TE demography simulation project write-up

Project Description

Zea mays subsp. mays, commonly known as maize or corn, was domesticated from the wild grass teosinte 9000 years ago. Over 80% of the maize genome is composed of transposable elements (TEs), which are genetic elements that move within a genome. During domestication, maize experienced significant population bottlenecks that shifted TE dynamics. Using the population genetics software SLiM, I simulated TE dynamics in maize to study their evolution during domestication and determine the effects of population bottlenecks on genetic differences between maize and teosinte. My simulation models TE dynamics in an ancestral population of teosinte through a population bottleneck and subsequent expansion. When designing my simulation, I considered differences in TE activity, biology, and transposition mechanisms. I also incorporated empirical estimates of TE mutation rates, disabling, and effects on fitness. I modeled epigenetic silencing to include realistic genomic responses to increased TE activity. My simulation reached equilibrium where TE diversity stabilized, with most TEs being deleterious. These results can be used to identify the most important characteristics of TE biology in plant genomes. My simulation can also provide insight into the evolutionary history of maize, an agriculturally important crop.

In this simulation, the following parameters were incorporated: TE disabling, epigenetic silencing and unsilencing, replication/propagation, and conversion of autonomous TEs to non-autonomous TEs.

Goal

Our goal is to simulate transposable element dynamics in the maize genome over the past 9,000 years of domestication. We also want to answer whether bottlenecks can explain some of the differences in transposable element (TE) composition between maize and teosinte by comparing our simulated results to differences in real life.

Model setup

To model the maize genome, we used recipe 14.12 in SLiM (v4) as a basis. This recipe was designed specifically for modeling transposable elements. The individuals in the population are diploid, hermaphroditic, and sexually reproducing to best represent the genomic structure of maize. The simulation runs for a total of 3000 generations/cycles. We first introduced 100 autonomous and neutral TEs into the simulation across a population of 1500 individuals. Alht population of 1500 was chosen as this was chosen due to computational constraints. Only 100 TEs were introduced to the population to see how TEs are able to populate a genome. Each TE that was inserted has a 0.001 probability of jumping across the genome (teJumpP), a 0.005

probability of being disabled (teDisableP), and a 0.05 probability of being epigenetically silenced (teSilenceP). These values are based on metrics found in the current literature on maize transposable element dynamics [1, 2, 3, 4]. In this simulation there are eight different types of TEs that will be specified in the next section. Although in general there are many types of TEs with different properties (LTRs, LINES, etc), we do not model the specific families of TEs in this simulation, however the TEs are assumed to be retrotransposons. This is because the majority of TEs in maize are Long Terminal Repeats (LTRS), which are retrotransposons. Retrotransposons are able to copy themselves through reverse transcription. As a result, in this simulation, a TE is able to propagate by replicating/copying itself. During the course of this simulation a bottleneck is introduced early on at generation/cycle 500 in order to model the bottleneck(s) that occurred in Maize during its domestication. After the simulation has finished running, various statistics generated from the simulation are printed.

Genome Architecture

We initialized the genome with one genomic element that includes 6 different mutation types. Each mutation type represents a different type of TE: mutation type m1 represents autonomous and neutral TEs, m2 represents autonomous and deleterious TEs, m3 represents autonomous and advantageous TEs, m4 represents non-autonomous and neutral TEs, m5 represents non-autonomous and deleterious TEs, m6 represents non-autonomous and advantageous TEs, m7 represents disabled TEs, and m8 represents epigenetically silenced TEs. Autonomous TEs are TEs with functioning transposition machinery, which means they are able to copy themselves to another position in the genome. Non-autonomous TEs are unable to transpose on their own, and need to use the transposition machinery of nearby autonomous TEs to copy themselves. Currently, non-autonomous TEs are unable to transpose at all in this simulation for simplicity, but that may change with further developments to the program. Neutral TEs do not affect the evolutionary fitness of individuals, deleterious TEs negatively affect the fitness of individuals, and advantageous TEs positively affect the fitness of individuals. In the beginning of the simulation, one genomic element is initialized with m1, m2, m3, m4, m5, and m6 mutations distributed proportionally (1.0, 1.0, 0.01, 1.0, 1.0, 0.01 respectively). Advantageous TEs are less prevalent in the genomic element because they are generally very rare. The recombination rate for this genomic element is 1e-8, and mutation rate for transposons is 1e-7. These rates are based on the literature [1, 2, 3, 4]. The nucleotide sequences of the TEs are not modeled in this simulation, instead each TE is depicted as an individual unit in the genome that occupies a single position.

