Identification, classification, and annotation of transposable elements in non-model organism: *Arabidopsis lyrata*

by

Natalia Aidé López Lugo

Department of Biology

McGill University, Montreal

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

" After all, every story has a story.”

Renée Ahdieh, *The Wrath & the Dawn*

TABLE OF CONTENTS

ABSTRACT

Studies have shown that transposable elements (TEs) play important roles in the systems they inhabit. The study of transposable elements is a fundamental part in understanding the genome, the evolution, and the effects in any organism. However, research has been limited since there is a lack of well-annotated TE libraries of non-model organisms which are essential to understand their role in biology. The annotation of TEs is a necessary step to understand their impact and role in biology. Currently, there's no method that can produce reliable results on the different classifications of TEs without the additional step of manual curation. It has been suggested that a combination of independent methods can be used to detect and classify TEs with more reliable results. This research aims to develop a pipeline that can be used in the identification, classification, and annotation of TEs in the non-model organism *Arabidopsis lyrata*. It uses a combination of open-source programs and independently created algorithms, in conjunction with multiple bioinformatic tools. The pipeline is described in enough detail that researchers with a basic understanding of programing can use the pipeline for their own research of interest. The TE pipeline was able to provide a list of 3,863 structurally conserved TEs for *A. lyrata*, a relatively young organism that has undergone little selection and has shown to have considerable TE activity. The list allowed for the analysis of TEs in *A. lyrata*, the length, size, location, and insertion in relation to the genes. The research addresses the lack of a pipeline for different types of TEs and furthers the research into the annotation of TEs and the characterisation of TEs in non-model organism *A. lyrata*.

RÉSUMÉ

Des études ont montré que les éléments transposables (ET) jouent un rôle important dans les systèmes qu'ils habitent. L'étude des éléments transposables est un élément fondamental pour comprendre le génome, l'évolution et les effets dans tout organisme. Cependant, la recherche a été limitée car il y a un manque de bibliothèques ET bien annotées d'organismes non-modèles qui sont essentielles pour comprendre leur rôle en biologie. L'annotation des ETs est une étape nécessaire pour comprendre leur impact et leur rôle en biologie. Actuellement, aucune méthode ne peut produire des résultats fiables sur les différentes classifications des ETs sans l'étape supplémentaire de curation manuelle. Il a été suggéré qu'une combinaison de méthodes indépendantes peut être utilisée pour détecter et classer les ET avec des résultats plus fiables. Cette recherche vise à développer un pipeline pouvant être utilisé dans l'identification, la classification et l'annotation des ETs dans l'organisme non-modèle *Arabidopsis lyrata*. Il utilise une combinaison de programmes open-source et d'algorithmes créés indépendamment, en conjonction avec de multiples outils bio-informatiques. Le pipeline est décrit avec suffisamment de détails pour que les chercheurs ayant une compréhension de base de la programmation puissent l'utiliser pour leurs propres recherches d'intérêt. Le ET pipeline a pu fournir une liste de 3,863 ET structurellement conservés pour *A. lyrata*, un organisme relativement jeune qui a subi peu de sélection et qui a montré une activité ET considérable. La liste a cédé la place à l'analyse des ET dans l'organisme, la longueur, la taille, l'emplacement et l'insertion par rapport aux gènes. La recherche aborde l'absence d'un pipeline pour différents types des ETs et approfondit la recherche sur l'annotation des ETs et la caractérisation des ETs dans l'organisme non-modèle *A. lyrata*.

ACKNOWLEDGEMENTS

First, I would like to thank my family. My parents, Aidé and Hugo, and my brother Hugo Antonio. They have given me their love and support all my life. I am thankful for everything, now and always.

My mini companions, Kitty and Fluffy. Kitty's constant warm presence will forever be etched in my memory and Fluffy's infinite energy keeps me going.

My adviser, Dr. Thomas Bureau, for his infinite patience and guidance in this whole endeavour.

Mr. Emilio Vello, his encouragement and advice in all things bioinformatic was invaluable.

Dr. Paul Harrison and Dr. Jaswinder Singh, my committee members, this would not have been possible without them.

Navei Cerda Hernández for all those bioinformatics days and her support.

Megan Letourneau for all her help, knowledge, and feedback.

Thank you. Thank you. Thank you.

And finally, thank you to CONACyT and McGill University for the funding that made this research possible.

PREFACE AND CONTRIBUTION OF AUTHORS

The dissertation for M.Sc. is composed of two chapters, I and II. I am the sole author.

Chapter I starts with an introduction and a literature review covering a comprehensive overview of transposable elements, the classification, detection methods, annotation, and the organism *Arabidopsis lyrata* and its importance as an organism of study.

Chapter II includes the methodology of the pipeline and the analysis of the transposable elements, the results, the discussion, and the concluding remarks of the research.

FIGURES AND TABLES

Figures

1. Pipeline for the detection of TEs.
2. Figure 2. Classification breakdown of the results of the pipeline steps: Run structural algorithm(A), Structural features for BLAT obtained sequences(B), Final library of intact TEs(C).
3. Locations of TEs in *Arabidopsis lyrata* obtained from the pipeline.
4. Locations of TEs in *Arabidopsis lyrata* obtained from the pipeline, for DNA: DTA/hAT(A), DTC/CACTA(B), DTH/PIF-Harbinger(C), DTT/Tc1-Mariner(D), DTM/Mutator(E), and LTR/COPIA/GYPSY(F).
5. TSD logo analysis for DTA/hAT(A), DTC/CACTA(B), DTM/Mutator(C), and LTR/COPIA/GYPSY(D).
6. TIR logo analysis for DTA/hAT(A), DTH/PIF-Harbinger(B), DTT/Tc1-Mariner(C).

Tables

1. Structurally conserved TEs obtained from step D, as shown in Figure1.
2. Location analysis of the TEs identified in the pipeline and the genes in *Arabidopsis lyrata*.

ABBREVIATIONS

A	Adenine
A.	*Arabidopsis*
BED/bed	Browser Extensible Data
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-Like Alignment Tool
C	Cytosine
DNA	Deoxyribonucleic Acid
DTA	hAT
DTC	CACTA
DTH	PIF-Harbinger
DTM	Mutator
DTT	Tc1-Mariner
DTX	DTA/DTC/DTH/DTT/DTM
EDTA	Extensive de novo TE Annotator
ET	Éléments Transposables
FASTA/fa	Fast All-purpose Scientific Sequence Analysis
G	Guanine

GB	Gigabyte
GFF/gff	General Feature Format
GRF	Generic Repeat Finder
LINE	Long Interspersed Nuclear Elements
LTR	Long Terminal Repeats
pblat	Parallelized BLAT
R	Purine
RAM	Random Access Memory
RNA	Ribonucleic Acid
SINE	Short Interspersed Nuclear Elements
T	Thymine
TB	Terabyte
TE	Transposable Elements
TIR	Terminal Inverted Repeats
TSD	Target Site Duplication
UNK/unk	Unknown

CHAPTER I

INTRODUCTION

Transposable elements can propagate in the genome and were once believed to be "junk DNA" for the organism (Dubin et al., 2018). However, studies have shown that transposable elements (TEs) play important roles in the systems they inhabit (Grotewold et al., 2015). They can be classified into Class I and Class II TEs, or most commonly known as retrotransposons or DNA transposons, respectively. Some of these TEs have structural features that make it possible to identify them by certain patterns in the sequences. (Wicker et al., 2007; Kapitonov & Jurka, 2008). TEs can have an effect on an organism when they are located in or nearby a gene, depending on the location of the insertion and on the type of TE. It can disrupt the expression or modify the response of a gene in multiple ways (Deneweth et al., 2022). However, research has been limited in non-model organisms since there is a lack of well-annotated TE libraries, which are essential to understand their role in biology. Currently, there's no method that can produce reliable results on the different classifications of TEs and a significant amount of manual curation is needed to annotate TEs in an organism (Goubert et al., 2022). However, it has been suggested that a combination of different methods and bioinformatic tools can be used to detect and classify TEs with more reliable results (Storer et al., 2022).

The aim of this project is to develop a bioinformatics pipeline created with a combination of open-source programs and independently created algorithms to accurately identify, classify, and annotate structurally conserved transposable elements in the non-model organism *Arabidopsis lyrata*. Ensure the pipeline is described with enough detail that researchers with basic programming knowledge will be able to use it, and even modify it to suit their research needs, organism of study, and the structural features for the different types of TEs. Finally, with the resulting list of curated TEs for *A. lyrata* from the pipeline dive into the TE characteristics, their

length, size, location, sequence characteristics, and place of insertion in relation to the genes of the organism.

LITERATURE REVIEW

Transposable Elements

Barbara McClintock first presented the term transposable element (TE) in 1947. It wasn't widely recognised in the scientific community despite several attempts. It wasn't until the late 1960s, after further research was done on gene structure and gene regulation that McClintock's research received the recognition it deserved (Kunze et al., 1997). She received the 1983 Physiology/Medicine Nobel Prize for her contributions. The recognition of this award gave way to further and expanded research in the field of TEs in both plants and animals (Grotewold et al., 2015).

Transposable elements can change positions and propagate in the genome and are also known as mobile DNA (Quadrana, 2020). At first, they were believed to be junk or 'selfish' DNA and were even considered parasitic for the organism (Makarevitch et al., 2015). However, more recent studies have shown that TEs can have a major impact on genome evolution (Quesneville, 2020). They can also have the ability to modify gene expression and create novel genes through their transposition in the genome, which can be viewed as a utility for genetic improvement in biotechnology, as well as a tool for unraveling basic physiological, biochemical, and genetic properties (Grotewold et al., 2015).

Even though TEs occur in all living organisms, they only accumulate in eukaryotes. In bacteria, they are removed by recombination. While that process also exists in eukaryotes, TEs

still persist in the genome due to the epigenetic mechanisms that are used in TEs, effectively silencing them (Grotewold et al., 2015). McClintock first suggested that the activity of TEs may be increased due to stressful environments. Plants have been exposed to environmental stress due to climate change, and there is evidence that the epigenetic controls are reversed (Cui & Cao, 2014).

TE effects in the organism

TEs are one of the many causes of genome instability in an organism (Bhat et al., 2022). While most insertions will have little to no effect on an organism, in some cases, the effects will be beneficial and in others, it will be detrimental. TE insertions can cause disruption in genes and regulatory regions, among others (Ramakrishnan et al., 2021). However, there are epigenetic mechanisms in place, particularly in plants, to silence the TE in the organism in order to maintain stability (Grotewold et al., 2015). These mechanisms can be DNA methylation, histone modifications, and other unknown pathways (Cui & Cao, 2014). Genome stability is essential for the transmission of genetic information and can ensure the fitness of an organism (Dion-Côté & Barbash, 2017). In contrast, genomic instability can also provide benefits to the organism, by aiding in gene regulation, improving genetic diversity, and accelerating evolution (Ramakrishnan et al., 2021).

