Observations of Transposable Element Richness and Diversity in Embryophytes

Andrew Lindsay, Honours Thesis, Summer 2017

Under Supervision of Dr. Michael Deyholos
University of British Columbia, Okanagan

Outline

- What are transposable elements (TEs)? Why are they important?
- Analysis of existing libraries, and methods of discovery
- Diversity of transposon content across green plants
- Comparison of *Arabidopsis thaliana* individuals
- Conclusions
- Future Research

What are transposons?

Transposons are:

- **Mobile genetic elements**
- **Often able to replicate during transposition**
- **Able to affect the genetic make-up of the host**

Where did they come from?

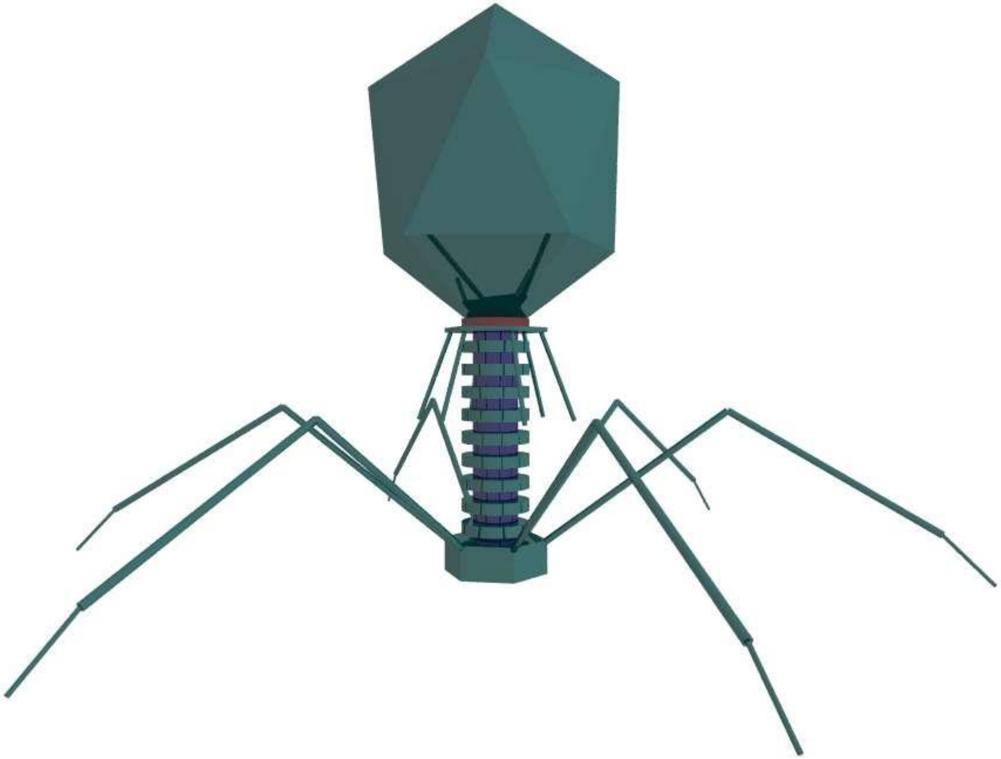


Figure 1. Bacteriaphage (http://cronodon.com, 2008)

How abundant are they?

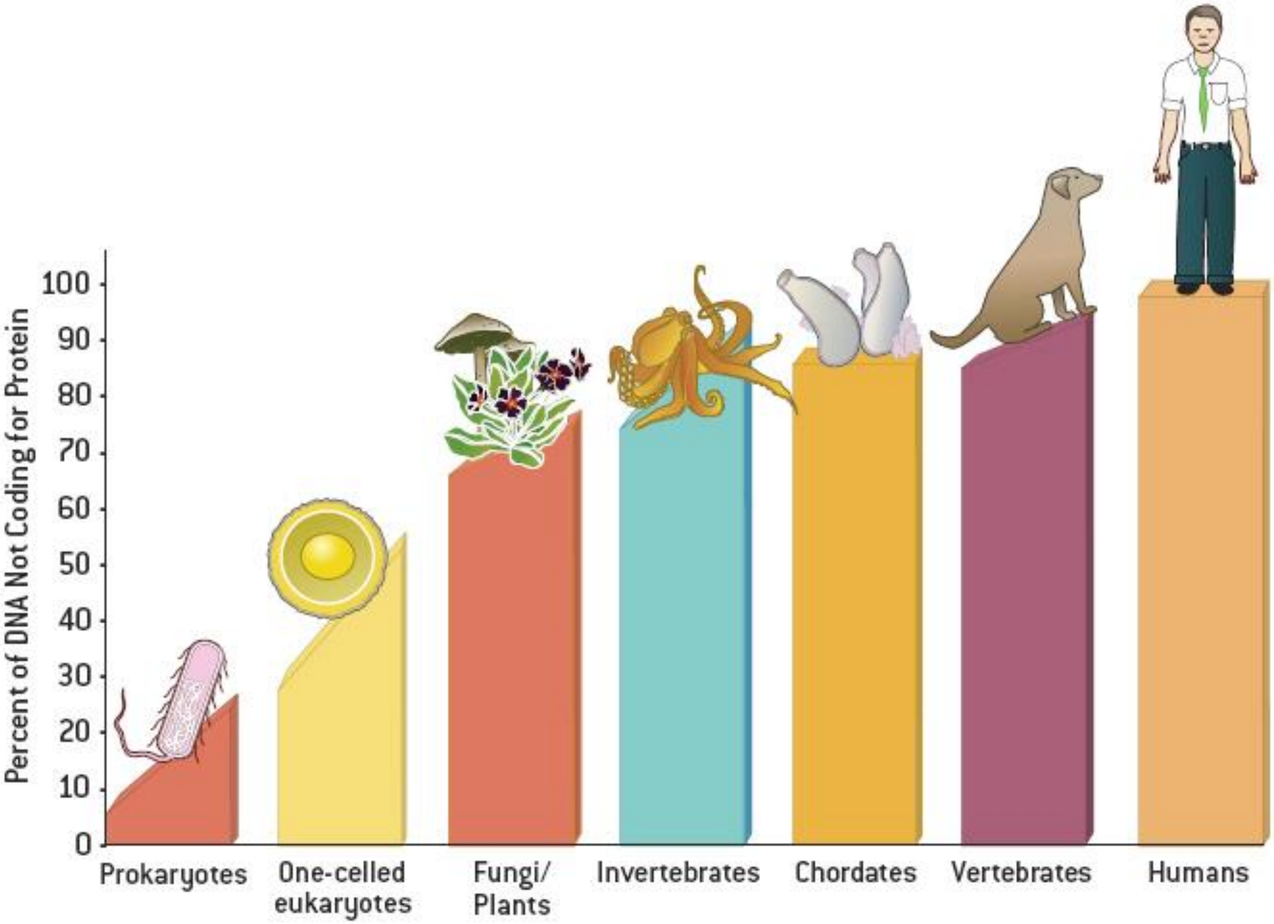


Figure 2. Non-coding DNA content species (Gregory, 2008)

What is their structure?

