Solanaceae Chloroplast tRNAs: Sequence Variation and Gene Transfer to the Nucleus

Dennis Mulligan
BME 230A
March 18, 2019

Abstract

Plant chloroplasts have a small, independent genome, that includes several transfer-RNAs (tRNAs) among its small set of genes. I compiled tRNA genes from chloroplasts of dozens of species within the Solanaceae family, and evaluated the sequences of the gene and the surrounding region. I compared sequences between species and found differential levels of variation along these gene regions. The two flanking regions on either side of the genes often had very separate levels of identity to a chloroplast consensus sequence, suggesting that these regions may be prone to significant sequence change events.

I then compared the chloroplast tRNA (cp-tRNA) sequences to tRNAs found in the nuclear genomes. There was evidence of homologous genes in the nucleus, likely the result of gene transfer. Interestingly, the species *Capsicum annuum* had a far higher proportion of these potentially homologous genes than another species, *Solanum lycopersicum*.

To observe the conservation across the chloroplast genome, I used minimizer seeds, which I compared between genomes. From this it was apparent that the tRNA genes often lay at the boundaries of regions with different levels of homology, and in many cases the immediate region on one or both sides had a low level of matched seeds.

Comparing seeds between chloroplast genomes and a nuclear chromosome, I observed sequences of nearly 7 kilobases which were clearly homologous. There was a possible relationship between tRNA gene locations and the these homologous regions.

Further work is needed to quantify the preliminary results observed here.

Introduction

Chloroplasts are an essential organelle in plant cells that contain the material to conduct photosynthesis. Evidence strongly supports the theory that they are derived from a prokaryotic cell which was incorporated into a eukaryote cell in at least one endosymbiotic event. The circular genome of the chloroplast remains distinct from the nuclear genome, and in modern plants is only around 150 kilobases long (Shinozaki et al. 1986). The majority of prokaryotic genes were transferred to the nucleus, with only those which are essential for chloroplast function remaining in the organelle's genome (Timmis et al. 2004).

The approximately 120 genes in the chloroplast include the proteins which make up the photosynthetic machinery, ribosomal RNA and some of the ribosomal, and transfer RNAs (tRNAs), which perform the important role of connecting specific amino acids with their corresponding codon sequence on a strand of messenger RNA in a ribosome. Aside from being needed in large amounts in the chloroplast, it is believed that RNA products like tRNAs and rRNAs cannot be imported into the plastid like proteins and other molecules.

Chloroplast genomes are well characterized; the first was sequenced from *Nicotiana tabacum* in 1986. These genomes have been found to be highly conserved in structure and organization, and are conserved in many gene sequences. Sequence variation in particular regions has been used for phylogenetic analyses, and can also be used to identify specific commercial breeds of a crop.

Recent studies have shown, that due both their importance and their high degree of transcription, tRNAs experience significant transcription-associated-mutagenesis, which along with selection against deleterious mutations, has uniquely shaped the loci of tRNA genes (Thornlow, et al., 2018). It is unknown how this will manifest in a genome like the chloroplast, with a small number of genes, all of which are essential.

Researchers have been able to identify 30 tRNA genes in the chloroplast since the first complete sequencing. Despite this, the variation of tRNA genes in these genomes has not been extensively analyzed.

A study on *N. tabacum* estimated the rate of gene transfer to be 1 event in 16,000 pollen grains (Huang et al. 2003). This suggests tRNA genes are likely to be repeatedly transferred to the nucleus, but the rate of transfer of these genes is unknown.

Analysis

Here, I present an analysis of chloroplast genomes from the Solanaceae family, a widely distributed taxon of flowering plants containing a variety of species. Many members of this family have been well studied, and it includes several significant agricultural crops and model organisms.

Genome Sequences

All sequence data was acquired from NCBI's GenBank database. I had sequences for the chloroplasts from 115 plant samples, 55 of which were from the genus Solanum, and 36 from Capsicum. For some species there were multiple samples (accessions), such as *Solanum lycopersicum* and *Capsicum annuum*, which each had seven.

I gathered the sequences for all 12 nuclear chromosomes for four species, *C. annuum*, *C. baccatum*, *S. lycopersicum*, and *S. pennellii*.