Due to the mobile nature of transposable elements, they can be disruptive in the genome, either in a positive or a negative way, or they can have little to no effect at all. (Werren, 2011) The effect these TE insertions may have in the organism can depend on the place of insertion. The different types of TEs have shown certain preferences of insertions in specific regions in the genome, however, these can vary between each organism (Ramakrishnan et al., 2022). Areas of

genes or regulatory regions that are not functionally important, such as introns, where TE insertions occur may not have any impact on the organism. (Hirsch & Springer, 2017). These insertions can potentially be maintained in the organism since they don't negatively affect the organism, however, they could accumulate mutations over time (Werren, 2011). The insertions in or next to genes with important functions may have some negative effects in the organism, such as gene disruption or gene inactivation (Casacuberta & González, 2013). Plants use epigenetic processes to reduce and attenuate the deleterious activity of these TEs in the organism (Sahebi et al., 2018).

When TEs are inserted into regulatory areas and genes, TE insertions can occasionally be advantageous (Casacuberta & González, 2013). The organism can benefit from TEs in the following ways. 1) Domestication: a TE acquires a new function related to its original function, which is then passed down over time (Jangam et al., 2017). 2) Exaptation: a TE insertion modifies a gene to provide an entirely new function and confer an additional benefit to the organism. These first two terms are sometimes used interchangeably (Joly-Lopez & Bureau, 2018). 3) Host gene regulation is impacted by TE insertions close to regulatory areas, which may change how gene expression is regulated downstream (Schrader & Schmitz, 2019).

These insertions are an important area of study in TEs since they can provide advantages in the organism, particularly in the presence of biotic and abiotic stresses. In the presence of these pressures, plants can grow more resilient and endure harsh environments (Ito, 2022). Like *MUSTANG*, a domesticated TE that has shown to provide important contributions to reproduction, plant growth, and flower development in angiosperms (Joly-Lopez et al., 2012). It has shown to increase salt resistance in *Arabidopsis thaliana* (Joly-Lopez et al., 2017) and more recently in *Camelina sativa* (Shao, 2022). *A. thaliana* is the universal model organism for plants and the

reference model in the field of crop science (Krämer, 2015), while *Camelina sativa* is an important oil seed crop with the potential for food and feed due to its resistance to abiotic and biotic stresses and its short period of maturity (Murphy, 2016). Knowledge of TEs that provide important functions to plants in adverse conditions is especially important in agriculture with the already known and increasing challenges that are presented by climate change and the threat to food insecurity all over the world, specially in third world countries (Newton et al., 2011).

TE Classification

There have been efforts to create universal classifications for TEs, with slight differences between them that become more apparent further down in the classification. (Wicker et al., 2007; Kapitonov & Jurka, 2008). While there is still a need to ensure one universal benchmarking approach for the classification of TEs (Storer et al., 2022), it is the general consensus that they can be categorized into two main classes depending on how they execute the transposition and the type of repeat sequence they leave behind: retrotransposons and DNA transposons (Wicker et al., 2007; Kapitonov & Jurka, 2008).

TEs can also be considered autonomous or non-autonomous. Autonomous elements possess the machinery necessary to cause their own movement, while non-autonomous elements require the machinery of an autonomous element to mobilize them in the genome. Usually, non-autonomous elements were once autonomous, however, they lost part or all the internal sequences necessary to move around the genome (Grotewold et al., 2015). This difference in structure has challenged the two-class classification system. There have been considerations made to account for this, creating a subclass or an order for the non-autonomous elements (Wicker et al., 2007; Kapitonov & Jurka, 2008).

Retrotransposons, or Class I transposons, are transcribed into RNA and are then reverse transcribed to DNA to be inserted into the genome. These elements are mostly responsible for the increase in genome size since they use the "copy and paste" method (Quesneville, 2020). They are classified into two major groups according to the DNA sequences on their ends: long terminal repeats (LTRs) and non-LTRs. LTR retrotransposons are present in plants and animals, but much more so in plants, and they have been better structurally characterized. They have identical DNA sequences from hundreds of bp to several thousands (Grotewold et al., 2015).

There are two types of LTR retrotransposons that present identifiable structural features: i) gypsy elements are known for having 5′-TG-3′ and 5′-CA-3′ ends and ii) Copia elements present a 5′-TG-3′ and 5′-C/G/TA-3′ structure (Sahebi et al., 2018). And once they are inserted into the genome, they produce a target site duplication (TSD) of 4-6bp (Wicker et al., 2007). Non-LTR elements, on the other hand, don't have these identical repeated sequences. They are known as long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). These have been more widely studied in animals but can also occur in plants. LINEs are long enough to be autonomous retrotransposons, while SINEs are non-autonomous retrotransposons, they are less than 500 bp long. When they insert themselves into the DNA, they create small duplications in the genome (Grotewold et al., 2015).

DNA or Class II transposons cut themselves out of their positions and reinsert themselves somewhere else in the genome, also known as the "cut and paste" method (Grotewold et al., 2015). They are usually composed of short, inverted repeat sequences at their front and rear ends. In between, there is a sequence that encodes a transposase protein that recognizes the inverted repeats and cuts the transposon out of the genome. The transposase then holds the sequence together until it finds a new place in the genome to insert the transposon (Quesneville, 2020). This method makes

DNA TEs unlikely to have high copy numbers of the same TE and therefore, are also not large contributors to the genome size (Lee & Kim, 2014).

There are five main types of DNA transposons that present identifiable structural features: i) hAT (DTA) with a TSD of 8bp and terminal inverted repeats (TIRs) of 5-27bp, ii) CACTA (DTC) with a TSD of 2-3bp and a TIR of 12-28bp that presents with 5′- CACTA/G - 3′ at the start, iii) PIF-Harbinger (DTH) with a TSD of 3bp (TAA/TTA) and a TIR of variable length, iv) Mutator (DTM) with a TSD of 7-11bp and a TIR of variable length, and v) Tc1 Mariner (DTT) with a TSD of 2bp (TA) and a TIR of variable length (Sahebi et al., 2018).

Helitrons are a subclass of Class II transposons. They replicate using a rolling-circle mechanism and do not generate TSDs (Wicker et al., 2007). Some of these TEs present 5′-TC-3′ at the start of the sequence and 5′-CTRR-3′ (where R is a purine) at the end (Yang & Bennetzen, 2009). However, only some Helitrons present these structural characteristics in their sequence (Wicker et al., 2007).

TE Annotation

In order to study transposable elements, there needs to be a high-quality TE library of the organism of interest. Currently, most, of the high-quality TE libraries available are only for the model organisms. Non-model organisms have been largely ignored in the research of TEs since a high-quality TE library is essential in order to study them. Given that model organisms were some of the first genomes to be sequenced, they were, understandably, the first ones to be annotated using a combination of manual methods, and most recently, using computational tools (Hoen et al., 2015). However, new whole genome sequencing technologies are much faster, increasing access to whole genome sequencing for non-model organisms (Villanueva-Cañas et al., 2017).

The annotation of non-model organisms will be able to provide more information into the characteristics and the behaviour of TEs. Some initial annotation of TEs in non-model organisms have provided insight into TE features that haven't been previously found in model organisms. Much larger TEs than the largest previously identified TEs in model organisms have been identified in non-model organisms (Storer et al., 2022).

TE identification methods

There are three main methods that can be used for the discovery of TEs. Each one has its own benefits and downsides. The use of each one depends on the needs of the study and the overall aim of the research.

1) *De novo*: looks for repeated sequences on the DNA, without using any prior information, by either using similarity in sequences or structural features to other previously known TE sequences. This method is not specific to TEs and therefore can also find other kinds of repeats in the genome (satellites, tandem repeats, etc.) (Bergman & Quesneville, 2007). Also, this method makes it difficult to identify those TEs that have gone through selection and have been degraded in the organism (Hoen et al., 2015).

2) Homology-based: uses prior knowledge on previously detected TEs to detect TEs in the genome with the aid of alignment programs. It can help detect numerous copies of one TE in the genome and even TEs that only present one copy. This method relies on the quality of previously identified TEs, can only detect TEs that have been previously identified, and those that have been active enough that retain enough of their sequence. Also, homology-based methods don't detect the boundaries of the TE, therefore, further analysis is needed to obtain the full sequence (Arkhipova et al., 2017).

3) Structural-based: also uses previous knowledge of identified TEs, however, it focuses on the structural features shared by the different types of TEs. It can detect the boundaries, therefore, providing a full sequence. However, it can only detect the types of TEs that have structural characteristics and that have maintained them through selection and degradation in the organism (Bergman & Quesneville, 2007).

Different computational methods and bioinformatic tools have been used to develop pipelines to detect and classify TEs. They are tested against existing high-quality TE libraries of model organisms (Bergman & Quesneville, 2007). The following programs use different methods or a combination of them to detect different types of transposons. There are more programs besides the ones mentioned, however, these were the ones that had the best performance among programs intended for the same type of transposable element (Ou et al., 2019).

For LTR retrotransposons, LTR_FINDER is intended to detect full length LTR retrotransposons with the use of *de novo*, homology, and structural-based methods (Xu & Wang, 2007), LTR_retriever uses a combination of structural and homology-based methods (Ou & Jiang, 2018), and LTRharvest also uses a combination of homology and structural-based methods (Ellinghaus et al., 2008). For transposons with structural features with TSD and TIR, GRF (Generic Repeat Finder) uses structural and homology-based methods to classify them (Shi & Liang, 2019) and TIR-Learner also uses structural and homology-based methods to detect these types of transposable elements (Su et al., 2019).

HelitronScanner uses a combination of all three methods, *de novo*, homology and structural-based to detect helitrons in an organism (Xiong et al., 2014). Two well known programs that are used for different types of TEs, RepeatModeler which uses de novo, homology, and

structural-based methods (Flynn et al., 2020) and RepeatMasker which uses homology and structural-based methods to detect transposable elements (Tarailo-Graovac & Chen, 2009)

Despite the advances in TE identification and the development of new programs, there is currently no single program that can provide high-quality results across the different types of TEs (Orozco-Arias et al., 2019). Manual curation still remains the most trusted method to annotate TEs and to ensure a curated TE library for an organism (Storer et al., 2022). However, it is too time consuming, considering the speed at which genomes of new organisms are getting sequenced (Goubert et al., 2022). Since every method has limitations, high quality results in TE annotation can be produced by combining different independent computational methods, to strengthen the robustness and confidence of the results in order to obtain high-quality TE library annotation for newly sequenced organisms (Bennetzen et al., 2004).