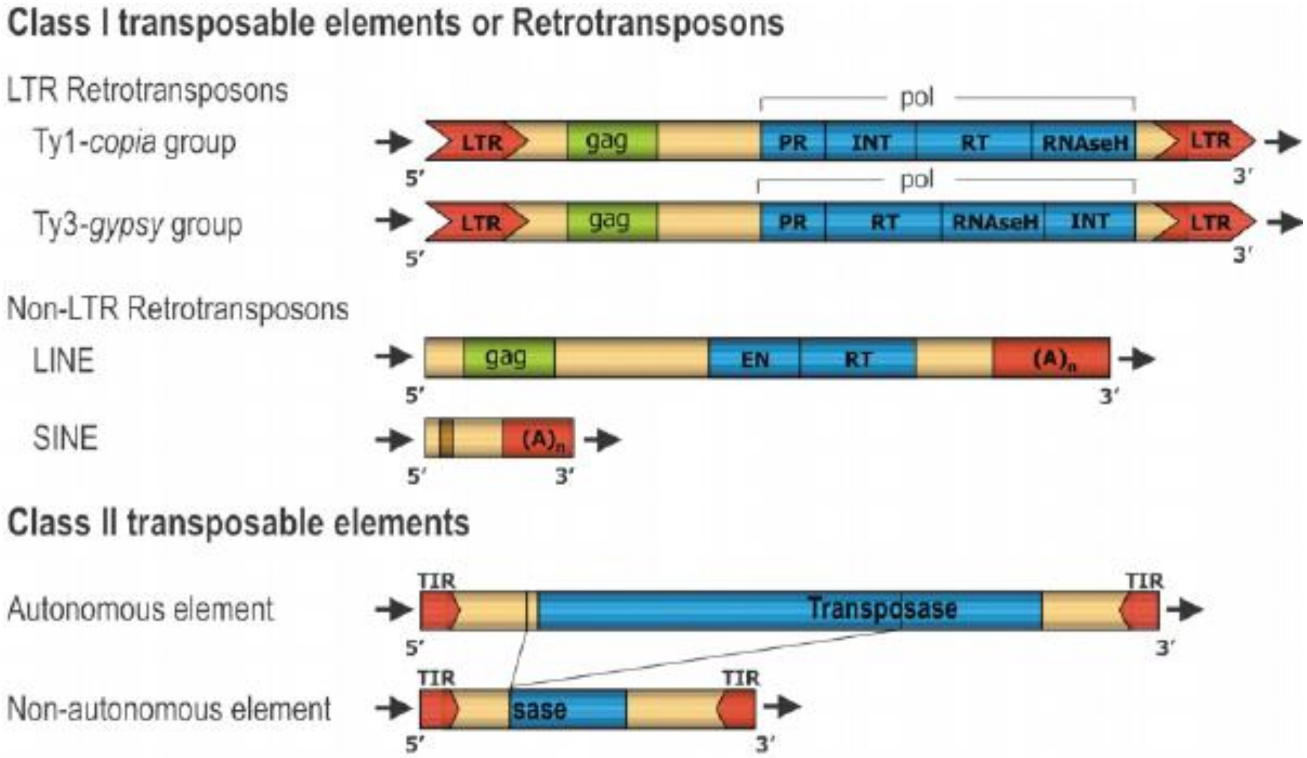
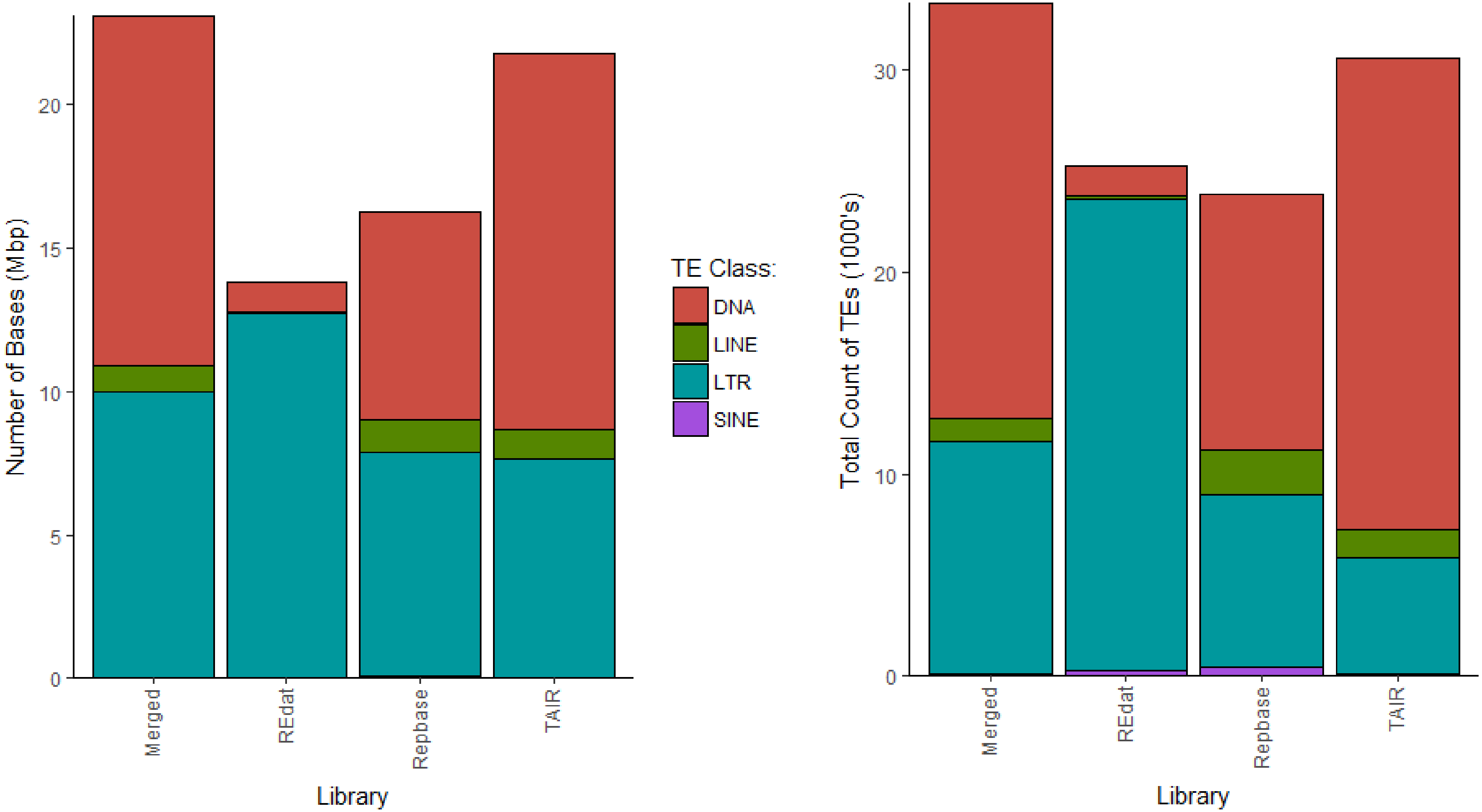


Figure 3. Structure of the different types of plant transposable elements (Casacuberta & Santiago, 2013).

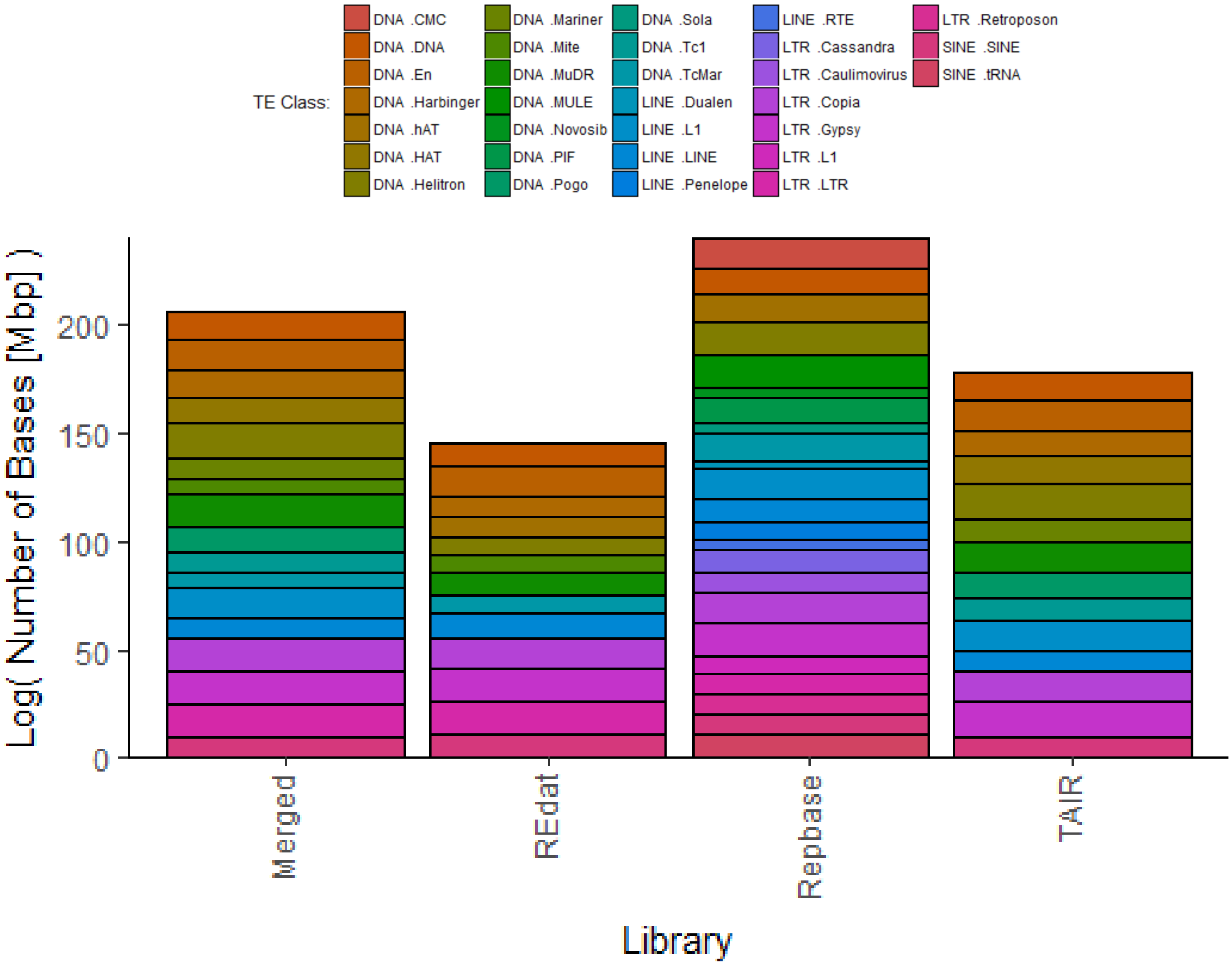
Comparison of TE Libraries

- **REdat (Plant Genome and System Biology Group, Germany)**
  - **Combination of TREP, TIGR repeats, PlantSat and Genbank libraries**
  - **450 Mbp, 61K sequences**
- **Repbase (Genetic Information Resource Institute, USA)**
  - **36 Mbp, 12k sequences**
- **TAIR (The Arabidopsis Information Resource, USA)**
  - **23 Mbp, 31k sequences**
- **Merged Library**
  - **Contains all above sequences**

