Genome Annotation of LTR Retrotransposons in *Trifolium repens*

Henry Wittich and Howard Laten

Background

Since the Human Genome Project launched in 1990, the cost and time it takes to sequence a species' genome has decreased dramatically. With the influx of genomic data being published on public databases comes the need for computational tools to analyze it.

One important feature of genomic sequences that warrants computational analysis is the location of repetitive DNA sequences. Transposable elements (TEs) are genetic sequences that have the ability to migrate throughout the genome. Retrotransposons, a class of transposable elements, accomplish this via a copy and paste mechanism, replicating to high copy numbers throughout many organisms' genomes. This is especially true in plants; for example, around 70% of the maize genome is estimated to consist of repetitive DNA sequences.[1]

Long-terminal repeating (LTR) retrotransposons are a specific class of TEs characterized by their long terminal repeating regions, a 100-500bp sequence that is repeated at the start and end of every element that can be identified computationally. All LTR-TEs contain protein-coding domains for protease, reverse transcriptase, and integrase, and the order of these domains can be used to classify elements into two superfamilies: *copia* and *gypsy*.

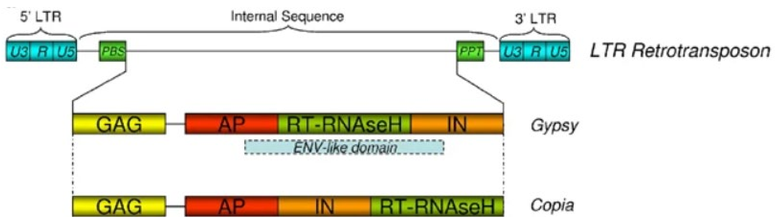

Fig. 1. Structure of largest superfamilies of LTR retrotransposons.[2]

Research has demonstrated that the location of insertion of a retrotransposon can influence gene expression, and they can even be responsible for domestication events in plants. *Trifolium repens* (white clover) is an allotetraploid that is theorized to be the result of a hybrid speciation event between *T. pallescens* and *T. occidentale*.[3] *T. repens* is also the subject of a current sequencing project, and thus in need of genome-wide sequence annotations. This project involves the genome annotation of LTR-TEs in *T. repens* for the benefit of future study of the ancestry of the species, and the potential role that retrotransposon insertions might have played in its speciation.

Methodology

Below I present the bioinformatics pipeline I developed for genome annotation of LTR-TEs in *Trifolium repens*. The pipeline combines several tools to accomplish three tasks: identify the locations of these repetitive sequences in the *T. repens* genome, classify them by superfamily, and compare insertion sites between *Trifolium* species to infer ancestry.

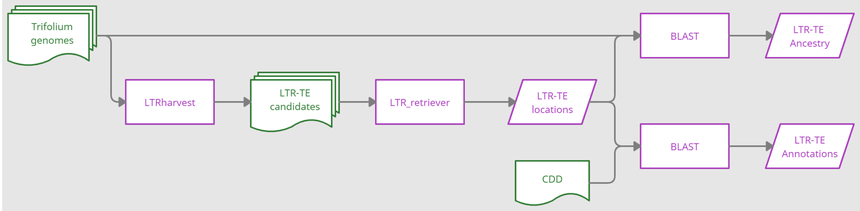

Fig. 3. Locations of LTR-TEs in the *Trifolium repens* genome.

The first step is to process the genomes of several *Trifolium* species (*T. repens, pallescens, and occidentale*) through the programs, LTRharvest and LTR_retriever, to identify all LTR-TEs. LTRharvest computationally searches for repeated sequences within around 15,000 base pairs of each other, and thus can be used to identify potential LTR retrotransposons. LTR_retriever is a program designed to filter down the output from LTRharvest using more stringent criteria – it searches for other common features of LTR-TEs like the target site duplication and the start/end motifs.

This list of LTR-TEs can then be searched against the protein conserved domains database (CDD) with the BLAST algorithm to identify the order of protein domains in the sequence, which is needed to classify them by family.

Finally, the list of LTR-TEs can be searched against the retrotransposons found in other related species to determine potential ancestral relationships.

Discussion

One limiting factor in this analysis is that the *Trifolium repens* assembly is still largely incomplete. There are significant gaps in the assembled chromosomes, and large scaffolds that have not been assembled. Given that LTR-TEs tend to cluster in non-coding regions like the centromeres, which are usually the most difficult to assemble, the low count of LTR-TEs discovered through this analysis makes sense.

Additionally, further analysis into the relationships between the LTR-TEs in *T. repens, T. pratense*, and *T. pallescens* is required to draw any conclusions about white clover's ancestry.

