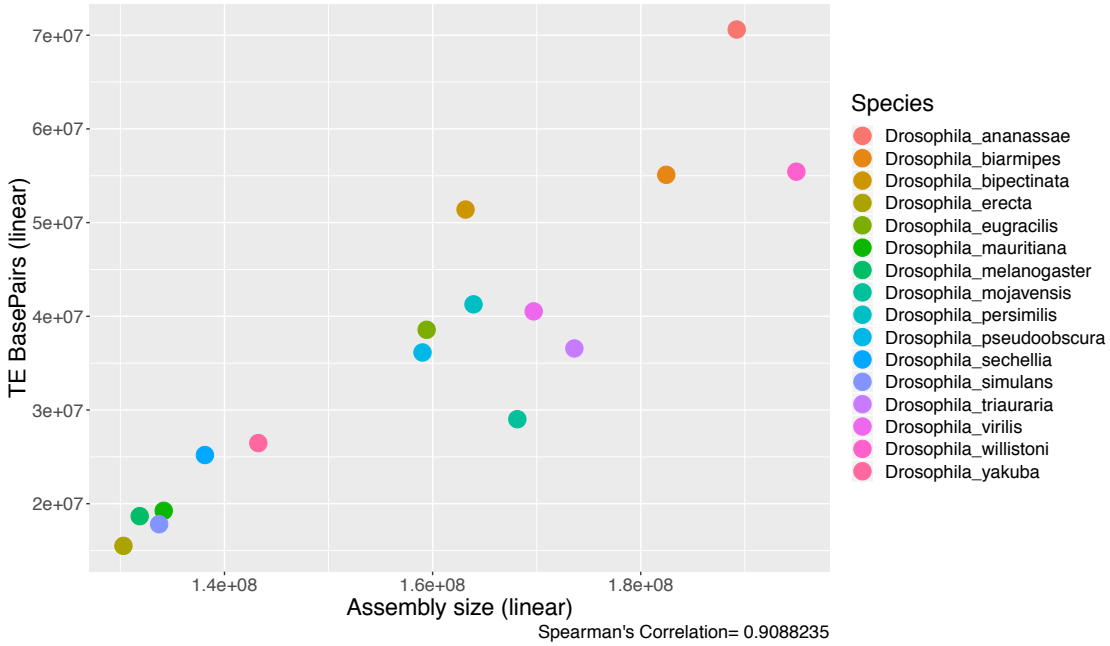
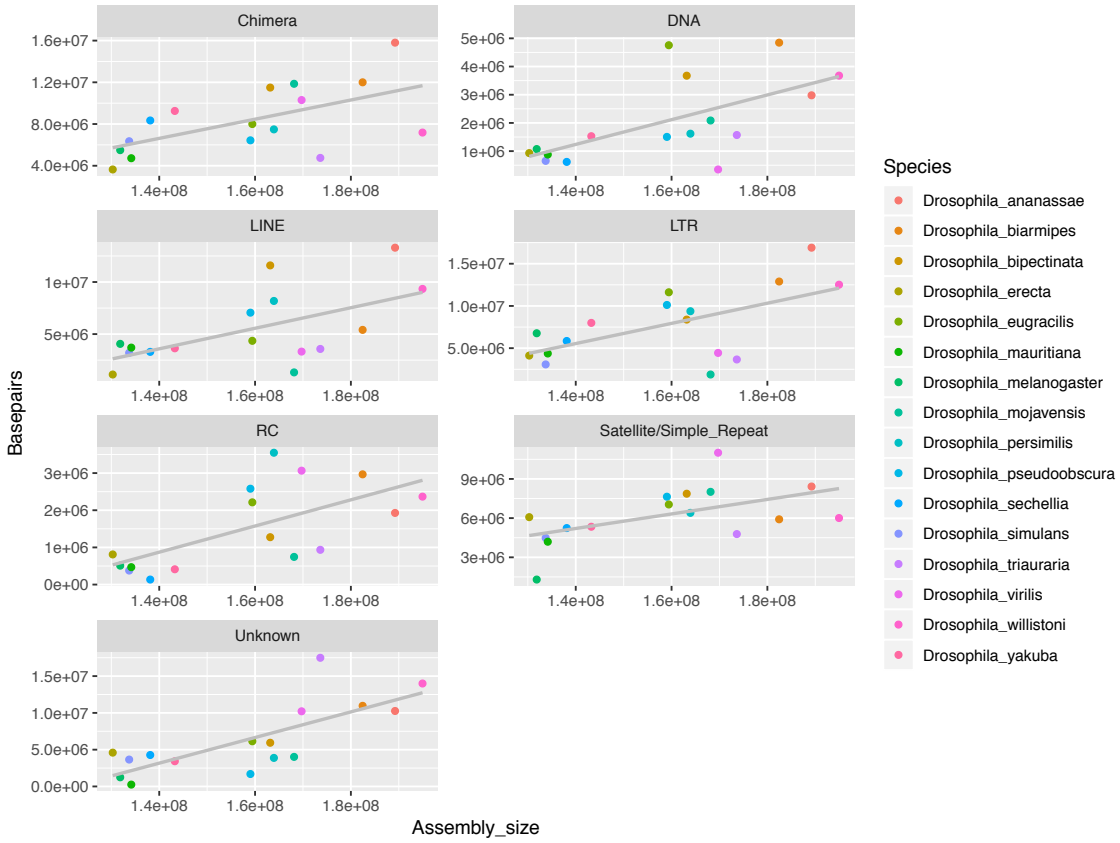


FIGURE 3.7: Genome Size and Relative Abundance of Repeats



Basepair correlation with assembly size, faceted by class of each repeating element analyzed.

It is key to note that *Chimera* and *Unknown* are both artificial classes created to simplify study.

Element	r_s	p	**Element**	r_s	p
Chimera	0.594 118	0.017 250	DNA	0.611 765	0.013 640
LINE	0.508 824	0.046 440	LTR	0.479 412	0.062 390
RC	0.629 412	0.010 660	Sat/SR	0.502 941	0.049 350
Unknown	0.761 765	0.000 927	–	–	–

FIGURE 3.8: TE class vs. Assembly Size

Chapter 4

Discussion

4.1 Recap of Work

In summary, we used RepeatModeler to identify repeating sequences from Drosophila genome assemblies that have been created using the same protocols and parameters *de novo*. After this, we ran UCLUST in order to cluster repeating sequences that had been annotated more than once by RepeatModeler. We then ran blastX to isolate repeating gene sequences, after which we ran RepeatMasker to: (1) assign TE classes to identified TEs, and (2) to find the frequency of the TEs within the genome assembly. We also ran TRF to better account for Sat/SR sequences.

4.1.1 Results Summary

Across all the species, we noticed TE annotation frequency range from *D. mojavensis*'s 220 annotations to *D. ananassea*'s 679 annotations. Across all the species, the basepairs of TE annotation and Sat/SR ranged from *D. erecta*'s $18,964,984$ bps on the low end to *D. ananassae*'s $66,360,109$ on the high end [Figure 3.5]. TE content as a fraction of genome size ranges from 0.12 for *D. erecta* to 0.37 for *D. ananassae* [Figure 3.6].

4.2 Comparison with Previous Studies

The number of basepairs we identified to be repeats within these species were larger compared to previous studies. We also identified a larger percent of genome size. This is likely because we used long-read Nanopore sequences, which allow for better identification of repeats.

The previous studies [1], [105], [112] used a combination of:

1. short reads: which decreases repeat identification due to excessive read overlap of similar sequences; and
2. variable sequencing strategies and parameters: which, though might increase TE annotations, does not allow for appropriate comparison across the species due to inconsistency in acquiring genome assemblies.

We have accounted for both shortcomings using Nanopore assemblies generated using the same protocols by *Miller et. al* [106]. Since we used assemblies generated using the same protocols, we can be confident in comparing our data across species.

4.3 Consensus Sequences absent from RepBase

We identified 3958 Unknown elements in total for all the species. These Unknown repeating sequences are absent from the RepBase library, the most commonly used database of repetitive DNA elements[1]. The frequency of Unknowns per species range from 51 in *D. melanogaster* to 568 in *D. triarauria*.

Their absence from RepBase implies novel sequences that have not been identified before, and shows that our pipeline can identify under-described, novel TEs.

4.4 Genome Size and Repeat Abundance

We noticed a correlation between the abundance of repeating elements and the genome size of the corresponding species [Figures 3.7 and 3.8].

There is a high correlation between the sum all the TE classes we identified, but not for each individual class. This leads us to believe that there is not a single class of element that contributes to genome size, but rather that multiple classes contribute to the size of the genome of the species.

4.5 Genome Size Contraction in *melanogaster* group

There is a genome size contraction in the *melanogaster* subgroup that includes *D. melanogaster*, *D. sechellia* and *D. simulans* relative to other species. This effect also seems to be expanded in *D. mauritiana*, *D. erecta* and *D. yakuba*.

The *melanogaster* subgroup shows a reduction in **all** TE classes, suggesting that those species evolved a way to more efficiently control TEs.

From our data, reduction in genome size seems to be due to DNA and RC elements. A reduction in amount of DNA and RC elements seems to have caused a reduction in genome size.

Previous studies also show a reduction in genome size of the *melanogaster* subgroup, which they also attribute to a reduction in TE content.

[1] RepBase

Chapter 5

Future Directions

5.1 Different Sequencing Strategies

It would be interesting to see the difference in repeat sequence identification, not just TE identification, that different sequencing strategies would provide.

We would like to include data on the pure versions of the following strategies:

1. Nanopore
2. PacBio SMRT
3. Sangar
4. Illumina

When we say pure versions, we imply that the genome assembly that was retrieved would be created the same for every species, without the aid of another sequencing method to accommodate for its inabilities.

It would be interesting to see the difference in correlations between the amount of annotated TEs and the size of the genome assembly using short read sequencing strategies vs. long read sequencing strategies

5.2 More Species

It would be beneficial to bolster our results using sequences from more than the 16 species we currently have; but there is no resource online that has all of these sequences, where all species have been sequenced using the same strategy and parameters.

5.3 Investigating Genome Size Contribution in *melanogaster* group

We can investigate the evolution of known TE control genes (such as piRNA pathways) to see if there are any evolutionary changes in the amino acid sequence or the copy number change that is unique to the *melanogaster* group.

We think that there could be a few reasons that could contribute to the genome size contraction in the *melanogaster* group. These reasons may be working independently of each other, but could also be working together and also as a series, where one method started the reduction while the others maintained it.

5.3.1 Stochastic Deletion

Stochastic deletion involves deletion of certain classes of TEs across all 6 species relatively recently, so that there is a reduction of TEs of a certain class across these 6 species.

This is the most unlikely of the three theories we propose. This is because stochastic deletion selecting for certain TEs across 6 unrelated (wrt to geographical and sexual isolation) species is very rare.

5.3.2 Arrival of Gene

There could also be a gene that evolved, was re-activated, or a gene amplification or gene duplication event that could have resulted in stricter control of TEs and their propagation.

We can investigate the evolution of known TE control genes (such as piRNA pathways) to see if there are any evolutionary changes in the amino acid sequence or the copy number change that is unique to the *melanogaster* group.

5.3.3 Population Size

Population size can effectively control for TEs and other mobile genetic elements. Mobile genetic elements are most often deleterious as they insert near or into genes and deactivate them or deviate them from normal expression.

Larger the population size, more the variation in phenotypes caused by TE insertion events, and more likely those individuals with deleterious insertions will not have the ability to mate. This inly allows individuals without a lot of insertions to reproduce and pass on their genome (a genome without many deleterious TE insertions).

Chapter 6

Programs Used

6.1 Anaconda

Table 6.1: Anaconda Information

Block	Information
Version	4.6.2
Function	Program and Package manager
Website	http://www.repeatmasker.org/RepeatModeler/
Download Link	Link to Version 4.6.2

Anaconda is an open-source distribution of the Python and R programming languages, as well as other utilities used for scientific computational analysis. These include utilities for fields such as data science, machine learning applications, large-scale data processing, predictive analytics, etc ... Anaconda aims to simplify package management and deployment. Anaconda has a package management system, `conda`, which manages packages.

The Anaconda distribution is used by over 6 million users and includes more than 1400 popular data-science packages suitable for Windows, Linux, and MacOS[1]. Download information can be found for MacOS here, or for other systems via this link.

6.2 RepeatModeler

RepeatModeler is a *de novo* repeat family identification and modeling package. It is a pipeline that consists of two programs, recon and RepeatScout, which employ complementary computational methods for identifying repeat element boundaries and family relationships from sequence data. RepeatModeler assists in automating the runs of RECON and RepeatScout by managing intermediate files given a genomic database. It reports putative repeats.

Table 6.2: RepeatModeler Information

Block	Information
Version	Open-1.0.101
Function	*de novo* repeat family identification and modeling.
Website	https://www.anaconda.com
Download Link	Link to Version 1.0.101

Table 6.3: Python3 Information

Block	Information
Version	3.6.5
Function	General purpose file management and calculation; custom scripts.
Website	python.org
Download Link	Link to Version 3.6.5

6.3 Python3

Python is a high-level programming language with dynamic semantics. Its high-level built in data structures, makes it very attractive for use as a "glue" language to connect existing components together, which is what we used it for. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python encourages program modularity and code reuse by supporting the use of modules and packages. Python libraries can be freely distributed, and are usually handled by its internal manager (`pip`), or can be handled by another program like Anaconda.

6.4 BedTools

The Bedtools utilities are a one-stop-shop of tools for a wide-range of genomics analysis tasks. The most widely-used tools enable genome arithmetic: that is, set theory on the genome. While each individual tool is designed to do a relatively simple task (e.g., intersect two interval files), quite sophisticated analyses can be conducted by combining multiple Bedtools operations on the UNIX command line.

Bedtools is developed in the Quinlan laboratory at the University of Utah and benefits from fantastic contributions made by scientists worldwide.