Comparison of TE Libraries

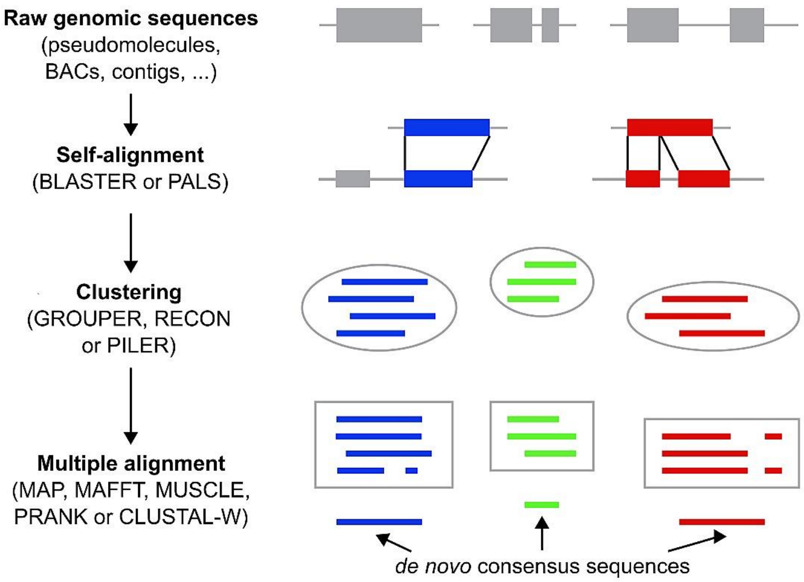


Comparison of TE Libraries



Discovery of Novel TEs

- **RepeatModeler**



Raw genomic sequences
(pseudomolecules,
BACs, contigs, ...)

Self-alignment
(BLASTER or PALS)

Clustering
(GROUPER, RECON
or PILER)

Multiple alignment
(MAP, MAFFT, MUSCLE,
PRANK or CLUSTAL-W)

de novo consensus sequences

- **Red (REpeat Detector)**
  - **Uses a machine learning algorithm to find repeats**

Discovery of Novel TEs

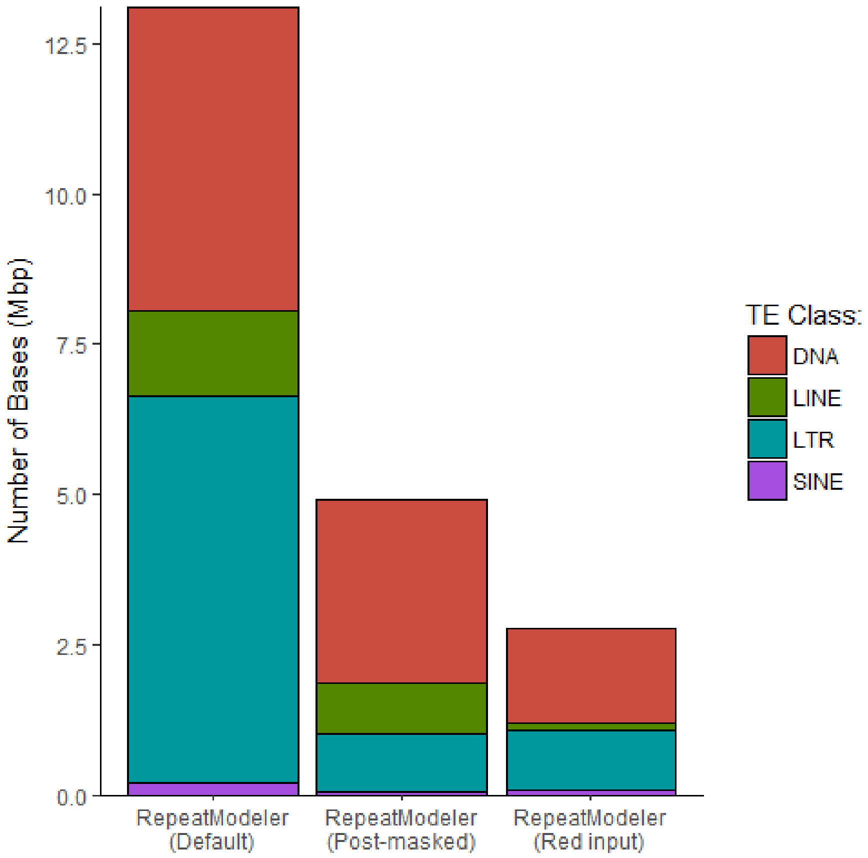
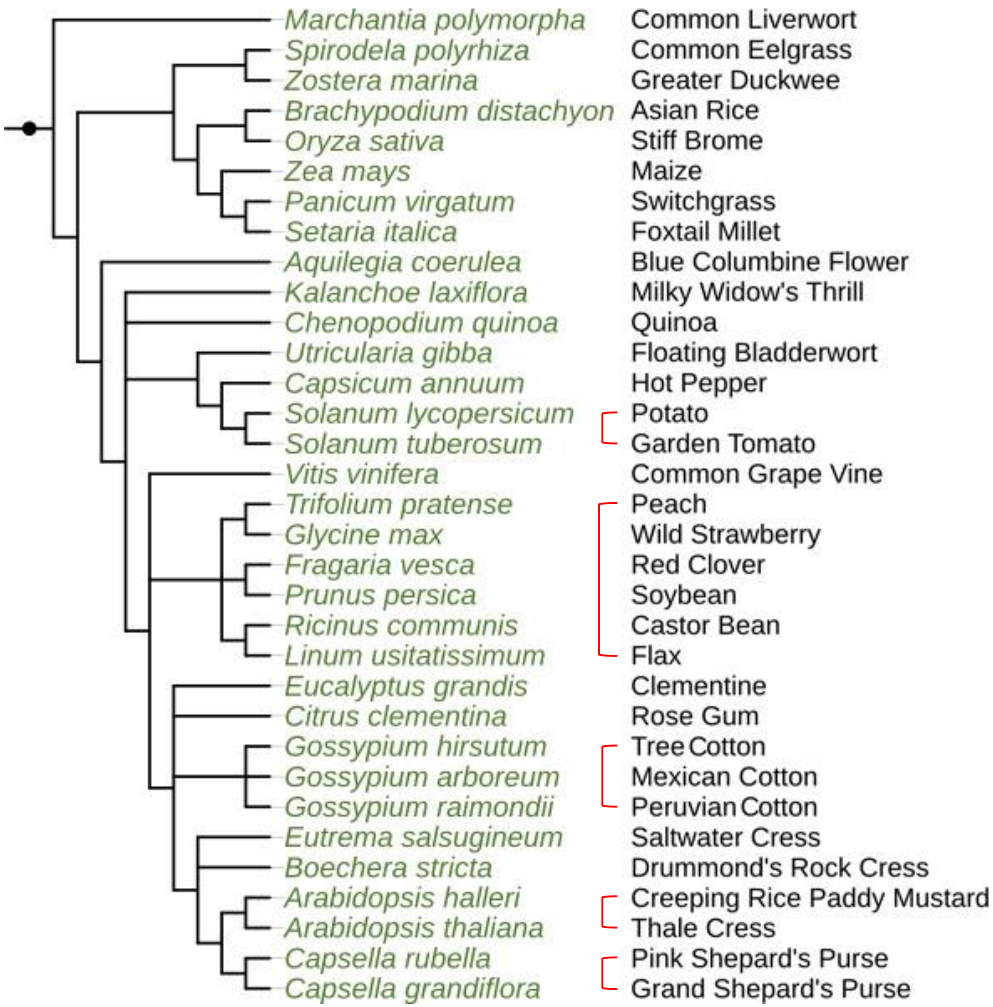