Arabidopsis lyrata

Arabidopsis lyrata diverged from the model organism *Arabidopsis thaliana* 10 million years ago. They are both part of the *Brassicaceae* family. The genome sequence of *A. thaliana* is 125Mb, making it one of the smallest angiosperm genomes, whereas the genome size of *A. lyrata* is 207Mb, which is closer to the average size in the family (Hu et al., 2011). *A. lyrata* has shown to be an ideal species for the study of pathogens, flowering time, and to further the study of evolution in angiosperms (Shmickl et al., 2010).

Like in other organisms, TEs contribute in large part to the size of the genome. In the case of *A. lyrata*, with the use of whole-genome alignment against *A. thaliana*, it is estimated that TEs compose 29.7% of the genome, with around 80,225 TEs. In comparison, *A. thaliana*'s genome is made up of 23.7% of TEs or 26,990 TEs (Hu et al., 2011). Two major processes can cause an

increase in the size of a genome: proliferation of TEs and polyploidization (Hollister et al., 2011). A decrease in size in the genome, however, can be due to the loss of whole chromosomes and deletion-biased mutations. In this case, the difference in genome size for these two species can be attributed to the high TE activity in *A. lyrata* (Rutter et al., 2012) and the hundreds of thousands of small deletions throughout the genome in *A. thaliana*, mainly in TEs and noncoding regions. Given the higher number of TEs and the higher activity they present (Hu et al., 2011), *A. lyrata* is an ideal organism to study TEs.

Arabidopsis lyrata does not currently have a well annotated TE library the way *Arabidopsis thaliana* does. There have been some attempts to quantify the number of TEs in the organism, however there is no current accurate annotation of TEs for this species (Hu et al., 2011). Therefore, there is a need to detect, annotate, and classify TEs in *A. lyrata* to study the effects they have in the organism.

CHAPTER II

METHODOLOGY

The following pipeline is intended for the accurate identification, classification, and annotation of transposable elements of non-model organisms. It makes use of *de novo*, homology, and structural-based methods for the identification of TEs. It can be used and modified according to the focus of the research and the organism of study. This pipeline is created with the use of open-source software and independently created code in Shell and algorithms in Python.

Environment

There is a series of programs that are needed to set up the working environment in order to run the pipeline. Below is the list of the programs, the version used, a brief explanation of the program, and the installation commands that were used to install them. It is important to note that the installation commands and the links used were the ones that worked for the system used to originally develop this pipeline. The mode of installation and the links for the programs may need to be changed based on the system used and the specifications of each system. Older versions of the programs could work with the pipeline, however, that will depend on each program and how much time has passed since the older version was created. It is recommended to use the versions specified below.

- Miniconda – v4.12.0

  Miniconda is the minimal installer for conda, which is an open-source package management system. It includes multiple packages that can be useful, allows for easy installation of other packages and can switch between different environments. (Wratten, et al., 2021)

  ```
  $ wget https://repo.anaconda.com/miniconda/Miniconda3-py39_4.12.0-Linux-x86_64.sh
  $ bash Miniconda3-py39_4.12.0-Linux-x86_64.sh
  ```

```
$ conda config --set auto_activate_base false
```

- EDTA – v2.0.1

  Extensive de novo TE Annotator (EDTA) is a program that gathers eight existing programs that utilise *de novo*, structural and/or homology methods, each one specialized in a certain type of transposon. The results of each program are grouped, put through several filters, and classified to obtain a TE library (Ou et al., 2019).

```
$ git clone https://github.com/oushujun/EDTA.git
$ cd EDTA
$ conda env create -f EDTA.yml
```

- Samtools – v1.15.1

  Samtools is used to work with alignment data. It can filter, arrange, and extract data of interest in the format required for different programs. It can also convert alignment files and provide reference sequences (Danecek et al., 2021).

```
$ wget https://github.com/samtools/samtools/releases/download/ 1.15.1/
samtools-1.15.1.tar.bz2
$ tar -xvjf samtools-1.1.tar.bz2
$ cd samtools-1.15.1
$ make
```

- Python – v3.9.12

  It is an open-source programming language that is well suited for the use in the field of bioinformatics. It's easy to write, has a fast execution, and most importantly, it has a large community that contributes to the development of new algorithms. There are multiple applications available for use in the field of biology and it also allows for the use of common file formats in bioinformatics (Kinser, 2010).

Python is one of the programs installed with Miniconda. It is only necessary to activate the default environment called base. However, it can also be installed separately and used without Miniconda.

```
$ conda activate
$ (base) python3 program.py
$ conda deactivate
```

Biopython – v1.80

It is an open-source tool for bioinformatics created for Python. It has multiple valuable commands for biology related analysis, including reading biological files and manipulating their data (Cock et al., 2009). It needs to be installed prior to using Python.

```
$ pip install biopython
```

- Bedtools – v2.30.0

It is an open-source software that uses multiple tools for use in genome datasets for analysis. It allows to analyze multiple datasets or extract specific information with a variety of commands to manipulate genomic features (Quinlan, 2014).

```
$ wget https://github.com/arq5x/bedtools2/releases/download/v2.30.0/bedtools-2.30.0.tar.gz
$ tar -zxvf bedtools-2.30.0.tar.gz
$ cd bedtools2
$ make
```

- pblat – v2.5

BLAT (BLAST-Like Alignment Tool) is an alignment tool that is faster and able to put together sequences with large or multiple gaps in them, whereas BLAST (Basic Local Alignment Search Tool) will only put together sequences with one or two gaps (Kent, 2002). The program pblat (parallel blat) works the same way as BLAT, except it provides the option to use multiple threads to process multiple sequences at the same time to reduce the run time. (Wang & Kong, 2019)

```
$ wget https://github.com/icebert/pblat/tarball/master
$ tar -xzvf icebert-pblat-2.5-2-ge26bf6b.tar.gz
$ cd icebert-pblat-2.5-2-ge26bf6b
$ make
```

Each program needs to be added to PATH as it allows access to the programs from the command prompt. The same line at the beginning and at the end is to display the change before and after the program is added to PATH.

```
$ echo $PATH
$ export PATH=$PATH:~/path/to/program/
$ echo $PATH
```

Pipeline setup

The pipeline requires a few files that need to be obtained and arranged in directories. One Shell script, three Python scripts, and the genome of the organism of interest.

It's recommended to create a directory where to run the pipeline.

```
$ mkdir run_organism
$ cd run_organism
```

The pipeline needs one input file, the genome fasta file of the organism of interest. In this case, the genome of *Arabidopsis lyrata* is downloaded, unzipped, and stored in a directory called genome.

```
$ mkdir genome

$ cd genome

$ wget http://ftp.ensemblgenomes.org/pub/plants/release-54/fasta/
arabidopsis_lyrata/dna/Arabidopsis_lyrata.v.1.0.dna.toplevel.fa.gz

$ gunzip Arabidopsis_lyrata.v.1.0.dna.toplevel.fa.gz
```

It also requires three python scripts (`blatfeatures.py` `edtasectionsandfilter.py` `tsdtirfeatures.py`) in a directory called python that is inside a directory called scripts.

```
$ mkdir scripts

$ cd scripts

$ wget https://github.com/natalialplg/TE-
pipeline/tree/main/TE%20pipeline/scripts/python

$ cd ../
```

And finally, it requires one bash script, `run.sh`, in the same location where the genome and the scripts directories are located, `run_organism`.

```
$ wget https://github.com/natalialplg/TE-
pipeline/blob/main/TE%20pipeline/run.sh
```

A few things to consider before running the pipeline:

1) In the bash file, run.sh, in line 12, add the path in your system to the file to run EDTA, `EDTA.pl`, as seen in bolded section below.

```
$ perl ../../../../software/EDTA/EDTA.pl --genome ../../genome/*.fa --
anno 1 --threads 45
```

2) The file `edtasectionsandfilter.py`, assigns a code to each annotation in the input file from EDTA, in the format GS0TE10000. The first two letters can be changed, line 47, to match the initials of your organism of interest.
3) The specifications of the system used are Dell R910 server with 512 GB of RAM and two MD1200 storage devices 72 TB at McGill University. The number of threads need to be adjusted according to the capacity of each system.
4) The files required to run this pipeline are available in a public repository at `https://github.com/natalialplg/TE-pipeline`.

Once everything is set up, the pipeline is ready to run with the following command.

```
$ bash run.sh
```

Pipeline

The TE pipeline makes use of multiple open-source programs and several algorithms created by the author, three in Python and one in Shell. In Python: 1) to arrange and filter the results from EDTA, `edtasectionsandfilter.py`, 2) to detect structural features in potential TEs and classify them, `tsdtirfeatures.py`, 3) to detect structural features in the sequences obtained from pblat, `blatfeatures.py`. In Shell, `run.sh`. is used to sequentially run all the necessary steps and programs in the pipeline for the identification, classification, and annotation of TEs.

A. The *A. lyrata* genome runs first through the EDTA program for a list of prospective TEs (Figure1.A). The EDTA program has shown reliable results with a pre-existing list of TEs to add to the program, however, the program's accuracy decreases without the input of a curated list of TEs (Ou et al., 2019).

B. The scaffolds are removed, only the sequences in the chromosomes are of interest. The file is then run in `edtasectionsandfilter.py` where each data point from the EDTA file is separated in its own column, and if the line does not include one of the data points, the space in the column is replaced by a dash. Each annotation is also labeled with a code GS0TE00000, for convenient reference and look up between files. The sequences smaller than 100bp are filtered and put into a separate file. Then it removes sequences, that have the same start or end as another entry, only keeping the longer sequence. Helitrons are also removed, they have no set structural characteristics (TSD or TIR) (Wicker et al., 2007) and are, therefore, not the target of the pipeline (Figure1.B).

