Observations of Transposable Element Richness and Diversity in Embryophytes

Andrew Lindsay[1]

[1] I.K. Barber School of Arts & Sciences, University of British Columbia, Okanagan Campus, Kelowna, BC V1V 1V7, Canada

Abstract

Transposable elements (TEs) are ubiquitous among organisms, and have shown to have significant effect on gene expression and evolution. To best identify these elements, a comparison of existing libraries was performed, as well as a test of software that generates novel TEs. To characterize patterns in both inter and intra-species relationships involving TEs, we analyzed 33 genomes of Embryophytes, and 18 accessions of *Arabidopsis thaliana*. We found relationships between genome size and TE richness, and a similar pattern with genome size and LTR elements. Evidence of TE horizontal gene transfer (HGT) events was observed with the presence of several TE subfamilies unique to a small subset of species, as well as a potential TE extinction event. For Arabidopsis, a potential flaw in sequencing methodology may have been discovered to reduce the divergence of repetitive sequences. Taken together, the results of this study demonstrate the hidden relationships and activities between TEs and the host genome both within a species, and across the evolutionary tree.

Key words: Transposable element, genome size, DNA diversity, horizontal gene transfer

Introduction

Transposable elements (TEs) are DNA sequences that can move and potentially replicate within a genome. These elements are ubiquitous in living organisms, and make-up nearly half of the human genome, but can reach as high as 90% some organisms, such as maize (Mills et al., 2007; SanMiguel et al., 1996). Furthermore, these figures likely underestimate the total contribution of TEs to genome, as many ancient TEs may be unrecognizable due to accumulated mutations (Forterre, Filée and Myllykallio, 2004). For many decades after their discovery in the 1940's by Barbara McClintock, TEs were thought to be 'junk' DNA with no functional role in the host genome (Pennisi, 2012). Some even considered them to be parasitic, as their persistence and proliferation in the genome can disrupt the host genome (Orgel and Crick, 1980). However, recent research has shown that some TEs have roles in gene regulation (Wang et al., 2013), rearrangement (Xuan et al., 2016) and recombination (Hallet, 1997), as well as acting as mutagens to disrupt gene function (Martienssen, 1998), or generate new functionality (Leprince et al., 2001), even from very early embryogenesis (Ge, 2016).

There are two classes of TEs – Class 1 and Class 2 – based on their mechanism of transposition. Retrotransposons make up Class 1, and are characterized by using an RNA intermediate to facilitate replication and movement in the genome (often referred to as a 'copy-and-paste mechanism'). This gave rise to the original notion of TEs being 'parasitic' (Orgel and Crick, 1980). Retrotransposons can be classified into LTR and non-LTR retrotransposons. LTR and non-LTR retrotransposons. LTR retrotransposons are characterized by the presence of Long Terminal Repeats (LTRs) flanking the coding region of the transposon. *Ty3/Gypsy* and *Ty1/Copia* LTRs are typically the dominate TEs in plant

genomes (Friesen, 2001). Non-LTR transposons are divided into *L*ong *IN*terspersed *E*lements (LINEs) and *S*hort *IN*terspersed *E*lements (SINEs). LINEs are autonomous – they contain all genes necessary to transpose – while SINEs have lost these mechanisms. However, they can still transpose if an autonomous element is present by hijacking its transposition proteins.

Class 2 transposons differ from Class 1 because they lack and RNA intermediate during their transposition process. This is known as a 'cut-and-paste' mechanism because the TE is typically excised from the genome, and leaving little trace. It can then reinsert itself elsewhere in the genome. Replication of the TE occurs during nuclear replication, if the TE excised itself after being copied, and re-inserted downstream of the replication fork.

The presence of TEs can have a drastic effect of genome function, and there is strong evidence that they may influence the evolutionary trajectory of the species (Galindo-González et al., 2017). Population processes ultimately determine the fate of the TE. Insertion and excision of TEs can affect allelic diversity, which is a key metric for determining potential for adaptation to future environmental changes (Caballero, 2013). Location of insertion and removal can drastically change the phenotype of the individual, but is not limited to coding sequences. Promoter regions, introns and untranslated regions are all insertion targets that can have effects ranging from no change, to subtle regulatory changes, to the creation of a novel protein via alternative splicing, and to the complete loss of function for the gene (Kidwell & Lisch, 2007).

Despite these benefits, if TE replication is left unchecked, they may overwhelm the genome, and reduce the fitness of the host. Once a TE exceeds a reasonable copy number, it will elicit a host response to reduce the transposition activity to a manageable level. This will inactivate the TE, removing its transposition activity. These methods are effective, and most genomes show a minimal level of TE activity (< 0.05% in humans; Mill et al., 2007). Methylation and phosphorylation are common methods to reduce TE activity (McDonald et al., 2005). Silencing pathways also exist, are often under control of epigenetic mechanisms, and can lead to sudden changes in transposition activity – Stresses can include infection, injury, hunger, or even temperature (Lisch, 2009). This could be particularly important for speciation events.

Transposable elements remain a difficult field of research. They compose a significant portion of most genomes, and evidence has shown they evolve over time and react to their environment, as well as work in unison with the host to optimize survival (Zhao et al., 2016). Proliferation and mutation can cause rapid evolutionary changes in the host, but later deletion events may leave no trace of this. We know TEs are a significant evolutionary force, and able to create new novel genes (Kaessmann, 2009). With a wide breadth of sequencing data available, we can take the opportunity to perform a systematic comparative analysis of TE richness and diversity across many different plant species. As well, we can look more in-depth at a group of individuals from a single species. These searches are not limited to existing TE libraries; tools exist to examine genomes for novel elements. In this study, we used this data to uncover patterns in TE richness and content that can help our understanding of potential genomic factors that can lead to speciation and diversity among plants.

Materials and Methods

Genomic Datasets

Full genomes for Embryophyte species were collected from Phytozome: Colorado Blue Columbine (*Aquilegia coerulea*), Thale Cress (*Arabidopsis thaliana*), *Arabidopsis halleri*, Stiff Brome (*Brachypodium distachyon*), Drummond's rockcress (*Boechera stricta*), Clementine (*Citrus clementina*), Grand Shepherd's-Purse (*Capsella grandiflora*), Pink Shepherd's-Purse (*Capsella rubella*), Rose Gum (*Eucalyptus grandis*), Saltwater Cress (*Eutrema salsugineum*), Wild Strawberry (*Fragaria vesca*), Soybean (*Glycine max*), Peruvian Cotton (*Gossypium raimondii*), Milky Widow's Thrill (*Kalanchoe laxiflora*), Flax (*Linum usitatissimum*), Common Liverwort (*Marchantia polymorpha*), Asian Rice (*Oryza sativa*), Switchgrass (*Panicum virgatum*), Peach (*Prunus persica*), Castor Bean (*Ricinus communis*), Foxtail Millet (*Setaria italica*), Garden Tomato (*Solanum lycopersicum*), Potato (*Solanum tuberosum*), Greater Duckweed (*Spirodela polyrhiza*), Red Clover (*Trifolium pratense*), Common Grape Vine (*Vitis vinifera*), Common Eelgrass (*Zostera marina*), Maize (*Zea mays*) (https://phytozome.jgi.doe.gov/pz/portal.html, V12, last accessed June 15, 2017). Further genomes were collected from NCBI: Hot Pepper (*Capsicum annuum*), Quinoa (*Chenopodium quinoa*), Tree Cotton (*Gossypium arboreum*) and Mexican Cotton (*Gossypium hirsutum*) (ftp://ftp.ncbi.nih.gov/genomes/, last accessed July 4, 2017) and Comparative Genomics: Floating Bladderwort (*Utricularia gibba*) (https://genomevolution.org/wiki/index.php/GenomeInfo, last accessed June 5, 2017). All genomes were complete, containing minimal ambiguous bases or masking (< 0.1% of bases are N or X). Calculations for genomes sizes were taken from count of GCAT bases.