For a comparison with bacterial genomes that might resemble the ancestral prokaryote of the chloroplast, I used four genomes; two from photosynthetic cyanobacteria: *Arthrospira platensis* and *Cyanobacterium apponium*; one from a photosynthetic bacteria from a related phylum: *Chloroflexus aurantiacus*; and one from *Escherichia coli* as an unrelated speices.

Detection of tRNAs

I identified the tRNA genes in all the genomes of my dataset using a tool called tRNAscan-SE 2.0, (Lowe and Chan, 2016). This was run in "bacterial" mode for all genomes, to encourage the identification of tRNAs of prokaryotic origin. Through testing I found that the mode had a small effect on the tRNAs identified in nuclear genomes, and affects the amino acid the program assigns to some tRNAs; important for proper grouping.

In nearly every chloroplast, tRNAscan-SE identified 29 tRNAs, the positions of which were very similar between chloroplasts. A small number of chloroplasts had additional low-scoring tRNAs, or had a missing tRNA near position-zero.

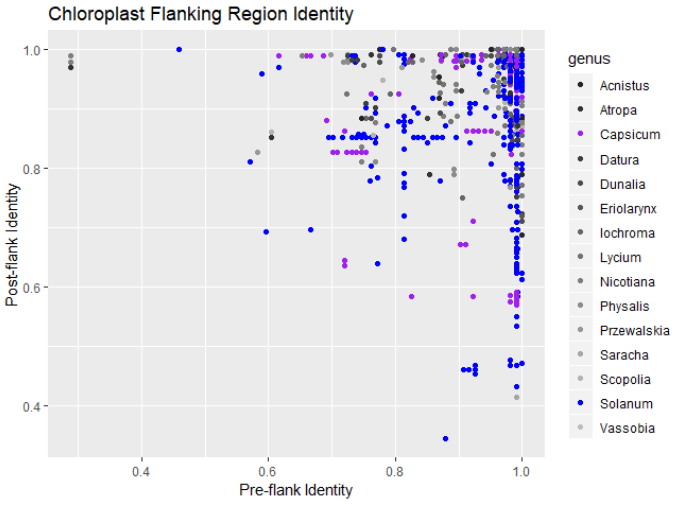
In the photosynthetic bacteria, I identified 40 to 50 tRNAs, and in E. Coli there were 89. In the nuclear genomes, there were between 1000 and 1500 tRNAs for each of the four species.

For each tRNA, I collected the gene sequence, and the sequence of 100 nucleotides on either side of the gene. I stored these, along with information about the tRNA, in a SQLite database.

Sequence identity in cp-tRNAs

To measure conservation of sequences, I grouped tRNA genes by the anti-codon and amino acid, as determined by tRNAscan-SE. I then used Clustal Omega (Sievers et al., 2011) to perform alignments on the total sequence, and then each of the three regions separately - the tRNA gene, the pre-flank, and the post-flank. From each of the three regions, I found a consensus sequence from the chloroplast tRNAs. I then quantified the level of identity for every tRNA, using the proportion of bases where that tRNA matches the chloroplast consensus.

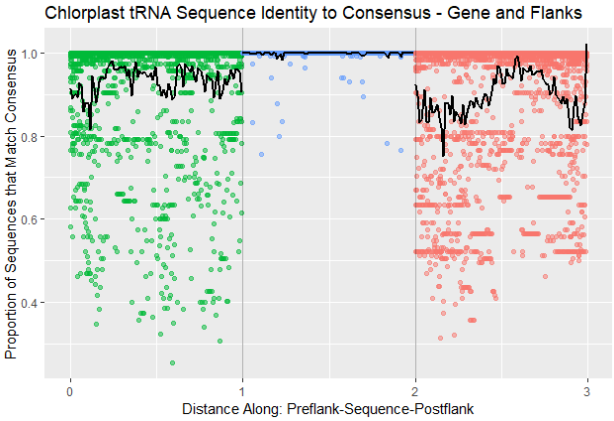
Across all chloroplast genomes, the 29 tRNAs gene sequences were highly conserved, with a mean identity of 99.89%. The pre-flanking region was less conserved, with mean identity of 94.89%, and the post-flanking region had even lower mean identity, at 90.48%. In the figure below, the level of identity is compared for the pre-flanking and post-flanking region for each tRNA gene.



Within a genus, there are strata of identity, where one region has differential identity while the other remains constant. There are also a large number of tRNAs with high pre-flanking identity, and lower post-flanking identity. When viewing the alignments, some

of these are clear cases where the end of the tRNA gene represents a distinct delineation.