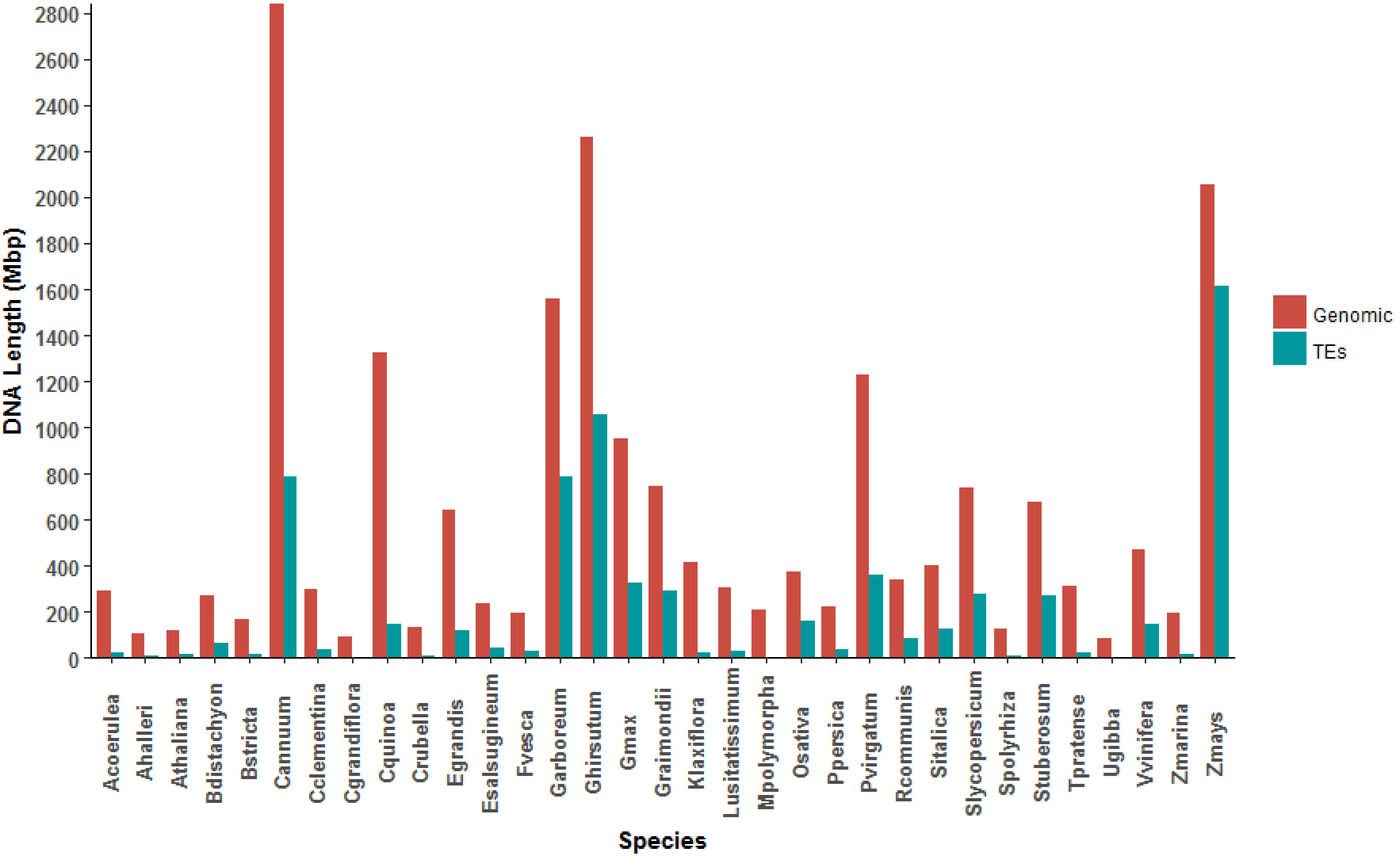
Table 1. Run-time for RepeatModeler with Embryophyte genomes