For Arabidopsis, further genomes for 18 individuals were obtained from the Wellcome Trust Center for Human Genetics (http://mtweb.cs.ucl.ac.uk/mus/www/19genomes/, last accessed June 5, 2017). Specimens were collected predominantly from European countries: Germany (No-0, Po-0, Wu-0, Zu-0), Ireland (Bur-0), Italy (Ct-1), Lithuania (Kn-0), Netherlands (Hi-0), Norway (Oy-0), Poland (Ler-0), Russia (Rsch-4, Wil-2, Ws-0), Scotland (Edi-0), Spain (Can-0, Sf-2). Asian and African specimens were also collected: Japan (Tsu-0), Libya (Mt-0). There was no data available on exact location, so Russian samples will be considered European.

Transposable Element Libraries

Libraries of TEs were obtained from the Genetic Information Research Institute (Repbase and Repbase-derived Repeat Masker library; http://www.girinst.org/server/RepBase/index.php, last accessed June 14, 2017) and the Plant Genome and Systems Biology Group (PGSB-REdat; ftp://ftpmips.helmholtz-muenchen.de/plants/REdat/, last accessed May 20, 2017). Repbase libraries were used exclusively with the parameter "-species viridiplantae". For Arabidopsis, a species specific library was obtained from The Arabidopsis Information Resource (https://www.arabidopsis.org/download/, v10, last accessed June 5, 2017). A merged library was also produced using FASTA sequences for Repbase, instead of the RepeatMasker-specific library, and included only those from Viridiplantae.

Novel TE libraries were generated by RepeatModeler (http://www.repeatmasker.org/RepeatModeler/, v1.0.10, last accessed May 18, 2017) and REpeat Detector

(RED; https://omictools.com/red-tool, last accessed July 15, 2017). RepeatModeler was run using the NCBI engine. Library creation via RED ran the following pipeline: TE detection with RED, a reverse BLAST search with NCBI blastn to return sequences with no hit in the Repbase database, then masking of low complexity and simple repeats with RepeatMasker (http://www.repeatmasker.org/RMDownload.html, v4.0.7, last accessed May 18, 2017), removal of sequences with > 50% masked bases via a custom script. Processing of formats was performed using bedtools (http://bedtools.readthedocs.io/en/latest/, v2.26.0, last accessed July 20, 2017) and samtools (http://www.htslib.org/, v1.5, last accessed July 20, 2017).

Potential novel repeats generated by RepeatModeler and Red were used to mask a reference Arabidopsis genome. RepeatModeler was also used with *O.sativa*, *S.italica* and *U.gibba*, but these genomes were not processed further. Known repeats were removed from the RepeatModeler TEs by a reverse BLAST against Repbase, and the remaining sequences were used to mask a reference Arabidopsis genome. RED novel TEs were also tested as input for RepeatModeler instead of a reference genome for analysis as a potential pre-processing step.

Genome Masking and Analysis

All masking of genomes was performed with RepeatMasker using the '-qq' parameter for rush jobs (10% reduced sensitivity). Repbase is the default library used with RepeatMasker, and works natively. Non-Repbase specific libraries were loading with the '-lib' parameter, and alignment files (.align) were generated with the '-a' parameter. TE richness was generated from RepeatMasker output file (.out), and processed with parseRM scripts (https://github.com/4ureliek/Parsing-RepeatMasker-Outputs, last accessed June 15, 2017).

Diversion from consensus sequences was calculated from Kimura distance as generated in RepeatMasker alignment files. Kimura distance is defined as: $K = -1/2\ ln(1-2p-q) - sqrt(1-2q)$, where q and p are the proportion of sites with transversions and transitions, respectively (Kimura,1980). The alignment files were parsed into a repeat landscape using parseRM.

Phylogeny Construction of Viridiplantae

Phylogenetic tree construction for Viridiplantae was generated with phyloT (http://phylot.biobyte.de/, last accessed August 5, 2017). The dendrogram is based on the recognized phylogeny from NCBI. Generating the phylogenetic tree based on TE content was done via RStudio, using hclust() with a complete-linkage hierarchical clustering. Dendroscope was used to compare the two dendrograms (http://dendroscope.org/, last accessed August 1, 2017).

Results

Comparison of TE Libraries

Three different libraries were used during this project – Repbase, REdat, and TAIR. Each was tested against a reference *A. thaliana* genome (Fig. 1). TAIR and Repbase showed similar execution time, but TAIR masked a higher proportion of TEs. REdat showed a significantly longer execution time, but was inferior to TAIR in masking. The merged library showed the highest masking, but a poor execution time.

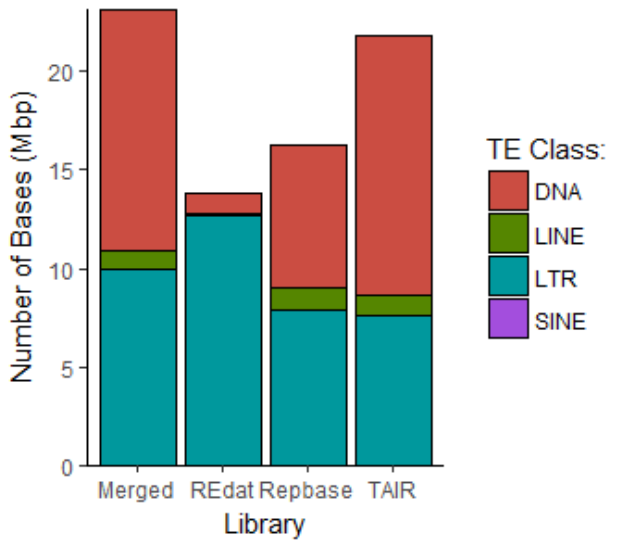

Figure 1. Length of respective TE class as masked against *A.thaliana* using various libraries.

When comparing TE richness of masked sequences, there was a wide difference in the families matched. REdat was particularly sensitive for LTR sequences, but was drastically inferior to the other libraries for other TEs. When looking at subfamilies, Repbase was the most diverse, with about 2.5x the subfamilies as TAIR, REdat and the merged library (Fig. 2).

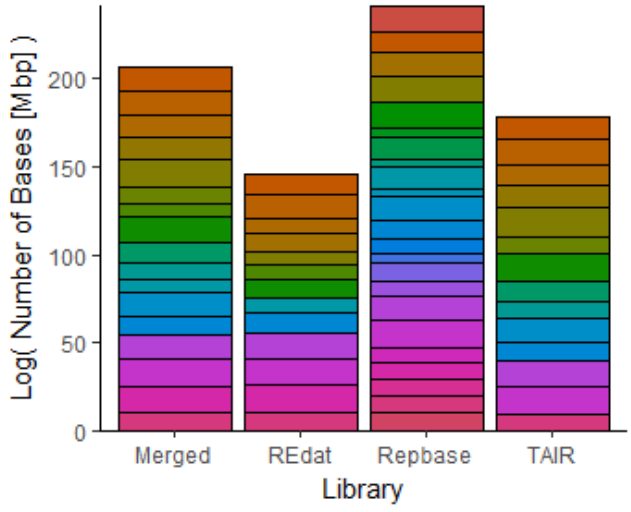

Figure 2. Log length of TE subfamilies. A total of 27 subfamily were found in the Arabidopsis genome. The number above each bar indicates the number of subfamilies detected with that library. Each color represents a single subfamily.

Based on these results, Repbase was chosen due to low search times and better subfamily data. As well, it is the native library for RepeatMasker, and would allow for use of command line arguments specific to the library.

Generation of Novel TE Libraries

RED returned 48.2 Mbp of novel sequences in approximately 12 minutes. After reverse BLASTing, the collection was reduced to 34.9 Mbp. Sequences with > 50% low complexity or simple repeats were removed, leaving 33.7 Mbp. This was too numerous to analyze, as it is still 28.8% of the genome. The resulting sequence was instead used as a preprocessing filter for RepeatModeler. This provided a slight time increase, but had a noticeable drop in sensitivity (Table 1, Fig. 3).

