Josh Halprin,
Princeton Joseph,
Larry Mason,
Lucas Pereira
CGS 4144
Nov. 22, 2022

Final Project Report

Github link: https://github.com/slinky55/BioInfo-Project

Abstract

Zea mays can have non-essential B chromosomes which do not have any obvious effect on plant development. The nondisjunction of these chromosomes occurs during *Zea mays*' second round of pollen mitosis, and contains an accumulation mechanism. Samples with B chromosomes have been used in a wide variety of genetic studies including gene dosage effect and the engineering of artificial chromosomes, but the precise effects of B chromosomes on gene expression is not yet known. Here we show that the number of B chromosomes affects gene expression to a measurable degree. After analyzing a set of *Zea mays* counts data that contained 1-7 B chromosomes and a control group via the construction of a PCA plot, we found that samples with at least 1 B chromosome were distinct from the control group; however, sample clusters with specific numbers of B-chromosomes were not as clearly defined. After conducting further analysis with K-means, hierarchical, PAM, and GMM (Gaussian Mixture Modeling) clustering algorithms, we found that the clusters these algorithms created closely mirrored the original sample groupings we found. This suggests that the presence of B chromosomes affects the distribution of gene expression across samples, as the samples with different numbers of chromosomes were relatively identifiable. While we are still unsure of which genetic pathways the presence of B chromosomes affects, the results in our experiment indicate that the study of gene expression and B chromosomes in *Zea mays* is not irrelevant and requires further exploration. More data sets will need to be examined to tell if this experiment can yield reproducible results, and investigation into whether the B-chromosome works in conjunction with other chromosomes may be necessary to distinguish any unique effect it might have. Samples with more B chromosomes may also be needed to determine if there is a limit to the effects the chromosome has on gene expression. If the exact effects of B chromosomes on gene expression are found, they may aid in the genetic modification of *Zea mays* as a crop; plants that consume resources efficiently and that have higher fruit yields are possible outcomes of such modification.

Introduction

Our group used RNAseq data from the W22 line of *Zea mays*, which is a cultivar that is commonly used in genetics research. Multiple lines containing 1-7 B chromosomes and a control line containing no B chromosomes were used; researchers used fluorescence in SITU hybridization to identify the chromosome count in each line. Using this data, we sought to determine how the B-chromosome count in each line affected gene expression in *Zea mays*, if at all. We predicted that the presence of a B-chromosome would influence gene expression.

While the B-chromosome is non-vital, research has shown that it may impact several parts of the *Zea mays* genome and phenotype expression; specifically, the transcription and expression of A chromosomes [1], [2], root meristem nuclear phenotype expression [8], leaf striping [7], and knob number [6] seem to be impacted by the B-chromosome. The B-chromosome may also play a role in survivability at higher altitudes, as B-chromosome count scales positively with increases in altitude [3], [6]. Furthermore, B-chromosomes are positioned non-randomly within *Zea may*'s gamete form, suggesting that it may influence gene expression [4]. In other species, the B chromosome can be harmful when occurring in high numbers, having the potential to affect germination and fertility [9]. Since the inheritance of the B-chromosome is non-mendelian and arises from A chromosome fragments that are erroneously produced during meiosis [10], [5], tracking its effects on the genome of a given organism has proven difficult [5]; the use of modern computational genomics tools may help solve this problem.

To that end, we initially used DESeq2 to extract differentially expressed genes from our dataset and clustered our samples based on chromosome count via PCA and UMAP plots. To examine genes across samples, heatmaps and volcano plots were created. Several highly differentially expressed genes were found, and the PCA and UMAP plots showed fairly distinct clusters. After we established that there were clear distinctions between samples containing different numbers of B chromosomes, we used several clustering algorithms on the top 5000 differentially expressed genes to verify the variance we found while clustering on our samples. Namely, the hierarchical clustering, Gaussian mixture modeling, K-means clustering, and PAM clustering algorithms were used; the resulting clusters aligned with the initial sample-based groups.