Bedtools has many utilities (summarized in [Table 6.5]). Though there are so many utilites provided, we only had the need of a few in our pipeline and a few for debugging.

[1] What is Anaconda?

Table 6.4: Bedtools Information

Block	**Information**
Version	2.27.0
Function	Genome Arithmatic
Website	Bedtools
Download Link	Link to Version 2.27.0

Table 6.5: Bedtools Utilities

Utility	**Utility**	**Utility**
annotate	bamtobed	bamtofastq
bed12tobed6	bedpetobam	bedtobam
closest	cluster	complement
coverage	expand	flank
genomecov	`getfasta`	`groupby`
igv	`intersect`	jaccard
links	makewindows	map
maskfasta	`merge`	multicov
multiinter	nuc	overlap
pairtobed	pairtopair	random
reldist	shift	shuffle
slop	sort	subtract
tag	unionbedg	window

We used the following utilities from the BedTools suite:

1. Bedtools merge or `mergeBed`: Many of the identified sequences have a lot of overlaps; this means that a single base-pair within multiple repeats might be accounted for multiple times. `mergeBed` combines overlapping or "book-ended" features in an interval file into a single feature which spans all of the combined features. This allows for a single base-pair to be only accounted for once.

2. Bedtools grouby or `groupBy`: `groupBy` is a tool that mimics the *group by* clause in database systems. Given a file or stream that is sorted by the appropriate "grouping columns" (`-g`), `groupBy` will compute summary statistics on another column (`-c`) in the file or stream. This will work with output from all Bedtools as well as any other tab-delimited file or stream. As such, this is a generally useful tool for all command-line analyses, not just genomics related research. We used it to be able to compress information about similar repeating elements within a file. This was done in a "tidy" format as we would need to pipe this data into R.
   Because our data was rather large, we had to sort it for easier computation using `<upstream analysis> | sort -k1,1 -k2,2n | <downstream analysis>` .

3. BEDTOOLS INTERSECT or `intersectBed`: `intersectBed` inputs two BED files and finds the intersection (overlap) between any of the sequences present in it and the other file(s). This was useful to be able to extract overlapping sequences and annotate them correctly.

4. BEDTOOLS GETFASTA or `getFastaFromBed`: `getFastaFromBed` extracts sequences from a FASTA file for each of the intervals defined in a BED/GFF/VCF file. This was particularly useful in debugging our code where we only needed particular sequences to `BLAST` against the assembly to verify our methods.

6.5 Perl5

TABLE 6.6: PERL Information

Block	**Information**
Version	5.26.2
Function	REPEATMODELER dependency; programming language.
Website	perl.org
Download Link	Link to Version 5.26.2

PERL5 is a highly capable, feature-rich programming language with over 30 years of development. Our use-case for PERL was as a REPEATMODELER dependency.

"Perl" is a family of languages, "PERL6" is part of the family, but it is a separate language which has its own development team. Its existence has no significant impact on the continuing development of "PERL5".

6.6 Tandem Repeat Finder (TRF)

TABLE 6.7: TRF Information

Block	**Information**
Version	4.0.4
Function	REPEATMODELER dependency; public database of Tandem repeats.
Website	TANDEM REPEAT FINDER
Download Link	Link to Version 4.0.4

A tandem repeat in DNA is a sequence of two or more bps repeated in such a way that the consensus repeats lie adjacent to each other. TRF is a program that helps us

locate these repeats in the DNA sequence. TRF outputs two files – an alignment file and a repeat table file. The repeat tabe file contains information such as locus, bp count, number of copies and dNTP content for each repeat.

TRF is very fast at analysing repeating elements as it only needs to look for adjacent repeats. Sequence information sent to the server is confidential and deleted after program execution.

6.7 blastX

Table 6.8: BlastX Information

Block	**Information**
Version	2.5.0
Function	Finds possible gene-protein alignments from all sequences.
Website	BLAST
Download Link	FTP Link to Version 2.5.0

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

6.8 RepeatScout

Table 6.9: RepeatScout Information

Block	**Information**
Version	1.0.5
Function	Discovers repetitive substrings form DNA.
Website	RepeatScout
Download Link	Link to Version 1.0.5

The purpose of the RepeatScout software is to identify repeat family sequences from genomes where hand-curated repeat databases are not available. In fact, the output of

this program can be used as input to RepeatMasker as a way of automatically masking newly-sequenced genomes.

6.9 RepeatMasker

Table 6.10: RepeatMasker Information

Block	Information
Version	Open-4.0.7
Function	Includes TRF libraries; library of all annotated repeats in multiple species.
Website	RepeatMasker
Download Link	Link to Version 4.0.7

RepeatMasker is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences. The output of the program is a detailed annotation of the repeats that are present in the query sequence as well as a modified version of the query sequence in which all the annotated repeats have been masked (default: replaced by Ns). Currently over 56% of human genomic sequence is identified and masked by the program. Sequence comparisons in RepeatMasker are performed by one of several popular search engines including `nhmmer`, `cross_match`, `ABBlast/WUBlast`, `RMBlast` and `Decypher`. RepeatMasker makes use of curated libraries of repeats and currently supports Dfam (profile HMM library derived from RepBase sequences) and RepBase, a service of the Genetic Information Research Institute.

6.10 RECON

Table 6.11: RECON Information

Block	Information
Version	1.08
Function	RepeatModeler dependency; automatic *de novo* identification.
Website	RECON
Download Link	Link to Version 1.05

RECON is required for proper identification of repetitive sequences is an essential step in genome analysis.

The RECON package performs *de novo* identification and classification of repeat sequence families from genomic sequences. The underlying algorithm is based on extensions

to the usual approach of single linkage clustering of local pairwise alignments between genomic sequences. Specifically, our extensions use multiple alignment information to define the boundaries of individual copies of the repeats and to distinguish homologous but distinct repeat element families. RECON should be useful for first-pass automatic classification of repeats in newly sequenced genomes.

6.11 R

TABLE 6.12: R Information

Block	Information
Version	3.5.2
Function	Downstream analysis and Graphical viewer
Website	The R Project for Statistical Computing
Download Link	University of California, Berkeley | CRAN

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MACOS[2]. R and its libraries implement a wide variety of statistical and graphical techniques, including classical statistical tests, classification, clustering, and others. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages. Many of R's standard functions are written in R itself, which makes it easy for users to follow the algorithmic choices made [116].

R can be called from the bash BASH command line using '`R`', which brings up an integrative environment within the command line. It is important to note that though we can allocate data and create graphs using the command line, we cannot view graphs there. We used RSTUDIO as an IDE to develop with R.

6.12 RStudio

TABLE 6.13: RStudio Information

Block	Information
Version	1.1.463
Function	IDE for R
Website	RSTUDIO
Download Link	Download RSTUDIO Desktop

[2] The R Project for Statistical Computing

RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, checking history, debugging and workspace management[3]. More information about the particulars of this IDE can be found here, at RStudio IDE features.

6.13 GitHub and Atom

We used Atom `[v1.36.0]` to develop our code and used GitHub in order to share our code.

The repository is titled `de_novo-identification`, and is under my name, `Dr-Drosophila`.

[3] RStudio/products

Appendix A

Diagrams.rmd

````
---
title: "Diagrams"
author: "Rele, Chinmay"
date: "2/22/2019"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

# Results

Creating a Stacked barplot:
More condensed
```{r}
# creating stacked barplot
library("tibble")
library(ggplot2)
temp = read.table( "./more_condensed.tab", header=TRUE, sep="\t" )
fly_data = as_tibble(temp)
fly_data
fly_data$species = factor(fly_data$species,levels = c( "dvir", "dmoj", "dwil", "dper", "dpse", "dana", "dbip", "dbia", "dtri", "deug", "dmel", "dsec", "dsim", "dmau", "dere", "dyak" ))
ggplot(fly_data, aes(fill=class, y=percent, x=species)) + geom_bar( stat="identity", position="fill") + labs( x = "Species", y = "Percent composition of Repeat",
       title = "Percent composition of each TE class within respective Nanopore assembly" )

ggsave( "accounted_unknowns_small.pdf", width=9, height=7 )


```


Creating a Stacked barplot:
### More condensed -- RAW BASE-PAIRS

```{r}
````
````

```
# creating stacked barplot
library("tibble")
library(ggplot2)

temp = read.table( "./more_condensed.tab", header=TRUE, sep="\t" )
fly_data = as_tibble(temp)
fly_data
fly_data$species = factor(fly_data$species,levels = c( 'dvir', 'dmoj', 'dwil', 'dper', 'dpse', 'dana', 'dbip', 'dbia', 'dtri', 'deug', 'dmel', 'dsec', 'dsim', 'dmau', 'dere', 'dyak' ))

ggplot( fly_data, aes(fill=class, y=bp, x=species) ) + geom_bar( stat = "identity") + labs( y = "Basepairs of Repeat", x = "Species", title = "Basepair composition of each TE class within respective Nanopore assembly") + theme_minimal(base_size = 15) # + theme( axis.title.y=element_blank(),  axis.ticks.y=element_blank(), axis.text.y=element_blank(), legend.position = c(0.8, 0.2) , legend.background = element_rect(fill="white",  size=0.5, linetype="solid"))  + coord_flip()

ggsave( "accounted_unknowns_raw_bp.pdf", width =10, height = 12)
```

Genome Size and Relative Abundance of Repeats

```{r}
library("tibble")
library(ggplot2)
library("dplyr")

temp = read.table( "./simple_data.tab", header=TRUE, sep="\t" )
simple_data = as_tibble( temp )
simple_data

cor.test(x=simple_data$assembly, y=simple_data$rmod_bp, method = 'spearman')


simple_data %>%
    mutate( TE_percent = rmod_bp/assembly ) %>%
    select( Species, rmod_bp, TE_percent )

ggplot( simple_data, aes( x = assembly,
                          y = rmod_bp,
                          color = Species ) ) %>%
    + theme(text = element_text(size=18) ) %>%
    + geom_point( size = 6 ) %>%
    + labs( x = "Assembly size (linear)",
                        y = "TE BasePairs (linear)",
                        color = "Species",
                        caption = "Spearman's Correlation= 0.9088235")
```

```
79
80 ggsave( "assembly_vs_reps.pdf", width=12, height=7 )
81
82
83 ```
84
85 ### Assembly size vs. all sequences (except Sat/SR) (including
      unknowns)
86
87 ```{r}
88 library("tibble")
89 library(ggplot2)
90 library(dplyr)
91
92 temp = read.table( "./simple_data.tab", header=TRUE, sep="\t" )
93 simple_data = as_tibble( temp )
94 simple_data
95
96 temp = read.table( "./more_condensed.tab", header=TRUE, sep="\t" )
97 fly_data = as_tibble(temp)
98 fly_data
99 colnames(fly_data)[colnames(fly_data)=="species"] <- "short"
100 fly_data
101
102 temp0 = merge( fly_data, simple_data, by=c("short") )
103 temp0
104 temp0 = temp0[!(temp0$class=="Satellite/Simple_Repeat"),]
105 temp0
106
107 keeps = c( "short", "assembly", "bp" )
108 keeps
109 total = temp0[ keeps ]
110 total
111 total$short = factor(total$short,levels = c( "dvir", "dmoj", "dwil
      ", "dper", "dpse", "dana", "dbip", "dbia", "dtri", "deug", "
      dmel", "dsec", "dsim", "dmau", "dere", "dyak" ))
112 total
113
114 temp1 = group_by(total, short)
115 temp1
116
117 temp2 = aggregate(temp1$bp, by=list(short=temp1$short, assembly=
      temp1$assembly), FUN=sum)
118 temp2
119
120 # summed = summarise( temp1,  )
121 # summed
122
123 ggplot( temp2, aes( x = assembly,
124                                 y = x,
125                                 color = short ) ) + theme(text =
      element_text(size=18) ) + geom_point(size =  6) + labs( x = "
      Assembly size (linear)",
```

```
126                          y = "All sequences (except Sat/SR)(
    including unknowns)",
127                          color = "Species",
128                          caption = "Spearman's Correlation
    =0.5029412")
129
130 cor.test(x=temp2$x, y=temp2$assembly, method = 'spearman')
131
132
133 ggsave( "assembly_vs_all_seq(except_SatSR).pdf", width=9, height=7
     )
134
135 ```
136
137 ### Assembly size vs. Just Sat/SR
138
139 ```{r}
140 library("tibble")
141 library(ggplot2)
142 library(dplyr)
143
144 temp = read.table( "./simple_data.tab", header=TRUE, sep="\t" )
145 simple_data = as_tibble( temp )
146 simple_data
147
148 temp = read.table( "./more_condensed.tab", header=TRUE, sep="\t" )
149 fly_data = as_tibble(temp)
150 fly_data
151 colnames(fly_data)[colnames(fly_data)=="species"] <- "short"
152 fly_data
153
154 temp0 = merge( fly_data, simple_data, by=c("short") )
155 temp0
156 temp0 = temp0[(temp0$class=="Satellite/Simple_Repeat"),]
157 temp0
158
159 keeps = c( "short", "assembly", "bp" )
160 keeps
161 total = temp0[ keeps ]
162 total
163 total$short = factor(total$short,levels = c( "dvir", "dmoj", "dwil
    ", "dper", "dpse", "dana", "dbip", "dbia", "dtri", "deug", "
    dmel", "dsec", "dsim", "dmau", "dere", "dyak" ))
164 total
165
166 temp1 = group_by(total, short)
167 temp1
168
169 temp2 = aggregate(temp1$bp, by=list(short=temp1$short, assembly=
    temp1$assembly), FUN=sum)
170 temp2
171
172 # summed = summarise( temp1,  )
```

```
# summed

cor.test(x=temp2$x, y=temp2$assembly, method = 'spearman')

ggplot( temp2, aes( x = assembly,
                              y = x,
                              color = short ) ) + theme(text = element_text(size=18) ) + geom_point(size =  6) + labs( x = "Assembly size (linear)",
                         y = "Only Satellites/Simple_Repeats",
                         color = "Species",
                         caption = "Spearman's Correlation =0.5764706",
                         title = "Assembly size vs. Just Sat/SR")



ggsave( "assembly_vs_SatSR.pdf", width=9, height=7 )