RepeatModeler was used on four different genomes – *A. thaliana*, *O. sativa*, *S. italica* and *U. gibba*. Time of computation was considerable (Table 1). Further analysis was only performed on *A.thaliana*.

Table 1. Run-time for RepeatModeler with Embryophyte genomes

Species	Input Library	Genome Size (Mbp)	Run-time (HHH:MM)
A.thaliana	Repbase	117	198:16
A.thaliana	RED pre-processed	38	154:52
O.sativa	Repbase	374	306:32
U.gibba	Repbase	68	76:38
S.italica	Repbase	407	314:28

For *A.thaliana*, RepeatModeler yielded 25,828 sequences equal to 13.0% of the genome (Fig. 3). Due to an unexpectedly high masking percentage, the sequences were reverse BLASTed against Repbase, which reduced the number of novel sequences to 5.35% of the genome, or 17,972 sequences.

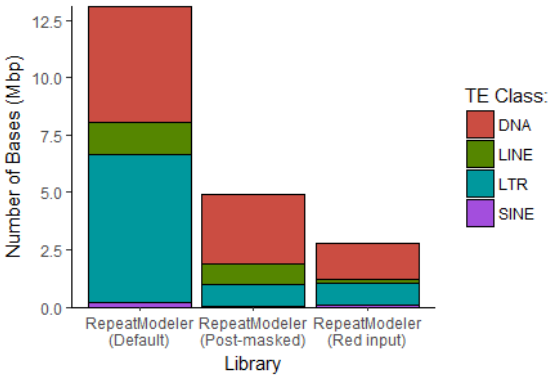

Figure 3. Length of TE classes for masking with RepeatModeler libraries. 'Default' is the RepeatModeler output using an Arabidopsis input. 'Post-masked' is the results of the 'Default' RepeatModeler output, but with sequences matching Repbase removed. 'RED input' is RepeatModeler output using the RED pre-processed Arabidopsis genome.

Arabidopsis TE Diversity

The genome size and total TE length was consistent between all 18 accessions. Genome size varied by only 0.83% and total TE length varied by 1.03%. DNA transposons, LTRs, and LINE retrotransposon showed relatively consistent proportions of the total genome. SINE retrotransposon length showed a high variation (20.1%), but occurred with a low copy number (< 200). The mean TE richness in the genome was 16.3%.

TE richness for class did not show significant difference (Fig. 5). The maximum difference was 1% between Wil-2 and Hi-0. Subfamily binned content showed homogenous richness, except fodr Sf-2 and Ler-0 with LINE/LINE (Fig. 6). These are a result of rounding, rather than significant difference.

Based on Kimura distances, Arabidopsis shows little deviation from consensus sequences (Fig. 7; Fig. 8). There was an exponential decrease in the TE proportion as Kimura Distance increased. For distances > 3, TE richness was less than 0.5%. There were no significant outliers. There is no obvious difference in Kimura distance for different geographical locations. Samples from the same country showed little difference in divergence compared to accessions from other countries.

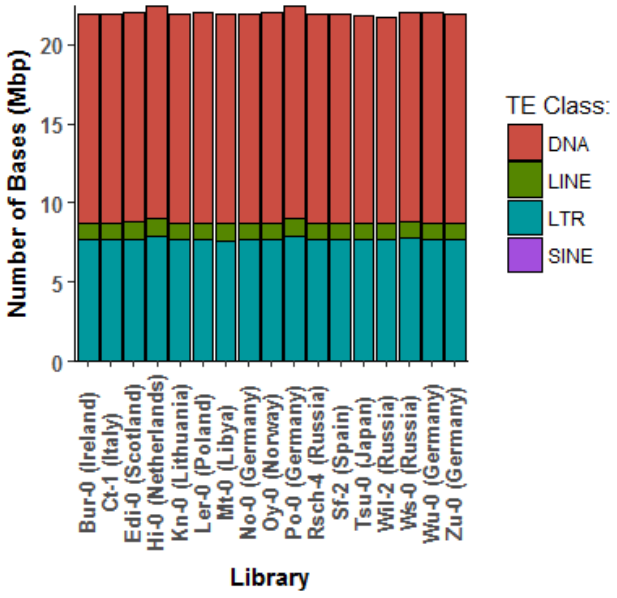

Figure 5. TE class richness across *A.thaliana* samples. SINEs were not shown due to low total base length (< 10 kbp).

Diversity of TE Richness in Viridiplantae

Genomes from 33 different species across 29 genera were analyzed (Fig. 9). Genome size drastically varied between species (*U. gibba* – 81 Mbp; *C. annuum* – 2.84 Gbp), and even between species in the same genus (*Gossypium raimondii* – 748 Mbp; *Gossypium hirsutum* – 2.26 Gbp). TE richness was highly variable between species, ranging from 1.8% in *M. polymorpha* to 78% in *Z. mays*. Close genera can show similar richness (*S. lycopersicum* and *S. tuberosum* with 37 and 40%, respectively), while other show a vast difference (*A. halleri* and *A. thaliana* with 7% and 13%).

There seems to be a strong, positive correlation between genome size and TE richness (Fig. 10). This statistically supported via Pearson's Correlation (t = 10.389, P << 0.0001). This correlation is not as strong when looking at closely related species. Members of

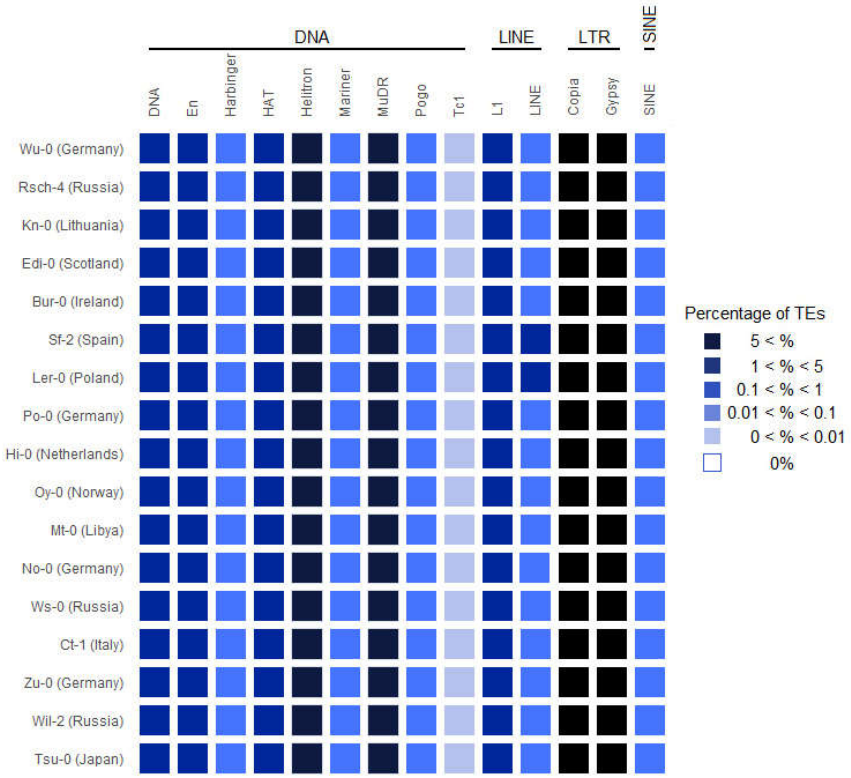

Figure 6. Subfamily richness for *A. thaliana* accessions. TE subfamilies content was determined as a percentage of the genome, using TAIR as the masking library.

