De Novo Discovery and Comparison of Transposable Element Families in *S. lycopersicum* and *S. pimpinellifolium*

Kristin Blacklock, Jeremy Edwards, Lukas Mueller

Boyce Thompson Institute for Plant Research

Introduction

Transposable elements (TEs), also known as "jumping genes", are sequences of DNA capable of changing their relative position in the genome of an organism, either autonomously or via an autonomous "activator" element[1]. Their discovery in the 1940s is credited to maize geneticist Barbara McClintock, whose suggestions of TE functionality were dismissed for decades thereafter. Recently, however, researchers have discovered several important aspects of TEs, including their roles in regulation[2], contributions to the structure of the genome, and ability to create, modify, and re-wire regulatory networks[3].

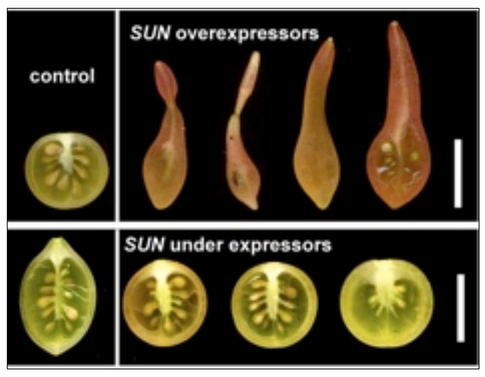

Fig. 1: Tomatoes with SUN gene turned on and knocked out. Photo courtesy of Ohio State University.

In the SUN gene of the domesticated tomato, an unusual retrotransposon, *Rider*, has been discovered to play an important role in fruit morphology phenotypes when activated[3]. Therefore, TEs may also have been integral in the speciation between the domesticated tomato (*Solanum lycopersicum*) and its wild ancestor (*Solanum pimpinellifolium*), and so the identification of putative new TEs absent from *S. lycopersicum* and *S. pimpinellifolium* reference databanks may be of particular interest for the advancement of tomato research .

The objectives of this study were to implement a pipeline to discover and classify transposable elements in the domesticated tomato genome, and to identify potentially active transposable element families by copy number, within-family sequence similarity, and indels in alignments with *S. pimpinellifolium*.

Materials

- A Lenovo ThinkPad laptop running Debian (Linux) was used throughout this project.
- The REPET Pipeline[4], a *de novo* TE discovery script, was used to find consensus sequences in the tomato genome.
- The genomes for *S. lycopersicum* and *S. pimpinellifolium* were obtained from the Tomato Genome Consortium[5].
- The Perl programming language was used in the filtering and comparison steps of this study.
- BLAST+[6] was used as the main local alignment tool, and FigTree[7] was used to visualize the LTR tree.

Methods

I. REPET Pipeline

The REPET package is an efficient *de novo* transposable element discovery pipeline that integrates several bioinformatics tools in order to answer biological questions at the genomic scale. Its two main components, TEdenovo and TEannot, are dedicated to the detection and analysis of repeats in genomic sequences, specifically designed for TEs. This study utilized only the TEdenovo pathway.

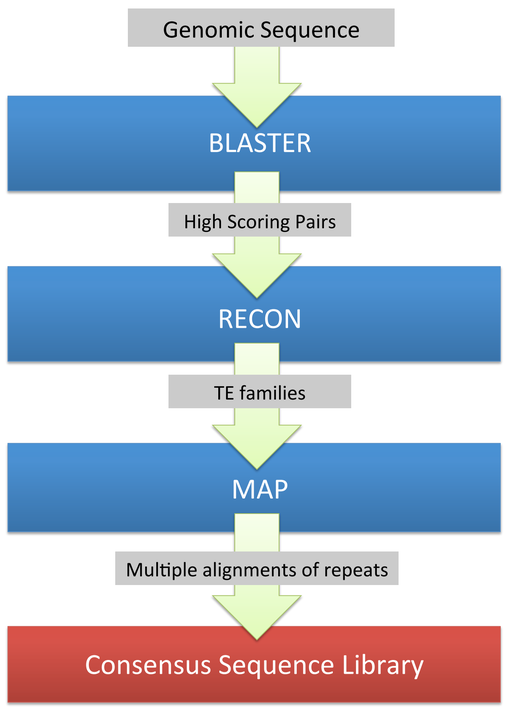

Figure 2: TEdenovo pipeline of the REPET package.

The TEdenovo pipeline began by comparing the genome with itself using BLASTER.

Then, it clustered matches with GROUPER, RECON, and PILER, which are clustering programs specific for interspersed repeats. This study implemented only RECON.

For each cluster, it built a multiple alignment from which a consensus sequence was derived.

Finally these consensus were classified according to TE features, and redundancy was removed. In the end, a library of non-redundant consensus sequences was obtained.

II. Filtering Pipeline

The consensus sequences were then run through a second pipeline, which determined TE families of interest.

First, the consensus sequences obtained for *S. lycopersicum* from the REPET pipeline were blasted against the *S. lycopersicum* genome.

Then, the blast results were filtered based on a percent identity greater than 98% and an aligned length greater than 200bp.

This resulted in a more specific TE family library representing TEs that are more active in the genome.

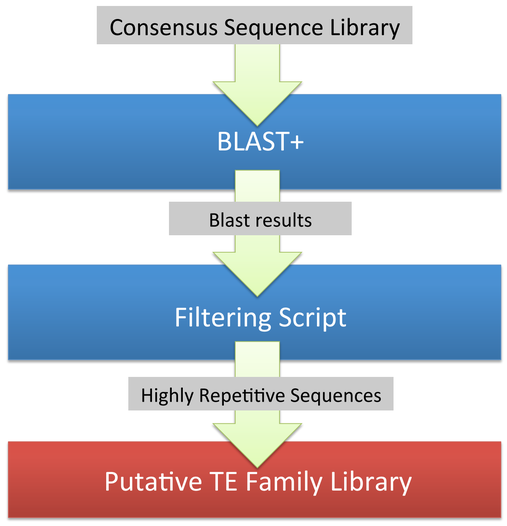

Figure 3: Filtering pipeline used to extract the most active TE families from the consensus sequence library.