For each alignment, I calculated the number of sequences that matched the consensus sequence for every position along each region. I scaled these positions to compare the identity across the whole length of the gene locus. The figure below shows the proportion of matching sequences across the pre-flanking region, the tRNA gene region, and the post-flanking region.

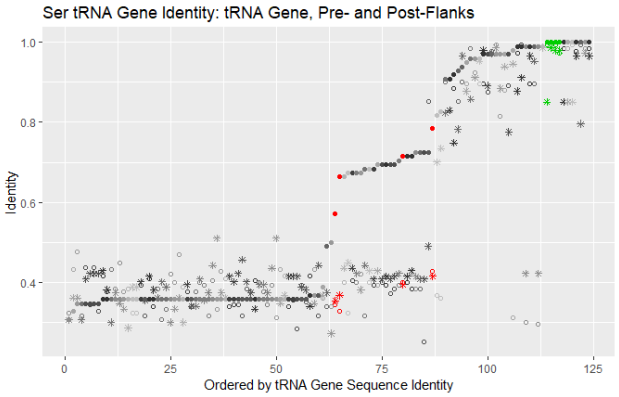


There is evidence for high levels of purifying selection that make the tRNA gene region extremely conserved. The variation in sequences in the two flanking regions suggests that these loci may experience transcription associated mutagenesis. Without good comparisons to other regions of the chloroplast genome, its hard to determine whether these regions have elevated substitution rates, or if they are simply not selected on to the degree that tRNA genes. The depth of the troughs in the flanking regions may indicate where polymorphisms are more likely to occur; or they may be influenced by how represented different lineages are in the data set, and when the polymorphisms first occurred relative to the divergence of different lineages.

Comparisons Within a tRNA Gene Group

For the four species which I had nuclear genomes, I compared their tRNA genes from their chloroplasts and all 12, and the genomes from the four bacterial species.

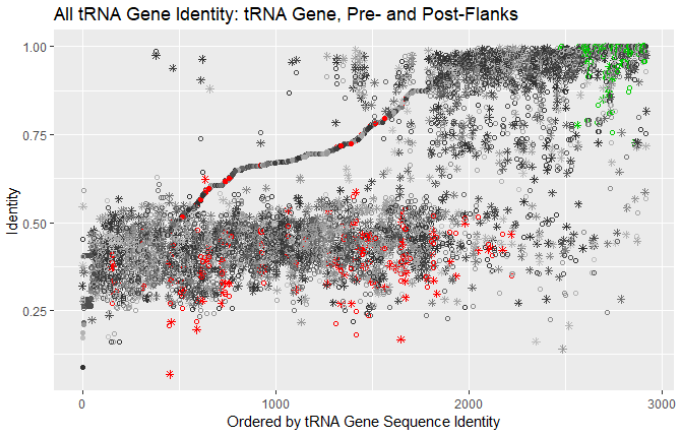
As before, I determined the identity with respect to the chloroplast consensus, with each region separate. For many tRNA groups, there were obvious homologous and non-homologous sequences. The figure below shows serine tRNAs, from chloroplasts (green), bacterial (red) and the nucleus (gray). They are plotted by quantile of the tRNA gene identity. There are distinct strata of homology, with genes that match the chloroplast tRNAs occurring at the top with a decline towards low homology. Then there is a band in the middle of sequences, possible a separately homologous group; and finally a band at the bottom, with a consistent level of identity.



● tRNA gene * Pre-flank O Post-flank

Chloroplast **Bacteria** **Nuclear**

In a plot of nearly all the tRNA genes and their identity to the consensus chloroplast, we see the different clusters of identity level. There is a steady change in identity for the actual tRNA gene sequences, due to conserved features of tRNAs, while the flanking sequences often fall within two ranges of high or low identity.

