Avery Chua

Emily Fryer

Nicole Rodrigues

**Analysis of Nuclear Genome size as it relates Chloroplast Genome Size in Angiosperms**

**Background:**

DNA is the biological blueprint of all that is living on our planet. The basis of life is heavily and intricately coded among millions of species of organisms all done with the basic chemical composition that is nucleotide sequence. The understanding of these simple and yet essential aspects of life, can lead us to make discoveries about the underpinnings of the physiological and evolutionary processes of life. While increased rates of transpiration are largely attributed angiosperm dominance, Simonin (2017) posits that a decrease in genome size may play a key role in evolutionary fitness. Researchers have looked into scaling of cell size; however, little work has been done to investigate the relationship between cell size and genome size (Chan 2010). Simonin’s work has led to conclusions suggesting that angiosperms did in fact undergo a reduction in genome size, correlating to a decrease in cell size, and thus displaying an ability to increase stomata and vein cell density in leaves.  The increase in cell density of these tissues can correlate to increases the organisms rate of evapotranspiration (Simonin 2017). From these conclusions arise questions regarding the genome size of angiosperms relative to their chloroplast genome size.

**Hypothesis:**

The findings in Simonin’s work has led us to wonder if there is a relationship between plant genome size and chloroplast genome size. Based on our initial understanding of plant physiology and genomics, we expect to see an inverse relationship between chloroplast size and nuclear genome size in angiosperms. This should appear as plants with larger genomes having smaller chloroplast genomes and angiosperms with smaller genomes having larger chloroplast genomes.

**Description of methodology: Description of methodology – code written, publicly available data obtained (100-300 words, 15 points)**

The data we used in this project was extracted from two different websites to obtain data for the chloroplast genomes and nuclear genomes at data.kew.org and NCBI respectively. The data from each site was saved into separate .csv files. In the case of the chloroplast genome data, we need to merge two columns containing separate genus and species name data; the whole genome data had these two strings in the same field already. We created separate dataframes for each .csv files. The length of each dataframe was different, preventing us from merging the two without any conflict. We found that the merge() function could be utilized to easily cross reference our two dataframes by name, and dropping any species that not referenced in both lists. This yielded our working dataframe from which we were able to generate some statistical analyses in R. Basic statistical data was obtained including mean, standard deviation, variance and Pearson’s correlation coefficient. Generation of all graphs in this document was also done using the R programming language. We utilized the R packages ggplot2 and QuantPsyc to generate our statistical values.

**Results:**

General descriptive statistics about whole genome size and chloroplast size were generated as shown in Figure 1a. We also tested the normality of the data which can be found in Figure 1b. A test for normality of the raw genomic data yielded poor results suggesting that transformation of the data was needed. We performed a log transformation on both the whole genome data and the chloroplast data. The results yielded acceptable values of skewness and kurtosis for the whole genome data (p = 6.81 E-4, t-val = 3.2), suggesting normalization of the data. Log transformation of the chloroplast genome did not yield results reflective of normalized data. The overall normality of each log transformed data set can be observed in Figure 2. One can observe that the transformed chloroplast data is extremely skewed to the right as well as displaying a large peak at -1.9.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Descriptive Statistics for Genome Data | | |
| Mean | Standard Deviation | Variance |
| Whole Genome | 3668.412 | 8531.308 | 72783211 |
| Chloroplast Genome | 0.1497353 | 0.01652292 | 0.000273007 |

