Exploring the predictive power of crossover hotspots in maize

Nikita Sajai[1] **and Ruth Epstein**[2]

[1]Biological Sciences, 2022, ns623
[2]Plant Breeding and Genetics, 3rd year, rke27

ABSTRACT

In maize and other plant species, mechanisms of crossover (CO) hotspots, locations across the maize genome where there is an enrichment of COs, remain elusive. Human and mouse studies have revealed CO hotspots are dictated by a zinc-finger protein, PRDM9, which is completely absent in plants. Nevertheless, CO hotspots in plant species exist and are maintained across evolutionary history. Meiotic recombination, the process by which genetic material is exchanged, is being explored as an avenue for generating new forms of genetic diversity, so understanding what factors control the formation of CO hotspots in maize and other important crop species will give breeders the power to unlock new and favorable combinations of alleles that lead to productive varieties. We constructed a labeled dataset incorporating several known features of COs in maize: site-specific DNA methylation, nucleosome occupancy, and distance to various chromosomal features as CO formation is biased towards the telomeric regions of many plant species' chromosomes. We also tested unconfirmed features such as indel density and SNP density. We visualized this data using the program based on non-linear dimensionality reduction, and we then use this dataset to implement a feature-based supervised machine learning model to answer if CO hotspot occurrence in maize can be predicted.

Keywords: recombination, machine-learning, crossover hotspots, maize

Project type: Research

Project repository: `https://github.com/16nikita/CO_hotspots_final`

1 Introduction

Meiotic recombination plays a large role in the diversification and maintenance of complex genomes over evolutionary time by generating allelic diversity and ensuring proper chromosome segregation. Arguably, the most important part of meiotic recombination is cross-over (CO) events, in which there is a reciprocal exchange of genetic material between homologous chromosomes. Plant breeding heavily relies upon CO events to create novel combinations of alleles between two crop varieties, however, the low frequency of recombination and the biased distribution of CO events has led to recombination being the limiting factor in releasing new plant varieties (Taagen et al., 2020). There are known features of COs in plants such as lower nucleosome occupancy levels, decreases in CG and CHG site methylation, and increases in the number of short indels (Wang et al., 2022). CO distribution in maize assumes a U-shape across each chromosome, with increased CO frequency towards sub-telomeric regions and suppression in the pericentromeric region.

The CO pathway is fundamental to the reproduction process. In the earliest stages of this pathway, double-stranded breaks (DSBs) form along chromosomes during prophase I of meoisis, which are catalyzed by the topoisomerase-like protein, SPO11 (Keeney et al., 1997). DSBs can either proceed to form or suppress COs, depending on the proteins involved in the homologous recombination-mediated repair of the DSB (Keeney et al., 1997). The recombinases RAD51 and disrupted meiotic cDNA1 (DMC1) promote the single-stranded ends of the DSBs to undergo invasion of the sister chromatid or homologous chromosome, forming a D-loop intermediate that is primed for DNA synthesis. From this point, non-crossovers may form through synthesis-dependent strand annealing (SDSA), by which the extended invading strand breaks off and reanneals to the original DSB. Otherwise, COs controlled by the proteins MLH3 or HEI10, may form as DNA synthesis occurs and then second-end capture leads to the formation of double Holliday junction (dHJ) intermediates that induce DNA exchange between homologous chromosomes (Taegen et al., 2020).

With the advent of novel environmental and pest pressures being exacerbated through climate change, there is a great need

to accelerate the breeding process such as in *Zea mays ssp. mays* (maize). In particular, there is interest in generating consistently high recombining regions in the genome, called CO hotspots. Past studies have successfully demonstrated the use of deep-learning neural networks to detect CO hotspot regions, but these studies were performed with mammalian genome data, in which PRDM9 binding motifs are known to be associated with hotspot formation (Li et al., 2022). It remains unknown which features dictate the recombination rate in maize. This study investigated if a machine learning approach could predict CO hotspot occurrence or CO rate in maize.

2 Methods

2.1 Labeling datasets

A cohesive CO dataset that incorporated single CO events and CO hotspots was created. Feature labels for CG and CHG (where H can be any base except G) site methylation, small insertion and deletion (InDel) density (1-50 bp), nucleosome occupancy, and SNP density were added using the modules bedtools intersect and bedmap (v2.29.2) (Quinlan et al., 2010). The small indels were taken from a dataset produced by the Hufford Lab and identified through a de novo analysis of 26 diverse inbred maize lines (Hufford et al., 2022). The methylation data was averaged over each CO region. The indel density was calculated by summing the number of intersecting indels within each region and dividing the total by the size of the region. The distance to the centromere, telomere, and nearest promoter was assigned by finding and calculating the features with the smallest absolute distance away from the CO site.

The single CO dataset and the CO hotspot dataset were taken from Rodgers-Melnick et al., 2016. Single COs and CO hotspot intervals that were greater than 15 kb were filtered out. All feature data were taken from the Pawlowski Lab, and are unpublished.

The final labeled dataset had 8 features; nucleosome occupancy, % of CG methylation, % of CHG methylation, indel density, distance to centromere, distance to telomere, distance to nearest promoter, distance to the nearest MLH3 binding site, and SNP density. The data was log transformed to create uniformly scaled data.