```

Repeats vs. Genome Assembly

```{r}
library("tibble")
library(ggplot2)

temp = read.table( "./simple_data.tab", header=TRUE, sep="\t" )
simple_data = as_tibble( temp )
simple_data

simple_data$sum = simple_data$trf + simple_data$rmod_bp

cor.test(x=simple_data$assembly, y=simple_data$sum , method = 'spearman')

ggplot( simple_data, aes( x = assembly, y = rmod_bp + trf, color = Species ) ) + theme(text = element_text(size=18) )  + geom_point(size = 6) + labs( x = "Assembly size (linear)",  y = "All Repeats BasePairs (linear)",  caption = "Spearman's Correlation= 0.8617647")

ggsave( "assembly_size_vs_all_reps.pdf", width = 9, height = 7 )

```

Simple Repeats vs. TEs

```{r}
library("tibble")
library(ggplot2)
temp = read.table( "./simple_data.tab", header=TRUE, sep="\t" )
simple_data = as_tibble( temp )
```

```
simple_data

simple_data$sum = simple_data$trf + simple_data$rmod_bp

corr = cor.test(x=simple_data$trf, y=simple_data$rmod_bp , method = 'spearman')
corr

ggplot( simple_data, aes( x = trf,
                          y = rmod_bp,
                          color = Species ) ) + theme(text = element_text(size=18) ) + geom_point(size = 6) + labs( x = "Simple Repeats BasePairs (linear)",
                       y = "TEs BasePairs (linear)",
                       caption = "Spearman's Correlation = 0.3411765")

ggsave( "TE_vs_SR.pdf", width=9, height=7 )

```

(BARPLOT) percent of total unknown sequences of each species of all TE

```{r}
library("tibble")
library(ggplot2)
library(dplyr)

temp = read.table( "./more_condensed.tab", header=TRUE, sep="\t" )
fly_data = as_tibble(temp)
colnames(fly_data)[colnames(fly_data)=="species"] <- "short"
fly_data

temp = read.table( "./simple_data.tab", header=TRUE, sep="\t" )
simple_data = as_tibble( temp )
simple_data = simple_data %>%
    mutate( repeats = rmod_bp + trf ) %>%
    select( short, repeats, assembly )

keeps = c( "short", "class", "bp" )
total = fly_data[ keeps ]
total$short = factor(total$short,levels = c( "dvir", "dmoj", "dwil", "dper", "dpse", "dana", "dbip", "dbia", "dtri", "deug", "dmel", "dsec", "dsim", "dmau", "dere", "dyak" ))
total = total[ total$class == "Unknown", ]
keep = c( "short", "bp" )
total = total[ keep ]
total

merged = merge( total , simple_data , by="short")
merged

```

```
order = c( "dvir", "dmoj", "dwil", "dper", "dpse", "dana", "dbip", "dbia", "dtri", "deug", "dmel", "dsec", "dsim", "dmau", "dere", "dyak" )
unknown_percent = merged %>%
    slice( match ( order, short ) ) %>%
    mutate( percent_of_assembly = bp/assembly ) %>%
    mutate( percent_of_repeat = bp/repeats ) %>%
    select( short, percent_of_assembly, percent_of_repeat )
unknown_percent

# unknown_percent_assembly$short = factor( unknown_percent_assembly$short, levels = unknown_percent_assembly$short[order(desc(unknown_percent_assembly$percent_of_assembly))])
# unknown_percent_assembly

ggplot(unknown_percent, aes(short, percent_of_assembly)) + geom_bar(aes(fill = percent_of_repeat), position = position_dodge(), stat="identity")

ggplot(data=unknown_percent, aes(x=short, y=percent_of_assembly)) + geom_bar(stat="identity", fill="steelblue") + theme_minimal() + labs( x = "Species",
                        y = "Percent of Genome Size")

ggsave( "unknown_percent_assembly.pdf", width = 9, height = 12)

```

```
### Summaries of Assemblies
```{r}
library("tibble")
library(ggplot2)
library(dplyr)

temp = read.table( "./assembly_summary.tab", header=TRUE, sep="\t" )
ass_sum = as_tibble(temp)
ass_sum

ggplot(data=ass_sum, aes(x=assembly_size, y=contig_count, color=avg_contig_size, size=N_50)) + geom_point() + labs( x = "Assembly Size", y="Number of Contigs", color = "Average Contig Size", size = "N50", title = "Summaries of Assemblies") + geom_label(aes(label = species))


ggsave( "assemblies_summary.pdf", width=9, height=7 )

```


Faceted plot of correlation; facet along TE class
```

```
```{r}
library("tibble")
library(Rmisc)
library(ggplot2)
library(dplyr)
library(plyr)

temp = read.table( "./simple_data.tab", header=TRUE, sep="\t" )
simple_data = as_tibble( temp )
simple_data

temp = read.table( "./more_condensed.tab", header=TRUE, sep="\t" )
fly_data = as_tibble(temp)
fly_data
colnames(fly_data)[colnames(fly_data)=="species"] <- "short"
fly_data

temp0 = merge( fly_data, simple_data, by=c("short") )
temp0 = temp0[!(temp0$class=="Other"),]
temp0

temp0 = aggregate( cbind(bp)~Species+class+assembly, temp0, sum )
dna = temp0 %>%
    filter( class == "DNA" )
line = temp0 %>%
    filter( class == "LINE" )
ltr = temp0 %>%
    filter( class == "LTR" )
rc = temp0 %>%
    filter( class == "RC" )
sat_sr = temp0 %>%
        filter( class == "Satellite/Simple_Repeat" )
unknown = temp0 %>%
    filter( class == "Unknown" )
chimera = temp0 %>%
    filter( class == "Chimera" )
dna
cor.test(x=dna$assembly, y=dna$bp , method = 'spearman')
cor.test(x=line$assembly, y=line$bp , method = 'spearman')
cor.test(x=ltr$assembly, y=ltr$bp , method = 'spearman')
# ltr; dvir, dtri, dmoj; remove the following from LTR and check
   spearman corr for LTR also
cor.test(x=rc$assembly, y=rc$bp , method = 'spearman')
cor.test(x=sat_sr$assembly, y=sat_sr$bp , method = 'spearman')
cor.test(x=unknown$assembly, y=unknown$bp , method = 'spearman')
cor.test(x=chimera$assembly, y=chimera$bp , method = 'spearman')

labels = c( "DNA", "LINE", "LTR", "RC", "Sat/SR", "Unknown", "
   Chimera" )

ggplot( temp0 , aes( x=assembly, y=bp, color = Species ) ) %>%
    + geom_point() %>%
```
```

```
    + geom_smooth(aes( group = class), method = "lm", se = FALSE, size = 0.9, colour = "grey", stat="smooth") %>%
    # + stat_smooth( method = 'lm' ) %>%
    + facet_wrap( ~class, scales="free", ncol = 2) %>%
    + labs( x = "Assembly_size",
            y = "Basepairs",
            color = "Species",
            title = "BP Correlation with Assembly size per Class of Repeats") # %>%
    # + geom_text( data = temp0, mapping = aes( label = labels ) )

ggsave( "faceted_class_correlation.pdf", width=9, height=7 )


```

RepeatModeler Summary (faceted per class)
- How manys equences were identified as TEs?
- How many sequences were identified as Sat/SR?
- How many sequences were descriebd as Unknown?

```{r}
library("tibble")
library(ggplot2)
library(dplyr)
library(magrittr)

temp = read.table( "./simple_data.tab", header=TRUE, sep="\t" )
simple_data = as_tibble( temp )
simple_data

temp = read.table( "./more_condensed.tab", header=TRUE, sep="\t" )
fly_data = as_tibble(temp)
fly_data
colnames(fly_data)[colnames(fly_data)=="species"] <- "short"
fly_data

more_data = merge( fly_data, simple_data, by=c("short") )
more_data = more_data[!(temp0$class=="Other"),]
more_data

grouped_fly = more_data %>%
    dplyr::group_by( short, class ) %>%
    dplyr::summarise( count = n() )

grouped_fly

temp = grouped_fly
grouped_fly = temp %>%
    filter( class != "Satellite/Simple_Repeat" ) %>%
    filter( class != "Other" ) %>%
    dplyr::group_by( short, class ) %>%
```

```
    dplyr::summarize( total = sum(count) )
grouped_fly

ggplot( data = grouped_fly, aes( x = class, y = total, fill = class ) ) %>%
    + geom_col() %>%
    + facet_wrap( ~short, nrow=4 ) %>%
    + labs( x = "Species",
            y = "Counts of Repeat Identified",
            fill = "Classes",
            title = "Summary of number of sequences identified per class") %>%
    + coord_flip()



ggsave( "rmodeler_summary.pdf", width=9, height=7 )

```


Percent composition of TEs of the whole genome

```{r}
library("tibble")
library(ggplot2)
library(dplyr)

temp = read.table( "./simple_data.tab", header=TRUE, sep="\t" )
simple_data = as_tibble( temp )
simple_data

temp = read.table( "./more_condensed.tab", header=TRUE, sep="\t" )
fly_data = as_tibble(temp)
fly_data
colnames(fly_data)[colnames(fly_data)=="species"] <- "short"
fly_data

more_data = merge( fly_data, simple_data, by=c("short") )
more_data = more_data[!(temp0$class=="Other"),]
more_data
grouped_fly = more_data %>%
    group_by( Species ) %>%
    mutate( bp_percent = bp/assembly ) %>%
    select( Species, bp_percent, class )
grouped_fly

grouped_fly %>%
    group_by( Species, class ) %>%
    dplyr::summarize( total = sum(bp_percent) )

grouped_fly = grouped_fly[!(grouped_fly$class=="Other"),]

```

```
grouped_fly$Species = factor(grouped_fly$Species,levels = c( "Drosophila_yakuba", "Drosophila_erecta", "Drosophila_mauritiana", "Drosophila_simulans", "Drosophila_sechellia", "Drosophila_melanogaster", "Drosophila_eugracilis", "Drosophila_triauraria", "Drosophila_biarmipes", "Drosophila_bipectinata", "Drosophila_ananassae", "Drosophila_pseudoobscura", "Drosophila_persimilis", "Drosophila_willistoni", "Drosophila_mojavensis", "Drosophila_virilis" ))

grouped_fly

ggplot( grouped_fly , aes( y=bp_percent, fill=class, x = Species) ) + geom_bar(stat = "identity") + labs( fill = "Classes",  y = "Percent of Genome",
       title = "Percent composition of repeats within Genome" ) + coord_flip()

ggsave( "percent_composition_repeats_in_genome.pdf", width=9, height=7 )


```


Plotting Unknown percent and BP

```{r}
library("tibble")
library(ggplot2)
library(dplyr)

temp = read.table( "./unknown_bp_percent.tab", header=TRUE, sep="\t" )
unknowns = as_tibble( temp )

ggplot(unknowns, aes(x = species)) +
  geom_col(aes( y = bp_mb, fill="redfill")) +
  geom_text(aes(y = bp_mb, label = bp_mb), fontface = "bold", vjust = 0.5, hjust = 1.4, color = "black", size = 7) +
  geom_line(aes(y = percent * 300, group = 1, color = 'blackline') ) +
  geom_text(aes(y = percent * 300, label = round(percent, 2)), hjust = 1, color = "black", size = 6) +
  scale_y_continuous(sec.axis = sec_axis(trans = ~ . / 300)) +
  scale_fill_manual('', labels = 'Repeats content (Mb)', values = "#b3bcff") +
  scale_color_manual('', labels = 'Percent of Genome', values = 'black') +
  theme_minimal(base_size = 20) + theme( legend.position = "bottom", axis.title.y=element_blank() ) + coord_flip()

ggsave( "unknown_bp_percentp.pdf", width = 9, height = 12)

```

````
```


### BARPLOT -- Lengths of each class

```{r}
library("tibble")
library(ggplot2)
library(dplyr)

temp = read.table( "./lengths.tab", header=TRUE, sep="\t" )
lens = as_tibble( temp )

lens

## remove unneeded classes
lens = lens[ !( lens$class == "buffer" ), ]
lens = lens[ !( lens$class == "DNA?" ), ]
lens = lens[ !( lens$class == "Other" ), ]
lens = lens[ !( lens$class == "rRNA" ), ]
lens = lens[ !( lens$class == "SINE" ), ]
lens = lens[ !( lens$class == "SINE?" ), ]

lens



ggplot( lens , aes(x=class, y=length)) +
    geom_boxplot() +
    scale_y_log10() +
    ggtitle( "Lenght of each Class of Putative TE", subtitle = "in
    log_10" )



ggsave( "class_lengths.pdf", width = 9, height = 6)

```
````
````

Appendix B

Extra Results

This appendix encapsulates all data that could not be accommodated within the body of the thesis.

B.1 Family Identification

Along with only identifying classes of TEs, we also chose to analyze the family classifications of our RepeatMasker data. This led us to Figure B.1, which was excluded from the Results and from further analysis due to the merging of colors for each class and at the class boundaries.