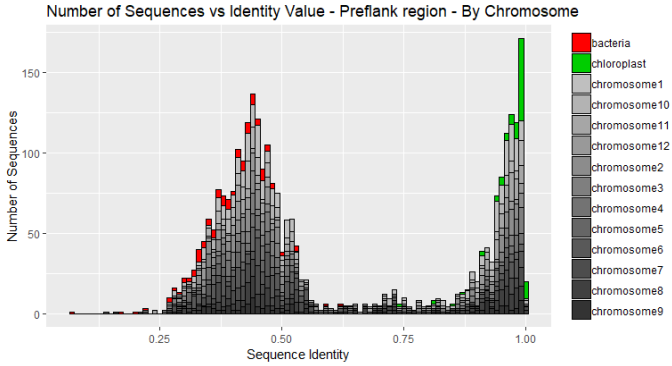


Levels of Identity: Gene and Flanking Regions

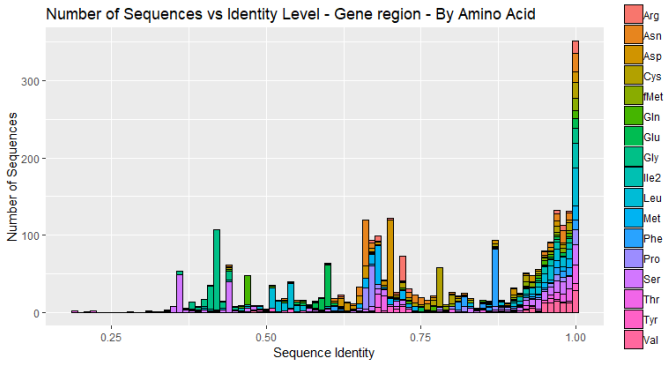
To explore some of these properties further, I created histograms of the number of sequences with a given sequence identity. Each of the three regions of

the loci are on separate histograms, with coloring to highlight particular features of the tRNA collection.

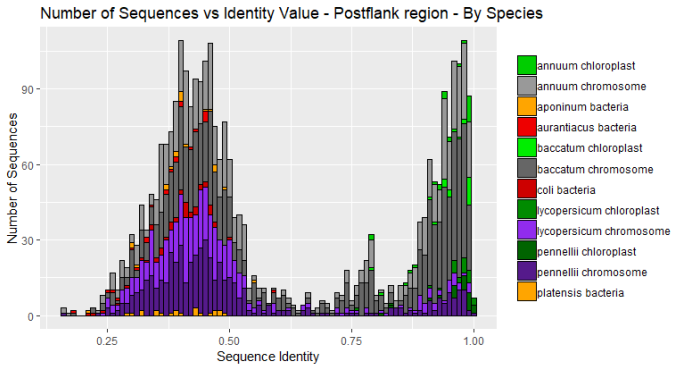
In the first histogram showing the pre-flanking region, the bars are colored by which chromosome the sequence originated from. We see the two distinct peaks: sequences that have the identity of a randomly aligned sequence (around 0.4), and sequences with very high identity(<0.9). The chloroplast and bacterial chromosomes appear where they would be expected, with the nuclear chromosomes exhibiting two types of tRNA genes, those that are homologous, and those that are non-homologous to the chloroplast genes.



The second histogram presents the number of sequences with a given sequence identity for the tRNA gene itself. The bars have been colored by the amino acid that tRNAscan assigned to the tRNA gene. It shows a large peak at high sequence identity, which includes an very high peak at near-perfect sequence identity. There are several small peaks at different levels of sequence identity, which correspond to specific amino acids. These likely show tRNA genes which are shared across chromosomes, but are not homologous to the chloroplast chromosome. The gentle peak around an identity of 0.7 may indicate an average identity for non-homologous tRNAs; this would be higher than the flanking region due to conserved general features of tRNAs.



The final histogram shows levels of sequence identity for the post-flanking region. This histogram is colored by both the chromosome type and the species of origin. It is in many ways similar to the histogram for the pre-flanking region, but there are some notable differences. The peaks are lower, with sequences having a wider range of identity. There is also a small but distinct peak near the identity value of 0.8; because this peak includes a number of chloroplast sequences, it might indicate a redundant tRNA gene in the chloroplast that was aligned to a different consensus sequence, and which has also been transferred to the nuclear chromosomes.



The post-flanking region histogram also shows the different behavior of the nuclear chromosomes in *Capsicum* and *Solanum*. The peak at a lower sequence identity appears to contain roughly equal amounts of tRNA genes from both genera. For the other peaks, however, the two *Capsicum* species – *annuum* and *baccatum* – have many sequences in their nuclear chromosomes with high identity to the consensus chloroplast sequence, while the *Solanum* species have relatively few. This suggests that gene transfer mechanism may be far more active in *Capsicum* species. The genomes of these *Capsicum* species are more than three times the size of the *Solanum* species, perhaps coinciding with a general tendency to duplicate genes more frequently.