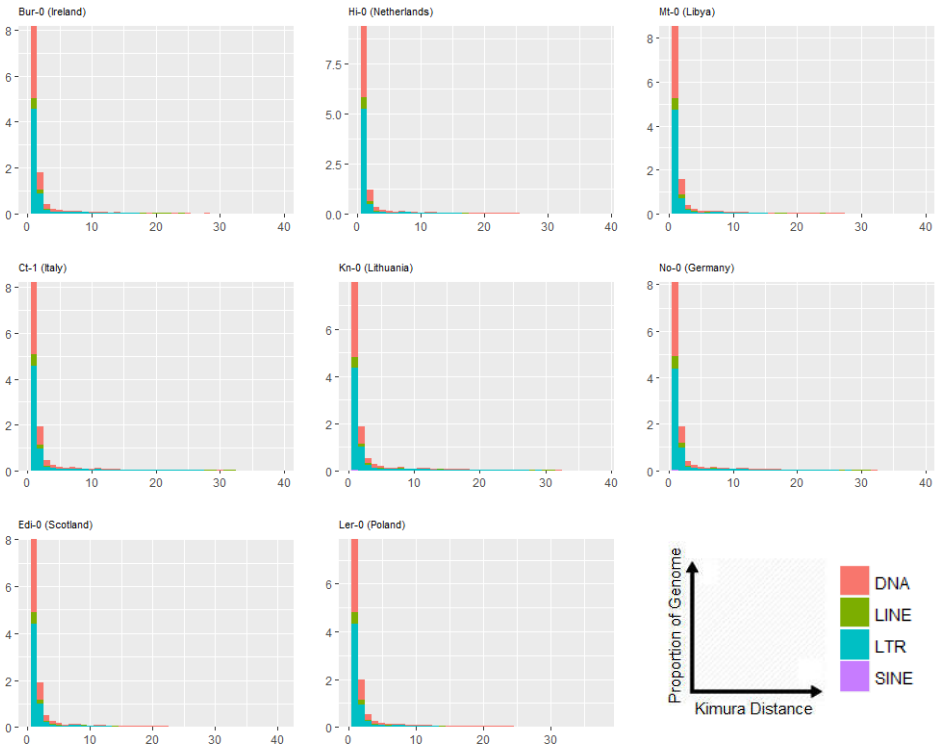

Figure 7. Kimura distance plots of *A. thaliana* accessions against Repbase consensus sequences as a proportion of the genome. The remaining 10 accessions showed no significant difference from these plots.

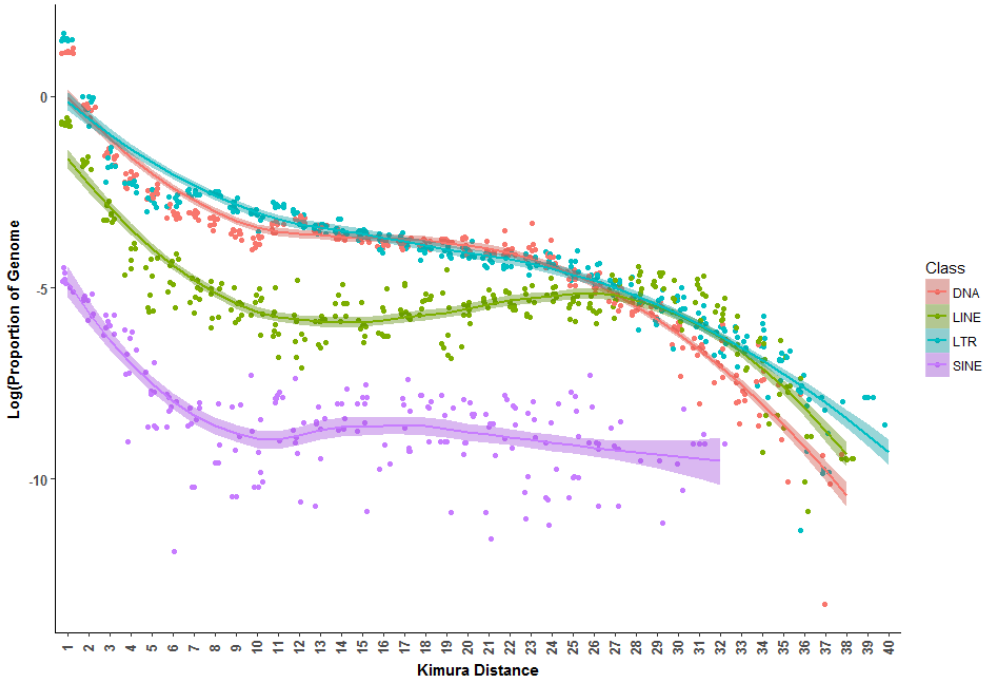

Figure 8. Scatterplot of Kimura distance of 18 *A. thaliana* accessions to Repbase consensus sequences as a proportion of the genome. Data points are jittered to improve visibility. Confidence intervals were generated by RStudio using Loess smoothing.

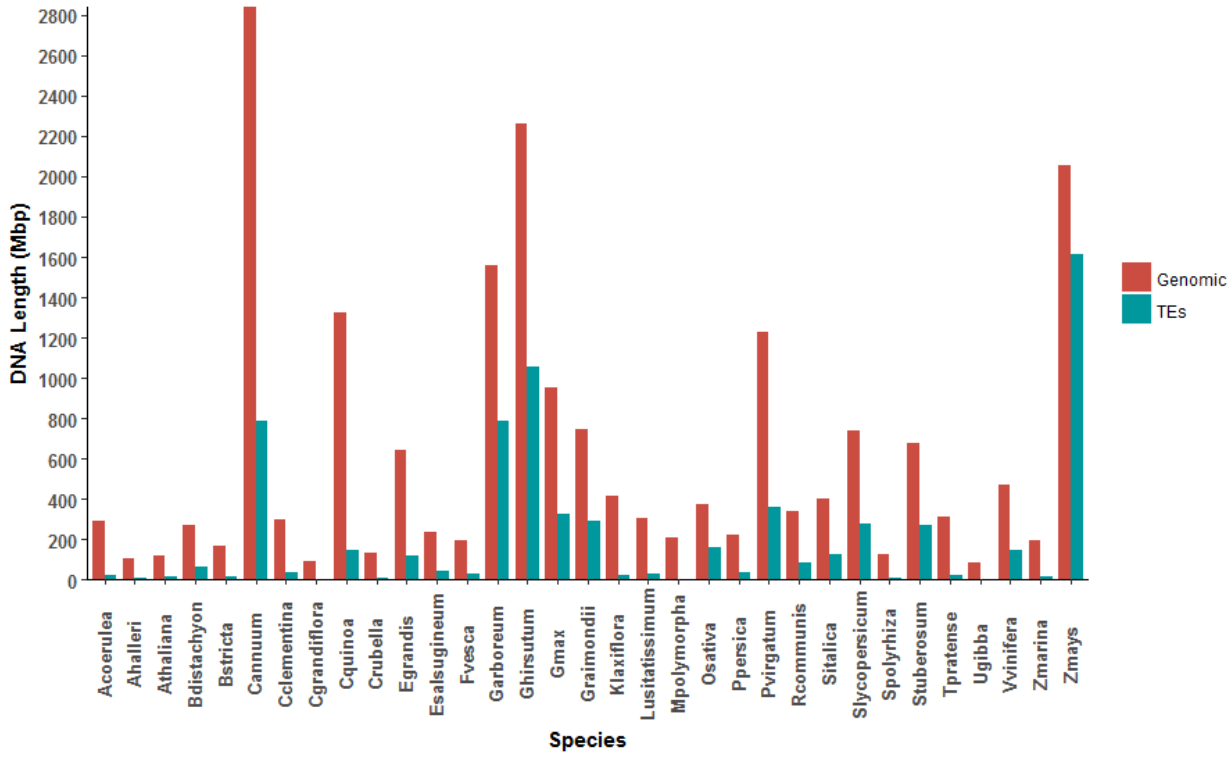

Figure 9. Genome and TE lengths for 33 Embryophyte species. Counts were generated based on only GCAT content of the respective species.

the order Fabid - *T. pratense*, *G. max*, *F. vesca*, *P. persica*, *R.communis* and *L.usitatissimum* - show a weaker correlation (t = 7.22, P < 0.002). The Gossypium samples show even weaker correlation (t = 8.49, P < 0.07).