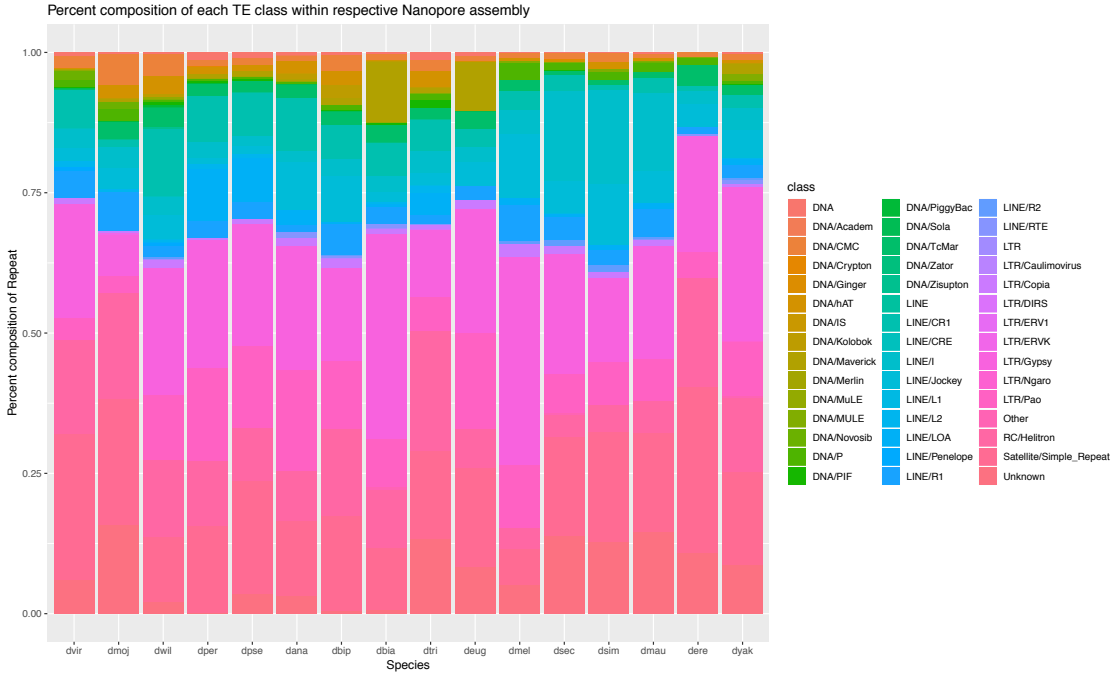


Figure B.1: Percentage composition of All TE Classes with Family identifications

B.2 Unknown + TE correlations

Here, we show the correlation of the base-pairs of all the TEs agains the assembly size for all the species of analysed *Drosophila*.

$$r_s = 0.5029412 \tag{B.1}$$

$$p < 2.2 \cdot 10^{-16} \tag{B.2}$$

With the inclusion of the Unknown sequences, we expected a decrease in the correlation of the base-pairs agains the genome size as "Unknown" and "Chimera" are umbrella categories that contains all types of elements. Some of these elements might be correlated with genome size, while others might not, or have negative correlation.

This correlation was highly statistically significant.

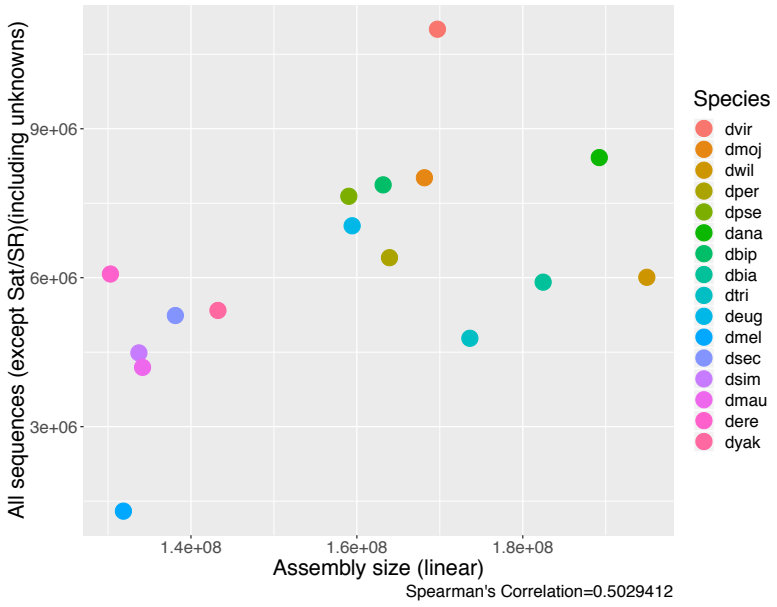


FIGURE B.2: Correlation of all identified TE classes with Unknowns

B.3 Sat/SR Correlations

We also wanted to test whether there was any correlation between Satellite/Simple Repeats with the assembly size.

We found a slight correlation of $r_s = 0.5764706$ for the amount of Satellite/Simple Repeat content with the assembly size.

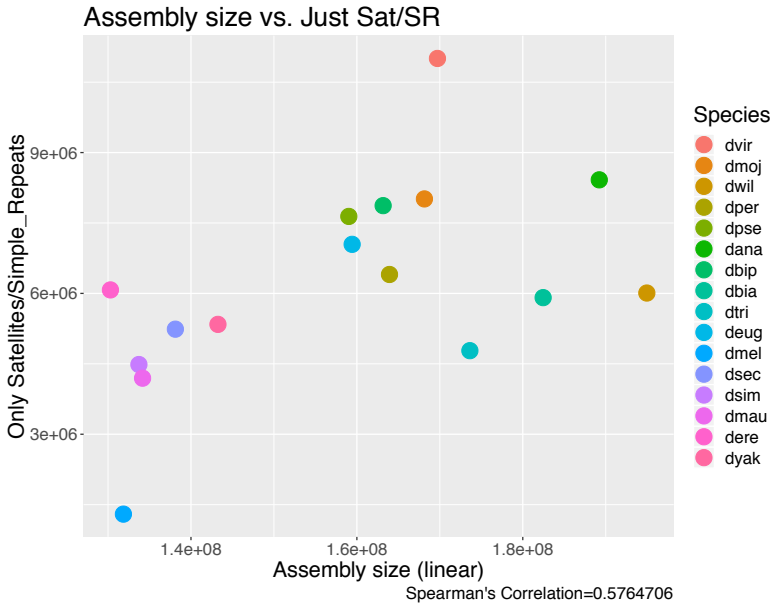


FIGURE B.3: Correlation of Sat/SR with Assembly Size

B.4 Phylogeny

We used Figure B.4 as our species phylogeny tree. We used data from Flybase.org as well as from the modENCODE project [117] to create this tree.

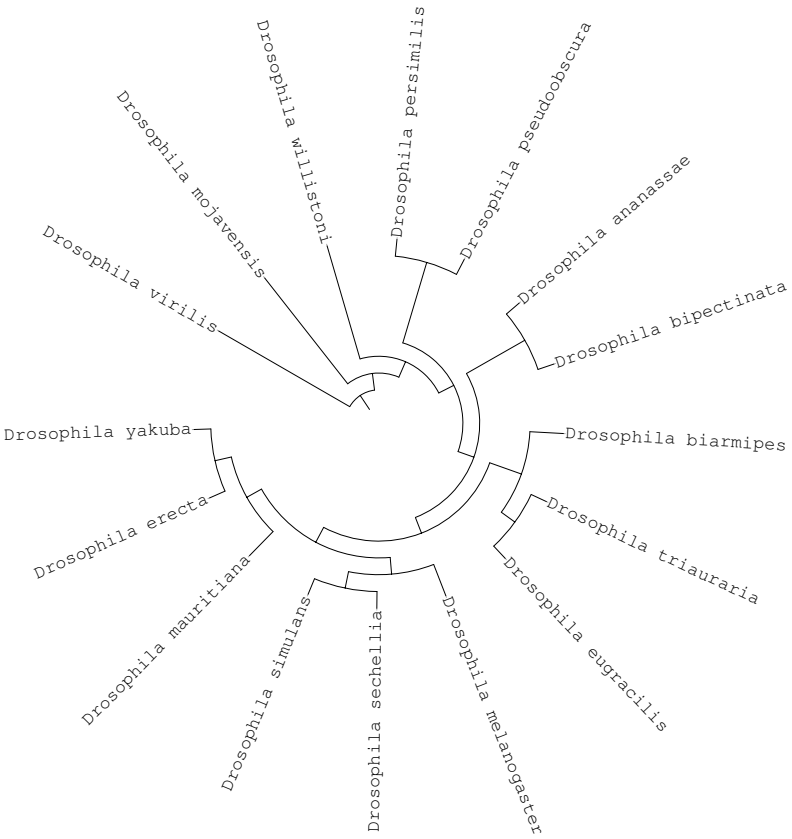


FIGURE B.4: Correlation of Sat/SR with Assembly Size

This tree was created on the interactive Tree of Life website using the `Newick format` presented below.

```
( ( ( ( ( ( ( ( Drosophila_sechellia, Drosophila_simulans ),
   Drosophila_melanogaster ), ( ( Drosophila_erecta,
   Drosophila_yakuba ), Drosophila_mauritiana ) ), ( (
   Drosophila_triauraria, Drosophila_eugracilis ),
   Drosophila_biarmipes ) ), ( Drosophila_ananassae,
   Drosophila_bipectinata ) ), ( Drosophila_persimilis,
   Drosophila_pseudoobscura ) ), Drosophila_willistoni ),
   Drosophila_mojavensis ), Drosophila_virilis )
```

LISTING B.1: Species Phylogeny

B.5 Grouping Species

We arbitrarily segregated species based on their lineages shown in B.4, and plotted their TE content with their assembly size.

In Figure B.5, *D. melanogaster* and its sister species have a grouping of 1, which moves to 9 for *D. virilis*, which is most removed from the phylogeny. As we can see, there is a clear correlation between the TE base-pairs and the genome size (as emphasized before), but there also seems to be a minor correlation between the grouping of the species with both of the above factors.

We notice the dots that represent species get substantially lighter from bottom-left to top-right as the TE base-pairs, the assembly size and the grouping category ID increases.

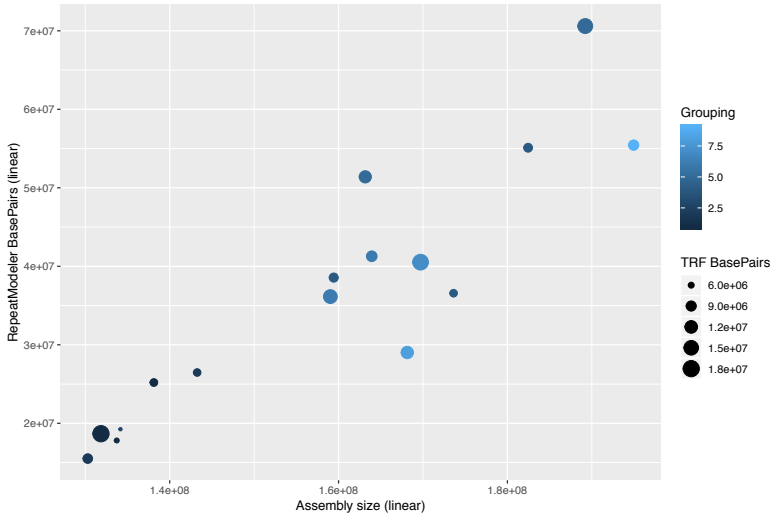


FIGURE B.5: TE content across species groups separated by Phylogenetic split

B.6 TEs vs. Simple Repeats

We also attempted to correlated TEs and Simple repeats, but a correlation of

$$r_s = 0.3411765 \tag{B.3}$$

thwarted any attempts or inclinations to analyze correlation further.