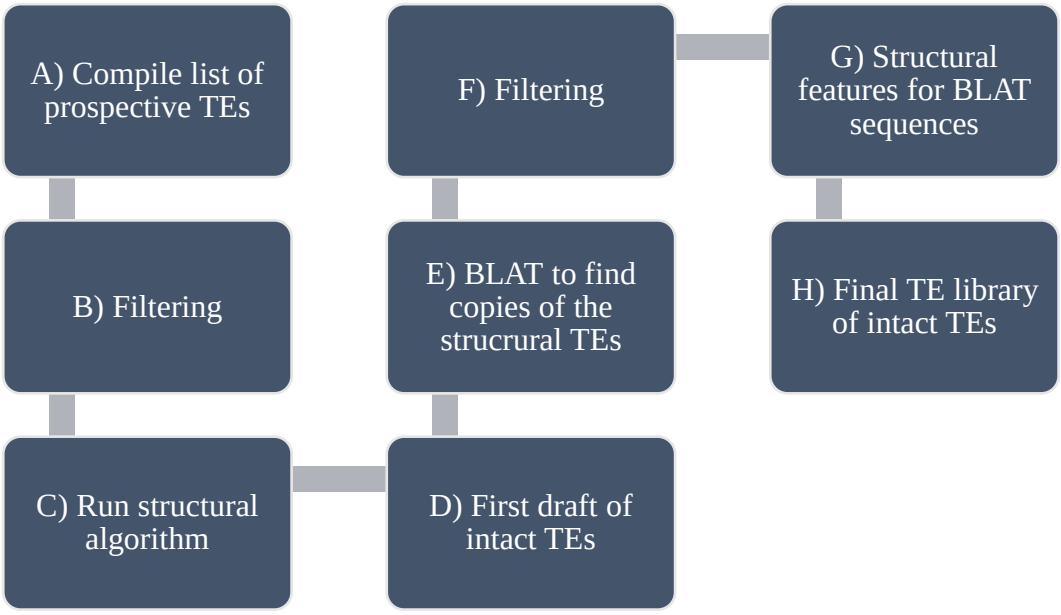

Figure 1. Pipeline for the detection of TEs.

C. The resulting sequences are increased by 40bp, 20bp on each side. Since sequences obtained by homology don't guarantee the exact boundaries (Arkhipova et al., 2017), this is done to obtain data on the TSD. The file is then run in the main algorithm, `tsdtirfeatures.py`, this step is in place of the manual curation needed to ensure the sequence is accurate (Platt et al., 2016). This independently created algorithm is used to

detect the patterns created by TSD and TIR, which will then be used to classify them based on the length and the data obtained. The program starts by finding repeated patterns, between the first 40bp and the last 40bp of the sequence. It then obtains the exact location of the potential TE that is found between the TSDs and extracts the sequence. The sequence is inversed and replaced by its complement bases (A-T, G-C) so that they can be compared and find the repeated sequence in both sides that would correspond to TIRs. Both features, TSD and TIR, are indicative of a TE sequence with structural characteristics and their length and sequence are used to classify them (Kapitonov & Jurka, 2008; Sahebi et al., 2018). The sequences that are unable to be accurately classified are labeled "unk" for unknown (Figure1.C). The program runs on iterations until it finds a potential TSD and a potential TIR that best fits the parameters for TEs with structural features. In order to ensure the reliability and consistency of the results in the algorithm, it can be run multiple times. The amount of run times can be modified, depending on the needs of the study, in line 52 in `run.sh`.

D. The location (chromosome:start-end), structural features, and classification of the resulting sequences are compared with the EDTA results. First, the sequences are matched by location. Then the structural features are compared, the match can be a full match or a partial one, since EDTA allows for an 80% match between TSDs and 85% between TIRs. The classification match first compares the classification (DNA or LTR), if it doesn't match, it is replaced with UNK/unk. However, if it does match, for LTR it adds LTR/COPIA/GYPSY. For DNA it compares the subclassification of hAT (DTA), CACTA (DTC), PIF-Harbinger (DTH), Mutator (DTM), and Tc1 Mariner (DTT). The three-letter classification is used for easy data manipulation and analysis, while the full common name

is used for easy user visualization. DNA subclassification matches are added as DNA/DTX/subclassification, the ones that don't match are reclassified as DNA/unk (Figure1.D). The resulting sequences are kept and added to the final curated TE library.

E. The fasta file of the sequences is obtained in order to run it through pblat (Figure1.E). This is done to find more instances of the TEs in the genome, with the method of homology (Platt et al., 2016).

F. The results are filtered with Shell script to remove the results in scaffolds. For *A. lyrata* this means removing all results from scaffold_9 through scaffold_1118. It also removes sequences smaller than 100bp, since none of the original query sequences were smaller than 100bp. It also removes the results that match the original query and any repeated sequences, keeping the one that has the highest match to the query, and sequences with the same start or end, keeping the longer sequence of the two (Figure1.F). The results are then ordered in descending order by chromosome, start values, and end values.

G. Then the structural features (TSD and TIR) are determined for the results obtained from BLAT with `blatfeatures.py` (Figure1.G). It is expected that the sequences obtained from BLAT will have similar characteristics from the sequence that they were obtained: The TSD will have the same length, but a different sequence, depending on the type of TE, since it will have been inserted in a different place in the genome. The algorithm, `blatfeatures.py`, will only accept a 100% match since the TSDs define the boundaries of the TE. The TIR should be similar, if not identical, to the TIR in the query sequence, depending on how selection and epigenetic mechanisms have affected the TE (Oliver & Greene, 2009). Once the boundaries are clearly set, the algorithm allows for an 80% match between the two TIRs. The resulting sequences are classified based on the length and

characteristics of the TSD and the TIR into the different types of TEs (Wicker et al., 2007; Kapitonov & Jurka, 2008).

H. The results from the structural characteristics from the BLAT sequences (6) and the list of curated TEs (4) are assembled into one list (Figure1.H). This curated list contains structurally intact TEs from the species *Arabidopsis lyrata*. The final output includes three files: 1) a GFF file with the annotations of the curated TEs, including the information on the classification, the TSD, the TIR and the length of the sequence, 2) a BED file with the locations, and 3) a FASTA file with the sequences.

1) `TEfinal.gff`

2) `TEfinal.bed`

3) `TEfinal.fa`

TE analysis

Rstudio is used to graph the percentage distribution on steps D, G, and H (Figure1) of the distinct types of TEs identified (Figure2). As well as the location of the TEs annotated with the pipeline, the overall distribution along the chromosomes (Figure3) and the distribution for each type of TE (Figure4).

*k*pLogo allows for the search of motifs with a logo visualization (Wu & Bartel, 2017). It was to look for sequence patterns in the TSDs (Figure5) and the TIRs of the different types of TEs (Figure6).

The genome of *A. thaliana* (`http://ftp.ensemblgenomes.org/pub/plants/release-54/fasta/arabidopsis_thaliana/dna/Arabidopsis_thaliana.TAIR10.dna.toplevel.fa.gz`) was used to test the TE pipeline and the annotated TE library of *A. thaliana*

(https://www.arabidopsis.org/download_files/Genes/TAIR10_genome_release/TAIR10_transposable_elements/TAIR10_Transposable_Elements.txt) to compare the results from the pipeline in *A. lyrata*.

RESULTS

Pipeline

Overall, the run time for EDTA with forty-two threads, is around 72 hours. It annotated 109,085 sequences of TEs in *Arabidopsis lyrata*. 35,964 LTR retrotransposons and 73,121 DNA transposons. Of those DNA TEs, 50,245 are helitrons and 22,876 are TEs with structural features (DTA/hAT, DTC/CACTA, DTH/PIF-Harbinger, DTT/Tc1-Mariner, DTM/Mutator). This list had multiple annotations with the same location, some with the same start or end, and some with a length of 4bp.

The run time for the second part of the pipeline next part of the pipeline until the final results was around 55 hours. The list needed to go through filtering to remove the sequences not of interest, with the aim of obtaining a list of prospective TEs. The 10,720 scaffolds were removed since the interest is in the eight chromosomes of *Arabidopsis lyrata*. Then, the filtering algorithm, `edtasectionsandfilter.py`, separated all the sequences with less than 100bp in length, 6,993, the helitrons were removed from the list, 44,397, and all sequences with the same start or end, 3,889. The list comes to a total of 43,086 TEs:

- 19,956 DNA
    - 2,936 DTA/hAT
    - 4,448 DTC/CACTA
    - 3,194 DTH/PIF-Harbinger
    - 2,372 DTT/Tc1-Mariner
    - 7,006 DTM/Mutator
- 23,130 LTR/COPIA/GYPSY

The total list of prospective TEs, 43,086, were run through the main algorithm, `tsdtirfeatures.py`, to determine the structural features, TSD and TIR, and to classify them. 26,544 sequences obtained structural features, that's over half of the prospective TEs, 61.6%.

The location (chromosome:start-end), structural characteristics and the classification are compared to the results from EDTA. The resulting matching sequences, 2,932, are now considered curated TEs (Figure2.A):

- 1,628 DNA
  - 616 DTA/hAT
  - 191 DTC/CACTA
  - 145 DTH/PIF-Harbinger
  - 49 DTT/Tc1-Mariner
  - 392 DTM/Mutator
  - 210 unk
- 1,094 LTR/COPIA/GYPSY
- 210 UNK/unk

A total of 82,315 sequences were obtained from pblat. After filtering, 13,630 sequences were submitted to the python program, `blatfeatures.py`, where 931 of the sequences had enough conserved features to be considered for the final TE library file (Figure2.B):

- 628 DNA
  - 98 DTA/hAT
  - 78 DTC/CACTA
  - 75 DTH/PIF-Harbinger

        - 59 DTT/Tc1-Mariner
        - 154 DTM/Mutator
        - 164 unk
    - 180 LTR/COPIA/GYPSY
    - 123 UNK/unk

Finally, the results from the main algorithm and the sequences obtained from pblat are joined for a total of 3,863 TEs from *Arabidopsis lyrata*(Figure2.C).

- 2,256 DNA
    - 714 DTA/hAT
    - 269 DTC/CACTA
    - 220 DTH/PIF-Harbinger
    - 108 DTT/Tc1-Mariner
    - 546 DTM/Mutator
    - 399 unk
- 1,274 LTR/COPIA/GYPSY
- 333 UNK/unk

The overall run time of the pipeline was around 127 hours, or 5.3 days.