Methods

To analyze data, gene counts were loaded from a comma-separated file downloaded from the GEO database, and processed using the R programming language. All code and analysis is available at the Git repository at *https://github.com/slinky55/BioInfo-Project*.
Gene counts data were cleaned and stored as a matrix, and then processed using the DESeq2 R package to produce a differential expression dataset, which was used in further analysis.

Visualization of differential expression

PCA
The differential expression dataset was filtered to only include genes with at least 10 total counts throughout samples, and was transformed using variance stabilizing transformations through the vst() function. Data was then dimensionally reduced using Principal Component Analysis (PCA) and scatterplotted using the plotPCA() function, with points colored based on the number of B-chromosomes contained.

UMAP
Vst-transformed data was transposed and dimensionally reduced using Uniform Manifold Approximation and Projection (UMAP), and scatterplotted via the ggplot package. Points were colored based on the number of B-chromosomes contained.

Density Plot
The raw counts data were log-scaled and the numerical range of log-scaled counts for each gene was determined. The frequencies of ranges were then density-plotted using the density() function of R.

Differential analysis

Differential analysis was performed on the expression dataset using the apeglm package and DESeq() function. Results from the comparison of the control group and 1 B-chromosome containing group were primarily used for further study, and only results for genes with $p < 0.05$ were included.

Volcano plot
A volcano plot of the results with $p < 0.01$ was created using the EnhancedVolcano R package.

Heatmap visualization
A heatmap was produced using the ComplexHeatmap package, featuring raw counts data for the top 30 most significantly expressed genes according to the p-values of the differential analysis

results. A sidebar was included to identify the number of B-chromosomes present in each sample.

Clustering

K-means clustering

The means and variances provided by the differential expression dataset were log-scaled, and then used as dimensions to perform K-means clustering, via the kmeans() R function. K=8 was used to correspond clusters to numbers of B-chromosomes, and data were scatter plotted through the ggplot2 fviz_cluster() function, with coloring based on each gene's cluster classification.

PAM clustering

The same log-scaled mean and variance data were also used to cluster genes via PAM, through the pamCluster() function of the R package cluster. K=8 was used, and data were plotted through the ggplot2 fviz_cluster() function, with coloring based on each gene's cluster classification.

Gaussian Mixture Modeling (GMM)

The R mclust package was used to produce a Bayesian Information Criterion (BIC) out of the log-scaled mean and variance data, and the BIC was used to classify genes. The log-scaled data were plotted using fviz_cluster() and genes were colored based on their classification. Eight clusters were naturally produced.

Hierarchical clustering

The log-scaled mean and variance data was used to produce a distance matrix, and this was passed into hclust(), a built-in R function for hierarchical clustering. The result from hclust() was then fed to cutree(), which delineated the 8 specified clusters in the plot and allowed for the extraction of a list of genes and their respective clusters.

Heatmap with clustering

To create the heatmaps, named vectors containing gene names and their clusters were taken from each clustering method and mapped to the counts data using the merge() function. The resulting mapped matrix (with clusters mapped against samples) was then fed into the Heatmap() function from the ComplexHeatmap package, and several helper functions were used to organize the data.

Alluvial diagram

To create the alluvial diagrams, named vectors containing genes and their clusters for the top n differentially expressed genes were first taken from the clustering methods by various means (typically by some getter function or decomposition of the clustering algorithm object). Then, the size of each cluster (based on percentage of genes belonging to the cluster) was calculated. The 4 lists for each input size were then combined into a matrix and used in ggplot() with the ggalluvial library.