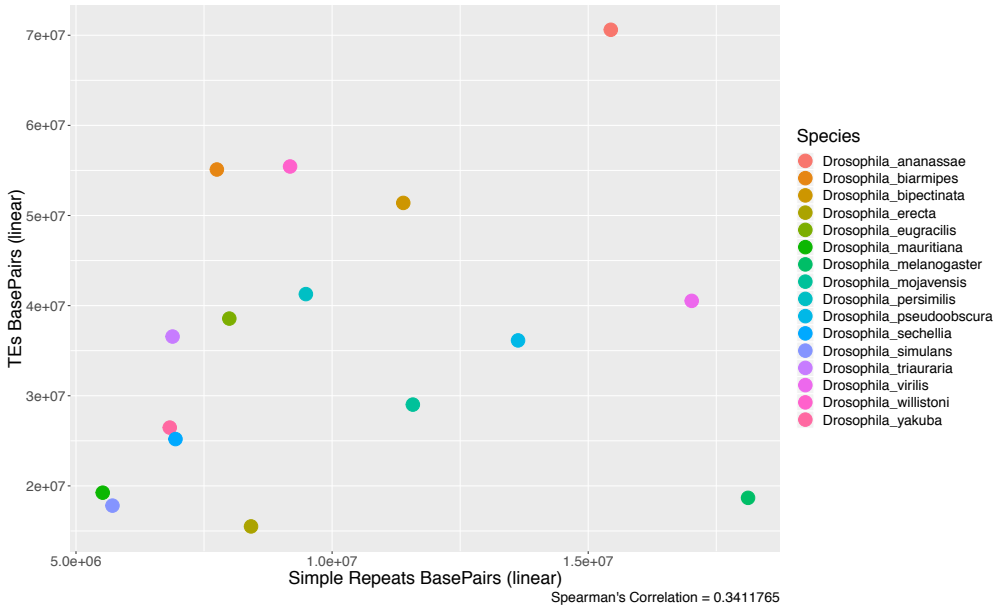


FIGURE B.6: TEs vs. Simple Repeats

Bibliography

[1] D. G. Consortium, A. G. Clark, M. B. Eisen, D. R. Smith, C. M. Bergman, B. Oliver, T. A. Markow, T. C. Kaufman, M. Kellis, W. Gelbart, V. N. Iyer, D. A. Pollard, T. B. Sackton, A. M. Larracuente, N. D. Singh, J. P. Abad, D. N. Abt, B. Adryan, M. Aguade, H. Akashi, W. W. Anderson, C. F. Aquadro, D. H. Ardell, R. Arguello, C. G. Artieri, D. A. Barbash, D. Barker, P. Barsanti, P. Batterham, S. Batzoglou, D. Begun, A. Bhutkar, E. Blanco, S. A. Bosak, R. K. Bradley, A. D. Brand, M. R. Brent, A. N. Brooks, R. H. Brown, R. K. Butlin, C. Caggese, B. R. Calvi, A Bernardo de Carvalho, A. Caspi, S. Castrezana, S. E. Celniker, J. L. Chang, C. Chapple, S. Chatterji, A. Chinwalla, A. Civetta, S. W. Clifton, J. M. Comeron, J. C. Costello, J. A. Coyne, J. Daub, R. G. David, A. L. Delcher, K. Delehaunty, C. B. Do, H. Ebling, K. Edwards, T. Eickbush, J. D. Evans, A. Filipski, S. Findeiss, E. Freyhult, L. Fulton, R. Fulton, A. C. L. Garcia, A. Gardiner, D. A. Garfield, B. E. Garvin, G. Gibson, D. Gilbert, S. Gnerre, J. Godfrey, R. Good, V. Gotea, B. Gravely, A. J. Greenberg, S. Griffiths-Jones, S. Gross, R. Guigo, E. A. Gustafson, W. Haerty, M. W. Hahn, D. L. Halligan, A. L. Halpern, G. M. Halter, M. V. Han, A. Heger, L. Hillier, A. S. Hinrichs, I. Holmes, R. A. Hoskins, M. J. Hubisz, D. Hultmark, M. A. Huntley, D. B. Jaffe, S. Jagadeeshan, W. R. Jeck, J. Johnson, C. D. Jones, W. C. Jordan, G. H. Karpen, E. Kataoka, P. D. Keightley, P. Kheradpour, E. F. Kirkness, L. B. Koerich, K. Kristiansen, D. Kudrna, R. J. Kulathinal, S. Kumar, R. Kwok, E. Lander, C. H. Langley, R. Lapoint, B. P. Lazzaro, S.-J. Lee, L. Levesque, R. Li, C.-F. Lin, M. F. Lin, K. Lindblad-Toh, A. Llopart, M. Long, L. Low, E. Lozovsky, J. Lu, M. Luo, C. A. Machado, W. Makalowski, M. Marzo, M. Matsuda, L. Matzkin, B. McAllister, C. S. McBride, B. McKernan, K. McKernan, M. Mendez-Lago, P. Minx, M. U. Mollenhauer, K. Montooth, S. M. Mount, X. Mu, E. Myers, B. Negre, S. Newfeld, R. Nielsen, M. A. F. Noor, P. O'Grady, L. Pachter, M. Papaceit, M. J. Parisi, M. Parisi, L. Parts, J. S. Pedersen, G. Pesole, A. M. Phillippy, C. P. Ponting, M. Pop, D. Porcelli, J. R. Powell, S. Prohaska, K. Pruitt, M. Puig, H. Quesneville, K. R. Ram, D. Rand, M. D. Rasmussen, L. K. Reed, R. Reenan, A. Reily, K. A. Remington, T. T. Rieger, M. G. Ritchie, C. Robin, Y.-H. Rogers, C. Rohde, J. Rozas, M. J. Rubenfield, A. Ruiz, S. Russo, S. L. Salzberg, A. Sanchez-Gracia, D. J. Saranga, H. Sato, S. W. Schaeffer, M. C. Schatz, T. Schlenke, R. Schwartz, C. Segarra, R. S. Singh, L. Sirot, M. Sirota, N. B. Sisneros, C. D. Smith, T. F. Smith, J. Spieth, D. E. Stage, A. Stark, W. Stephan, R. L. Strausberg, S. Strempel, D. Sturgill, G. Sutton, G. G. Sutton, W. Tao, S. Teichmann, Y. N. Tobari, Y. Tomimura, J. M. Tsolas, V. L. S. Valente, E. Venter, J. C. Venter, S. Vicario, F. G. Vieira, A. J. Vilella, A. Villasante, B. Walenz, J. Wang, M. Wasserman, T. Watts, D. Wilson, R. K. Wilson, R. A. Wing, M. F. Wolfner, A. Wong, G. K.-S. Wong, C.-I. Wu, G. Wu, D. Yamamoto, H.-P. Yang, S.-P. Yang, J. A. Yorke, K. Yoshida,

E. Zdobnov, P. Zhang, Y. Zhang, A. V. Zimin, J. Baldwin, A. Abdouelleil, J. Abdulkadir, A. Abebe, B. Abera, J. Abreu, S. C. Acer, L. Aftuck, A. Alexander, P. An, E. Anderson, S. Anderson, H. Arachi, M. Azer, P. Bachantsang, A. Barry, T. Bayul, A. Berlin, D. Bessette, T. Bloom, J. Blye, L. Boguslavskiy, C. Bonnet, B. Boukhgalter, I. Bourzgui, A. Brown, P. Cahill, S. Channer, Y. Cheshatsang, L. Chuda, M. Citroen, A. Collymore, P. Cooke, M. Costello, K. D'Aco, R. Daza, G. De Haan, S. DeGray, C. DeMaso, N. Dhargay, K. Dooley, E. Dooley, M. Doricent, P. Dorje, K. Dorjee, A. Dupes, R. Elong, J. Falk, A. Farina, S. Faro, D. Ferguson, S. Fisher, C. D. Foley, A. Franke, D. Friedrich, L. Gadbois, G. Gearin, C. R. Gearin, G. Giannoukos, T. Goode, J. Graham, E. Grandbois, S. Grewal, K. Gyaltsen, N. Hafez, B. Hagos, J. Hall, C. Henson, A. Hollinger, T. Honan, M. D. Huard, L. Hughes, B. Hurhula, M. E. Husby, A. Kamat, B. Kanga, S. Kashin, D. Khazanovich, P. Kisner, K. Lance, M. Lara, W. Lee, N. Lennon, F. Letendre, R. LeVine, A. Lipovsky, X. Liu, J. Liu, S. Liu, T. Lokyitsang, Y. Lokyitsang, R. Lubonja, A. Lui, P. MacDonald, V. Magnisalis, K. Maru, C. Matthews, W. McCusker, S. McDonough, T. Mehta, J. Meldrim, L. Meneus, O. Mihai, A. Mihalev, T. Mihova, R. Mittelman, V. Mlenga, A. Montmayeur, L. Mulrain, A. Navidi, J. Naylor, T. Negash, T. Nguyen, N. Nguyen, R. Nicol, C. Norbu, N. Norbu, N. Novod, B. O'Neill, S. Osman, E. Markiewicz, O. L. Oyono, C. Patti, P. Phunkhang, F. Pierre, M. Priest, S. Raghuraman, F. Rege, R. Reyes, C. Rise, P. Rogov, K. Ross, E. Ryan, S. Settipalli, T. Shea, N. Sherpa, L. Shi, D. Shih, T. Sparrow, J. Spaulding, J. Stalker, N. Stange-Thomann, S. Stavropoulos, C. Stone, C. Strader, S. Tesfaye, T. Thomson, Y. Thoulutsang, D. Thoulutsang, K. Topham, I. Topping, T. Tsamla, H. Vassiliev, A. Vo, T. Wangchuk, T. Wangdi, M. Weiand, J. Wilkinson, A. Wilson, S. Yadav, G. Young, Q. Yu, L. Zembek, D. Zhong, A. Zimmer, Z. Zwirko, D. B. Jaffe, P. Alvarez, W. Brockman, J. Butler, C. Chin, S. Gnerre, M. Grabherr, M. Kleber, E. Mauceli, and I. MacCallum, "Evolution of genes and genomes on the Drosophila phylogeny.", *Nature*, vol. 450, no. 7167, pp. 203–218, 2007, ISSN: 1476-4687. [Online]. Available: http://ovidsp.ovid.com/ovidweb.cgi?T=JS{\&}CSC=Y{\&}NEWS=N{\&}PAGE=fulltext{\&}D=medc{\&}AN=17994087https://rutgers.primo.exlibrisgroup.com/discovery/openurl?institution=01RUT{_}INST{\&}vid=01RUT{_}INST:01RUT{\&}?sid=OVID:medline{\&}id=doi:10.1038{\%}2Fnature06341{\&}issn=0028-0836{\&}isbn=.

[2] A. P. J. de Koning, W. Gu, T. A. Castoe, M. A. Batzer, and D. D. Pollock, "Repetitive elements may comprise over two-thirds of the human genome.", eng, *PLoS genetics*, vol. 7, no. 12, e1002384, 2011, ISSN: 1553-7404 (Electronic). DOI: 10.1371/journal.pgen.1002384.

[3] J. Jurka, V. V. Kapitonov, O. Kohany, and M. V. Jurka, "Repetitive Sequences in Complex Genomes: Structure and Evolution", *Annual Review of Genomics and Human Genetics*, vol. 8, no. 1, pp. 241–259, 2007, ISSN: 1527-8204. DOI: 10.1146/annurev.genom.8.080706.092416. [Online]. Available: http://www.annualreviews.org/doi/10.1146/annurev.genom.8.080706.092416.

[4] M. C. Kiefer, R. A. Owens, and T. O. Diener, "Structural similarities between viroids and transposable genetic elements.", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 80, no. 20, pp. 6234–8, 1983, ISSN: 0027-8424. DOI: 10.1073/PNAS.80.20.6234. [Online]. Available: http:

//www.ncbi.nlm.nih.gov/pubmed/6312450http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC394270.

[5] F. J. M. Mojica, C. Diez-Villasenor, E. Soria, and G. Juez, "Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria", *Molecular Microbiology*, vol. 36, no. 1, pp. 244–246, 2000, ISSN: 0950-382X. DOI: 10.1046/j.1365-2958.2000.01838.x. [Online]. Available: http://doi.wiley.com/10.1046/j.1365-2958.2000.01838.x.

[6] M. A. Biscotti, E. Olmo, and J. S. Heslop-Harrison, "Repetitive DNA in eukaryotic genomes", *Chromosome Research*, vol. 23, no. 3, pp. 415–420, 2015, ISSN: 0967-3849. DOI: 10.1007/s10577-015-9499-z. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26514350http://link.springer.com/10.1007/s10577-015-9499-z.

[7] Z. Lippman, A.-V. Gendrel, M. Black, M. W. Vaughn, N. Dedhia, W. Richard McCombie, K. Lavine, V. Mittal, B. May, K. D. Kasschau, J. C. Carrington, R. W. Doerge, V. Colot, and R. Martienssen, "Role of transposable elements in heterochromatin and epigenetic control", *Nature*, vol. 430, no. 6998, pp. 471–476, 2004, ISSN: 0028-0836. DOI: 10.1038/nature02651. [Online]. Available: http://www.nature.com/doifinder/10.1038/nature02651.

[8] R. K. Slotkin and R. Martienssen, "Transposable elements and the epigenetic regulation of the genome", *Nature Reviews Genetics*, vol. 8, no. 4, pp. 272–285, 2007, ISSN: 1471-0056. DOI: 10.1038/nrg2072. [Online]. Available: http://www.nature.com/articles/nrg2072.

[9] C Wicky, A. M. Villeneuve, N Lauper, L Codourey, H Tobler, and F Müller, "Telomeric repeats (TTAGGC)n are sufficient for chromosome capping function in Caenorhabditis elegans.", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 17, pp. 8983–8, 1996, ISSN: 0027-8424. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8799140http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC38581.

[10] A. Ruiz-Herrera, S. Nergadze, M. Santagostino, and E. Giulotto, "Telomeric repeats far from the ends: mechanisms of origin and role in evolution", *Cytogenetic and Genome Research*, vol. 122, no. 3-4, pp. 219–228, 2008, ISSN: 1424-8581. DOI: 10.1159/000167807. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19188690https://www.karger.com/Article/FullText/167807.

[11] B Starcich, L Ratner, S. F. Josephs, T Okamoto, R. C. Gallo, and F Wong-Staal, "Characterization of long terminal repeat sequences of HTLV-III.", *Science (New York, N.Y.)*, vol. 227, no. 4686, pp. 538–40, 1985, ISSN: 0036-8075. DOI: 10.1126/SCIENCE.2981438. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/2981438.

[12] C. P. Witte, Q. H. Le, T Bureau, and A Kumar, "Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes.", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 24, pp. 13 778–83, 2001, ISSN: 0027-8424. DOI: 10.1073/pnas.241341898. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11717436http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC61118.

[13] R Meloni, V. Albanèse, P. Ravassard, F. Treilhou, and J. Mallet, "A tetranucleotide polymorphic microsatellite, located in the first intron of the tyrosine hydroxylase gene, acts as a transcription regulatory element in vitro", *Human Molecular Genetics*, vol. 7, no. 3, pp. 423–428, 1998, ISSN: 14602083. DOI: 10.1093/hmg/7.3.423. [Online]. Available: https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/7.3.423.

[14] C. R. Boland and A. Goel, "Microsatellite Instability in Colorectal Cancer", *Gastroenterology*, vol. 138, no. 6, 2073–2087.e3, 2010, ISSN: 0016-5085. DOI: 10.1053/J.GASTRO.2009.12.064. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0016508510001691.

[15] P. Dimitri, "Revising the selfish DNA hypothesis new evidence on accumulation of transposable elements in heterochromatin", *Trends in Genetics*, vol. 15, no. 4, pp. 123–124, 1999.

[16] G. Benson, "Tandem repeats finder: a program to analyze DNA sequences", Tech. Rep. 2, 1999, pp. 573–580. [Online]. Available: http://nar.oxfordjournals.org/.

[17] R. Frothingham and W. A Meeker-O'Connell, "Genetic diversity in the Mycobacterium tuberculosis complex based on variable numbers of tandem DNA repeats", *Microbiology*, vol. 144, no. 5, pp. 1189–1196, 1998, ISSN: 1350-0872. DOI: 10.1099/00221287-144-5-1189. [Online]. Available: http://mic.microbiologyresearch.org/content/journal/micro/10.1099/00221287-144-5-1189.