TE transposition

Autonomous TEs have a jumping probability, teJumpP, of 0.001. Although this parameter is specified as the jumping probability, this parameter is akin to a TE copying itself to another

position within the genome. This value is fixed for the entire simulation. Only autonomous TEs (mutation types m1, m2, and m3) are able to jump to another position in the genome as they have their own transposition machinery. The jumping process works by first determining jumpCount, or the number of TEs that will copy itself to another position in the genome. This value is based on a Poisson distribution where $\lambda = teCount * tejumpP$. teCount is the total number of autonomous TEs. We then take a sample of a jumpCount number of autonomous TEs. For each TE in the sample, we add a new TE of the same mutation type at a new position in the genome. This models the original TE copying itself to a different part of the genome.

TE disabling

Disabling a TE in this simulation removes the ability for a TE to copy itself and renders a TE completely inactive. Both non-autonomous and autonomous TEs can be disabled. TE disabling works similarly to TE transposition; it takes a sample of TEs based on a Poisson distribution with lambda equal to teCount times teDisableP. For each TE in the sample, the current mutation type is removed, and is replaced by the m7, or disabled mutation type.

Autonomous to non-autonomous TE conversion

Converting a TE from autonomous to non-autonomous in this model works almost exactly the same as TE disabling. The only difference is that removing transposition machinery takes a sample of TEs of only autonomous TEs and converts them to a corresponding non-autonomous TE. For example, we could take the m3 mutation type (autonomous and advantageous) and convert it to the m6 mutation type (non-autonomous and advantageous). It is important to note that a non-autonomous TE is still active, and is different from a disabled TE.

TE silencing

In this model we epigenetically silence a portion of the TEs in the genome. A silenced TE is indicated by the silenced mutation type, m8. TE silencing also works nearly identical to TE disabling. A TE of any mutation type can be silenced. The major difference is that epigenetic silencing only begins once the number of TEs in an individual exceeds a threshold. This is similar to how epigenetic silencing works in real life. In this model we determined that threshold to be when teCount for an individual exceeds 40. Additionally, the simulation checks if 20 generations/cycles have passed before silencing each TE, as newer TEs are less likely to be silenced. If these conditions are met, then a sample of TEs are taken based on a Poisson distribution where lambda equals the total TE count in an individual multiplied by the probability of being silenced (0.005). The TEs in this sample are then silenced and are given a new mutation type of m8. Additionally, there is a 20 percent chance that an epigenetically silenced TE is reactivated with each cycle, which mirrors observed patterns in TE biology.

Bottleneck

To model the demography of maize, this simulation incorporates a bottleneck towards the beginning of the simulation at generation/cycle 500. This is due to the fact that maize underwent a population bottleneck early on in its domestication. The mechanism for this bottleneck was adapted from Beissinger et al. 2016's demographic model. It is downscaled 10x and does not consider gene flow with teosinte. For the bottleneck, the population is set to 5 percent of the original population of 12278, or 646. Then exponential population growth is incorporated into the model with the formula: $646 \times (1 + 0.0025943665)^{t}$. The values represent exponential growth in maize and are adapted from Beissinger et al. 2016, and where $t = current\ simulation\ cycle - 1$. In this simulation, exponential occurs from cycle 501 to cycle 1500. For more information on the model and the values used, please refer to Beissinger et al. 2016 [1].