Chi-squared testing of clustering methods

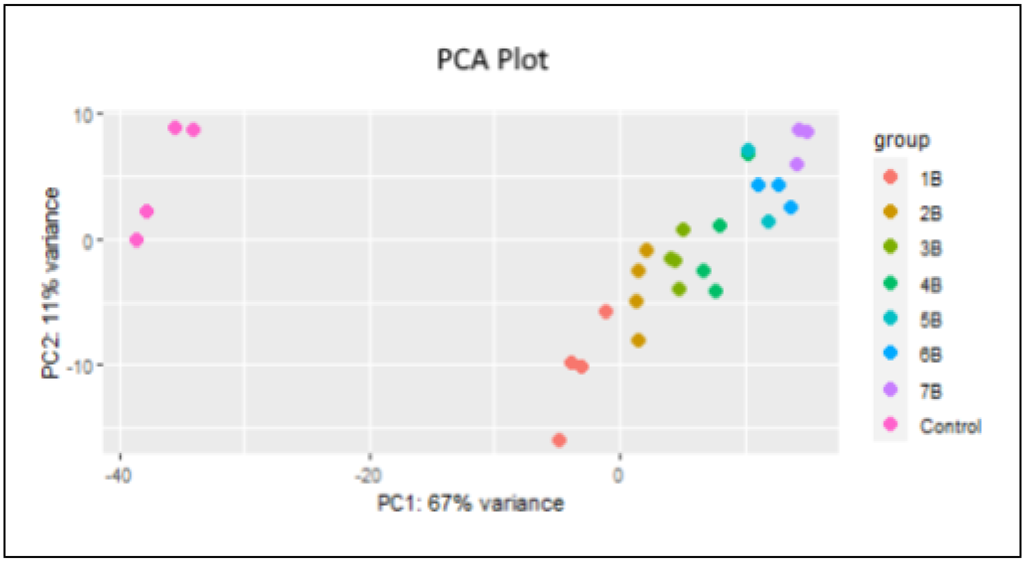
Tables were produced for each clustering method (K-means, PAM, GMM, hierarchical), with rows representing clusters, and columns representing testing groups based on the number of B-chromosomes in a sample. Each cell was populated with the log-scaled total sum of genes within the cluster of the cell's row, and the testing group of the cell's column. A chi-squared test of independence was performed on each table to determine whether testing groups and clusters assigned by a clustering method were independent of each other.

Chi-squared testing was also performed to compare the gene counts of clusters produced by different clustering methods, by performing pairwise comparisons between clustering methods. P-values produced were adjusted using the Bonferroni correction.

Results

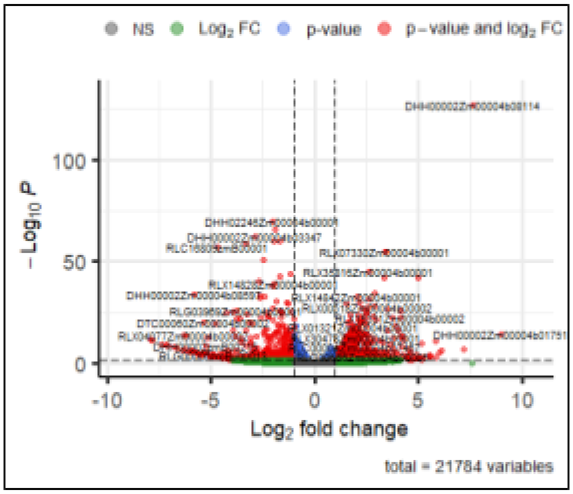
We were able to determine that the presence of B-chromosomes in cells indeed affects the gene expression of those cells.

PCA scatter plotting showed a large distinction between samples with B-chromosomes and control samples without B-chromosomes, as well as slighter distinctions between samples with differing numbers of B-chromosomes. This suggests a difference in gene counts and variances influenced by B-chromosome count, but with diminishing results beyond one B-chromosome.



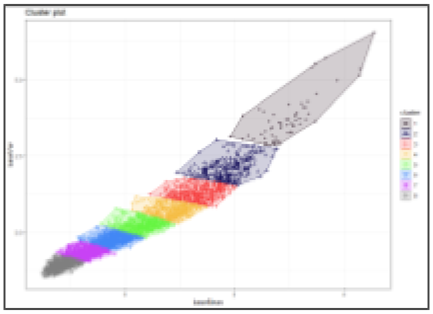
PCA plot depicting variance among varying B chromosome counts.