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

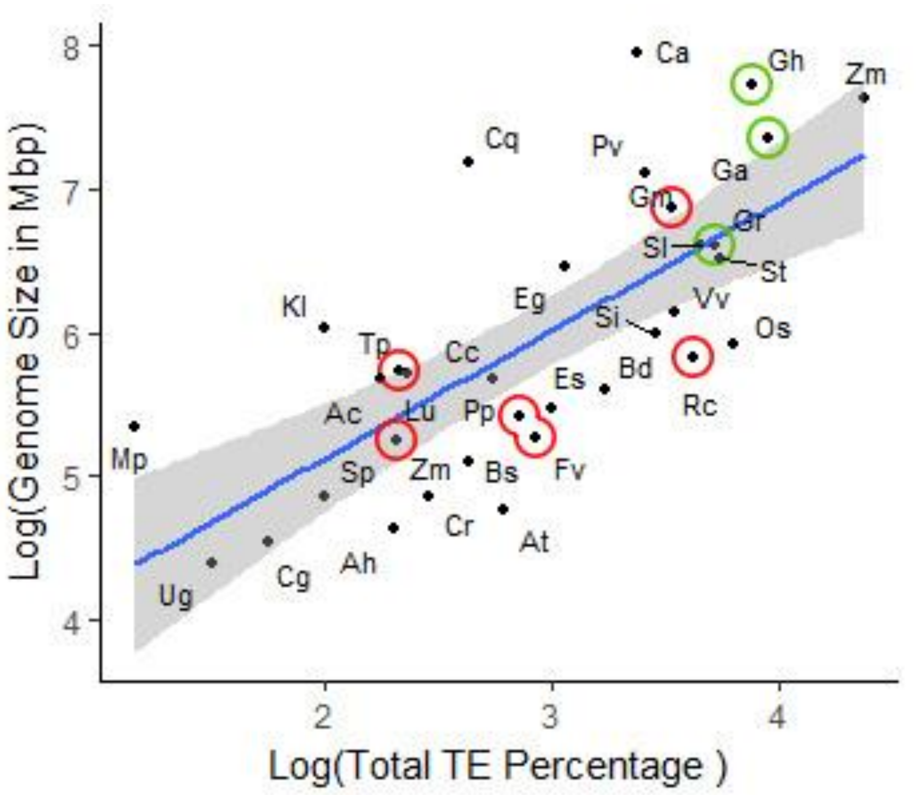
Transposons in Green Plants



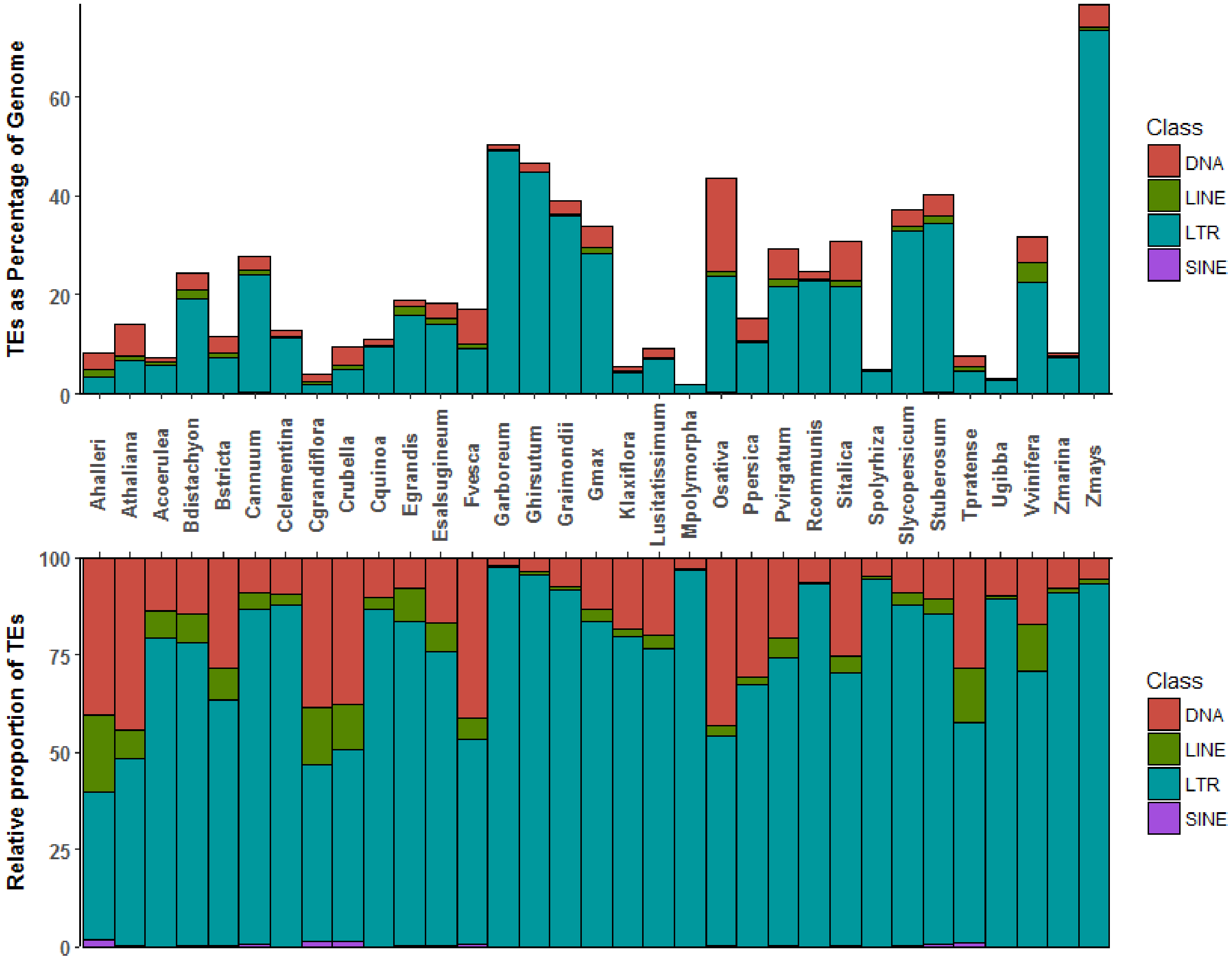
Transposons in Green Plants



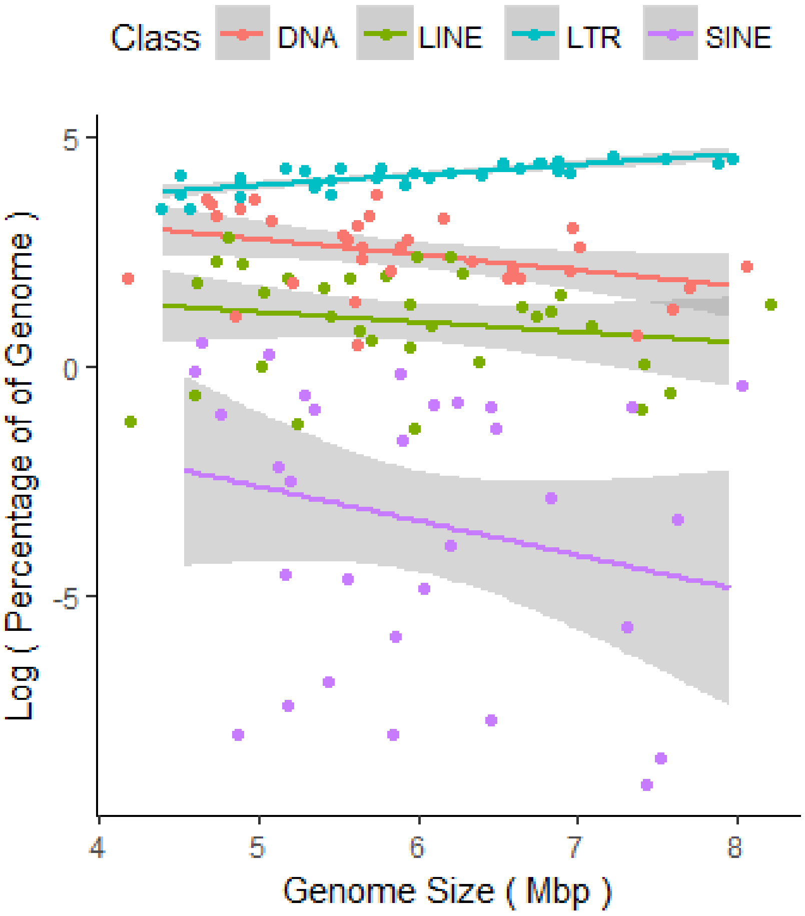
Transposons in Green Plants



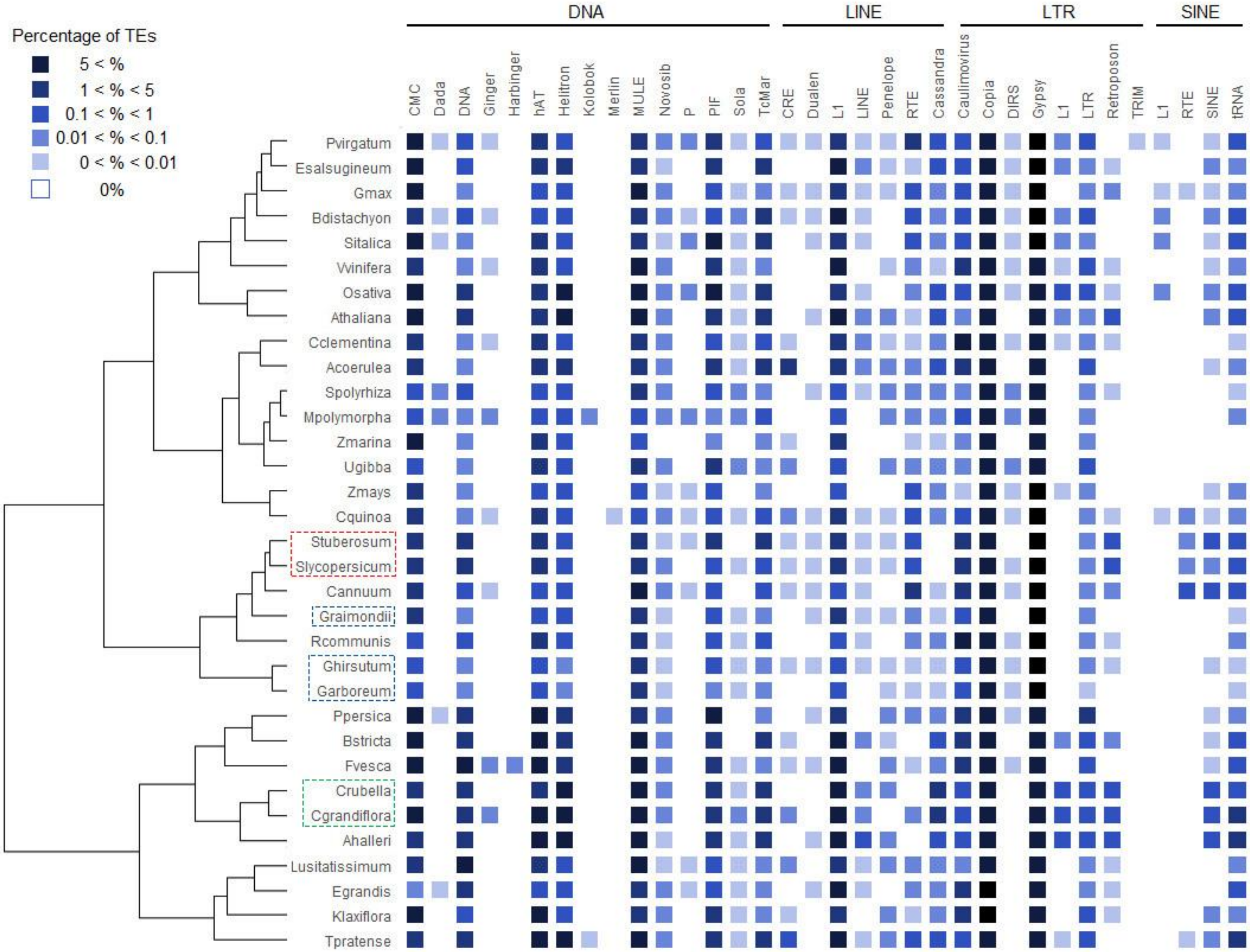
Transposons in Green Plants