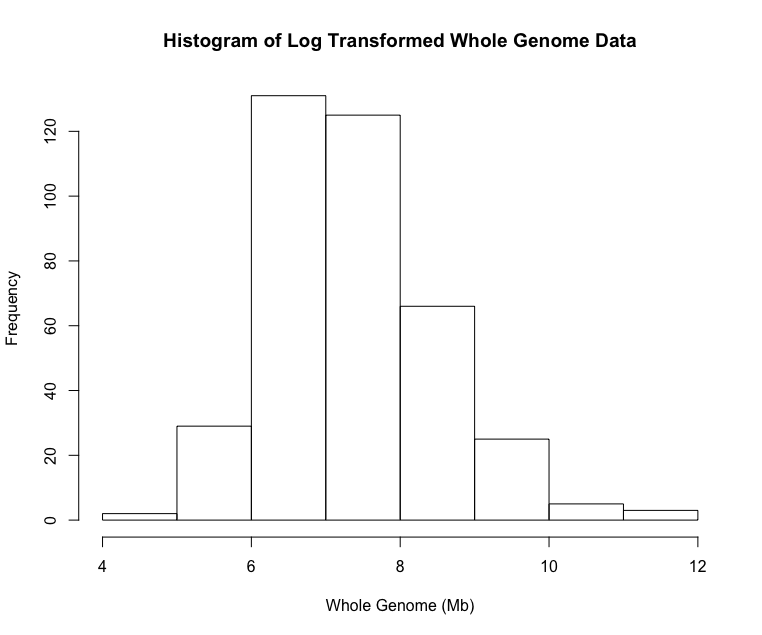
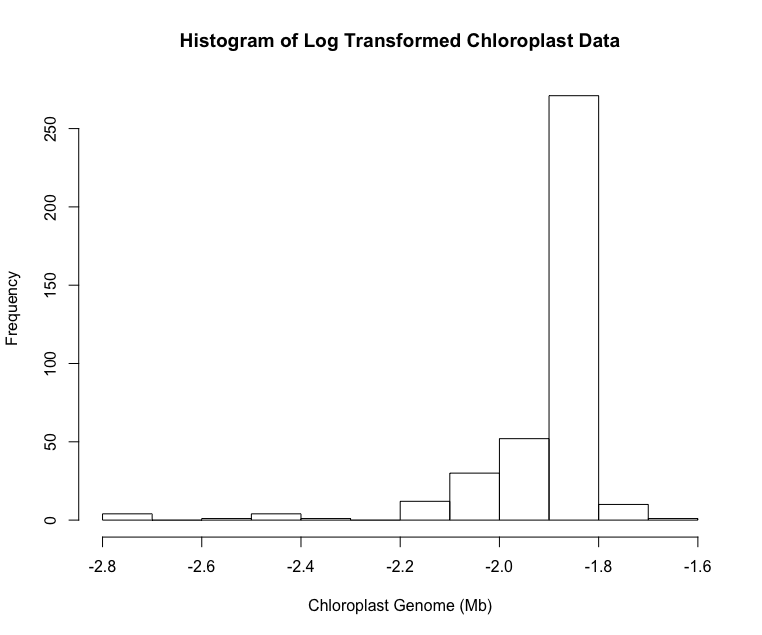
**Figure 1.** We see large means for whole genome size, high standard deviation and

large amount of variance. Chloroplast size yields low values for mean, variance,

and standard deviation.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Normality Statistics for Raw Genome Data | | | | | |
| Skewness | t-val (Skewness) | p (Skewness) | Kurtosis | t-val (Kurtosis) | p (Kurtosis) |
| Whole Genome | 5.999708 | 48.12249 | 0 | 41.92896 | 168.15203 | 0 |
| Chloroplast Genome | -2.554816 | -20.49169 | 0 | 9.116381 | 36.56036 | 0 |
|  | Normality Statistics for Log Transformed Genome Data | | | | | |
| Skewness | t-val (Skewness) | p (Skewness) | Kurtosis | t-val (Kurtosis) | p (Kurtosis) |
| Whole Genome | 0.6933214 | 5.560996 | 1.34E-08 | 0.7985176 | 3.202377 | 6.81E-04 |
| Chloroplast Genome | -3.617487 | -29.01516 | 0 | 17.023024 | 68.26919 | 0 |

**Figure 1b.** Normality statistics of raw and log transformed data.

**Figure 2:** Histograms of the log transformed genome data. Whole genome data is displaying normality, while transformed chloroplast data has remained skewed.

Since we were unable to achieve complete normality with the data, we choose to run the Spearman’s correlation coefficient, as it is a non-parametric test, to understand whether these two variables have any correlation with each other. The Spearman test yielded values that display a weak negative correlation between chloroplast and nuclear genome size (*rho = -0.2537499, p-value = 4.369e-07*).

A scatterplot of the chloroplast genome size versus nuclear genome size was also generated in an attempt to observe any trends as shown in Figure 3 in the Appendix. From this plot it can be observed that organisms with smaller genomes tend to have larger chloroplasts relative to their overall genome size. For instance, an organism with a nuclear genome size under 5000 Mb will have a chloroplast genome size around the mean of .149 Mb. Conversely, an organism with a large genome size of approximately 80,000 Mb will retain a chloroplasts genome size around the mean of .049. There does appear to be a gentle trend in the data showing a number of organisms with small nuclear genome sizes and small chloroplast genome sizes. Lastly, it can be observed from the data that chloroplast genome size levels off around the mean of .149 Mb. It should be noted that outliers with nuclear genome sizes larger than 30,000 Mb were removed from the plot as they skewed the graph, however their correlating chloroplast genome size was reflective of the mean.

**Conclusion:**

Our findings did not yield statistically significant results regarding the correlation between chloroplast and nuclear genome size, thus leading us to accept the null hypothesis. This is not to say that our analyses did not yield findings of interest. The observation that chloroplast genome size “levels off” around the mean of 0.149 Mb as nuclear genome size increases is of interest. Perhaps further exploration into how chloroplast genome size relates to chloroplast size and/or chloroplast density will be of interest to future researchers.

The biggest challenge to this form of research is the amount of data acquired. To further solidify our findings, we would need a much larger dataset with a wider range of species. This will help us analyze the trends of organism genome size vs chloroplast genome size. Of the ~2000 chloroplast genome sequences we had, only ~400 of them had their entire genomes sequenced. The sequencing of these plants and their chloroplasts are still in its infantile stages and so data is still relatively sparse. Further sequencing and archiving of this information is needed before to generate a more robust dataset for analyses.