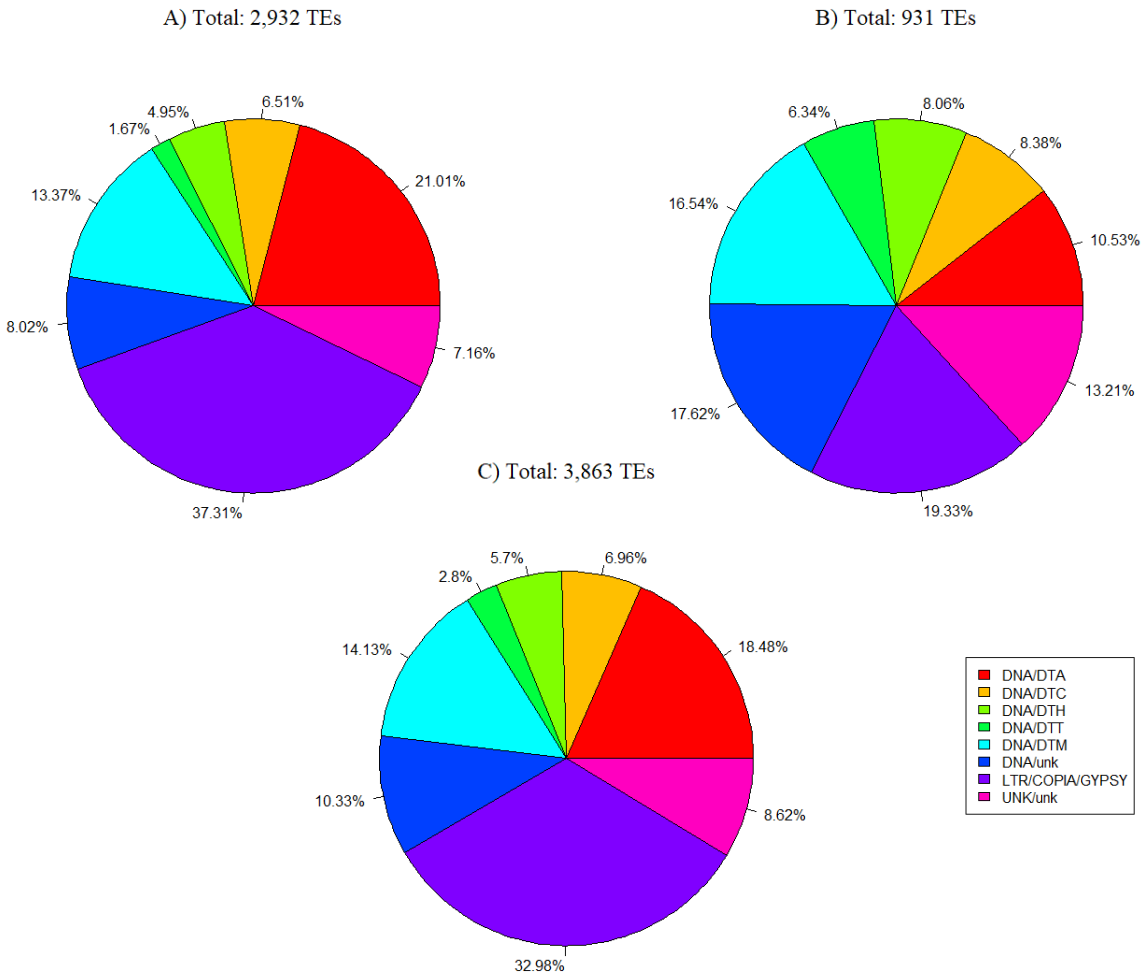

Figure 2. Classification breakdown of the results of the pipeline steps: Run structural algorithm(A), Structural features for BLAT obtained sequences(B), Final library of intact TEs(C).

TE characteristics

The TEs from the main algorithm (Figure1.D) are used to look at the characteristics in length, size, and how much they contribute to the genome size.

Classification	Subclassification	Number of TEs	Shortest TE (bp)	Longest TE (bp)	Average size of TE (bp)	Total length (bp)	% Of the genome
DNA	DTA/hAT	616	156	4,939	538	331,191	0.17%
	DTC/CACTA	191	300	5,015	2,924	558,533	0.29%
	DTH/PIF-Harbinger	145	150	4,888	1,252	181,476	0.09%
	DTT/Tc1-Mariner	49	136	4,955	1,886	92,418	0.05%
	DTM/Mutator	392	118	5,018	1,777	696,562	0.36%
	unk	235	102	4,808	1,180	277,394	0.14%
LTR	COPIA/GYPSY	1,094	114	17,729	6,374	6,973,241	3.59%
UNK	unk	210	114	13,708	1,616	339,316	0.17%
Total		2,932				9,450,131	4.87%

Table 1. Structurally conserved TEs obtained from step D, as shown in Figure1.

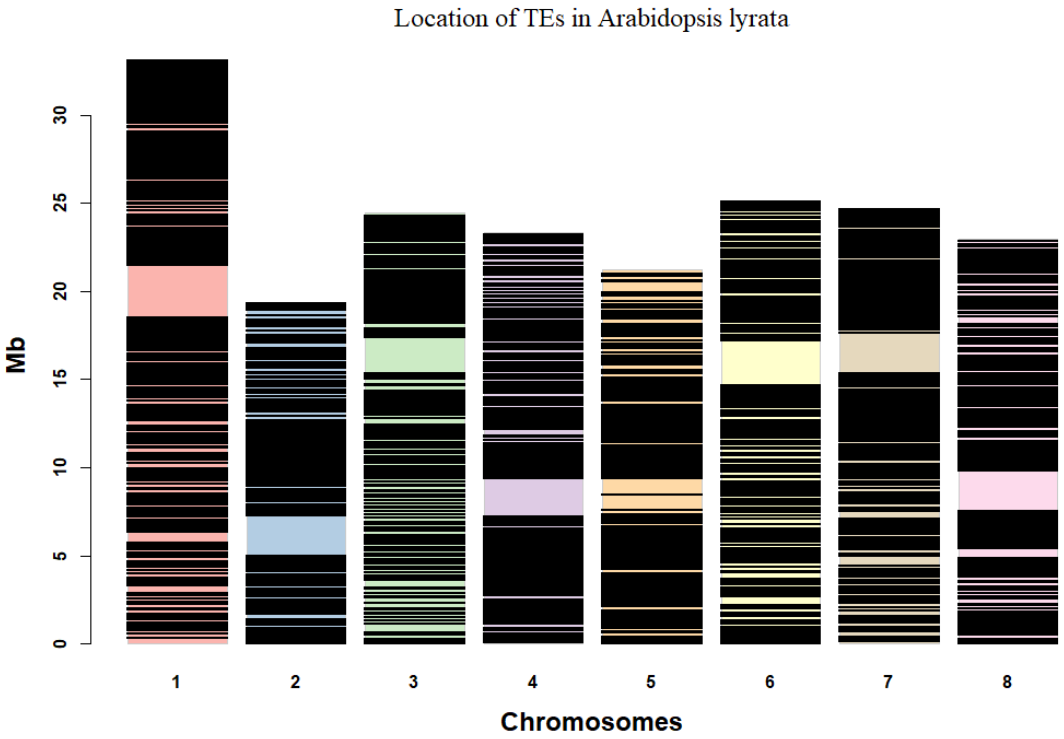

Figure 3. Locations of TEs in *Arabidopsis lyrata* obtained from the pipeline.

The TEs that were identified in the pipeline were found all over the genome (Figure3). Each bar in the graph represents each of the chromosomes for a total of 8. Each black line in the graph represents a TE identified in the pipeline. The small square gap seen on each of the bars is the

centromere for each of the chromosomes. The sequencing for which is located in the scaffolds (Hu et al., 2011) and were not included in the pipeline.

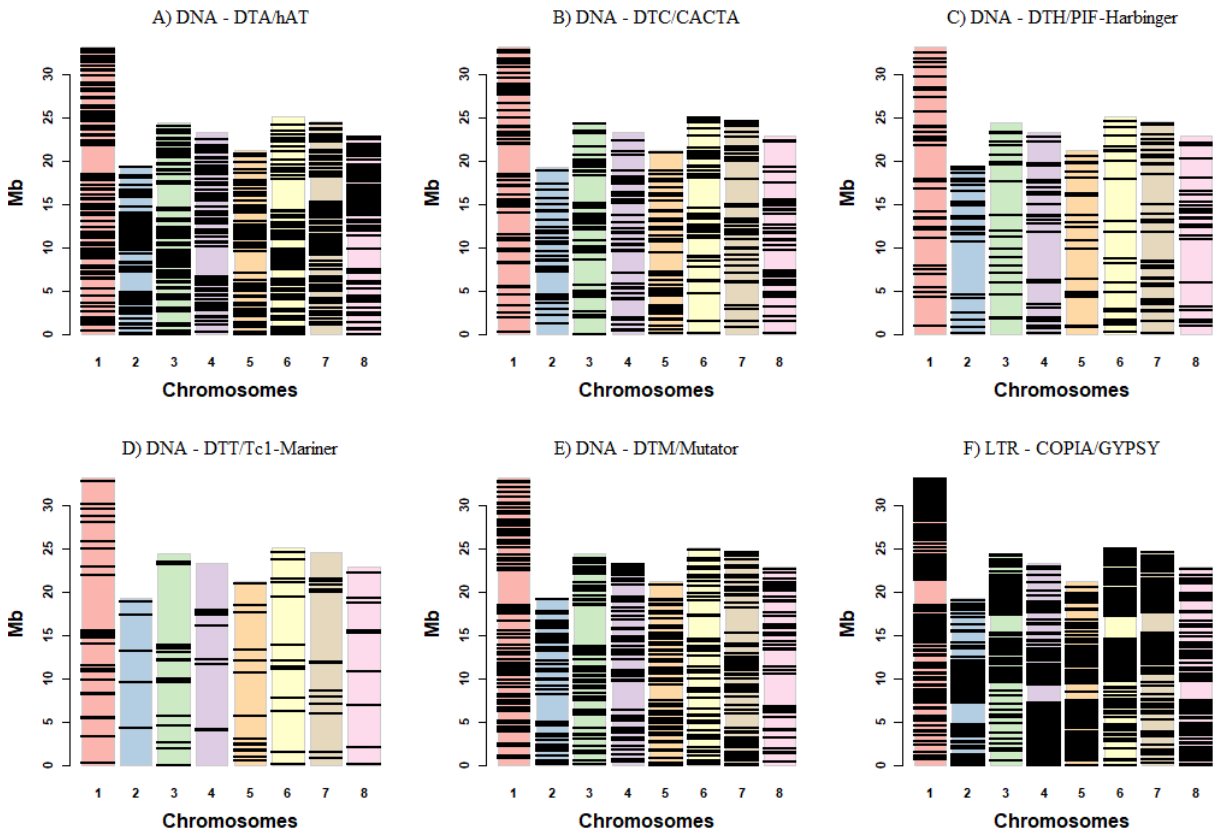

Figure 4. Locations of TEs in *Arabidopsis lyrata* obtained from the pipeline, for DNA: DTA/hAT(A), DTC/CACTA(B), DTH/PIF-Harbinger(C), DTT/Tc1-Mariner(D), DTM/Mutator(E), and LTR/COPIA/GYPSY(F).

The locations of the TEs were separated into the 5 DNA subclassifications and LTR for better visualization (Figure4).

TE structural features

The TSDs and the TIRs can sometimes present certain patterns, as seen on DTH/PIF-Harbinger and DTT/Tc1-Mariner for TSDs, and DTC/CACTA and COPIA/GYPSY for TIRs (Wicker et al., 2007; Kapitonov & Jurka, 2008). Therefore, a logo analysis was run on DTA/hAT, DTC/CACTA, DTM/Mutator, and LTR/COPIA/GYPSY for TSDs (Figure5) and DTA/hAT, DTH/PIF-Harbinger, DTT/Tc1-Mariner for TIRs (Figure6).