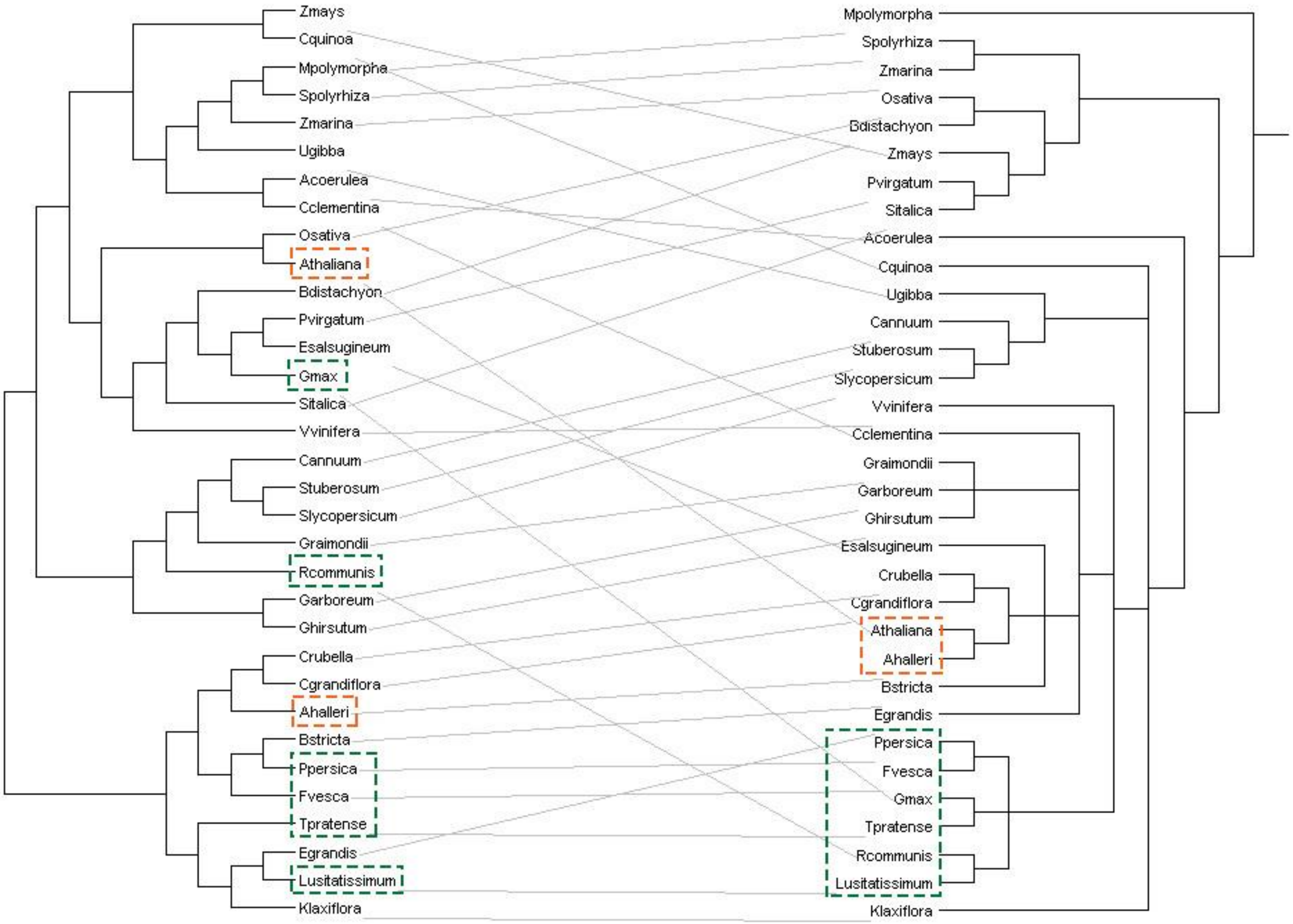
Transposons in Green Plants



Transposons in Green Plants



Transposons in Green Plants



Transposons in Arabidopsis

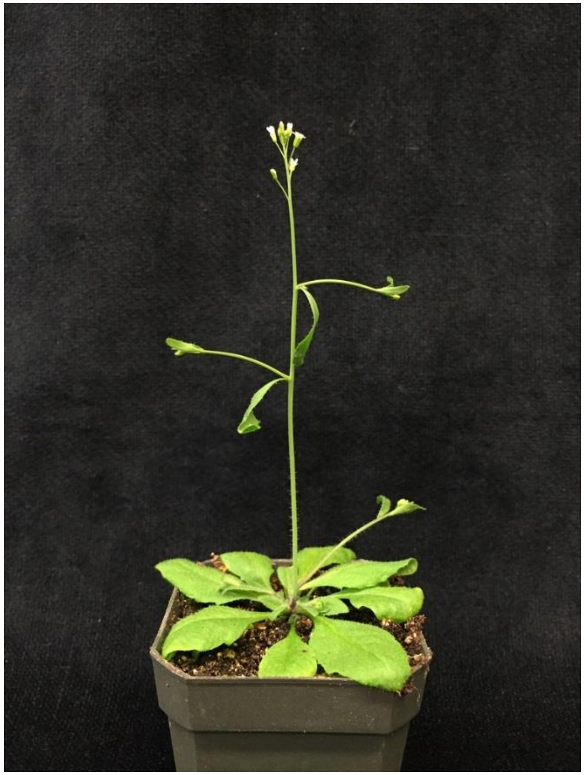
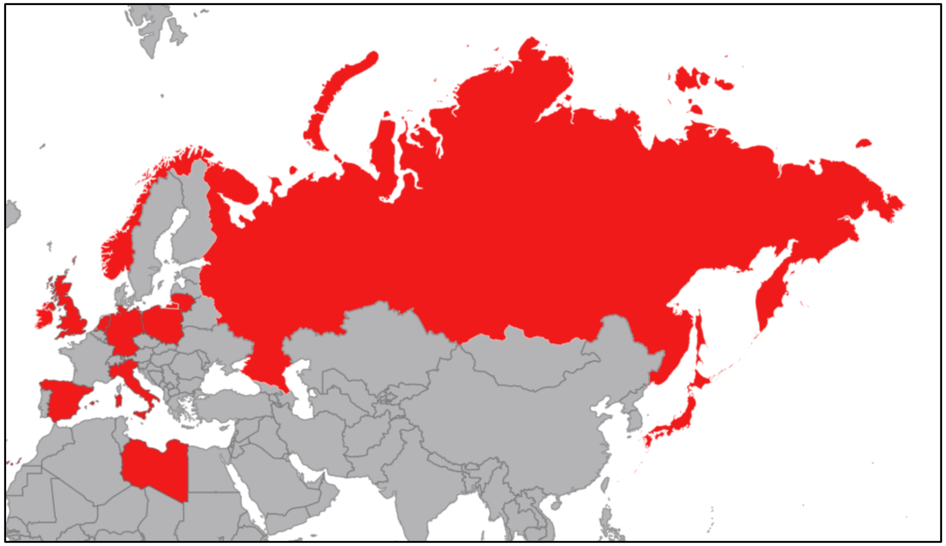
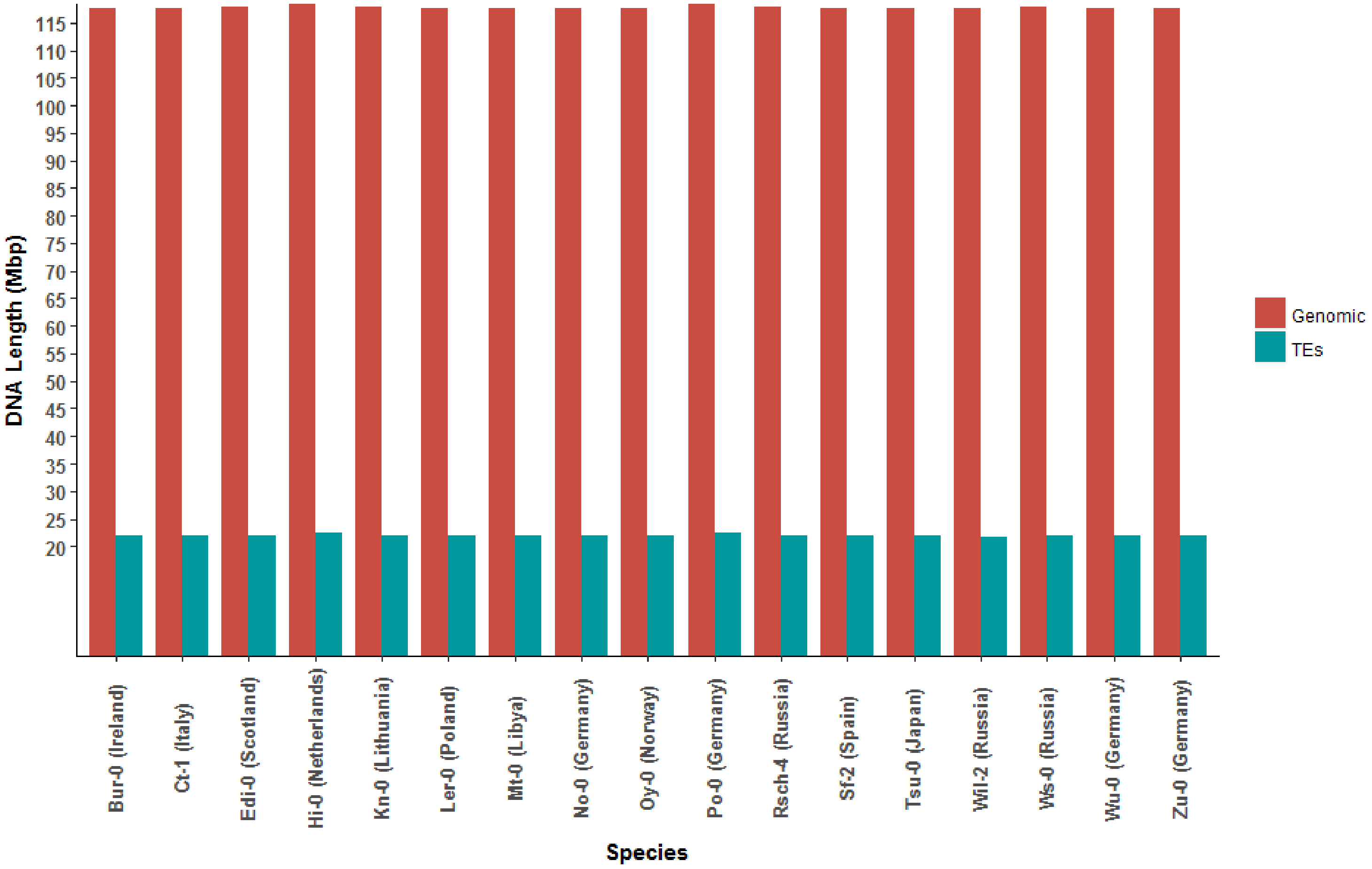


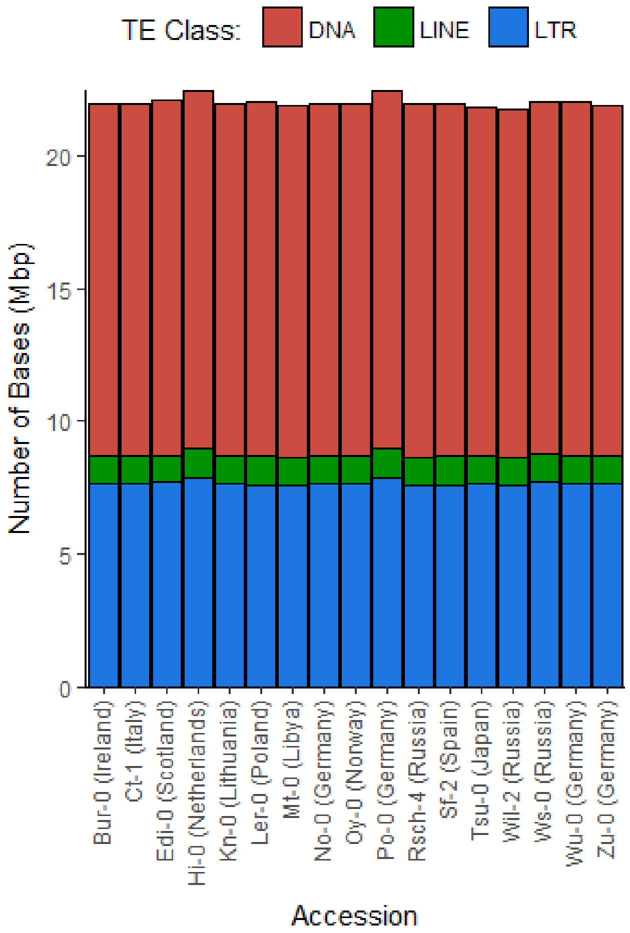
Figure 3. Arabidopsis thaliana (Roepers, 2004)