A volcano plot showing log2fold changes and p-values produced by differential analysis between control samples and 1-B chromosome containing samples showed a number of significantly differentially expressed genes, suggesting that the expression of some genes are highly dependent on the presence of a B-chromosome.

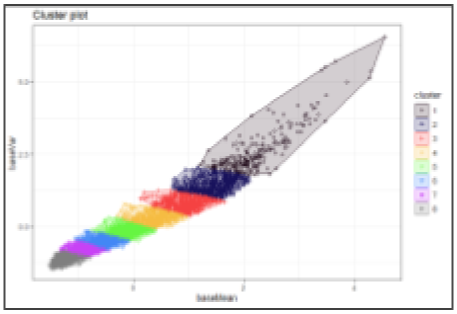


Volcano plot displaying statistically significant genes.

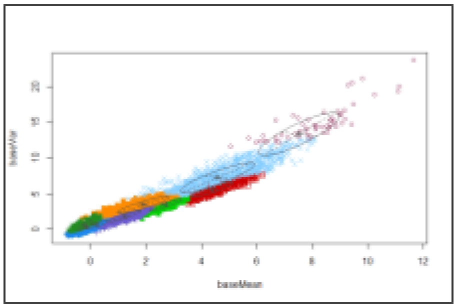
K-means, PAM and GMM-based clustering showed consistent clustering results based on the means and variances of gene counts between samples, indicating a potential relationship between B-chromosome count and the distribution of a gene's expression among samples. However, Chi-squared testing to quantitatively compare the similarities between clustering algorithms proved inconclusive, due to a high number of observations producing abnormally small p-values.



K-means clustering plot

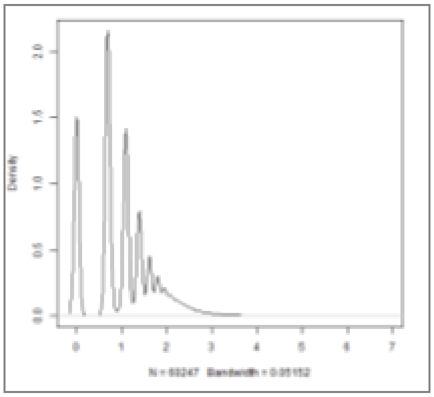


PAM clustering plot



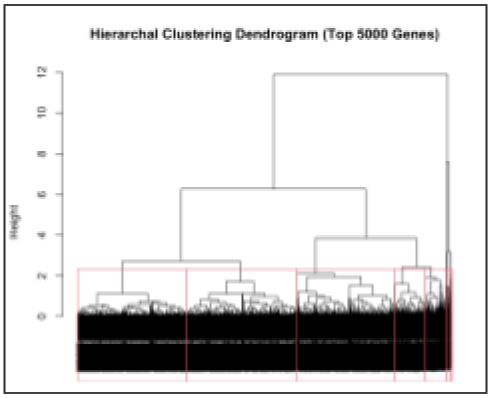
GMM plot

A density plot showing the prevalence of ranges of gene counts between samples showed that the density of ranges was highly periodic, with the potential ranges of gene counts almost appearing to be quantized and discrete.



A density plot comparing gene counts between samples.

Hierarchical clustering showed significant variation in the sizes of clusters, and a clear distinction between one small family of genes and a much wider range of related genes.



Dendrogram depicting 8 clusters found among the top 500 differentially expressed genes.

To avoid bioethical issues, we made sure to use a freely accessible database for our counts data and selected an organism that is non-endangered and widely used in studies.