Whole genome comparison

My next stage of analysis sought to evaluate the extent to which the mutation rate of the regions around the cp-tRNAs was unusual relative to other places within the chloroplast genome. I also wanted to know the general rate of gene/sequence transfer from the chloroplast to the nucleus.

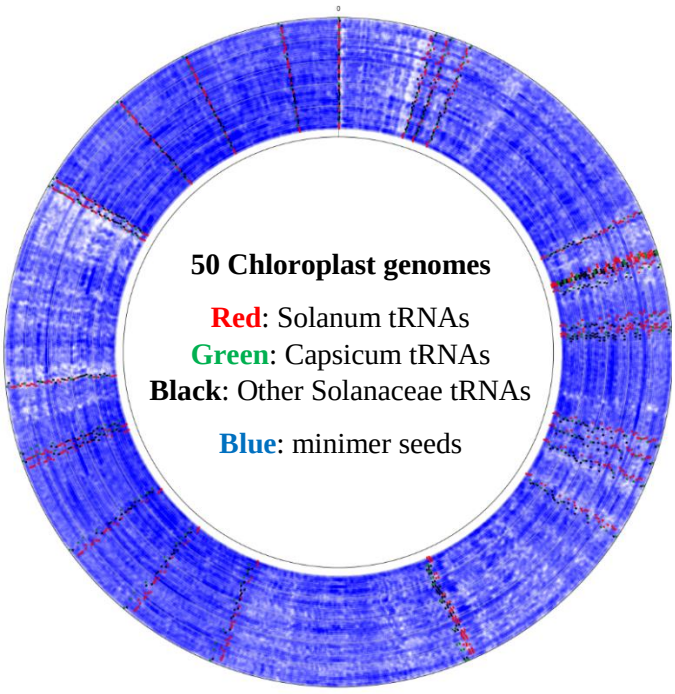
Though chloroplast genomes are relatively short, it is still a large task to compare the sequences across the entire genomes of many chloroplasts. And

investigating transfer to multiple nuclear genomes means aligning with several gigabases worth of sequence data.

I chose to use minimizer seeds, in which the lexicographically lowest kmer in each window of a given length is stored (Roberts, et al. 2004). This reduced the amount of sequence data that must be held in memory, and these minimer seeds can be compared between two sequences relatively quickly.

I used a fairly small kmer size of 13, hoping to capture sequences that may have mutated over time without purifying selection to keep the sequences conserved. This includes unused genes that have transferred from the chloroplast to the nucleus; and regions of the chloroplast that are not essential, such as the flanking regions of tRNA genes.

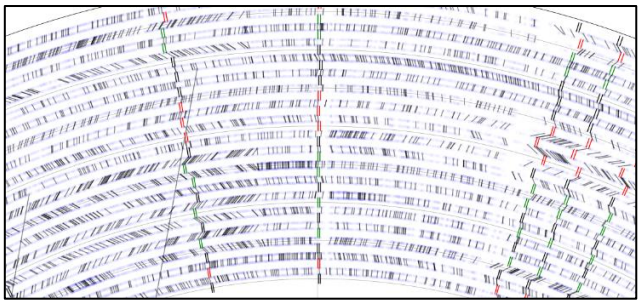
I used the minimer seeds to first make comparisons between whole chloroplast genomes. I randomly selected a list of 50 genomes from my dataset, and compared each genome to the ones above and below it in the list, recording minimer seeds that are shared between the two genomes. I plotted each genome in concentric circles in the figure below, showing the minimer seeds and locations of tRNAs.



The effect of this plot is to show a rough density of conserved elements within the Solanaceae family. Darker blue generally corresponds to regions that are more highly conserved with features shared between species, while in white you can see areas where there are frequent changes.

The location of tRNAs appear to be correlated with this conservation density, marking the boundaries between higher and lower conservation regions. And the lowest density of seeds occurs in regions immediately adjacent to some of the tRNAs; though other are in regions . It is unclear whether the tRNA genes themselves are causing this effect, or if the general structure of the genome has tRNAs placed in these regions.