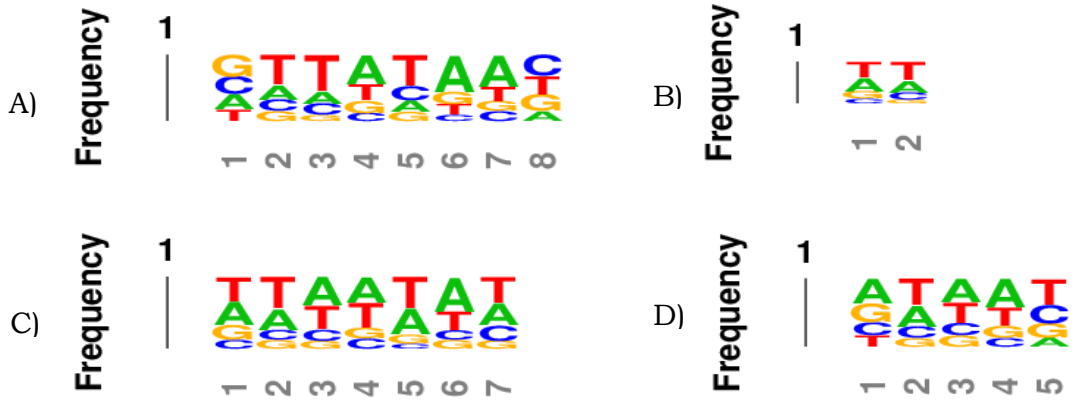

Figure 5. TSD logo analysis for DTA/hAT(A), DTC/CACTA(B), DTM/Mutator(C), and LTR/COPIA/GYPSY(D).

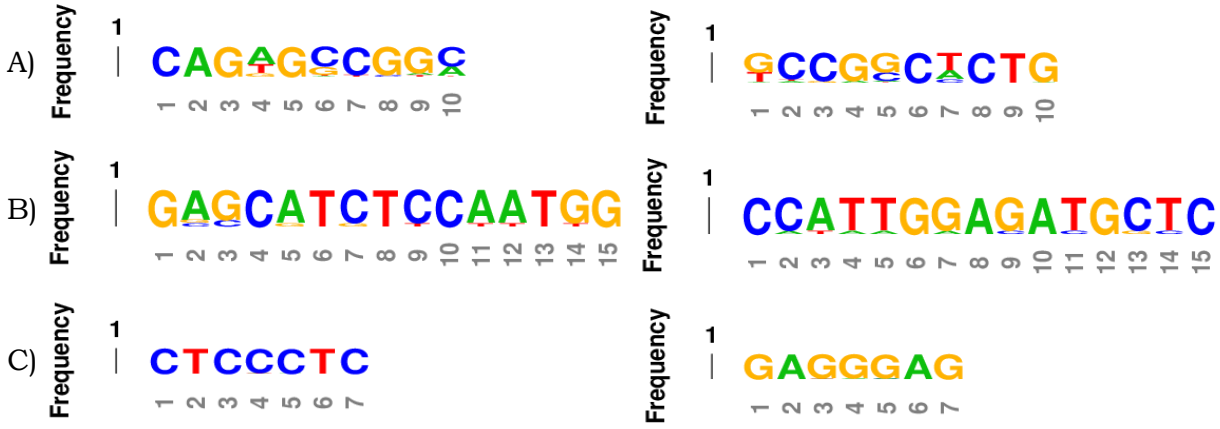

Figure 6. TIR logo analysis for DTA/hAT(A), DTH/PIF-Harbinger(B), DTT/Tc1-Mariner(C).

Impact of TE insertions on genes

The gff file of the pipeline, `TEfinal.gff`, was added to the gff file of *Arabidopsis lyrata* (`http://ftp.ensemblgenomes.org/pub/plants/release-51/gff3/arabidopsis_lyrata/Arabidopsis_lyrata.v.1.0.51.gff3.gz`) in order to locate their position in relation to the annotated genes in the organism when they are inserted in the 5' region of the gene (Casacuberta & González, 2013).

Classification	Subclassification	Inserted in one gene	Inserted near a gene (<2kb)	Inserted far from a gene (>=2000)
DNA	DTA/hAT	64	237	413
	DTC/CACTA	81	55	133
	DTH/PIF-Harbinger	31	58	131
	DTT/Tc1-Mariner	29	29	50
	DTM/Mutator	94	162	290
LTR	COPIA/GYPSY	81	380	813
Total		380	921	1,830

Table 2. Location analysis of the TEs identified in the pipeline and the genes in *Arabidopsis lyrata*.

The TE pipeline was run with the *A. thaliana* genome. The final results annotated a total of 612 TEs:

- 358 DNA
  - 70 DTA/hAT
  - 66 DTC/CACTA
  - 21 DTH/PIF-Harbinger
  - 8 DTT/Tc1-Mariner
  - 128 DTM/Mutator
  - 65 unk

- 217 LTR/COPIA/GYPSY
- 37 UNK/unk

DISCUSSION

TE Pipeline

As TEs evolve and go through selection, it becomes more difficult to characterize them for genome annotation. (Caspi & Patcher, 2006) Despite the challenges that TEs present in the first step for their study, their identification in an organism, they have proven to be an important area of study. They can provide various benefits to the organism, environmental adaptation, as well as valuable insight in the study of genome stability, evolution, and genetic expression (Grotewold et al., 2015). This pipeline is intended for TEs that have structural characteristics and that have gone through little to no selection in the genome. It's intended for the non-model organism *A. lyrata*, since the organism doesn't have a curated TE library since it is first necessary to annotate the TEs before being able to study them in the organism.

Regarding the main algorithm in the pipeline, it is expected that it will be unable to identify sequences from EDTA that have not conserved the structural features through selection or epigenetic mechanisms. The main objective of the pipeline is to obtain conserved TE sequences with intact structural features, which is why it is so restrictive. Only making sure that the sequences that have a 100% match in the TSD and the TIR for the main algorithm, `tsdtirfeatures.py`, and then a 100% match with the results obtained with EDTA will be included in the final results. Once these sequences are confirmed, the sequences obtained from blat are less restrictive for TIR, with a match of at least 80%. However, TSD match is still maintained at 100% since that is what defines the boundaries, and that is one of the benefits of TE identification with structural methods (Bergman & Quesneville, 2007).

The pipeline is set with the overall structural parameters from the information previously obtained from TEs (Wicker et al., 2007; Kapitonov & Jurka, 2008). However, it can be modified to suit different TE structural characteristics, different classification parameters, and different organisms (Kidwell, 2002). Most of the knowledge gathered for the structure of TEs has been from model organisms. As more TEs are identified and the knowledge in the field expands, the average parameters may need to be adjusted (Storer et al., 2022). This makes it necessary to adapt the broad parameters of the TE pipeline for other organisms for more accurate and inclusive identification of TEs. Also, parameters may differ between different organisms of study in case the pipeline needs to be adjusted for a specific one. Therefore, those parameters can be adjusted as well to fit the needs of the study.

The pipeline is useful for the detection, classification, and annotation of different types of structurally intact TEs. It uses several algorithms, in particular the main algorithm, `tsdtirfeatures.py`, is in place of the manual curation to ensure the accuracy of the potential TEs. Once the environment is set up it is easy to use, it only requires one file as the input, the genome of the organism of interest. The final annotation includes information on the classification with the three-letter classification (DTX) and the common name along with the structural features, TSD and TIRs, for each annotated TE. Finally, the results output three different files, gff, fasta, and bed, that have different aspects of the TE data. They are easily available for the multiple ways the data can be analysed.

Pipeline results

A. lyrata

The pipeline is intended for structurally conserved TEs. Therefore, it can only be used for TEs with structural features, LTR retrotransposons and DNA TEs (DTA/hAT, DTC/CACTA, DTH/PIF-Harbinger, DTT/Tc1-Mariner, and DTM/Mutator,). These TEs have very well-defined structural features that make it convenient to code for them. Of the 43,086 of the potential TEs obtained from EDTA, around 6% were characterized in the pipeline. The main algorithm showed better results with DNA transposons DTA/hAT and DTM/Mutator, which have the longest TSDs among the 6 types of TEs for the pipeline, 8bp and 9-11bp, respectively. From the total of the filtered EDTA results, 21% of hAT and 6% of Mutator TEs were accurately characterized. Compared with 4% for CACTA, 5% for PIF-Harbinger, and 2% for Tc1-Mariner. For LTR retrotransposons, 5% of them were able to be structurally characterized. For the TEs characterized from homology with the use of pblat, even though more TEs were found for DTM/Mutator and DTA/hAT, 154 and 98, respectively, considering the number of TEs that were in the input, DTT/Tc1-Mariner obtained the most, 59 from the 49 in the input, followed by DTT/PIF-Harbinger, 75 from the 145 in the input (Figure2.B).

The pipeline is the first step in the task of obtaining a curated TE library for non-model organism *A. lyrata*. It is able to obtain structurally conserved TEs with a good degree of confidence since it uses a combination of independently created programs and different algorithms (Bennetzen et al., 2004) and the subsequent analysis of the structural features for classification. The development of new algorithms that allows for the detection of less conserved sequences is needed, as well as programs for the detection of TEs with no structural features. However, this pipeline is

designed to be modified and adjusted as needed, including the addition of even more algorithms and methodologies.

There are some mismatches with classification between EDTA and the structural algorithm (Figure1.C), DNA/unk and UNK/unk. They will require further analysis into the content of the sequence to ensure the TEs are being accurately classified, and also to classify the LTRs into COPIA or GYPSY. Also, a further look into the content of each sequence, if they have the features necessary to transpose in the genome, is necessary to determine whether each sequence is autonomous or non-autonomous.

The results from the pipeline in *A. lyrata* were aligned with the genome of *A. thaliana*. With the use of pblat and an identity of 65%, the minimum threshold required for a TE sequence to be considered in the annotation of another genome (Quesneville, 2020). Of the 3,863 results in *A. lyrata*, 1,313 found at least one result in the *A. thaliana* genome, and 1,011 found at least two results with a minimum identity of 65%. The 2,550 of the TE sequences that did not find an alignment with a minimum threshold of 65% in *A. thaliana* is possible that have undergone significant changes over time to the extent that they become undetectable (Quadrana, 2020) or the TE sequences found in *A. lyrata* could indicate new instances of TEs, since *A. lyrata* has shown more activity in comparison with its closely related organism (Hu et al., 2011).