**Future Research:**

As the research was conducted, only more questions arose from the results. With the limitations of our current databases on genomic information, the next step would be to compare mitochondrial DNA, an essential to almost all cells, with genomic DNA and see if the relationship is the same. As mentioned previously, an exploration into whole genome size as it relates to chloroplast density as well how chloroplast size relates to its genome size. Another form of research is to repeat the analyses of this paper but with a much larger list of data along with a wider range of phylogeny to see if the phenomena are isolated to just the angiosperms. The results we had are definitely eye opening, but we are only at the tip of the iceberg when it comes to understanding the complexities of DNA and how it plays a role in our biology. Further research could be correlations between microbiota and mitochondria size, due to their abundance in well documented databases, as well as pursuing the hypothesis of “are cells integrating the organelle DNA (chloroplast/mitochondria) resulting in a larger genome?” as proposed by our bioinformatics professor.

**Productive use of R**

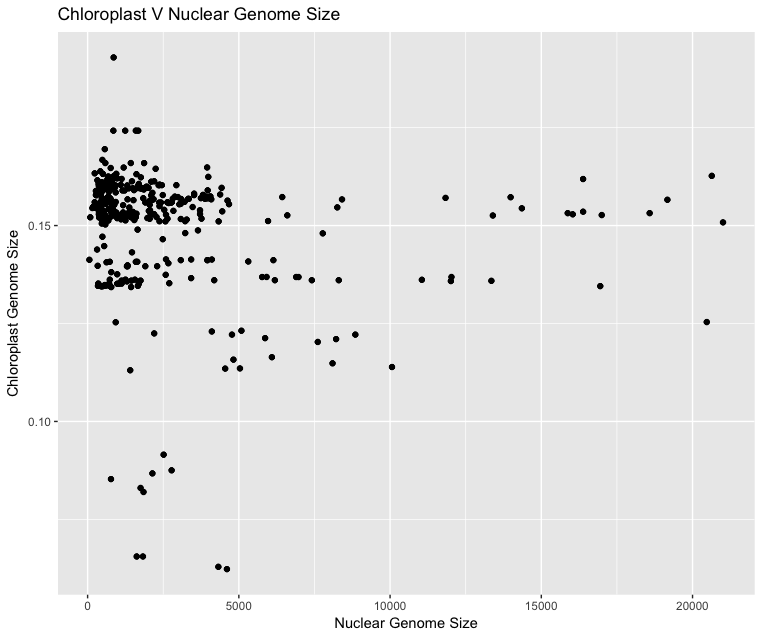
The general use of R was definitely helpful with the production of the data. With of course the limitations of the actual data itself, our overall use of R was necessary. The two CVS files are the data we downloaded from each of the given websites. With a bit of manipulation through the use of excel we were able to delete excess data as well as rename our column titles for R use. Mostly what made R so helpful was the data crunching as well as being able to manipulate 7000 and more entrees. Albeit our code for the production of the data is fairly short, it was intuitive to being able to use the merge() to do most of the heavy work in terms of cross referencing our data frames. Another way we would have done it, if merge() was not an option, was to create a new dataframe and compare each of the database frames by organism name. With convention naming being very predictable we could compare and then take the corresponding information of each frame and insert them into our new dataframe. Luckily, with how R is so intuitively made, we didn’t have to go about it the long way with that. R was further used to generate the graphs and perform the statistical analyses to interpret the significance of our findings.

**References:**

Chan, Yee-Hung M, and Wallace F Marshall. “Scaling Properties of Cell and Organelle Size.” *Organogenesis* 6.2 (2010): 88–96. Print.

Simonin, K.A., Roddy, A.B. (2017). Genome downsizing, physiological novelty, and the global dominance of flowering plants. *Manuscript: ResearchGate.* Retrieved from https://www.researchgate.net/publication/319087912\_Genome\_downsizing\_physiological\_novelty\_and\_the\_global\_dominance\_of\_flowering\_plants?enrichId=rgreq2079bed389faf9d1e9e3c9bb3f482e87XXX&enrichSource=Y292ZXJQYWdlOzMxOTA4NzkxMjtBUzo1MjY2NTY1NTgxMjUwNTZAMTUwMjU3NjExNjk1OQ%3D%3D&el=1\_x\_3&\_esc=publicationCoverPdf

**Appendix:**



**Figure 3:** The relationship of untransformed Nuclear genome size versus chloroplast genome size. The “leveling of” of chloroplast genome size can be observed.

**Github Repository:**

<https://github.com/efryer333/GenomeVChloroplast>

**Description of files in repositoryL**

DS\_Store: This is a file that gets automatically added whenever I create a new repo. It has no value.

Biol638\_Final.R: The R scripts used to generate figures and perform analyses with descriptive comments.

ChloroplastGenomes.csv: The data downloaded for the chloroplast genomes.

GenomeVChlorPlot.pdf: A pdf of the plot of chloroplast versus nuclear genome size

PlantGenomes.csv: The nuclear genome data that we downloaded to analyze.

Simonin\_.pdf: Our inspiration.