Conclusion

We determined that B-chromosome count in Zea mays affects gene expression within cells. Through differential analysis of the gene counts of samples with zero and one B-chromosome, we found a number of genes with significant log2fold changes and low p-values; this suggests a significant influence from the addition of the B-chromosome in gene expression. The results of PCA dimensionality reduction and plotting further suggest that the expression of genes is significantly influenced by the presence of at least one B-chromosome, but that this influence diminishes somewhat with the addition of more B-chromosomes. A heatmap also showed a distinct cut between the gene counts of samples with and without B-chromosomes for certain differentially expressed genes, and some smaller visible cuts among samples with at least one B-chromosome, further suggesting that the influence of B-chromosomes on gene expression diminishes with the number of B-chromosomes. Clustering genes based on the means and variances of their counts via PAM-K, K-Means, and Gaussian mixture modeling also found visually distinct clusters of genes when accounting for only the top 1000 more differentially expressed genes, but chi-squared testing to determine the statistical significant of these groupings was inconclusive. As such, while we can conclude that the expression of certain genes in Zea Mays may be influenced by the presence of B-chromosomes, as well as quantitatively determine the extent that individual genes are differentially expressed based on B-chromosome count, more investigation is required to quantitatively determine the likelihood that any given sample contains a certain number of B-chromosomes based only on gene expression data.

In future studies, a different method of quantitatively comparing the results of clustering methods may be used instead, such as through recording the number of disagreements on a gene's clustering. Having a larger number of samples could also allow future studies to more properly test its hypotheses by separating samples into an analysis group and a testing group, blinding the data from the testing group, and then attempting to classify them based on the results of analyzing the analysis group.

References

[1] Huang, Wei, et al. "B chromosome contains active genes and impacts the transcription of A chromosomes in maize (Zea mays L.)." *BMC Plant Biology* 16.1 (2016): 1-14. https://link.springer.com/article/10.1186/s12870-016-0775-7

[2] Shi, Xiaowen, et al. “Effect of Aneuploidy of a Non-Essential Chromosome on Gene Expression in Maize.” *The Plant Journal*, vol. 110, no. 1, (2022): 193–211. https://onlinelibrary.wiley.com/doi/10.1111/tpj.15665

[3] Rosato, Marcela, et al. "Genome size and numerical polymorphism for the B chromosome in races of maize (Zea mays ssp. mays, Poaceae)." *American Journal of Botany* 85.2 (1998): 168-174. https://bsapubs.onlinelibrary.wiley.com/doi/abs/10.2307/2446305

[4] Rusche, Maxine Losoff, et al. "B chromosomes of maize (Zea mays L.) are positioned non randomly within sperm nuclei." *Sexual Plant Reproduction* 13.4 (2001): 231-234. https://link.springer.com/article/10.1007/s004970000055

[5] Kao, KW., Lin, CY., Peng, SF., et al. Characterization of four B-chromosome-specific RAPDs and the development of SCAR markers on the maize B-chromosome. *Mol Genet Genomics* 290 (2015): 431–441. https://doi.org/10.1007/s00438-014-0926-1

[6] Rayburn, A. Lane, and J. A. Auger. "Genome size variation in Zea mays ssp. mays adapted to different altitudes." *Theoretical and Applied Genetics* 79.4 (1990): 470-474. https://link.springer.com/article/10.1007/BF00226155

[7] Staub, Rick W. "Leaf striping correlated with the presence of B chromosomes in maize." *Journal of Heredity* 78.2 (1987): 71-74. https://academic.oup.com/jhered/article-abstract/78/2/71/768004

[8] Ayonoadu, U. W., and H. Rees. "The effects of B chromosomes on the nuclear phenotype in root meristems of maize." *Heredity* 27.3 (1971): 365-383. https://www.nature.com/articles/hdy1971101

[9] Jones, R. Neil. "B chromosomes in plants." *New Phytologist* 131.4 (1995): 411-434. https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8137.1995.tb03079.x

[10] Jones, Neil, and Andreas Houben. "B chromosomes in plants: escapees from the A chromosome genome?." *Trends in plant science* 8.9 (2003): 417-423. https://www.sciencedirect.com/science/article/pii/S1360138503001870?casa_token=xIiyOTd_L2kAAAAA:MLLWx8i503jf8F3EwtmWx6C7OHsmUDErWvJhoI86wKAKUNcb-6oidGDLhC5EWmTtRE1NSm_pPw