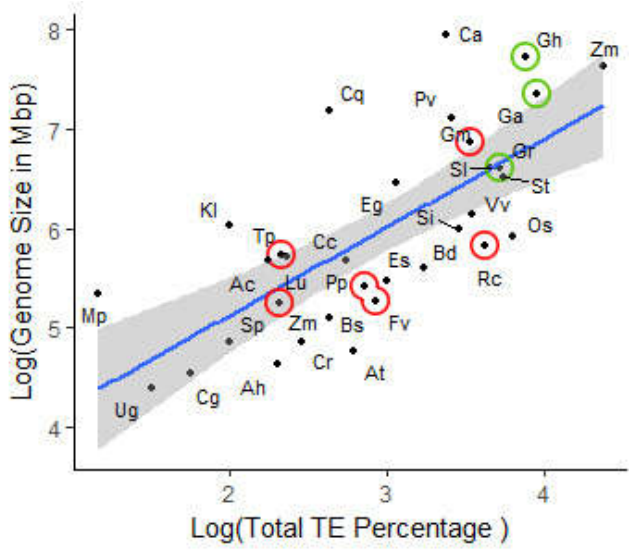

Figure 10. Correlation of genome size and TE richness for species in Viridiplantae. Labels indicate the species based on the first letter of the Genus and species – i.e. At: Arabidopsis thaliana. Confidence intervals were generated by RStudio using Loess smoothing. Red circles indicate the Order Fabid, and green circles indicate the Genus Gossypium.

For *U. gibba*, an inconsistency was noted in this study versus other research. With Repbase, the genome showed a 4.50% masking. This contrasts with studies showing the ncDNA of *U. gibba* being 3% (Ibarra-Laclette et al., 2013). By definition, ncDNA contains all TEs, so the length of TEs cannot be greater than the length of ncDNA. Further analysis of aligned TEs should be undertaken to investigate this.

TE Family Richness in Viridiplantae

The richness of TE families (DNA, LINEs, SINEs and LTRs) was highly variable between species (Fig. 12). Most species TE content is LTR-centric, with only *A. halleri* and *C. grandiflora* being < 50% LTR. DNA transposons were the next most common, followed by LINEs. SINE content was minimal in all species. Samples from the same genus – Arabidopsis, Capsella, Gossypium and Solanum – show relatively similar proportion. *U. gibba* and *Z. marina* show no SINE elements.

The trend of TE richness seems to be correlated with genome size for LTRs, LINEs and DNA transposons (Fig. 11). However, by Pearson's correlation test, this is only true for LTRs; they have a strong, positive correlation (t = 5.18, P < 0.0005). This is supported by most large genomes having a high LTR content (*Z. mays* – 2.05 Gbp, 93.1%; *G. arboreum* – 1.56 Gbp, 95%; *G. hirsutum*, 92%). DNA transposons are not correlated (t = 1.13, p < 0.27), nor are LINEs (t = 1.40, P < 0.17). SINEs have such a low proportion of the genome, with *A.halleri* having the highest proportion of just 1.7%, that no correlation was expected.

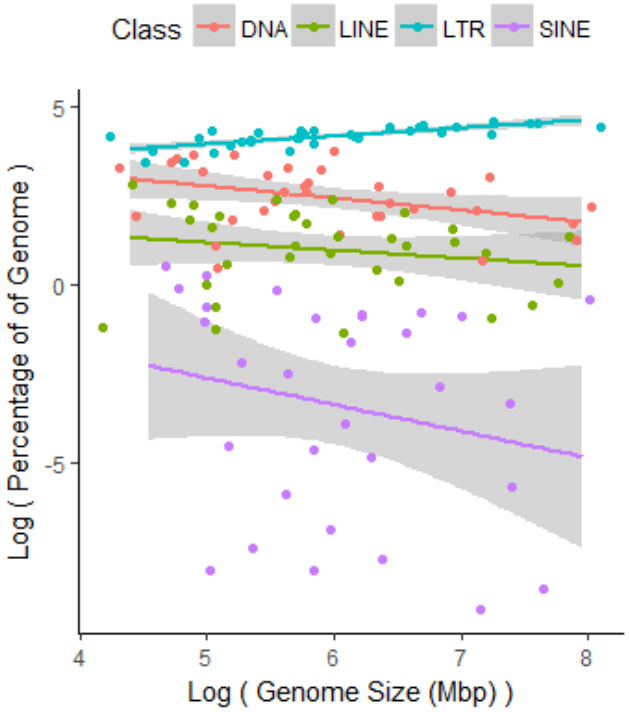

Figure 11. Richness of TE classes compared to genome size. LTRs showed a strong positive correlation via Pearson's correlation test (t = 5.18, P < 0.0005). All others show a slight negative correlation, but is not statistically significant. Confidence intervals were generated by RStudio using Loess smoothing.

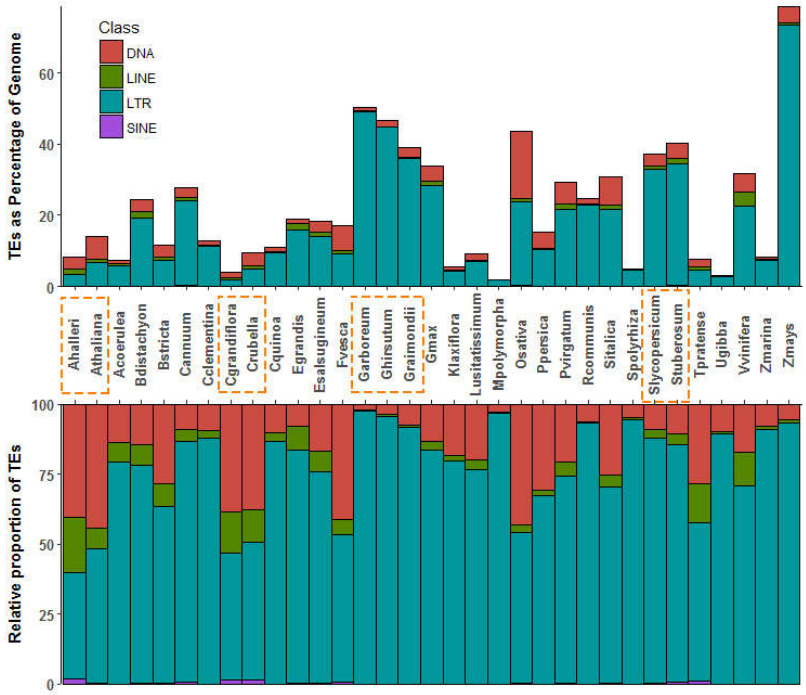

Figure 12. Diversity of TE classes as a percentage of the genome and of TE total length. Species in the same genus are grouped in orange boxes. SINEs are present on both graphs, but are indistinguishable in the top, and limited to the lower edge of the bottom graph.