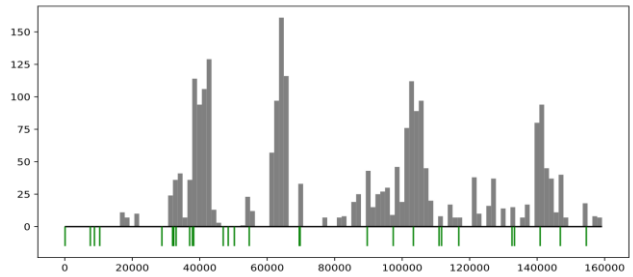
I also made a similar plot, with lines connecting randomly selected seed locations to the location in the neighboring genome. This makes it clear where insertion/deletion events may have occurred, depicted as a change in angle of these lines along a genome.



Comparison to nuclear chromosome

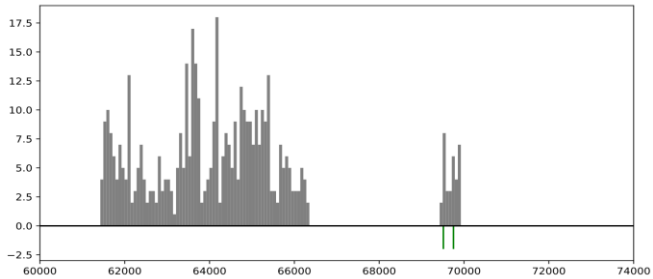
For *S. lycopersicum* and *C. annuum*, I compared the chloroplast genome with a nuclear chromosome (chromosome 8) using the same minimizer seeding method. For this I clustered the seeds

Below is a histogram of seed pairs, by their location in the C. annuum chloroplast genome. Locations of the cp-tRNAs are indicated with green lines below the histogram.

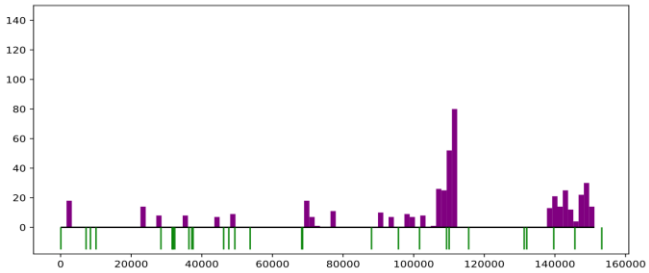


Certain regions have far greater numbers of seed matches, while other locations have nearly none. There

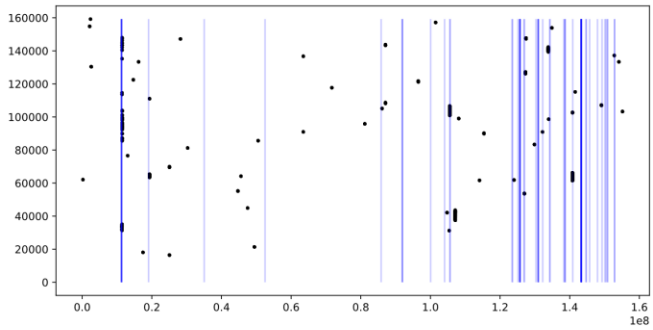
appears to be some relationship with tRNA locations, with some isolated peaks occuring above a tRNA or multi-tRNA area, though whether this effect is real is not obvious. A closer view at a region with one of these peaks is shown below.



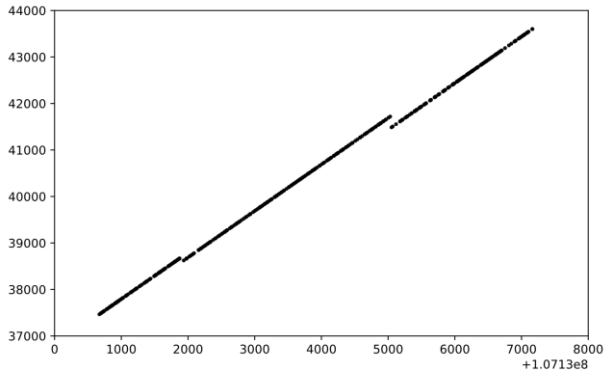
When S. lycopersicum is plotted this way (below), there are far fewer seed pairs, though s ome of this can be explained by smaller chromosome size (150 vs 65 megabases),



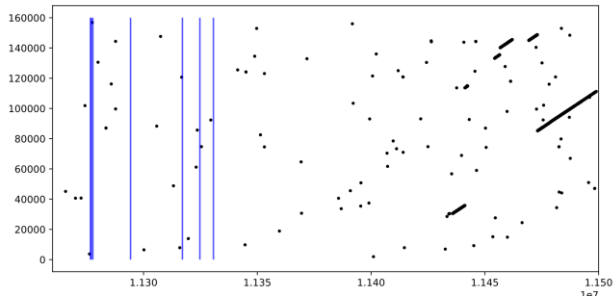
I next plotted the positions of the seed pairs in both nuclear chromosome 8 (x axis) and in the chloroplast (y axis), shown as black points. The nuclear tRNAs are plotted as blue lines in the plot.