[18] J. K. Colbourne, M. E. Pfrender, D. Gilbert, W. K. Thomas, A. Tucker, T. H. Oakley, S. Tokishita, A. Aerts, G. J. Arnold, M. K. Basu, D. J. Bauer, C. E. Cáceres, L. Carmel, C. Casola, J.-H. Choi, J. C. Detter, Q. Dong, S. Dusheyko, B. D. Eads, T. Fröhlich, K. A. Geiler-Samerotte, D. Gerlach, P. Hatcher, S. Jogdeo, J. Krijgsveld, E. V. Kriventseva, D. Kültz, C. Laforsch, E. Lindquist, J. Lopez, J. R. Manak, J. Muller, J. Pangilinan, R. P. Patwardhan, S. Pitluck, E. J. Pritham, A. Rechtsteiner, M. Rho, I. B. Rogozin, O. Sakarya, A. Salamov, S. Schaack, H. Shapiro, Y. Shiga, C. Skalitzky, Z. Smith, A. Souvorov, W. Sung, Z. Tang, D. Tsuchiya, H. Tu, H. Vos, M. Wang, Y. I. Wolf, H. Yamagata, T. Yamada, Y. Ye, J. R. Shaw, J. Andrews, T. J. Crease, H. Tang, S. M. Lucas, H. M. Robertson, P. Bork, E. V. Koonin, E. M. Zdobnov, I. V. Grigoriev, M. Lynch, and J. L. Boore, "The ecoresponsive genome of Daphnia pulex.", *Science (New York, N.Y.)*, vol. 331, no. 6017, pp. 555–61, 2011, ISSN: 1095-9203. DOI: 10.1126/science.1197761. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21292972http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3529199.

[19] S. J. J. Brouns, M. M. Jore, M. Lundgren, E. R. Westra, R. J. H. Slijkhuis, A. P. L. Snijders, M. J. Dickman, K. S. Makarova, E. V. Koonin, J. van der Oost, and F. Zhang, "Small CRISPR RNAs guide antiviral defense in prokaryotes.", *Science (New York, N.Y.)*, vol. 321, no. 5891, pp. 960–4, 2008, ISSN: 1095-9203. DOI: 10.1126/science.1159689. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18703739http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5898235http://www.sciencemag.org/cgi/doi/10.1126/science.1159689.

[20] A. Bailey, "Exploratory analysis of pathogen resistance responsible genetic elements in two eudicots from whole genome sequence", *Theses*, 2010. [Online]. Available: https://scholarworks.rit.edu/theses/4070.

[21] E. Heard and R. Martienssen, "Transgenerational Epigenetic Inheritance: Myths and Mechanisms", *Cell*, vol. 157, no. 1, pp. 95–109, 2014, ISSN: 0092-8674. DOI: 10.1016/J.CELL.2014.02.045. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092867414002864.

[22] B. McClintock, "Controlling elements and the gene.", *Cold Spring Harbor symposia on quantitative biology*, vol. 21, pp. 197–216, 1956, ISSN: 0091-7451. DOI: 10.1101/SQB.1956.021.01.017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/13433592.

[23] G Bryan, D Garza, and D Hartl, "Insertion and excision of the transposable element mariner in Drosophila.", *Genetics*, vol. 125, no. 1, pp. 103–14, 1990, ISSN: 0016-6731. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1203992/.

[24] S. I. Wright and D. J. Schoen, "Transposon dynamics and the breeding system", *Genetica*, vol. 107, no. 1/3, pp. 139–148, 1999, ISSN: 00166707. DOI: 10.1023/A:1003953126700. [Online]. Available: http://link.springer.com/10.1023/A:1003953126700.

[25] T. C. Osborn, J. Chris Pires, J. A. Birchler, D. L. Auger, Z. Jeffery Chen, H.-S. Lee, L. Comai, A. Madlung, R. Doerge, V. Colot, and R. A. Martienssen, "Understanding mechanisms of novel gene expression in polyploids", *Trends in Genetics*, vol. 19, no. 3, pp. 141–147, 2003, ISSN: 0168-9525. DOI: 10.1016/S0168-9525(03)00015-5. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168952503000155.

[26] R. Rebollo, B. Horard, B. Hubert, and C. Vieira, "Jumping genes and epigenetics: Towards new species", *Gene*, vol. 454, no. 1-2, pp. 1–7, 2010, ISSN: 0378-1119. DOI: 10.1016/J.GENE.2010.01.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378111910000296.

[27] M. Schwartz, J. Chen, M. Janda, M. Sullivan, J. den Boon, and P. Ahlquist, "A Positive-Strand RNA Virus Replication Complex Parallels Form and Function of Retrovirus Capsids", *Molecular Cell*, vol. 9, no. 3, pp. 505–514, 2002, ISSN: 1097-2765. DOI: 10.1016/S1097-2765(02)00474-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1097276502004744.

[28] M. G. Barrón, A.-S. Fiston-Lavier, D. A. Petrov, and J. González, "Population Genomics of Transposable Elements in Drosophila", *Annual Review of Genetics*, vol. 48, no. 1, pp. 561–581, 2014. DOI: 10.1146/annurev-genet-120213-092359. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25292358.

[29] M. T. Reilly, G. J. Faulkner, J. Dubnau, I. Ponomarev, and F. H. Gage, "The role of transposable elements in health and diseases of the central nervous system.", *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 33, no. 45, pp. 17 577–86, 2013, ISSN: 1529-2401. DOI: 10.1523/JNEUROSCI.3369-13.2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24198348http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3818539.

[30] T. Wicker, F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, and A. H. Schulman, "A unified classification system for eukaryotic transposable elements", *Nature Reviews Genetics*, vol. 8, no. 12, pp. 973–982, 2007, ISSN: 1471-0056. DOI: 10.1038/nrg2165. [Online]. Available: http://www.nature.com/articles/nrg2165.

[31] R. H. Plasterk, Z. Izsvák, and Z. Ivics, "Resident aliens: the Tc1/mariner superfamily of transposable elements", *Trends in Genetics*, vol. 15, no. 8, pp. 326–332, 1999, ISSN: 0168-9525. DOI: 10.1016/S0168-9525(99)01777-1. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168952599017771.

[32] H. H. Kazazian, "Mobile elements: drivers of genome evolution.", *Science (New York, N.Y.)*, vol. 303, no. 5664, pp. 1626–32, 2004, ISSN: 1095-9203. DOI: 10.1126/science.1089670. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15016989.

[33] M. W. Simmen, S Leitgeb, J Charlton, S. J. Jones, B. R. Harris, V. H. Clark, and A Bird, "Nonmethylated Transposable Elements and Methylated Genes in a Chordate Genome", *Science*, vol. 283, no. 5405, pp. 1164–1167, 1999, ISSN: 00368075. DOI: 10.1126/science.283.5405.1164. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10024242http://www.sciencemag.org/cgi/doi/10.1126/science.283.5405.1164.

[34] D. J. Finnegan, "Eukaryotic transposable elements and genome evolution", *Trends in Genetics*, vol. 5, pp. 103–107, 1989, ISSN: 0168-9525. DOI: 10.1016/0168-9525(89)90039-5. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0168952589900395.

[35] D. J. Finnegan, "Transposable elements", *Current opinion in genetics & development*, vol. 2, no. 6, pp. 861–867, 1992, ISSN: 0959-437X. [Online]. Available: http://ovidsp.ovid.com/ovidweb.cgi?T=JS{\&}CSC=Y{\&}NEWS=N{\&}PAGE=fulltext{\&}D=med3{\&}AN=1335807https://rutgers.primo.exlibrisgroup.com/discovery/openurl?institution=01RUT{_}INST{\&}vid=01RUT{_}INST:01RUT{\&}?sid=OVID:medline{\&}id=doi:10.1016{\%}2FS0959-437X{\%}2805{\%}2980108-X{\&}issn=09.

[36] E. B. Chuong, N. C. Elde, and C. Feschotte, "Regulatory activities of transposable elements: from conflicts to benefits.", *Nature reviews. Genetics*, vol. advance on, no. 2, pp. 71–86, 2016, ISSN: 1471-0064. DOI: 10.1038/nrg.2016.139. [Online]. Available: http://dx.doi.org/10.1038/nrg.2016.139.

[37] M. Bushell and P. Sarnow, "Hijacking the translation apparatus by RNA viruses.", *The Journal of cell biology*, vol. 158, no. 3, pp. 395–9, 2002, ISSN: 0021-9525. DOI: 10.1083/jcb.200205044. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12163463http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2173839.

[38] R. Marquet, C. Isel, C. Ehresmann, and B. Ehresmann, "tRNAs as primer of reverse transcriptases", *Biochimie*, vol. 77, no. 1-2, pp. 113–124, 1995, ISSN: 0300-9084. DOI: 10.1016/0300-9084(96)88114-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0300908496881144.

[39] J. H. Lin and H. L. Levin, "Reverse transcription of a self-primed retrotransposon requires an RNA structure similar to the U5-IR stem-loop of retroviruses.", *Molecular & Cellular Biology*, vol. 18, no. 11, pp. 6859–6869, 1998, ISSN: 0270-7306. [Online]. Available: http://ovidsp.ovid.com/ovidweb.cgi?T=JS{\&}CSC=Y{\&}NEWS=N{\&}PAGE=fulltext{\&}D=med4{\&}AN=9774699https://rutgers.primo.exlibrisgroup.com/discovery/openurl?institution=01RUT{_}INST{\&}vid=01RUT{_}INST:01RUT{\&}?sid=OVID:medline{\&}id=doi:10.1128{\%}2FMCB.18.11.6859{\&}issn=0270-7306{\&}isb.

[40] G. M. Rubin and A. C. Spradling, "Genetic transformation of Drosophila with transposable element vectors.", *Science (New York, N.Y.)*, vol. 218, no. 4570, pp. 348–53, 1982, ISSN: 0036-8075. DOI: 10.1126/SCIENCE.6289436. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/6289436.

[41] J. Dostie, T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green, and J. Dekker, "Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements.", *Genome research*, vol. 16, no. 10, pp. 1299–309, 2006, ISSN: 1088-9051. DOI: 10.1101/gr.5571506. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16954542http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1581439.

[42] Bonev Boyan and Cavalli Giacomo, "Organization and function of the 3D genome", *Nature Review Genetics*, vol. 17, pp. 661 –678, 2016, ISSN: 1471-0056. DOI: 10.1038/nrg.2016.112. [Online]. Available: http://www.nature.com/nrg/journal/v17/n11/full/nrg.2016.112.html?WT.mc{_}id=FBK{_}NatureReviews.

[43] M. Lynch and J. S. Conery, "The origins of genome complexity.", *Science*, vol. 302, no. 5649, pp. 1401–1404, 2003, ISSN: 1095-9203.

[44] W. de Laat and D. Duboule, "Topology of mammalian developmental enhancers and their regulatory landscapes.", *Nature*, vol. 502, no. 7472, pp. 499–506, 2013, ISSN: 1476-4687. DOI: 10.1038/nature12753. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24153303.

[45] S Jeon and P. F. Lambert, "Integration of human papillomavirus type 16 DNA into the human genome leads to increased stability of E6 and E7 mRNAs: implications for cervical carcinogenesis.", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 5, pp. 1654–8, 1995, ISSN: 0027-8424. DOI: 10.1073/PNAS.92.5.1654. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/7878034http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC42578.

[46] A Ludwig, V. L.d. S. Valente, and E. L. S. Loreto, "Multiple invasions of Errantivirus in the genus Drosophila.", *Insect molecular biology*, vol. 17, no. 2, pp. 113–124, 2008, ISSN: 1365-2583. DOI: //dx.doi.org/10.1111/j.1365-2583.2007.00787.x. [Online]. Available: http://ovidsp.ovid.com/ovidweb.cgi?T=JS{\&}CSC=Y{\&}NEWS=N{\&}PAGE=fulltext{\&}D=med6{\&}AN=18353101https://rutgers.primo.exlibrisgroup.com/discovery/openurl?institution=01RUT{_}INST{\&}vid=01RUT{_}INST:01RUT{\&}?sid=OVID:medline{\&}id=doi:10.1111{\%}2Fj.1365-2583.2007.00787.x{\&}issn=09.

[47] P. Neumann, D. Požárková, and J. Macas, "Highly abundant pea LTR retrotransposon Ogre is constitutively transcribed and partially spliced", *Plant Molecular Biology*, vol. 53, no. 3, pp. 399–410, 2003, ISSN: 0167-4412. DOI: 10.1023/B:PLAN.0000006945.77043.ce. [Online]. Available: http://link.springer.com/10.1023/B:PLAN.0000006945.77043.ce.

[48] M. S. Lalonde and W. I. Sundquist, "How HIV finds the door.", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 46, pp. 18 631–2, 2012, ISSN: 1091-6490. DOI: 10.1073/pnas.1215940109. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23118338http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3503163.

[49] J. S. Saad, J. Miller, J. Tai, A. Kim, R. H. Ghanam, and M. F. Summers, "Structural basis for targeting HIV-1 Gag proteins to the plasma membrane for virus assembly.", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 30, pp. 11 364–9, 2006, ISSN: 0027-8424. DOI: 10.1073/pnas.0602818103. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16840558http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1544092.

[50] E. M. McCarthy and J. F. McDonald, "Long terminal repeat retrotransposons of Mus musculus", *Genome Biology*, vol. 5, no. 3, R14, 2004, ISSN: 14656906. DOI: 10.1186/gb-2004-5-3-r14. [Online]. Available: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2004-5-3-r14.

[51] D. J. Griffiths, "Endogenous retroviruses in the human genome sequence", *Genome Biology*, vol. 2, no. 6, reviews1017.1, 2001, ISSN: 14656906. DOI: 10.1186/gb-2001-2-6-reviews1017. [Online]. Available: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2001-2-6-reviews1017.

[52] A. J. Flavell, E. Dunbar, R. Anderson, S. R. Pearce, R. Hartley, and A. Kumar, "Ty1–copia group retrotransposons are ubiquitous and heterogeneous in higher plants", *Nucleic Acids Research*, vol. 20, no. 14, pp. 3639–3644, 1992. DOI: 10.1093/nar/20.14.3639. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/20.14.3639.