Results

LTRharvest identified 2,730 LTR-TE candidates in the *T. repens* assembly, which was filtered down to around 649 unique elements by LTR_retriever. 380 were identified as *copia* elements, 125 were identified as *gypsy* elements, and 144 could not be identified by BLASTing against the CDD.

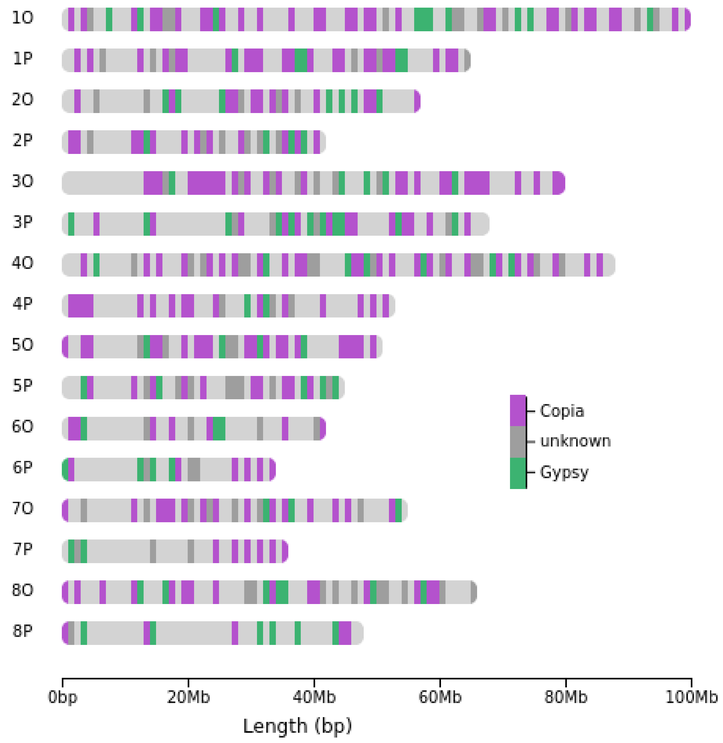

Fig. 3. Locations of LTR-TEs in the *Trifolium repens* genome.

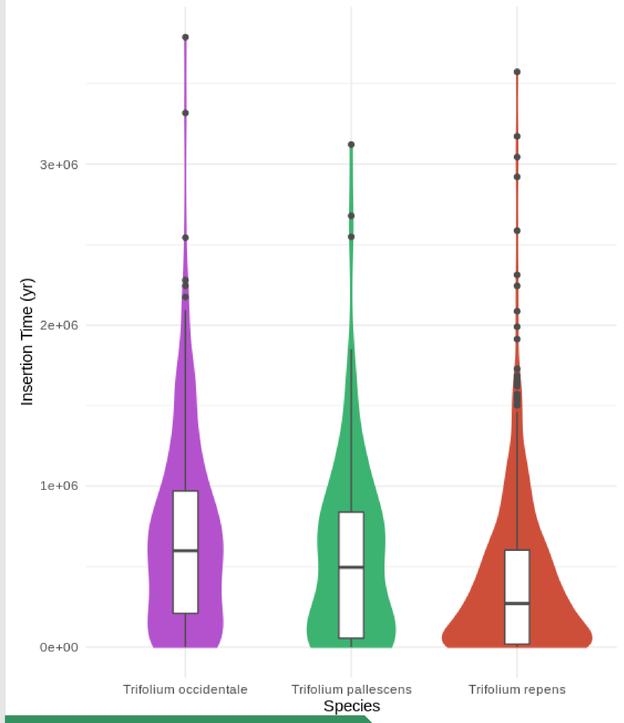

Fig. 4. Estimated insertion time in years of the LTR-TEs found in *T. occidentale*, *T. pallescens*, and *T. repens*.

Literature Cited

1. Biémont, C. (2010). A Brief History of the Status of Transposable Elements: From Junk DNA to Major Players in Evolution: Figure 1.—. *Genetics*, 186(4), 1085-1093.
2. Sabot, F., Schulman, A. Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. *Heredity* 97, 381–388 (2006).
3. Ellison, N. W., Liston, A., Steiner, J. J., Williams, W. M., & Taylor, N. L. (2006). Molecular phylogenetics of the clover genus (Trifolium—Leguminosae). *Molecular Phylogenetics and Evolution*, 39(3), 688-705.

Acknowledgements

- Members of the Laten Lab: Jasen Jackson, Zain Anwar, Abdullah Mazher, and Rohan Rajagopal
- Dr. Howard Laten, my research mentor
- Dr. Andrew Griffiths and the *T. repens* sequencing lab
- Provost fellowship program

Contact

hwittich@luc.edu