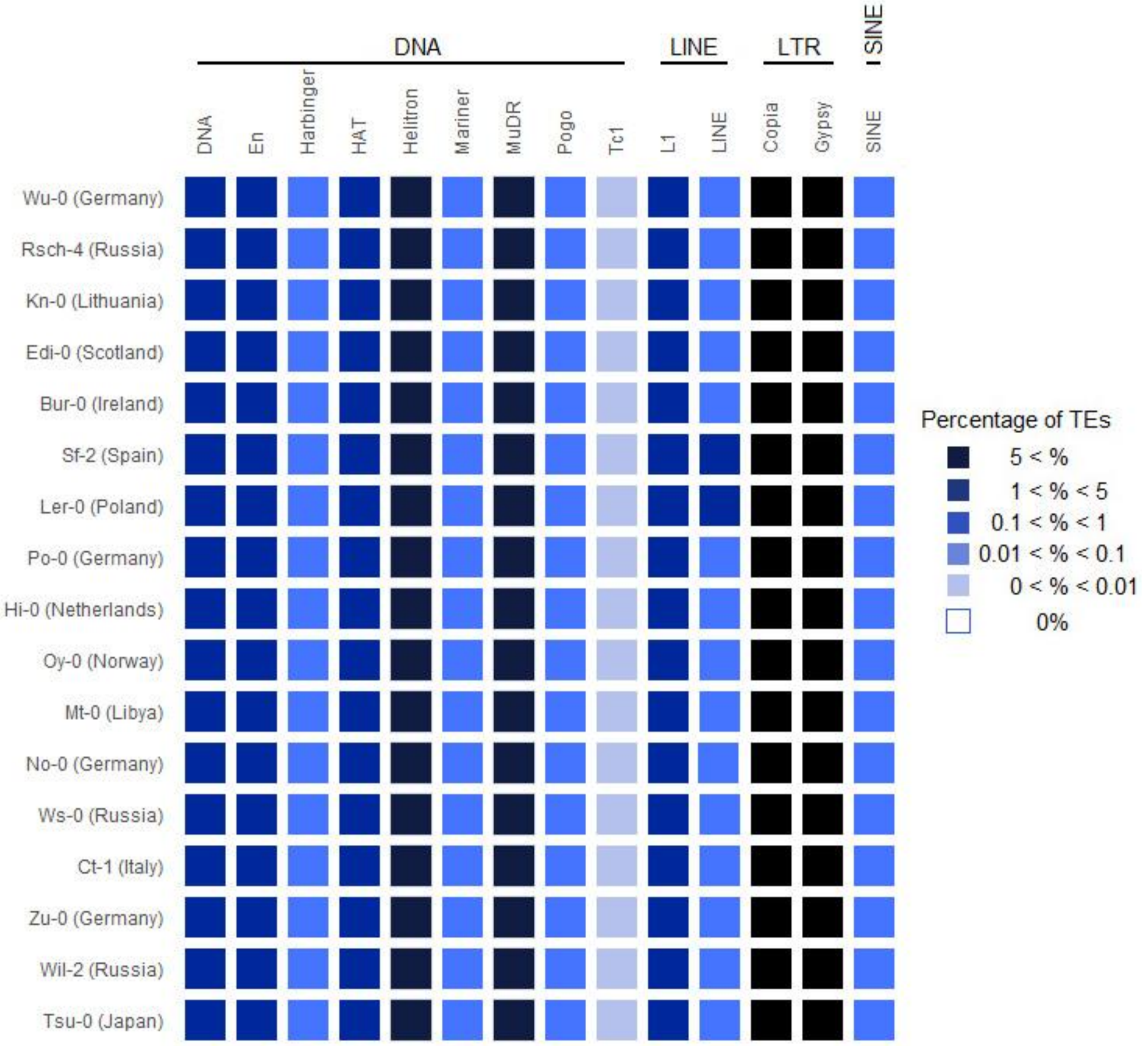
Transposons in Arabidopsis



Transposons in Arabidopsis



Transposons in Arabidopsis



Transposons in Arabidopsis

Kimura Distance:

$$K = -\frac{1}{2}\ln(1-2p-q)\sqrt{1-2q}$$

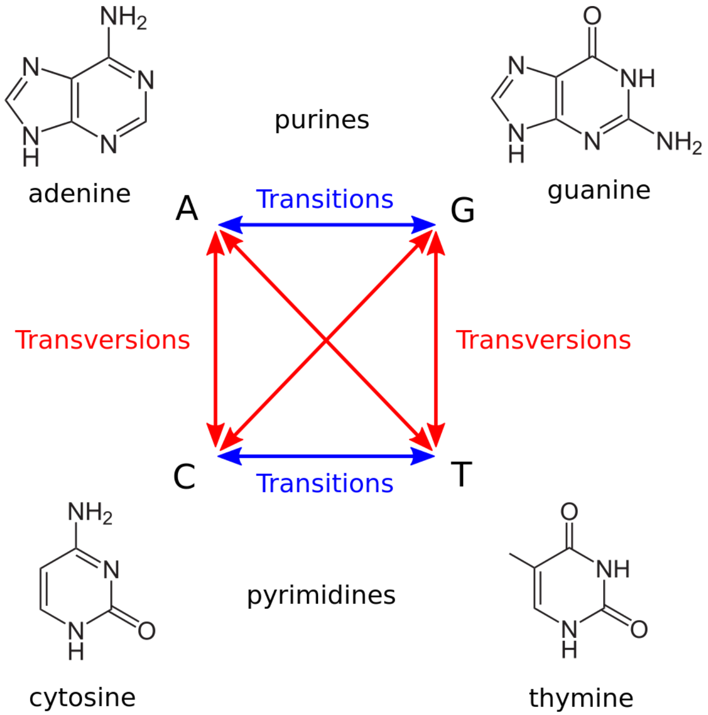
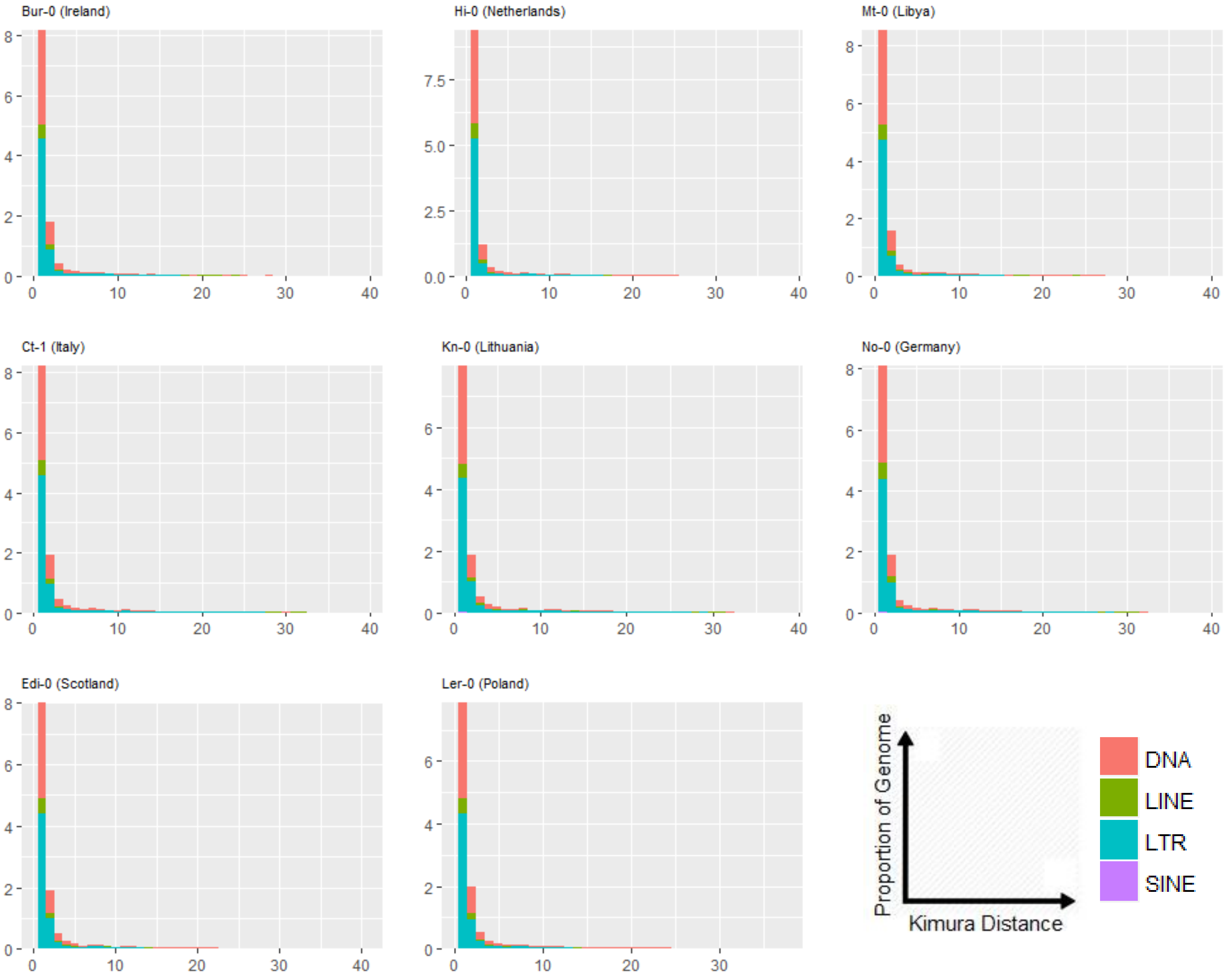
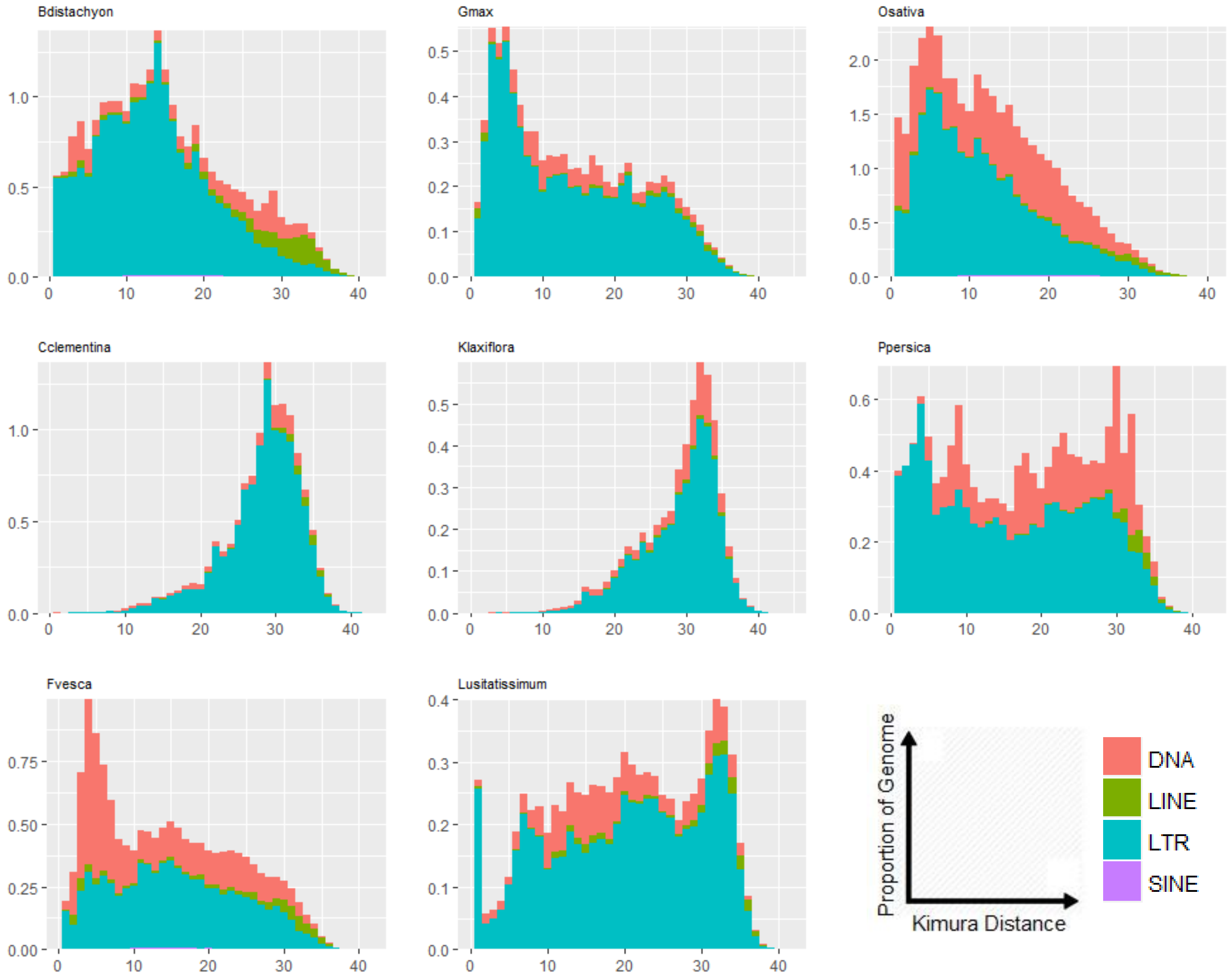