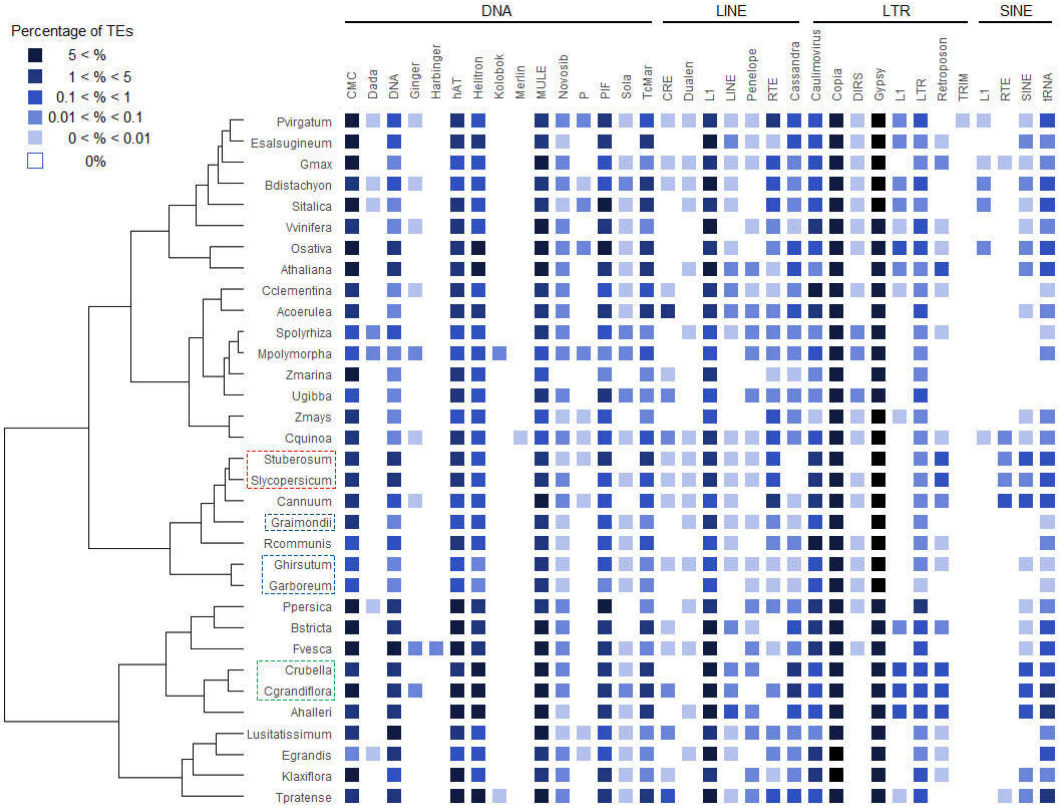

Figure 13. Subfamily richness for Embryophyte genomes. TE subfamilies content was determined as a percentage of the genome, using TAIR as the masking library. Subfamilies binned as 0% show no elements in the genome. The dendrogram was derived with complete-linkage hierarchical clustering based on true proportion, rather than binned data. Species with a common genus are highlighted in boxes.

To determine phylogeny based on TEs, subfamilies were examined (Fig. 13). Repbase recognized 34 different subfamilies across all species examined. The relative proportion of the family was binned into a logarithmic scale to better show low copy number TEs. Most subfamilies had widespread expression, but a few were unique to a small number of species. Of the DNA transposons, *DNA/Harbinger* elements appeared only in *P. persica*, while *DNA/Kolobok* showed up exclusively in *F. vesca*. For LTR's, *P.virgatum* was the only host of LTR/TRIM TEs. Neither LINEs or SINEs contained any TEs that are particularly rare. Another notable exception is the lack of *LINE/Cassandra* in Solanum species while it is present in all others.

Using the data on subfamily richness, a phylogenetic tree can be constructed. Species with a similar TE landscape would show as closely related. The constructed tree showed close relationships between species in the same Genus. The members of Capsella and Solanum both shared terminal tree branches, while Gossypium members only slightly removed. Arabidopsis was an exception, with *A.thaliana* and *A. halleri* being far removed.

The phylogenetic tree was compared to the constructed tree by subfamily content (Fig. 14). The top and bottom sides of each tree showed a great deal of consistency, with 12 species being in the same terminal branch order. The Order Fabid was mostly consistent, with nearly 4 of 6 species being clustered to-

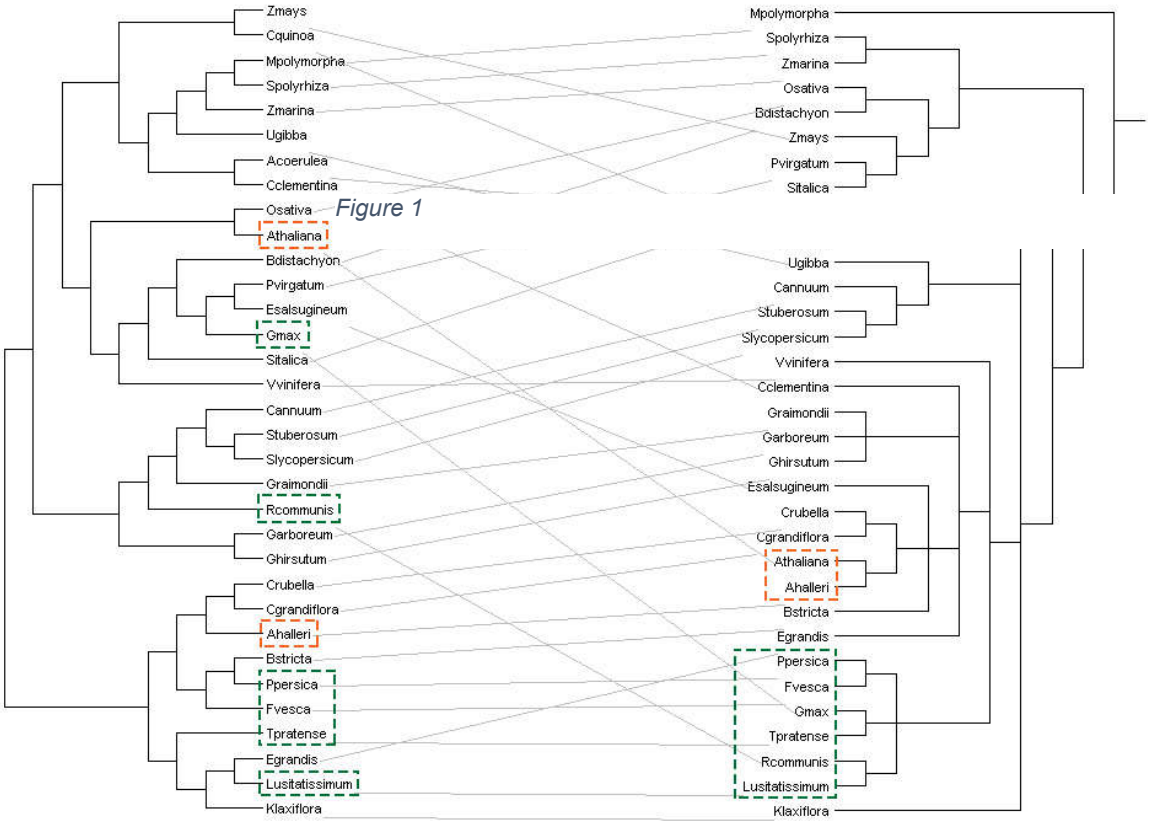

Figure 1

Figure 14. Tanglegram of a NCBI-generated phylogenetic tree for Embryophyte species compared to the constructed dendrogram based on TE subfamily content. Species the same genus that are separated in the final branches are indicated with dashed boxes.

gether. *G. max* and *R. communis* were exceptions, and showed moderate re-arrangement in the constructed tree.

Discussion

Comparison of Existing TE Libraries

Due to a species-specific library being present (TAIR), the merged library predominantly masked sequences common between them. A higher Smith-Waterman score, and hence alignment, is expected in this situation. The merged library did mask a greater proportion of the genome than TAIR, which shows that TAIR's library is not comprehensive for Arabidopsis, though the difference was low (18.6 and 19.7%, for TAIR and the merged library, respectively).

Each library showed an optimum situation for use. Repbase showed the greatest diversity for subfamilies; TAIR showed a high species-specific masking coverage; REdat showed a strong affinity to LTR TEs; and the merged library showed the greatest overall masking coverage, so long as computation time is not a factor.

Discovery of Novel TEs

For RED, the speed of novel element detection seems to be accurate. This was far faster than RepeatModeler, but the sensitivity was very poor. The resulting sequences made up about 1/3 of the genome, so classification of all elements would be computationally expensive. As well, the results from using these sequences as an input for RepeatModeler show that doing so may be inefficient due to the much lower masking percentage of Arabidopsis.

RepeatModeler showed substantial computation time, but yielded a manageable number of sequences to analyze. However, after masking against Repbase, most sequences were lost. When RepeatMasker is used in the pipeline for RepeatModeler, this simple step should have been included.

The usage of RepeatModeler is best suited for a cluster system. Utilizing multiple clusters would drastically reduce the considerable computation time, and would better take advantage of the recovery mechanisms for a crashed session.

Further research could be done on discovered TEs. A proposed pipeline for verification has been suggested, but was not attempted based on the results seen here (Girgis, 2015).