[53] J. S. Han, "Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions", *Mobile DNA*, vol. 1, no. 1, p. 15, 2010, ISSN: 1759-8753. DOI: 10.1186/1759-8753-1-15. [Online]. Available: http://mobilednajournal.biomedcentral.com/articles/10.1186/1759-8753-1-15.

[54] J. González, K. Lenkov, M. Lipatov, J. M. Macpherson, and D. A. Petrov, "High Rate of Recent Transposable Element–Induced Adaptation in Drosophila melanogaster", *PLoS Biology*, vol. 6, no. 10, M. A. F. Noor, Ed., e251, 2008, ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0060251. [Online]. Available: http://dx.plos.org/10.1371/journal.pbio.0060251.

[55] T. H. Bestor, "Transposons Reanimated in Mice", *Cell*, vol. 122, no. 3, pp. 322–325, 2005, ISSN: 0092-8674. DOI: 10.1016/J.CELL.2005.07.024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092867405007555.

[56] M. F. Singer, "SINEs and LINEs: Highly repeated short and long interspersed sequences in mammalian genomes", *Cell*, vol. 28, no. 3, pp. 433–434, 1982, ISSN: 00928674. DOI: 10.1016/0092-8674(82)90194-5. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/0092867482901945.

[57] S. V. Nuzhdin, E. G. Pasyukova, and T. F. Mackay, "Accumulation of transposable elements in laboratory lines of Drosophila melanogaster", *Genetica*, vol. 100, no. 1-3, pp. 167–175, 1997, ISSN: 0016-6707. [Online]. Available: http://ovidsp.ovid.com/ovidweb.cgi?T=JS{\&}CSC=Y{\&}NEWS=N{\&}PAGE=fulltext{\&}D=med4{\&}AN=9440270https://rutgers.primo.exlibrisgroup.com/discovery/openurl?institution=01RUT{_}INST{\&}vid=01RUT{_}INST:01RUT{\&}?sid=OVID:medline{\&}id=doi:10.1023{\%}2FA{\%}3A1018381512384{\&}issn=0016-6707{\&}.

[58] H. Ha, J. W. Loh, and J. Xing, "Identification of polymorphic SVA retrotransposons using a mobile element scanning method for SVA (ME-Scan-SVA)", *Mobile DNA*, vol. 7, no. 1, p. 15, 2016, ISSN: 1759-8753. DOI: 10.1186/s13100-016-0072-x. [Online]. Available: http://mobilednajournal.biomedcentral.com/articles/10.1186/s13100-016-0072-x.

[59] C. W. Schmid and P. L. Deininger, "Sequence organization of the human genome", *Cell*, vol. 6, no. 3, pp. 345–358, 1975, ISSN: 0092-8674. DOI: 10.1016/0092-8674(75)90184-1. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0092867475901841?via{\%}3Dihub.

[60] R. Cordaux, S. Pichon, A. Ling, P. Pérez, C. Delaunay, F. Vavre, D. Bouchon, and P. Grève, "Intense transpositional activity of insertion sequences in an ancient obligate endosymbiont.", *Molecular biology and evolution*, vol. 25, no. 9, pp. 1889–96, 2008, ISSN: 1537-1719. DOI: 10.1093/molbev/msn134. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18562339http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2515875.

[61] R. Cordaux and M. A. Batzer, "The impact of retrotransposons on human genome evolution", *Nature Reviews Genetics*, vol. 10, no. 10, pp. 691–703, 2009, ISSN: 1471-0056. DOI: 10.1038/nrg2640. [Online]. Available: http://www.nature.com/articles/nrg2640.

[62] S. E. Holt, W. E. Wright, and J. W. Shay, "Regulation of telomerase activity in immortal cell lines.", *Molecular and cellular biology*, vol. 16, no. 6, pp. 2932–9, 1996, ISSN: 0270-7306. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8649404http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC231287.

[63] V. Lundblad and N. Kleckner, "Mismatch Repair Mutations Of Escherichia Coli K12 Enhance Transposon Excision", *Genetics*, vol. 109, no. 1, pp. 3–19, 1985.

[64] Y. C. G. Lee and C. H. Langley, "Transposable elements in natural populations of Drosophila melanogaster", *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*, vol. 365, no. 1544, pp. 1219–1228, 2010, ISSN: 1471-2970. DOI: //dx.doi.org/10.1098/rstb.2009.0318. [Online]. Available: http://ovidsp.ovid.com/ovidweb.cgi?T=JS{\&}CSC=Y{\&}NEWS=N{\&}PAGE=fulltext{\&}D=med6{\&}AN=20308097https://rutgers.primo.exlibrisgroup.com/discovery/openurl?institution=01RUT{_}INST{\&

}vid=01RUT{_}INST:01RUT{\&}?sid=OVID:medline{\&}id=doi:10.1098{\%}2Frstb.2009.0318{\&}issn=0962-8436{\&}is.

[65] W. Deng, J. W. Rupon, I. Krivega, L. Breda, I. Motta, K. S. Jahn, A. Reik, P. D. Gregory, S. Rivella, A. Dean, and G. A. Blobel, "Reactivation of developmentally silenced globin genes by forced chromatin looping.", *Cell*, vol. 158, no. 4, pp. 849–60, 2014, ISSN: 1097-4172. DOI: 10.1016/j.cell.2014.05.050. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25126789http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4134511.

[66] J. Thomas and E. J. Pritham, "Helitrons, the Eukaryotic Rolling-circle Transposable Elements", *Microbiology Spectrum*, vol. 3, no. 4, 2015, ISSN: 2165-0497. DOI: 10.1128/microbiolspec.MDNA3-0049-2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26350323http://www.asmscience.org/content/journal/microbiolspec/10.1128/microbiolspec.MDNA3-0049-2014.

[67] M. W. Bruford and R. K. Wayne, "Microsatellites and their application to population genetic studies", *Current Opinion in Genetics & Development*, vol. 3, no. 6, pp. 939–943, 1993, ISSN: 0959-437X. DOI: 10.1016/0959-437X(93)90017-J. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0959437X9390017J.

[68] S. Richards, Y. Liu, B. R. Bettencourt, P. Hradecky, S. Letovsky, R. Nielsen, K. Thornton, M. J. Hubisz, R. Chen, R. P. Meisel, O. Couronne, S. Hua, M. A. Smith, P. Zhang, J. Liu, H. J. Bussemaker, M. F. van Batenburg, S. L. Howells, S. E. Scherer, E. Sodergren, B. B. Matthews, M. A. Crosby, A. J. Schroeder, D. Ortiz-Barrientos, C. M. Rives, M. L. Metzker, D. M. Muzny, G. Scott, D. Steffen, D. A. Wheeler, K. C. Worley, P. Havlak, K. J. Durbin, A. Egan, R. Gill, J. Hume, M. B. Morgan, G. Miner, C. Hamilton, Y. Huang, L. Waldron, D. Verduzco, K. P. Clerc-Blankenburg, I. Dubchak, M. A. F. Noor, W. Anderson, K. P. White, A. G. Clark, S. W. Schaeffer, W. Gelbart, G. M. Weinstock, and R. A. Gibbs, "Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution.", *Genome research*, vol. 15, no. 1, pp. 1–18, 2005, ISSN: 1088-9051. DOI: 10.1101/gr.3059305. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15632085http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC540289.

[69] H. Tabara, M. Sarkissian, W. G. Kelly, J. Fleenor, A. Grishok, L. Timmons, A. Fire, and C. C. Mello, "The rde-1 Gene, RNA Interference, and Transposon Silencing in C. elegans", *Cell*, vol. 99, no. 2, pp. 123–132, 1999, ISSN: 0092-8674. DOI: 10.1016/S0092-8674(00)81644-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S009286740081644X.

[70] R. A. Waterland and R. L. Jirtle, "Early nutrition, epigenetic changes at transposons and imprinted genes, and enhanced susceptibility to adult chronic diseases.", *Nutrition (Burbank, Los Angeles County, Calif.)*, vol. 20, no. 1, pp. 63–8, 2004, ISSN: 0899-9007. DOI: 10.1016/J.NUT.2003.09.011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14698016.

[71] B. Burgess-Beusse, C. Farrell, M. Gaszner, M. Litt, V. Mutskov, F. Recillas-Targa, M. Simpson, A. West, and G. Felsenfeld, "The insulation of genes from external enhancers and silencing chromatin.", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99 Suppl 4, no. Dc, pp. 16 433–16 437, 2002, ISSN: 00278424. DOI: 10.1073/pnas.162342499.

[72] M. Gause, P. Morcillo, and D. Dorsett, "Insulation of Enhancer-Promoter Communication by a Gypsy Transposon Insert in the Drosophila cut Gene: Cooperation between Suppressor of Hairy-wing and Modifier of mdg4 Proteins", *Molecular and Cellular Biology*, vol. 21, no. 14, pp. 4807–4817, 2001, ISSN: 0270-7306. DOI: 10.1128/MCB.21.14.4807-4817.2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11416154http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC87172http://mcb.asm.org/cgi/doi/10.1128/MCB.21.14.4807-4817.2001.

[73] A. S. Wilson, B. E. Power, and P. L. Molloy, "DNA hypomethylation and human diseases", *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, vol. 1775, no. 1, pp. 138–162, 2007, ISSN: 0304-419X. DOI: 10.1016/J.BBCAN.2006.08.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304419X06000564.

[74] G. F. Bouvet, V. Jacobi, K. V. Plourde, and L. Bernier, "Stress-induced mobility of OPHIO1 and OPHIO2, DNA transposons of the Dutch elm disease fungi", *Fungal Genetics and Biology*, vol. 45, no. 4, pp. 565–578, 2008, ISSN: 1087-1845. DOI: 10.1016/J.FGB.2007.12.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1087184507002411.

[75] V. P. Belancio, P. L. Deininger, and A. M. Roy-Engel, "LINE dancing in the human genome: transposable elements and disease", *Genome Medicine*, vol. 1, no. 10, p. 97, 2009, ISSN: 1756-994X. DOI: 10.1186/gm97. [Online]. Available: http://genomemedicine.biomedcentral.com/articles/10.1186/gm97.

[76] L. S. Collier, C. M. Carlson, S. Ravimohan, A. J. Dupuy, and D. A. Largaespada, "Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse", *Nature*, vol. 436, no. 7048, pp. 272–276, 2005. DOI: 10.1038/nature03681. [Online]. Available: http://www.nature.com/articles/nature03681.

[77] A. J. Dupuy, L. M. Rogers, J. Kim, K. Nannapaneni, T. K. Starr, P. Liu, D. A. Largaespada, T. E. Scheetz, N. A. Jenkins, and N. G. Copeland, "A modified sleeping beauty transposon system that can be used to model a wide variety of human cancers in mice.", *Cancer research*, vol. 69, no. 20, pp. 8150–6, 2009, ISSN: 1538-7445. DOI: 10.1158/0008-5472.CAN-09-1135. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19808965http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3700628.

[78] A. V. Spirov, A. B. Kazansky, L. Zamdborg, J. J. Merelo, and V. F. Levchenko, "Forced Evolution in Silico by Artificial Transposons and their Genetic Operators: The John Muir Ant Problem", 2009. arXiv: 0910.5542. [Online]. Available: http://arxiv.org/abs/0910.5542.

[79] G. F. Barry, "Permanent Insertion of Foreign Genes into the Chromosomes of Soil Bacteria", *Nature Biotechnology*, vol. 4, no. 5, pp. 446–449, 1986, ISSN: 1087-0156. DOI: 10.1038/nbt0586-446. [Online]. Available: http://www.nature.com/doifinder/10.1038/nbt0586-446.

[80] A. Ling and R. Cordaux, "Insertion Sequence Inversions Mediated by Ectopic Recombination between Terminal Inverted Repeats", *PLoS ONE*, vol. 5, no. 12, M. A. Batzer, Ed., e15654, 2010, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0015654. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21187977http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3004938https://dx.plos.org/10.1371/journal.pone.0015654.

[81] A. Sanchez-Gracia, X. Maside, and B. Charlesworth, "High rate of horizontal transfer of transposable elements in Drosophila.", *Trends in Genetics*, vol. 21, no. 4, pp. 200–203, 2005, ISSN: 0168-9525.

[82] M. M. Magwire, F. Bayer, C. L. Webster, C. Cao, and F. M. Jiggins, "Successive increases in the resistance of Drosophila to viral infection through a transposon insertion followed by a Duplication.", *PLoS Genetics*, vol. 7, no. 10, e1002337, 2011, ISSN: 1553-7404. DOI: //dx.doi.org/10.1371/journal.pgen.1002337. [Online]. Available: http://ovidsp.ovid.com/ovidweb.cgi?T=JS{\&}CSC=Y{\&}NEWS=N{\&}PAGE=fulltext{\&}D=med7{\&}AN=22028673https://rutgers.primo.exlibrisgroup.com/discovery/openurl?institution=01RUT{_}INST{\&}vid=01RUT{_}INST:01RUT{\&}?sid=OVID:medline{\&}id=doi:10.1371{\%}2Fjournal.pgen.1002337{\&}issn=1553-7.

[83] H. M. Robertson, C. R. Preston, R. W. Phillis, D. M. Johnson-Schlitz, W. K. Benz, and W. R. Engels, "A stable genomic source of P element transposase in Drosophila melanogaster.", *Genetics*, vol. 118, no. 3, 1988.

[84] J. B. S. Haldane, "The Rate of Mutation of Human Genes", *Hereditas*, vol. 35, no. S1, pp. 267–273, 1949, ISSN: 00180661. DOI: 10.1111/j.1601-5223.1949.tb03339.x. [Online]. Available: http://doi.wiley.com/10.1111/j.1601-5223.1949.tb03339.x.

[85] J. W. Drake, "A constant rate of spontaneous mutation in DNA-based microbes.", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 88, no. 16, pp. 7160–4, 1991, ISSN: 0027-8424. DOI: 10.1073/PNAS.88.16.7160. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/1831267http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC52253.