Examining one of the closer of the spots on the plot with a large number of seeds shows clear homology between the sequences. This stretch of nearly 7000 bases has had little disruption, with only a couple moderate breaks in the sequences. The lack of a completely filled in line suggests that some level of point or small mutations causes non-matching minimer seeds across this sequence.

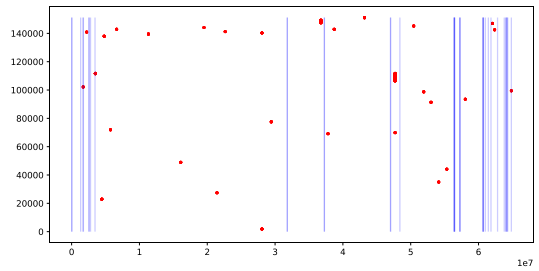


A closer view of the dark blue line on the left side of the graph, which has a number of seeds that overlap on this plot, shows some interesting structure.



Although on the larger plot the seeds appear to overlap with the tRNA lines, we see that they are separated by around 150 kilobases. The seeds appear to correspond to a range that spans most of the chloroplast genome, though segments have been rearranged in the nuclear genome. The tRNAs do not have many close seeds, but their proximity to the homologous sequences is interesting.

Below we see the full nuclear chromosome vs chloroplast genome seed plot for *S. lycopersicum*, which also has some clusters of seeds, but fewer than *C. annuum*, as expected.



Discussion

This work is the preliminary stages of an investigation into tRNA genes in plant chloroplasts. It shows potentially interesting results concerning these genes and their effect on sequence conservation. It also

shows the possibility that tRNA genes have an outsized effect on overall chloroplast genome archictecture, by serving as the boundaries of regions with distinct conservation levels.

The most surprising result was the levels of apparent tRNA gene transfer, with *Capsicum annuum* having a far higher number of homologous tRNAs than *Solanum lycopersicum*.

It is not clear whether these represent many gene transfer events, or a few events followed by duplication within the nuclear genome.

Next Steps

The results of this work are largely interpretation of visualizations of the data. The next steps are to quantify all the results, equivalently across all data in the dataset.

Seemingly homologous tRNAs in the nuclear genome were observed, but I need to classify each tRNA individually as to its likely homology to the chloroplast tRNA. Exact rates of gene transfer could then be estimated. It also might be possible to estimate how long ago each tRNA transferred to the nucleus.

Further studies could also account for phylogenetic relationships between samples. In many cases, all chloroplast tRNAs were treated equally, despite some species being over-represented.

The minimizer seeds give the potential food obtaining more precise alignments, which can also give alignment scores. This information would more accurately measure conservation across the chloroplast genome, and also measure the overall rate of homologous sequences between nuclear genomes and chloroplast genomes.

Conclusion

For this work, I developed a pipeline to go from plant a bacterial genome files, to identify and compile tRNA genes. The pipeline also processes them to prepare for a more in depth analysis, for which exploratory analysis has been done. Continuing this may help to understand an important family of plants, and could yield agricultural benefits.

Note: This project was initiated for a course, BME 232: Evolutionary Genomics. This paper uses some lightly modified text from the paper I wrote for that class (especially in the introduction). New analysis was performed for this class, BME 230A.