Arabidopsis Genome and TE Richness

As a model organism, Arabidopsis has been long used as a genetic tool due to its low maintenance lifecycle, small and tractable genome, and close evolutionary relationship to many economically important crop plants. Many projects, such as the 1001Genomes Project, are dedicated to studying this plant. The accessions used in this project were collected by an affiliate of the project from across Europe and Asia (Fig. 12).

Figure 12. Collection sites of A.thaliana samples. All countries had one sample collected except for: Germany, 4; Russia, 3; Spain, 2.

Initial investigation into the sample genomes showed an unexpectedly similar genome size and TE richness. Previous studies

have indicated that accessions collected between countries may show up to a 10% difference in genome size (Schmuths et al., 2004). The difference of 0.8% found here, would be more typical of a small region up to a few hundred kilometers. Applying a Shapiro-Wilk normality test shows that the accessions were not drawn from a normal distribution ($p < 10^{-7}$), which is unusual for such a geographically wide sampling. This would suggest that these samples of *A. thaliana* may have had a recent evolutionary ancestor. There is strong evidence to support this through projects like 1001genomes. Approximately 95% of the current worldwide population is likely to have originated from a single group in the 17th century (Exposito-Alonso et al., 2016). 'Relict' populations are still present primarily in Northern Africa and Southern Spain, but the accessions in this collection from those regions were not members.

Genomes typically do not change quickly, but some elements within the genome can. Some TEs, especially retrotransposons, show a significantly higher mutation rate over normal DNA – studies have shown a 10^4 to 10^6-fold increase in mutation rate (Flavell, 1992). However, variation with TE families and subfamilies do not reflect this value. *A. thaliana* does not have a relatively active mobilome, as it contains a much lower level of 'young' TEs compared to others in its genus, but increased TE variation would still be expected (Quadrana et al., 2016).

Low Diversity in *A.thaliana* Kimura Distances

The results seen with *A. thaliana* may be due to the sequencing of the sample genomes. The sequencing used was a combination of iterative read-mapping and de novo assembly (Gon et al., 2011). This sequencing can have difficulty in assembling repeat sequences, but would not cause such a homogenous TE landscape (Treangen & Salzberg, 2012). Per Gon et al., "to report complex alleles consistently, we defined all variants against the multiple alignment consensus of Col-0 and the assembled genomes". For each accession, this resulted in about 45,000 ambiguous nucleotides residing primarily within regions of known TEs. Because of this, many SNPs in TEs would not be considered when determining the Kimura distance. This is consistent with our results showing a heavily left-skewed graph of Kimura distance. As such, this may be a potential flaw in the assembly procedure used by the Wellcome Trust Center.

Genome and TE richness of Embryophytes

The trend shown for genome size versus TE richness is consistent with the idea that essential genes are relatively conserved (Rubin et al., 2015). For species that do not closely follow the trend, this may be a limitation of the library being used, as other studies have shown significantly higher TE content in some organisms; *C. annuum* has shown TE richness of up to 81% in some studies (Qin et al., 2015; Kim et al., 2014).

Evidence of TE Horizontal Gene Transfer

Evidence has been found for horizontal gene transfer (HGT) events for multiple genes between unrelated species (Wallau, Ortiz & Loreto, 2012). Based on subfamily richness, this may have occurred 4 times within the species analyzed in this study – *F. vesca* with *DNA/Harbinger*, *M. polymorpha* and *T. pratense* with *DNA/Kolobok*, *C. quinoa* with *DNA/Merlin*, and *P. virgatum* with *LTR/TRIM*. In each of these instances, the TE is present only in those species. It would be unlikely to have gone extinct in all other samples, and re-

main just in these. It is possible that spontaneous mutation of an existing TE may have created a new family, but in each of these cases, the TE is present in other unrelated species.

DNA/Harbinger is not uncommon in plants. Out of the species in this study, Repbase also has entries for Harbinger elements detected in the genomes of *O. sativa* and *Sorghum bicolor* (Bao et al., 2015). As TEs can be highly transient, they may not have been present in the samples used in this study. A known mechanism for gene transfer between plants is anastomosis, a physical connection between cells (Qui, 2016). *F. vesca* shares habitat with both these species, so such a instance may have occurred.

C. quinoa, or quinoa, is an angiosperm. These flowering plants originated around 125 million years ago, in the Lower Cretaceous period. This contrasts greatly *Selaginella moellendorffii*, a lycophyte which arose from a species dated back 400 million years. The same *DNA/Harbinger* TE was matched between both species. The vast difference in time and divergence gives strong support for an HGT event.

Kolobok family TEs shows across multiple unrelated species; *Trichomonas vaginalis*, a protozoan parasite (Meng et al., 2011); *Hydra vulgaris*, a freshwater polyp (Bao et al. 2009); *Danio rerio*, a freshwater fish (Howe et al., 2013). Such diversity in species makes inheritance unlikely. Furthermore, the matched sequence in Repbase is attributed to *Micromonas pusilla*, which is a type of green algae. *M. polymorpha* grows predominately in damp habitats, such as the banks of river and ponds. The geographic proximity and an unknown mechanism may be responsible for the TE moving from *M. pusilla* to *M. polymorpha*.

For LTR transposons, evidence has shown HGT events for dozens of species (El Baidouri et al., 2014). LTR/TRIM transposons, or *T*erminal-*R*epeat retrotransposons *I*n *M*iniature, are a particularly small LTR. They vary from 300-800 bp, so this HGT events may be more common, due to the smaller size of the element being transferred.

TE Extinction in Solanum

Solanum lycopersicum and Solanum tuberosum show a lack of *LINE/Cassandra* TEs, indicating the extinction of the TE in that genus. This TE is common to all other species, so this suggests a recent extinction. Both species are widely used as food crops, which often faces significant artificial selection, resulting in shallow gene pools (Govindarag et al., 2015). This would affect allelic diversity, and genetic drift may cause some genes to go extinct (Star and Spencer, 2013). This may give evidence for concern towards the genetic future of our food crops, as a lower genetic diversity in a species may reduce the capacity to adapt to environmental change or disease (Hughes, 2008).

Phylogeny based on TE content

Compared to the NCBI phylogenetic tree, the constructed dendrogram shows a relatively close similarity. Related species are generally close together with those in the Orders of Solanum, Capsella and Gossypium being very close terminal branches. The Fabids remain relatively close, though *G. max* and *R. communis* do appear separated from the rest. These two species have the largest genome sizes, and at least double the TE richness as the other species in the Order. There is support in some species that TE activity increases as the TE richness increases, so a larger divergence would be expected (Gao et al., 2016).

To generate phylogeny based on TE content, R uses a hierarchical clustering method known as complete-linkage clustering by default. In this algorithm, the furthest matched samples are decided to be the furthest branched species. This progresses down the tree until no branches can be made. The final nodes of the dendrogram are generated forma common ancestor, and should be the evolutionary result of constantly diverging species. However, this is not always accurate, as evolution is not necessarily divergent. Insects, birds, and bats all evolved a mechanism of flight independently. Studies support the notion that environmental circumstances can select for similar traits (Losos, 2011). Further algorithms, such as single-link, average, median, or centroid, could be utilized to investigate the potential of using TE subfamilies in phylogenetic reconstruction (ETH Zürich, 2016; Senthilnath et al., 2016).

Conclusions

In this work, we have presented observations about TE richness and diversity across Embryophytes, as well as some potential avenues of discovery for novel elements. Analysis of inter- and intra-species TE content shows strongly supported trends and relationships. The identification of lone TE subfamilies give evidence towards a mechanism for horizontal transfer of TEs. Extinction of TE subfamilies in food crops may have cause for concern for reduced diversity and fitness of food crops. Further research into these avenues may reveal some of the mysteries behind these important and enigmatic elements.

References

Bao, W., Jurka, M. G., Kapitonov, V. V., & Jurka, J. (2009). New superfamilies of eukaryotic DNA transposons and their internal divisions. *Molecular biology and evolution*, 26(5), 983-993.