Figure 3. Transitions-transversions (Petulda, 2008)

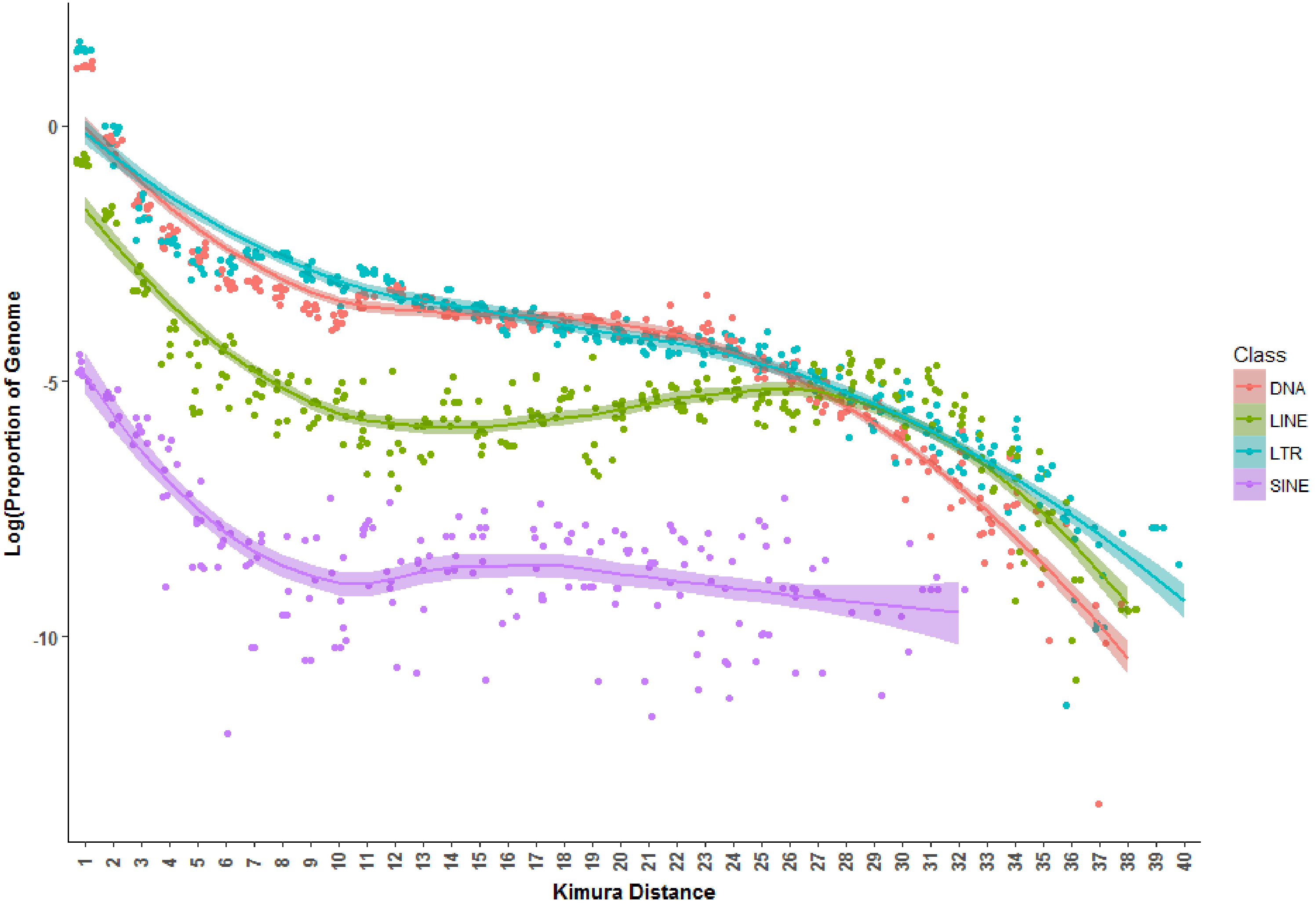
Transposons in Arabidopsis



Transposons in Green Plants



Transposons in Arabidopsis



Conclusions

Repbase has the widest library of TE subfamilies

- 23 found in A.thaliana, < 15 found in others

REdat has strong LTR content

- By copy number, REdat LTR > All Repbase

High amount of overlap is existing TE libraries

- 22% on merged; 19%, 15%, 14% on others

Conclusions

Correlation between genome size and TE content

- Pearson’s Test (*t* = 10.389, P << 0.005)

Correlation between genome size and LTR content

- Pearson’s Test (*t* = 5.18, P < 0.005)

Evidence for TE transfer

- *DNA/Kolobok* in *M. polymorpha*, *DNA/Merlin* in *C. quinoa*

Evidence for TE extinction

- *LINE/Cassandra* in *S. tuberosum*, *S. lycopersicum*

TE content is not reflective of phylogeny

- *G.max, R.communis* removed from other Fabids

Conclusions

Map distances may not reflect genomic differences

- Japan and Libya show minimal separation

Arabidopsis has recent common ancestor

- Equal genome sizes, minimal divergence for TEs

Cited mutation rates may be incorrect

- 10^{4} -10^{6} rate not reflected in Arabidopsis

Future Work

Develop TE discovery pipeline with Red

- Use existing libraries as training data

Analyze RepeatModeler TEs

- Compare against known sequences (TE, coding, ncDNA)

Create a tool to generate non-overlapping libraries

- Combine and cluster sequences to maximize efficiency

Future Work

Expand search for TE transfer mechanisms

- Search for individual subfamilies across unrelated species

Investigate TE contribution to speciation

- Detect bursts of transposons

Analyze extremely large genomes for trends

- *Paris japonica* (130 Gbp)

Future Work

Expand Arabidopsis data

- Find relict individuals (pre-1700’s)

Compare other equations for evolutionary change

- Include for insertions and deletions

Detect transposon activity

- Search parent/offspring for increased copy numbers