A. thaliana

The TE pipeline was run with the genome of *A. thaliana* in order to compare the results with its well curated TE library. The results from the TE pipeline showed that the number of annotated TEs is lower than the results from *A. lyrata*. Considering that the TEs in *A. thaliana* have undergone more selection and mutations and are, therefore, less structurally intact in

comparison (Hu et al., 2011). These results were compared with the curated TE library of *A. thaliana*. 444 TE sequences were found to have a total or partial match with sequences in the library, not all of the sequences annotated in the curated library have set structural boundaries. The other 168 TE sequences from the pipeline, were run in pblat with a minimum of 65% identity match with annotated sequences. 98 of the sequences had a partial match that could be potential copies of the TEs located somewhere else in the genome and could indicate activity in the TE (Lisch, 2013). However, that needs to be confirmed with additional methods. For the remaining 70 sequences, all of which were classified as DNA, manual curation confirmed that they all presented the structural features of the TSD and TIR, as annotated in the final file. Therefore, while they were not found in the curated TE library for *A. thaliana*, they could be potential TEs and further research into the sequence would be able to provide more information in their internal features (Wicker et al., 2007).

TE characteristics

The average length of each type of TE is 538 for DTA/hAT, 2,924 for DTC/CACTA, 1,252 for DTH/PIF-Harbinger, 1,886 for DTT/Tc1-Mariner, 1,777 for DTM/Mutator, and 6,374 for LTR/COPIA/GYPSY (Table1). As expected, the type with the longest length is LTR/COPIA/GYPSY, since they are, on average, the type with the longest sequences (Vitte & Panaud, 2005). It is estimated that 29.7% of the genome of *Arabidopsis lyrata* is made of TEs (Hu et al., 2011). Of the TEs that were accurately classified from the main algorithm, the DNA TEs represent 0.96% and the LTR retrotransposon TEs represent 3.59% of the genome in *Arabidopsis lyrata*, or an overall 4.55% (Table1). The TEs followed the overall size characteristics expected of each subclassification seen previously in *A. thaliana*. (Vitte & Panaud, 2005). Further annotation of TEs in *A. lyrata* will be able to provide more information to determine overall sizes in each

subclassification of this organism. TEs obtained from pblat are not included in this calculation since due to their length it is possible that they may have mutations and not all of the sequence belongs to the annotated TE, further sequence analysis is required for these sequences. The longest TEs of each TE is 16,040 for DTA/hAT, 396,586 for DTC/CACTA, 23,270 for DTH/PIF-Harbinger, 2,519,171 for DTT/Tc1-Mariner, 2,279,482 for DTM/Mutator, and 2,474,780 for LTR/COPIA/GYPSY, in comparison, the longest TEs from the main algorithm are much shorter (Table1).

TE structural features

According to the structural characteristics of the TEs obtained from the pipeline, the most common length for the TSD for the LTR retrotransposons is 5, 97.3%, while only 2% had a length of 3, 0.6% had a length of 4, and 0.1% had a length of 6. For DNA TEs, the most common TSD length for DTC/CACTA is 3, with 83%, and 17% had a length of 2. For DTM/Mutator, the most common TSD length is 10, with 40%, while 30% had a length of 7, 29% had a length of 9, and 1% a length of 11. The use of *k*pLogo program provided some insight into the most common sequence in the TSD and the TIR for the types of TEs without a set sequence. It first filtered by size, since the TSDs ranged from 3-11bp and the TIRs ranged from 2-150bp. For the TEs without an established sequence in the TSD, DTA/hAT, DTC/CACTA, DTM/Mutator, and LTR/COPIA/GYPSY, it showed that T and A are the most common bases, even though all four bases are present in the TSD. Preferences for T and A sequences in TSD in TEs have previously been found for different types of TEs (Le et al., 2000).

In the sequence logo for the TIRs, DTA/hAT showed the most common length is 10bp with some distinctive bases most commonly found in the TIRs (Figure6.A), for DTH/PIF-

Harbinger the most common length is 15bp and also showed very distinctive bases for multiple positions in the TIRs (Figure6.C). For DTT/Tc1-Mariner, the most common length found was 7bp and all seven of them showed very distinctive bases for each position (Figure6.C). A preliminary analysis on the DTT/Tc1-Mariner in the TE library of *Arabidopsis thaliana* showed that around 20% of them had this same sequence that presented in the DTT/Tc1-Mariner TEs identified in *A. lyrata*. Further identification and annotation of TEs is needed, along with an analysis on the sequences of the TSD and the TIRs, in order to gain more insight into the structural features for certain types of TEs. It is possible that the TEs discovered with the same sequence in the TIRs (Figure6), DTA/hAT, DTH/PIF-Harbinger, and DTT/Tc1-Mariner, have other sequences in their TIRs, however, the ones identified with the pipeline may be the ones that work better with the main algorithm.

TE location

The locations of the TEs identified in *A. lyrata* along the chromosomes, do not seem to show an insertion preference. All TEs seem distributed along the entire chromosome, except for the centromeres, which were not included in the analysis (Figure4). The different types of TEs are also distributed along the entire length of the chromosomes, with some clustering in certain areas (Figure5). However, in order to determine insertion patterns for each type of TE, further TE identification is needed, as well as further research in the characterisation of the regions of the chromosomes.

The location of the TEs obtained with the pipeline was also analysed in relation to the annotated genes of *A. lyrata*. Of the 3,131 of the TEs that were accurately classified, 1,830 were located farther than 2kb from the gene, 921 were located less than 2kb from the gene, and 380

were located inserted in one gene (Table2). Around 58% of the TEs were located farther than 2kb from the nearest gene. Interestingly, around 30% of the DTC/CACTA TEs were located inserted in a TE, the most out of all of the types of TEs. In *A. thaliana*, each type of TE has shown insertion preferences in the chromosomes (Underwood et al., 2017), and while closely related organisms could have similar patterns, it is necessary to first further research this organism. Also, the insertion of a TE near or in a gene, can have beneficial, detrimental, or no effects at all to the organism. More in depth study of the effects of these genes that have a TE insertion could be of interest to determine the effect they may have. In particular, it would be of interest to determine the genes which have a beneficial effect from the insertion of a TE.

CONCLUDING REMARKS

The study of transposable elements is a fundamental part of understanding the genome, the evolution, and the effects in any organism. The annotation of TEs is a necessary first step in this field of research. This pipeline manages to annotate TEs in the organism *Arabidopsis lyrata*. It uses a combination of open-source programs and independently created algorithms, in conjunction with multiple bioinformatic tools. The pipeline is described in enough detail that researchers with a basic understanding of programing can use the pipeline for their own research of interest and it is designed to be modified and improved to include a wider range of TEs. It can only detect conserved TEs with structural features, TEs without structural features or TEs that have gone through selection will need to be annotated with different methods. Even so, the TE pipeline was able to provide a list of 3,863 structurally conserved TEs for *A. lyrata*. The list gave way for the analysis of TEs in the organism, the length, size, sequences in the structural features, location in relation to the chromosome, and insertion in relation to the genes. The research addresses the lack of a pipeline for different types of TEs and furthers the research into the annotation of TEs and the characterisation of TEs in non-model organism *A. lyrata*.

REFERENCES

Arkhipova, I. R. (2017). Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mobile DNA*, *8*(1), 19. https://doi.org/10.1186/s13100-017-0103-2

Bennetzen, J. L., Coleman, C., Liu, R., Ma, J., & Ramakrishna, W. (2004). Consistent over-estimation of gene number in complex plant genomes. *Current Opinion in Plant Biology*, *7*(6), 732–736. https://doi.org/10.1016/j.pbi.2004.09.003

Bergman, C. M., & Quesneville, H. (2007). Discovering and detecting transposable elements in genome sequences. *Briefings in Bioinformatics*, *8*(6), 382–392. https://doi.org/10.1093/bib/bbm048

Bhat, A., Ghatage, T., Bhan, S., Lahane, G. P., Dhar, A., Kumar, R., Pandita, R. K., Bhat, K. M., Ramos, K. S., & Pandita, T. K. (2022). Role of Transposable Elements in Genome Stability: Implications for Health and Disease. *International Journal of Molecular Sciences*, *23*(14), Article 14. https://doi.org/10.3390/ijms23147802

Casacuberta, E., & González, J. (2013). The impact of transposable elements in environmental adaptation. *Molecular Ecology*, *22*(6), 1503–1517. https://doi.org/10.1111/mec.12170

Caspi, A., & Pachter, L. (2006). Identification of transposable elements using multiple alignments of related genomes. *Genome Research*, *16*(2), 260–270. https://doi.org/10.1101/gr.4361206

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: Freely

available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Cui, X., & Cao, X. (2014). Epigenetic regulation and functional exaptation of transposable elements in higher plants. *Current Opinion in Plant Biology*, *21*, 83–88. https://doi.org/10.1016/j.pbi.2014.07.001

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Deneweth, J., Van de Peer, Y., & Vermeirssen, V. (2022). Nearby transposable elements impact plant stress gene regulatory networks: A meta-analysis in A. thaliana and S. lycopersicum. *BMC Genomics*, *23*(1), 18. https://doi.org/10.1186/s12864-021-08215-8

Dion-Côté, A.-M., & Barbash, D. A. (2017). Beyond speciation genes: An overview of genome stability in evolution and speciation. *Current Opinion in Genetics & Development*, *47*, 17–23. https://doi.org/10.1016/j.gde.2017.07.014

Dubin, M. J., Mittelsten Scheid, O., & Becker, C. (2018). Transposons: A blessing curse. *Current Opinion in Plant Biology*, *42*, 23–29. https://doi.org/10.1016/j.pbi.2018.01.003

Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, *9*(1), 18. https://doi.org/10.1186/1471-2105-9-18

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element

families. *Proceedings of the National Academy of Sciences*, *117*(17), 9451–9457. https://doi.org/10.1073/pnas.1921046117

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

Grotewold, E., Chappell, J., & Kellogg, E. A. (2015). *Plant genes, genomes, and genetics*. John Wiley & Sons Inc.

Hirsch, C. D., & Springer, N. M. (2017). Transposable element influences on gene expression in plants. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1860*(1), 157–165. https://doi.org/10.1016/j.bbagrm.2016.05.010

Hoen, D. R., Hickey, G., Bourque, G., Casacuberta, J., Cordaux, R., Feschotte, C., Fiston-Lavier, A.-S., Hua-Van, A., Hubley, R., Kapusta, A., Lerat, E., Maumus, F., Pollock, D. D., Quesneville, H., Smit, A., Wheeler, T. J., Bureau, T. E., & Blanchette, M. (2015). A call for benchmarking transposable element annotation methods. *Mobile DNA*, *6*(1), 13. https://doi.org/10.1186/s13100-015-0044-6

Hollister, J. D., Smith, L. M., Guo, Y.-L., Ott, F., Weigel, D., & Gaut, B. S. (2011). Transposable elements and small RNAs contribute to gene expression divergence between Arabidopsis thaliana and Arabidopsis lyrata. *Proceedings of the National Academy of Sciences*, *108*(6), 2322–2327. https://doi.org/10.1073/pnas.1018222108

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J.-F., Clark, R. M., Fahlgren, N., Fawcett, J. A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J. D., Ossowski, S., Ottilar, R. P., Salamov, A. A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., … Guo, Y.-L.