Model data

After the simulation has completed running, TE statistics of the population are printed. These statistics include the total number of all types of TEs as well as the number of autonomous, non-autonomous, disabled, epigenetically silenced, advantageous, neutral, and deleterious TEs. Additional statistics that are printed are divided by pre-bottleneck and post-bottleneck for comparison purposes. These statistics include the percentage of silenced, disabled, active, neutral, advantageous, deleterious, autonomous, and non-autonomous TEs in the population.

Additionally, for autonomous TEs, we print its position in the genome, frequency, and the percentage of autonomous TEs that are active (not disabled). For disabled TEs, the position and frequency are listed.

Additional statistics calculated include the mean and standard deviation of the age of all TEs, the mean age of autonomous TEs, and the mean age of non-autonomous TEs.

Model Outcomes

For our preliminary results, we found that pre-bottleneck, there are more active and neutral TEs present in the population. There are also less epigenetically silenced and disabled TEs compared to post-bottleneck statistics. This aligns with what I read in the literature as over time, TEs are expected to be disabled in order to minimize the deleterious impact of TEs on the population. More neutral TEs could be explained by the fact that the TEs introduced to the population were all neutral. Because the bottleneck was early on in the simulation, deleterious TEs may not have had enough time to arrive in the population. We also found that post-bottleneck, the majority of

TEs are epigenetically silenced, and the majority of active TEs are deleterious. This aligns with what is seen in maize, as most TEs are epigenetically silenced and deleterious. I ensured that the parameter values for epigenetic silencing and deleterious TEs were much higher than other TE types to yield this outcome. One fault with the simulation may be the fact that the bottleneck was placed too early on in the simulation. This could have caused there to be less epigenetically silenced and more active TEs due to the fact that not enough generations had passed to introduce other TE types. It is unclear whether this accurately models the domestication of maize.

Next Steps

Next steps would be to incorporate more realistic TE dynamics. For example, in this simulation TEs do not occupy base pairs but rather have positions in the genome. Incorporating the length of TEs in base pairs may yield more accurate results. Additionally, in this simulation non-autonomous TEs are unable to transpose at all, however non-autonomous TEs in real life are able to transpose when close to autonomous TEs by utilizing autonomous TE transposition machinery. Moreover, the statistics generated from my simulation can be compared to real-life data from maize to see the accuracy of the simulation. Based on the real-life data, parameters in the simulation can be adjusted.

Acknowledgements

Thank you to Natasha Dhamrait, Regina Fairbanks, and Dr. Jeffrey Ross-Ibarra for support and guidance on this project.

Works Cited

1. Beissinger, T., Wang, L., Crosby, K. *et al.* Recent demography drives changes in linked selection across the maize genome. *Nature Plants* 2, 16084 (2016). https://doi.org/10.1038/nplants.2016.84; downscaled 10x

2. Eyre-Walker, A., Keightley, P. The distribution of fitness effects of new mutations. *Nat Rev Genet* 8, 610–618 (2007). https://doi.org/10.1038/nrg2146

3. Gozukirmizi, N., Temel, A., Marakli, S., & Yilmaz, S. (2016). Transposon activity in plant genomes. *Plant Omics: Trends and Applications*, 83–108. https://doi.org/10.1007/978-3-319-31703-8_4

4. Le Rouzic, A., Boutin, T. S., & Capy, P. (2007). Long-term evolution of transposable elements. *Proceedings of the National Academy of Sciences*, *104*(49), 19375–19380. https://doi.org/10.1073/pnas.0705238104

5. Haller, B.C., and Messer, P.W. (2016). SLiM: An Evolutionary Simulation Framework. URL: http://benhaller.com/slim/SLiM_Manual.pdf

6. Haller, B.C., and Messer, P.W. (2023). SLiM 4: Multispecies eco-evolutionary modeling. The American Naturalist 201(5). DOI: https://doi.org/10.1086/723601