Bao, W., Kojima, K.K., Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*, 2015;6:11

Caballero, A., & García-Dorado, A. (2013). Allelic diversity and its implications for the rate of adaptation. *Genetics*, 195(4), 1373-1384.

El Baidouri, M., Carpentier, M. C., Cooke, R., Gao, D., Lasserre, E., Llauro, C., ... & Panaud, O. (2014). Widespread and frequent horizontal transfers of transposable elements in plants. *Genome research*, 24(5), 831-838.

ETH Zürich. (2016). R Documentation, hclust {stats}.

Exposito-Alonso, M., Becker, C., Schuenemann, V. J., Reiter, E., Setzer, C., Slovak, R., ... & Busch, W. (2016). The rate and effect of de novo mutations in a colonizing lineage of Arabidopsis thaliana. *bioRxiv,* 050203.

Flavell, A. J. (1992). Ty1-copia group retrotransposons and the evolution of retroelements in the eukaryotes. *Transposable Elements and Evolution Contemporary Issues in Genetics and Evolution*, 258-274. doi:10.1007/978-94-011-2028-9_19

Forterre, P., Filée, J., & Myllykallio, H. (2004). Origin and Evolution of DNA and DNA Replication Machineries. The Genetic Code and the Origin of Life, 145-168. doi:10.1007/0-387-26887-1_10

Friesen, Nikolai, Andrea Brandes, and John Seymour Heslop-Harrison. "Diversity, origin, and distribution of retrotransposons (gypsy and copia)" *Molecular Biology and Evolution* 18.7 (2001): 1176-1188.

Gan, X., Stegle, O., Behr, J., Steffen, J. G., Drewe, P., Hildebrand, K. L., ... & Kahles, A. (2011). Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*, *477*(7365), 419.

Gao, B., Shen, D., Xue, S., Chen, C., Cui, H., & Song, C. (2016). The contribution of transposable elements to size variations between four teleost genomes. *Mobile DNA*, 7(1), 4.

Girgis, Hani Z. "Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale*." BMC bioinformatics 16.1* (2015): 227.

Govindaraj, M., Vetriventhan, M., & Srinivasan, M. (2015). Importance of genetic diversity assessment in crop plants and its recent advances: an overview

of its analytical perspectives. *Genetics research international*, 2015.

Hallet, B. (1997). Transposition and site-specific recombination: adapting DNA cut-and-paste mechanisms to a variety of genetic rearrangements. *FEMS Microbiology Reviews*, 21(2), 157-178. doi:10.1016/s0168-6445(97)00055-7

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., ... & McLaren, S. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, 496(7446), 498.

Hughes, A. R., Inouye, B. D., Johnson, M. T., Underwood, N., & Vellend, M. (2008). Ecological consequences of genetic diversity. *Ecology letters*, 11(6), 609-623.

Ibarra-Laclette, E., Lyons, E., Hernández-Guzmán, G., Pérez-Torres, C. A., Carretero-Paulet, L., Chang, T. H., ... & Fernández-Cortés, A. (2013). Architecture and evolution of a minute plant genome. *Nature*, 498(7452), 94.

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. *Genome research*, 20(10), 1313-1326.

Kidwell, M. G. (1992). Horizontal transfer of P elements and other short inverted repeat transposons. *Genetica*, 86(1-3), 275-286. doi:10.1007/bf00133726

Kim, S., Park, M., Yeom, S. I., Kim, Y. M., Lee, J. M., Lee, H. A., ... & Jung, K. (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nature genetics*, 46(3), 270-278.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution,* 16(2), 111-120.

Leprince, A.S., Grandbastien, M.A., Meyer, C., (2001). Retrotransposons of the Tnt1B family are mobile in Nicotiana plumbaginifolia and can induce alternative splicing of the host gene upon insertion. *Plant Molecular Biology*, 47, 533–541. doi:10.1023/A:1011846910918.

Lisch, D. (2009). Epigenetic regulation of transposable elements in plants. *Annual review of plant biology*, 60, 43-66.

Losos, J. B. (2011). Convergence, adaptation, and constraint. *Evolution*, 65(7), 1827-1840.

Martienssen, R. A. (1998). Functional genomics: Probing plant gene function and expression with transposons. *Proceedings of the National Academy of Sciences*, 95(5), 2021-2026. doi:10.1073/pnas.95.5.2021

Meng, Q., Chen, K., Ma, L., Hu, S., & Yu, J. (2011). A systematic identification of Kolobok superfamily transposons in Trichomonas vaginalis and sequence analysis on related transposases. *Journal of Genetics and Genomics*, 38(2), 63-70.

Mills, R. E., Bennett, E. A., Iskow, R. C., & Devine, S. E. (2007). Which transposable elements are active in the human genome? *Trends in Genetics*, 23(4), 183-191. doi:10.1016/j.tig.2007.02.006

Orgel, L. E., & Crick, F. H. (1980). Selfish DNA: the ultimate parasite. *Nature*, 284(5757), 604-607. doi:10.1038/284604a0

Pennisi, E. (2012). "ENCODE Project Writes Eulogy for Junk DNA". *Science*, 337(6099), 1159–1161. doi:10.1126/science.337.6099.1159.

Qiu, H., Cai, G., Luo, J., Bhattacharya, D., & Zhang, N. (2016). Extensive horizontal gene transfers between plant pathogenic fungi. *BMC Biology*, 14(1). doi:10.1186/s12915-016-0264-3

Rubin, B. E., Wetmore, K. M., Price, M. N., Diamond, S., Shultzaberger, R. K., Lowe, L. C., ... & Golden, S. S. (2015). The essential gene set of a photosynthetic organism. *Proceedings of the National Academy of Sciences*, 112(48), E6634-E6643.

SanMiguel, P., Tikhonov, A., Jin, Y., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., . . . Bennetzen, J. L. (1996). Nested Retrotransposons in the Intergenic Regions of the Maize Genome. *Science*, 274(5288),765-768. doi:10.1126/science.274.5288.765

Schmuths, H., Meister, A., Horres, R., & Bachmann, K. (2004). Genome size variation among accessions of Arabidopsis thaliana. *Annals of Botany*, 93(3), 317-321.

Senthilnath, J., Kumar, D., Benediktsson, J. A., & Zhang, X. (2016). A novel hierarchical clustering technique based on splitting and merging. *International Journal of Image and Data Fusion*, 7(1), 19-41.

Star, B., & Spencer, H. G. (2013). Effects of genetic drift and gene flow on the selective maintenance of genetic variation. *Genetics*, 194(1), 235-244.

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature reviews. *Genetics*, 13(1), 36.

Quadrana, L., Silveira, A. B., Mayhew, G. F., LeBlanc, C., Martienssen, R. A., Jeddeloh, J. A., & Colot, V. (2016). The Arabidopsis thaliana mobilome and its impact at the species level. *Elife*, 5, e15716.

Qin, C., Yu, C., Shen, Y., Fang, X., Chen, L., Min, J., ... & Yang, Y. (2014). Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. *Proceedings of the National Academy of Sciences*, 111(14), 5135-5140.

Wallau, G. L., Ortiz, M. F., & Loreto, E. L. S. (2012). Horizontal transposon transfer in eukarya: detection, bias, and perspectives. *Genome biology and evolution*, 4(8), 689-699.

Wang, W., Zheng, H., Fan, C., Li, J., Shi, J., Cai, Z., ... & Lu, Z. (2006). High rate of chimeric gene origination by retroposition in plant genomes. The Plant Cell, 18(8), 1791-1802.

Xuan, Y. H., Peterson, T., & Han, C. D. (2016). Generation and Analysis of Transposon Ac/Ds-Induced Chromosomal Rearrangements in Rice Plants. *Chromosome and Genomic Engineering in Plants: Methods and Protocols*, 49-61.

Zhao, D., Ferguson, A. A., & Jiang, N. (2016). What makes up plant genomes: The vanishing line between transposable elements and genes. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1859(2), 366-380.