III. Comparison Protocol

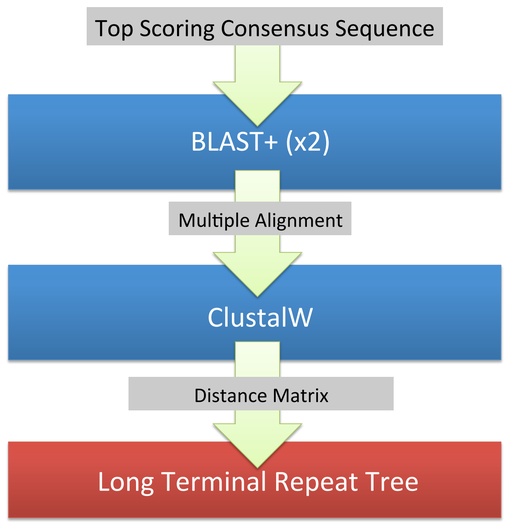

Figure 4: Comparison pipeline used to visualize significance of TE family library.

The consensus sequence with the greatest number of passable blast hits was then blasted against itself to find LTRs, which were then blasted against the *S. lycopersicum* and *S. pimpinellifolium* genomes to create a multiple alignment.

A distance matrix was created from the multiple alignment.

A tree (Figure 5) was constructed from the distance matrix using the neighbor-joining method.

Results and Discussion

	S. lycopersicum Consensus Sequences	Number of hits matching criteria	TE Classification
1	tomato_Blaster_Recon_418_Map_20	317	Gypsy-39_STu-I:ClassI:LTR:Gypsy
2	tomato_Blaster_Recon_4998_Map_7	115	Copia-2_SL-LTR:ClassI:LTR:Copia
3	tomato_Blaster_Recon_1385_Map_20	97	Inconclusive
4	tomato_Blaster_Recon_1171_Map_20	84	Gypsy-7_CP-I:ClassI:LTR:Gypsy
5	tomato_Blaster_Recon_63_Map_20	79	Copia-2_SL-I:ClassI:LTR:Copia

Table 1: Top five consensus sequences in *S. lycopersicum* with passable blast hits.

The top consensus sequences derived from *S. lycopersicum* were found to contain mainly LTR-type transposons, one of which, …63_Map_20, matches the retrotransposon *Rider*. The four sequences outranking Rider indicate that more active transposons may be present in the tomato genome, which may also have contributed to the highly differential fruit morphologies between the tomato and its wild ancestor.

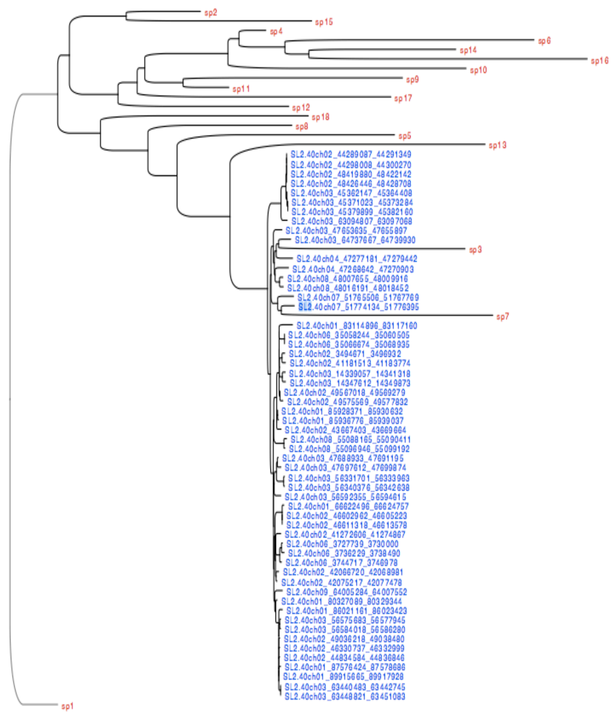

Figure 5: Tree depicting LTRs in *S. lycopersicum* (blue) and *S. pimpinellifolium* (red).

The comparison of the LTR in the top consensus sequence in Table 1 to both the *S. lycopersicum* and *S. pimpinellifolium* genomes indicates a clear burst of amplification of this LTR after the speciation or domestication of *S. lycopersicum*.

References and Acknowledgements

[1] Pray, L, Zhaurova, K (2008) Barbara McClintock and the discovery of jumping genes (transposons). *Nature Education* 1(1).

[2] Pray, L (2008) Transposons, or jumping genes: Not junk DNA? *Nature Education* 1(1).

[3] Jiang N, Gao D, Xiao H (2009) Genome organization of the tomato *sun* locus and characterization of the unusual retrotransposon *Rider*. *The Plant Journal* 60(1).

[4] Flutre T, Duprat E, Feuillet C, Quesneville H (2011) Considering Transposable Element Diversification. *De Novo* Annotation Approaches.PLoS ONE 6(1): e16526. doi:10.1371/journal.pone.0016526.

[5] Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485(7400).

[6] Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215: 403-410.

For support in this project, we would like to thank Boyce Thompson Institute for Plant Research, the Sol Genomics Network, and the National Science Foundation.

Special thanks to Joyce Van Eck for coordinating the Bioinformatics internship.