(2011). The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature Genetics*, *43*(5), Article 5. https://doi.org/10.1038/ng.807

Ito, H. (2022). Environmental stress and transposons in plants. *Genes & Genetic Systems*, *97*(4), 169–175. https://doi.org/10.1266/ggs.22-00045

Jangam, D., Feschotte, C., & Betrán, E. (2017). Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. *Trends in Genetics*, *33*(11), 817–831. https://doi.org/10.1016/j.tig.2017.07.011

Joly-Lopez, Z., & Bureau, T. E. (2018). Exaptation of transposable element coding sequences. *Current Opinion in Genetics & Development*, *49*, 34–42. https://doi.org/10.1016/j.gde.2018.02.011

Joly-Lopez, Z., Forczek, E., Hoen, D. R., Juretic, N., & Bureau, T. E. (2012). A Gene Family Derived from Transposable Elements during Early Angiosperm Evolution Has Reproductive Fitness Benefits in Arabidopsis thaliana. *PLOS Genetics*, *8*(9), e1002931. https://doi.org/10.1371/journal.pgen.1002931

Joly-Lopez, Z., Forczek, E., Vello, E., Hoen, D. R., Tomita, A., & Bureau, T. E. (2017). Abiotic Stress Phenotypes Are Associated with Conserved Genes Derived from Transposable Elements. *Frontiers in Plant Science*, *8*. https://www.frontiersin.org/articles/10.3389/fpls.2017.02027

Kapitonov, V. V., & Jurka, J. (2008). A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature Reviews Genetics*, *9*(5), Article 5. https://doi.org/10.1038/nrg2165-c1

Kent, W. J. (2002). BLAT—The BLAST-Like Alignment Tool. *Genome Research*, *12*(4), 656–664. https://doi.org/10.1101/gr.229202

Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, *115*(1), 49–63. https://doi.org/10.1023/A:1016072014259

Kinser, J. (2010). *Python for Bioinformatics*. Jones & Bartlett Publishers.

Krämer, U. (2015). Planting molecular functions in an ecological context with Arabidopsis thaliana. *ELife*, *4*, e06100. https://doi.org/10.7554/eLife.06100

Kunze, R., Saedler, H., & Lönnig, W.-E. (1997). Plant Transposable Elements. In J. A. Callow (Ed.), *Advances in Botanical Research* (Vol. 27, pp. 331–470). Academic Press. https://doi.org/10.1016/S0065-2296(08)60284-0

Le, Q. H., Wright, S., Yu, Z., & Bureau, T. (2000). Transposon diversity in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, *97*(13), 7376–7381. https://doi.org/10.1073/pnas.97.13.7376

Lee, S.-I., & Kim, N.-S. (2014). Transposable Elements and Genome Size Variations in Plants. *Genomics & Informatics*, *12*(3), 87–97. https://doi.org/10.5808/GI.2014.12.3.87

Lisch, D. (2013). How important are transposons for plant evolution? *Nature Reviews Genetics*, *14*(1), Article 1. https://doi.org/10.1038/nrg3374

Makarevitch, I., Waters, A. J., West, P. T., Stitzer, M., Hirsch, C. N., Ross-Ibarra, J., & Springer, N. M. (2015). Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress. *PLOS Genetics*, *11*(1), e1004915. https://doi.org/10.1371/journal.pgen.1004915

Murphy, E. J. (2016). Chapter 8—Camelina (Camelina sativa). In T. A. McKeon, D. G. Hayes, D. F. Hildebrand, & R. J. Weselake (Eds.), *Industrial Oil Crops* (pp. 207–230). AOCS Press. https://doi.org/10.1016/B978-1-893997-98-1.00008-7

Newton, A. C., Johnson, S. N., & Gregory, P. J. (2011). Implications of climate change for diseases, crop yields and food security. *Euphytica*, *179*(1), 3–18. https://doi.org/10.1007/s10681-011-0359-4

Oliver, K. R., & Greene, W. K. (2009). Transposable elements: Powerful facilitators of evolution. *BioEssays*, *31*(7), 703–714. https://doi.org/10.1002/bies.200800219

Orozco-Arias, S., Isaza, G., Guyot, R., & Tabares-Soto, R. (2019). A systematic review of the application of machine learning in the detection and classification of transposable elements. *PeerJ*, *7*, e8311. https://doi.org/10.7717/peerj.8311

Ou, S., & Jiang, N. (2018). LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiology*, *176*(2), 1410–1422. https://doi.org/10.1104/pp.17.01310

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R. A., Hellinga, A. J., Lugo, C. S. B., Elliott, T. A., Ware, D., Peterson, T., Jiang, N., Hirsch, C. N., & Hufford, M. B. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology*, *20*, 275. https://doi.org/10.1186/s13059-019-1905-y

Platt, R. N., Blanco-Berdugo, L., & Ray, D. A. (2016). Accurate Transposable Element Annotation Is Vital When Analyzing New Genome Assemblies. *Genome Biology and Evolution*, *8*(2), 403–410. https://doi.org/10.1093/gbe/evw009

Quadrana, L. (2020). The contribution of transposable elements to transcriptional novelty in plants: The FLC affair. *Transcription*, *11*(3–4), 192–198. https://doi.org/10.1080/21541264.2020.1803031

Quesneville, H. (2020). Twenty years of transposable element analysis in the Arabidopsis thaliana genome. *Mobile DNA*, *11*(1), 28. https://doi.org/10.1186/s13100-020-00223-x

Quinlan, A. R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Current Protocols in Bioinformatics*, *47*(1), 11.12.1-11.12.34. https://doi.org/10.1002/0471250953.bi1112s47

Ramakrishnan, M., Satish, L., Kalendar, R., Narayanan, M., Kandasamy, S., Sharma, A., Emamverdian, A., Wei, Q., & Zhou, M. (2021). The Dynamism of Transposon Methylation for Plant Development and Stress Adaptation. *International Journal of Molecular Sciences*, *22*(21), Article 21. https://doi.org/10.3390/ijms222111387

Ramakrishnan, M., Satish, L., Sharma, A., Kurungara Vinod, K., Emamverdian, A., Zhou, M., & Wei, Q. (2022). Transposable elements in plants: Recent advancements, tools and prospects. *Plant Molecular Biology Reporter*, *40*(4), 628–645. https://doi.org/10.1007/s11105-022-01342-w

Rutter, M. T., Cross, K. V., & Van Woert, P. A. (2012). Birth, death and subfunctionalization in the Arabidopsis genome. *Trends in Plant Science*, *17*(4), 204–212. https://doi.org/10.1016/j.tplants.2012.01.006

Sahebi, M., Hanafi, M. M., van Wijnen, A. J., Rice, D., Rafii, M. Y., Azizi, P., Osman, M., Taheri, S., Bakar, M. F. A., Isa, M. N. M., & Noor, Y. M. (2018). Contribution of

transposable elements in the plant's genome. *Gene*, *665*, 155–166. https://doi.org/10.1016/j.gene.2018.04.050

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

Shao, Y. (n.d.). *Functional characterization and application of exapted transposable element gene family MUSTANG-A*. McGill University. Retrieved 10 February 2023, from https://escholarship.mcgill.ca/concern/theses/0v8385592

Shi, J., & Liang, C. (2019). Generic Repeat Finder: A High-Sensitivity Tool for Genome-Wide De Novo Repeat Detection. *Plant Physiology*, *180*(4), 1803–1815. https://doi.org/10.1104/pp.19.00386

Storer, J. M., Hubley, R., Rosen, J., & Smit, A. F. A. (2022). Methodologies for the De novo Discovery of Transposable Element Families. *Genes*, *13*(4), Article 4. https://doi.org/10.3390/genes13040709

Su, W., Gu, X., & Peterson, T. (2019). TIR-Learner, a New Ensemble Method for TIR Transposable Element Annotation, Provides Evidence for Abundant New Transposable Elements in the Maize Genome. *Molecular Plant*, *12*(3), 447–460. https://doi.org/10.1016/j.molp.2019.02.008

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Current Protocols in Bioinformatics*, *25*(1), 4.10.1-4.10.14. https://doi.org/10.1002/0471250953.bi0410s25

Underwood, C. J., Henderson, I. R., & Martienssen, R. A. (2017). Genetic and epigenetic variation of transposable elements in Arabidopsis. *Current Opinion in Plant Biology*, *36*, 135–141. https://doi.org/10.1016/j.pbi.2017.03.002

Villanueva-Cañas, J. L., Rech, G. E., de Cara, M. A. R., & González, J. (2017). Beyond SNPs: How to detect selection on transposable element insertions. *Methods in Ecology and Evolution*, *8*(6), 728–737. https://doi.org/10.1111/2041-210X.12781

Vitte, C., & Panaud, O. (2005). LTR retrotransposons and flowering plant genome size: Emergence of the increase/decrease model. *Cytogenetic and Genome Research*, *110*(1–4), 91–107. https://doi.org/10.1159/000084941

Wang, M., & Kong, L. (2019). pblat: A multithread blat algorithm speeding up aligning sequences to genomes. *BMC Bioinformatics*, *20*(1), 28. https://doi.org/10.1186/s12859-019-2597-8

Werren, J. H. (2011). Selfish genetic elements, genetic conflict, and evolutionary innovation. *Proceedings of the National Academy of Sciences*, *108*(supplement_2), 10863–10870. https://doi.org/10.1073/pnas.1102343108

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, *8*(12), Article 12. https://doi.org/10.1038/nrg2165

Wratten, L., Wilm, A., & Göke, J. (2021). Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature Methods*, *18*(10), Article 10. https://doi.org/10.1038/s41592-021-01254-9

Wu, X., & Bartel, D. P. (2017). kpLogo: Positional k-mer analysis reveals hidden specificity in biological sequences. *Nucleic Acids Research*, *45*(Web Server issue), W534–W538. https://doi.org/10.1093/nar/gkx323

Xiong, W., He, L., Lai, J., Dooner, H. K., & Du, C. (2014). HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proceedings of the National Academy of Sciences*, *111*(28), 10263–10268. https://doi.org/10.1073/pnas.1410068111

Xu, Z., & Wang, H. (2007). LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, *35*(suppl_2), W265–W268. https://doi.org/10.1093/nar/gkm286

Yang, L., & Bennetzen, J. L. (2009). Structure-based discovery and description of plant and animal Helitrons. *Proceedings of the National Academy of Sciences*, *106*(31), 12832–12837. https://doi.org/10.1073/pnas.0905563106