[86] E. Lerat, C. Rizzon, and C. Biémont, "Sequence divergence within transposable element families in the Drosophila melanogaster genome.", *Genome research*, vol. 13, no. 8, pp. 1889–96, 2003, ISSN: 1088-9051. DOI: 10.1101/gr.827603.

[87] M. W. Snyder, A. Adey, J. O. Kitzman, and J. Shendure, "Haplotype-resolved genome sequencing: experimental methods and applications.", *Nature reviews. Genetics*, vol. 16, no. 6, pp. 344–58, 2015, ISSN: 1471-0064. DOI: 10.1038/nrg3903. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25948246.

[88] V. Joardar, M. Lindeberg, R. W. Jackson, J. Selengut, R. Dodson, L. M. Brinkac, S. C. Daugherty, R. Deboy, A. S. Durkin, M. G. Giglio, R. Madupu, W. C. Nelson, M. J. Rosovitz, S. Sullivan, J. Crabtree, T. Creasy, T. Davidsen, D. H. Haft, N. Zafar, L. Zhou, R. Halpin, T. Holley, H. Khouri, T. Feldblyum, O. White, C. M. Fraser, A. K. Chatterjee, S. Cartinhour, D. J. Schneider, J. Mansfield, A. Collmer, and C. R. Buell, "Whole-genome sequence analysis of Pseudomonas syringae pv. phaseolicola 1448A reveals divergence among pathovars in genes involved in virulence and transposition.", *Journal of bacteriology*, vol. 187, no. 18, pp. 6488–98, 2005, ISSN: 0021-9193. DOI: 10.1128/JB.187.18.6488-6498.2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16159782http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1236638.

[89] J. González and D. A. Petrov, "The adaptive role of transposable elements in the Drosophila genome", *Gene*, vol. 448, no. 2, pp. 124–133, 2009, ISSN: 03781119. DOI: 10.1016/j.gene.2009.06.008. [Online]. Available: http://dx.doi.org/10.1016/j.gene.2009.06.008.

[90] M. Louwers, E. Splinter, R. van Driel, W. de Laat, and M. Stam, "Studying physical chromatin interactions in plants using Chromosome Conformation Capture (3C).", *Nature protocols*, vol. 4, no. 8, pp. 1216–29, 2009, ISSN: 1750-2799. DOI: 10.1038/nprot.2009.113. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19644461.

[91] H. Hagège, P. Klous, C. Braem, E. Splinter, J. Dekker, G. Cathala, W. de Laat, and T. Forné, "Quantitative analysis of chromosome conformation capture assays (3C-qPCR).", *Nature protocols*, vol. 2, no. 7, pp. 1722–33, 2007, ISSN: 1750-2799. DOI: 10.1038/nprot.2007.243. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17641637.

[92] E. Gómez-Díaz and V. G. Corces, "Architectural proteins: regulators of 3D genome organization in cell fate.", *Trends in cell biology*, vol. 24, no. 11, pp. 703–11, 2014, ISSN: 1879-3088. DOI: 10.1016/j.tcb.2014.08.003. arXiv: NIHMS150003. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0962892414001354http://www.ncbi.nlm.nih.gov/pubmed/25218583http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4254322.

[93] T. Nagano, Y. Lubling, T. J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E. D. Laue, A. Tanay, and P. Fraser, "Single-cell Hi-C reveals cell-to-cell variability in chromosome structure.", *Nature*, vol. 502, no. 7469, pp. 59–64, 2013, ISSN: 1476-4687. DOI: 10.1038/nature12593. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24067610http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3869051.

[94] O. Schwartzman, Z. Mukamel, N. Oded-Elkayam, P. Olivares-Chauvet, Y. Lubling, G. Landan, S. Izraeli, and A. Tanay, "UMI-4C for quantitative and targeted chromosomal contact profiling.", *Nature methods*, vol. 13, no. 8, pp. 685–91, 2016, ISSN: 1548-7105. DOI: 10.1038/nmeth.3922. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/27376768.

[95] G. E. Scordilis, H Ree, and T. G. Lessie, "Identification of transposable elements which activate gene expression in Pseudomonas cepacia.", *Journal of bacteriology*, vol. 169, no. 1, pp. 8–13, 1987, ISSN: 0021-9193. DOI: 10.1128/JB.169.1.

8-13.1987. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/3025189http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC211726.

[96] A. Caspi and L. Pachter, "Identification of transposable elements using multiple alignments of related genomes.", *Genome research*, vol. 16, no. 2, pp. 260–70, 2006, ISSN: 1088-9051. DOI: 10.1101/gr.4361206. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16354754http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1361722.

[97] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data", *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, 2014, ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btu170. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu170.

[98] J. G. Caporaso, C. L. Lauber, W. A. Walters, D. Berg-Lyons, J. Huntley, N. Fierer, S. M. Owens, J. Betley, L. Fraser, M. Bauer, N. Gormley, J. A. Gilbert, G. Smith, and R. Knight, "Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms", *The ISME Journal*, vol. 6, no. 8, pp. 1621–1624, 2012, ISSN: 1751-7362. DOI: 10.1038/ismej.2012.8. [Online]. Available: http://www.nature.com/articles/ismej20128.

[99] A. Rhoads and K. F. Au, "PacBio Sequencing and Its Applications", *Genomics, Proteomics & Bioinformatics*, vol. 13, no. 5, pp. 278–289, 2015, ISSN: 1672-0229. DOI: 10.1016/J.GPB.2015.08.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1672022915001345.

[100] A. S. Mikheyev and M. M. Y. Tin, "A first look at the Oxford Nanopore MinION sequencer", *Molecular Ecology Resources*, vol. 14, no. 6, pp. 1097–1102, 2014, ISSN: 1755098X. DOI: 10.1111/1755-0998.12324. [Online]. Available: http://doi.wiley.com/10.1111/1755-0998.12324.

[101] A. Sanyal, B. R. Lajoie, G. Jain, and J. Dekker, "The long-range interaction landscape of gene promoters.", *Nature*, vol. 489, no. 7414, pp. 109–13, 2012, ISSN: 1476-4687. DOI: 10.1038/nature11279. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22955621http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3555147.

[102] T. F. C. Mackay, S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, D. Zhu, S. Casillas, Y. Han, M. M. Magwire, J. M. Cridland, M. F. Richardson, R. R. H. Anholt, M. Barrón, C. Bess, K. P. Blankenburg, M. A. Carbone, D. Castellano, L. Chaboub, L. Duncan, Z. Harris, M. Javaid, J. C. Jayaseelan, S. N. Jhangiani, K. W. Jordan, F. Lara, F. Lawrence, S. L. Lee, P. Librado, R. S. Linheiro, R. F. Lyman, A. J. Mackey, M. Munidasa, D. M. Muzny, L. Nazareth, I. Newsham, L. Perales, L.-L. Pu, C. Qu, M. Ràmia, J. G. Reid, S. M. Rollmann, J. Rozas, N. Saada, L. Turlapati, K. C. Worley, Y.-Q. Wu, A. Yamamoto, Y. Zhu, C. M. Bergman, K. R. Thornton, D. Mittelman, and R. A. Gibbs, "The Drosophila melanogaster Genetic Reference Panel", *Nature*, vol. 482, no. 7384, pp. 173–178, 2012, ISSN: 0028-0836. DOI: 10.1038/nature10811. [Online]. Available: http://www.nature.com/articles/nature10811.

[103] B. R. Graveley, A. N. Brooks, J. W. Carlson, M. O. Duff, J. M. Landolin, L. Yang, C. G. Artieri, M. J. van Baren, N. Boley, B. W. Booth, J. B. Brown, L. Cherbas, C. A. Davis, A. Dobin, R. Li, W. Lin, J. H. Malone, N. R. Mattiuzzo, D. Miller, D. Sturgill, B. B. Tuch, C. Zaleski, D. Zhang, M. Blanchette, S. Dudoit, B. Eads, R. E. Green, A. Hammonds, L. Jiang, P. Kapranov, L. Langton, N. Perrimon, J. E. Sandler, K. H. Wan, A. Willingham, Y. Zhang, Y. Zou, J. Andrews, P. J. Bickel, S. E. Brenner, M. R. Brent, P. Cherbas, T. R. Gingeras, R. A. Hoskins, T. C. Kaufman, B. Oliver, and S. E. Celniker, "The developmental transcriptome of Drosophila melanogaster", *Nature*, vol. 471, no. 7339, pp. 473–479, 2011, ISSN: 0028-0836. DOI: 10.1038/nature09715. [Online]. Available: http://www.nature.com/articles/nature09715.

[104] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.", *Bioinformatics (Oxford, England)*, vol. 26, no. 1, pp. 139–40, 2010, ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp616. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19910308http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2796818.

[105] C. Sessegolo, N. Burlet, and A. Haudry, "Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies", *Biology Letters*, vol. 12, no. 8, p. 20 160 407, 2016, ISSN: 1744-9561. DOI: 10.1098/rsbl.2016.0407. [Online]. Available: http://rsbl.royalsocietypublishing.org/lookup/doi/10.1098/rsbl.2016.0407.

[106] D. E. Miller, C. Staber, J. Zeitlinger, and R. S. Hawley, "Highly Contiguous Genome Assemblies of 15 Drosophila Species Generated Using Nanopore Sequencing.", *G3 (Bethesda, Md.)*, vol. 8, no. 10, pp. 3131–3141, 2018, ISSN: 2160-1836. DOI: 10.1534/g3.118.200160. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/30087105http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6169393.

[107] Cooper GM, *The Cell: A Molecular Approach*, S. Associates, Ed. MA: Sundarland (MA), 2000.

[108] M. F. Lin, J. W. Carlson, M. A. Crosby, B. B. Matthews, C. Yu, S. Park, K. H. Wan, A. J. Schroeder, L. S. Gramates, S. E. St Pierre, M. Roark, K. L. Wiley, R. J. Kulathinal, P. Zhang, K. V. Myrick, J. V. Antone, S. E. Celniker, W. M. Gelbart, M. Kellis, and M. Kellis, "Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes.", *Genome research*, vol. 17, no. 12, pp. 1823–36, 2007, ISSN: 1088-9051. DOI: 10.1101/gr.6679507. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17989253http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2099591.

[109] L. Myers and M. J. Sirois, "Spearman Correlation Coefficients, Differences between", in *Encyclopedia of Statistical Sciences*, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2006. DOI: 10.1002/0471667196.ess5050.pub2. [Online]. Available: http://doi.wiley.com/10.1002/0471667196.ess5050.pub2.

[110] E. A. Solares, M. Chakraborty, D. E. Miller, S. Kalsow, K. Hall, A. G. Perera, J. J. Emerson, and R. S. Hawley, "Rapid Low-Cost Assembly of the Drosophila melanogaster Reference Genome Using Low-Coverage, Long-Read Sequencing.",

G3 (Bethesda, Md.), vol. 8, no. 10, pp. 3143–3154, 2018, ISSN: 2160-1836. DOI: 10.1534/g3.118.200162. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/30018084http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6169397.

[111] M. Lynch, *The Origins of Genome Architecture*. Sunderland, Mass. : Sinauer Associates, 1951, ISBN: 9780878934843.

[112] A. Stark, M. F. Lin, P. Kheradpour, J. S. Pedersen, L. Parts, J. W. Carlson, M. A. Crosby, M. D. Rasmussen, S. Roy, A. N. Deoras, J. G. Ruby, J. Brennecke, E. Hodges, A. S. Hinrichs, A. Caspi, B. Paten, S.-W. Park, M. V. Han, M. L. Maeder, B. J. Polansky, B. E. Robson, S. Aerts, J. van Helden, B. Hassan, D. G. Gilbert, D. A. Eastman, M. Rice, M. Weir, M. W. Hahn, Y. Park, C. N. Dewey, L. Pachter, W. J. Kent, D. Haussler, E. C. Lai, D. P. Bartel, G. J. Hannon, T. C. Kaufman, M. B. Eisen, A. G. Clark, D. Smith, S. E. Celniker, W. M. Gelbart, and M. Kellis, "Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures", *Nature*, vol. 450, no. 7167, pp. 219–232, 2007, ISSN: 0028-0836. DOI: 10.1038/nature06340. [Online]. Available: http://www.nature.com/articles/nature06340.

[113] W. Bao, K. K. Kojima, and O. Kohany, "Repbase Update, a database of repetitive elements in eukaryotic genomes", *Mobile DNA*, vol. 6, no. 1, p. 11, 2015, ISSN: 1759-8753. DOI: 10.1186/s13100-015-0041-9. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26045719http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4455052http://www.mobilednajournal.com/content/6/1/11.

[114] J. Jurka, V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz, "Repbase Update, a database of eukaryotic repetitive elements", *Cytogenetic and Genome Research*, vol. 110, no. 1-4, pp. 462–467, 2005, ISSN: 1424-8581. DOI: 10.1159/000084979. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16093699https://www.karger.com/Article/FullText/84979.

[115] A. L. Price, N. C. Jones, and P. A. Pevzner, "De novo identification of repeat families in large genomes", *Bioinformatics*, vol. 21, no. Suppl 1, pp. i351–i358, 2005, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti1018. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15961478https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bti1018.

[116] F. Morandat, B. Hill, L. Osvald, and J. Vitek, "Evaluating the Design of the R Language", in, Springer, Berlin, Heidelberg, 2012, pp. 104–131. DOI: 10.1007/978-3-642-31057-7_6. [Online]. Available: http://link.springer.com/10.1007/978-3-642-31057-7{_}6.

[117] F. Piano, "A Proposal for Comparative Genomics in Support the modENCODE Project Organizers", Tech. Rep. [Online]. Available: https://www.genome.gov/pages/research/sequencing/seqproposals/modencode{_}comparativegenomics{_}whitepaper.pdf.

This thesis was written with the help of a LaTeX template created by Steve Gunn and Sunil Patel with a heavy amount of modification in order to fulfill my vision of the thesis. The template (Version 2.5) was maintained by Vel as of 2017/08/27.