Citations

1. C Haberle, Rosemarie & Matthew Fourcade, H & Boore, Jeffrey & Jansen, Robert. (2008). Extensive Rearrangements in the Chloroplast Genome of Trachelium caeruleum Are Associated with Repeats and tRNA Genes. Journal of molecular evolution. 66. 350-61. 10.1007/s00239-008-9086-4.
2. Erixon, Per, and Bengt Oxelman. "Whole-gene positive selection, elevated synonymous substitution rates, duplication, and indel evolution of the chloroplast clpP1 gene." PLoS One 3.1 (2008): e1386.
3. Falcón, Luisa I., Susana Magallón, and Amanda Castillo. "Dating the cyanobacterial ancestor of the chloroplast." The ISME journal 4.6 (2010): 777.
4. Fujishima, Kosuke and Akio Kanai. “tRNA gene diversity in the three domains of life” Frontiers in genetics vol. 5 142. 26 May. 2014, doi:10.3389/fgene.2014.00142
5. Huang, Chun Y., Michael A. Ayliffe, and Jeremy N. Timmis. "Direct measurement of the transfer rate of chloroplast DNA into the nucleus." Nature 422.6927 (2003): 72.
6. Hubisz MJ, Pollard KS, Siepel A. PHAST and RPHAST: Phylogenetic Analysis with Space/Time Models. *Briefings in Bioinfomatics* 12(1):41-51, 2011. http://compgen.cshl.edu/phast/index.php
7. Jensen, Poul Erik and Dario Leister. “Chloroplast evolution, structure and functions” F1000prime reports vol. 6 40. 2 Jun. 2014, doi:10.12703/P6-40 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4075315/
8. Jo, Yeong Deuk, et al. "Complete sequencing and comparative analyses of the pepper (Capsicum annuum L.) plastome revealed high frequency of tandem repeats and large insertion/deletions on pepper plastome." Plant cell reports 30.2 (2011): 217-229.
9. Khan, Abdur Rahim, et al. "The whole chloroplast genome sequence of black nightshade plant (Solanum nigrum)." Mitochondrial DNA Part A 28.2 (2017): 169-170.
10. Lowe, T.M. and Chan, P.P. (2016) tRNAscan-SE On-line: Search and Contextual Analysis of Transfer RNA Genes. Nucl. Acids Res. 44: W54-57.
11. Michaud, M. , Cognat, V. , Duchêne, A. and Maréchal-Drouard, L. (2011), A global picture of tRNA genes in plant genomes. *The Plant Journal*, 66: 80-93. doi:10.1111/j.1365-313X.2011.04490.x https://onlinelibrary.wiley.com/doi/full/10.1111/j.1365-313X.2011.04490.x
12. Mohanta, Tapan K., et al. "Novel Genomic and Evolutionary Perspective of Cyanobacterial tRNAs." *Frontiers in genetics* 8 (2017): 200.
13. Provan, Jim, Wayne Powell, and Peter M. Hollingsworth. "Chloroplast microsatellites: new tools for studies in plant ecology and evolution." Trends in ecology & evolution 16.3 (2001): 142-147.
14. Roberts, Michael, et al. "Reducing storage requirements for biological sequence comparison." Bioinformatics 20.18 (2004): 3363-3369.
15. Rogalski, Marcelo, Daniel Karcher, and Ralph Bock. "Superwobbling facilitates translation with reduced tRNA

sets." Nature structural & molecular biology 15.2 (2008): 192.

16. Shinozaki, K., et al. "The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression." The EMBO journal 5.9 (1986): 2043-2049.
17. Siepel A and Haussler D. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol 21:468-488, 2004
18. Sievers F, Wilm A, Dineen DG, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 7:539 doi:10.1038/msb.2011.75
19. Thornlow, Bryan P et al. “Transfer RNA genes experience exceptionally elevated mutation rates” *Proceedings of the National Academy of Sciences of the United States of America* vol. 115,36 (2018): 8996-9001.
20. Timmis, J.N., Ayliffe, M.A., Huang, C.Y., Martin, W. (2004). “Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes”. Nat. Rev. Genet. 5: 123–135
21. Vogl, Claus, et al. "Probabilistic analysis indicates discordant gene trees in chloroplast evolution." Journal of molecular evolution 56.3 (2003): 330-340.
22. Wu, Zhiqiang. "The completed eight chloroplast genomes of tomato from Solanum genus." Mitochondrial DNA Part A 27.6 (2016): 4155-4157.
23. Yagi, Yusuke, and Takashi Shiina. "Recent advances in the study of chloroplast gene expression and its evolution." Frontiers in plant science 5 (2014): 61.
24. Zoschke, Reimo, and Ralph Bock. "Chloroplast Translation: Structural and Functional Organization, Operational Control and Regulation." *The Plant Cell* (2018): tpc-00016. http://www.plantcell.org/content/30/4/745