2.2 Data visualization

For preliminary visualization steps, we created pairwise plots of each feature against both classification (CO vs hotspot) and CO rate to gauge general trends. UMAP was used to cumulatively visualize the data (v0.5.3) (McInnes et al., 2018). UMAP utilizes a novel non-linear dimensionality reduction algorithm, creating a high-dimensional graph of the data and projecting a low-dimensional graph with the closest topology. The algorithm approximates the topology of the high-dimensional data under the assumption of the Nerve theorem: each point in the high-dimensional data is assigned a 0-simplex combinatorial building block which are connected with the simplices of neighboring points, a process that is guaranteed to yield a lower dimensional representation. This neighborhood graph can be constructed using fuzzy open balls and extending the radius of each 'ball' until it makes contact with the radii of neighboring balls. In projecting into a lower dimension, the loss function cross-entropy is calculated. In the formula for cross-entropy shown below, a balance between attractive and repulsive forces between the points e minimizes this loss function and ensures that high-probability edges stay together. Ultimately, this algorithm renders a projection that retains the topology of source data.

$$\sum_{e \varepsilon E} w_h log(\frac{w_h(e)}{w_l(e)}) + (1 - w_h(e)) log(\frac{1 - w_h(e)}{1 - w_l(e)})$$

The formula for cross-entropy (McInnes et al., 2018)

UMAP takes in a wide array of hyperparameters, but we assigned values to four: n_neighbors=30, min_dist=.25, n_components=2/3, and random_state=0. The hyperparameter k (n_neighbors) represents the kth nearest neighbor and defines the high-density and low-density regions of the projection. Larger values of k preserve global structure while smaller values retain local structure. The algorithm ensures that whatever topology between each point and its n_neighbors is reflected within the lower dimensional projection. The min_dist sets the smallest spacing permitted between points in the low dimensional representation. The n_components define the dimensionality of the reduction (i.e. 1-D, 2-D, 3-D). Lastly, we set random_state to ensure reproducibility. We also incorporated a scaling step for variance stabilization.

2.3 Machine Learning

The models were implemented in Python 3. The "xgboost" and "sklearn" Python packages were used to construct all models (v1.7.2) (Chen et al., 2015). All models were trained on 80% of the labeled data and tested on a 20% holdout of all labeled data.

We constructed two supervised machine learning models. The XGBOOST algorithm, which stands for eXtreme Gradient Boosted Trees, uses a combination of decision trees and gradient boosting (Chen et al., 2015). Gradient boosting merges predictions from multiple weaker models, decision trees, in order to predict the target variable. The new models boost features that produced inaccuracies in previous trees, while minimizing the loss from developing these new trees using a gradient descent algorithm.

A binary classification model was built to predict a categorical variable, the status of a region as either a single CO or a CO hotspot. The regression model predicted the continuous variable of CO rate. Before training these models, the CO dataset was cleaned and split into a training and testing dataset. We incorporated imbalanced dataset optimizations to tune the hyperparameters to best fit the training data and labels. The tuning steps and results are described in the next section.

3 Results

3.1 Binary Classification

The 2-dimension reduction of the data showed 2 distinct clusters, however, the single COs and CO hotspots were intermixed within the clusters (Figure 1). To further understand the clustering dynamics, a 3-dimension reduction was implemented but the 3-D projection also did not explain the labeled data well.

Despite the lack of clustering in the dimension reduction analysis, a binary logistic XGBoost model was trained in Python 3 using 5300 observations of either single COs or CO hotspots with the 8 above-mentioned features. Once the model was trained, the model was fed a confusion matrix to assess the performance of the model. Using the confusion matrix output, the model's accuracy was 91.8%, however, the model was only able to correctly classify one CO hotspot and misclassified 130 CO hotspots as single COs. To optimize the model, a grid search was performed to find the best hyperparameter fit. Scale position weight, which is related to an imbalance in the dataset sizes, was set to 0.1, the learning rate was set to 0.1, and the maximum tree depth was set to 0.3. Once the model's hyperparameters were optimized, the model was re-trained and fed a confusion matrix to evaluate model performance. The model had a lower prediction accuracy of 72.7% but was able to predict CO hotspots correctly 65% of the time (Figure 2). However, the misclassification rate was 27.2%. The unoptimized model had a high misclassification rate for CO hotspots, and thus through optimizing the model, sensitivity for the detection of single COs was reduced in favor of identifying more CO hotspots. Additionally, the model had precision, sensitivity, and specificity rates of 64.6%, 15.9%, and 96.4%, respectively. The precision rate refers to the ratio between predicted and true CO hotspots, sensitivity refers to the model's ability to detect true CO hotspots, and specificity refers to the model's ability to detect true single COs.

The final binary classification model identified that % of CG methylation, distance to telomere, and indel density were the top three defining features of the model (Figure 3).

3.2 Linear Regression

Since the binary classification model only had a prediction accuracy of 71.8% and a relatively high misclassification rate, an XGBoost regressor model was trained to predict a continuous variable instead, the recombination rate. The recombination rate for single CO observations was 1 CO divided by the length of the interval, while the rate for CO hotspots was more than 1 CO observation divided by the length of the interval.

A 2-D and 3-D reduction of the CO rate dataset was completed. Interestingly, two clusters were found in the two-dimensional reduction, while a third distinct cluster was identified in the three-dimensional reduction (Figure 4). However, there was no obvious segregation based on higher or lower CO rates, although one cluster did appear to contain a greater number of regions with high rates. The regressor model was trained and predicted CO rate in the test dataset. R^2 and residual squared mean error (rsme) was used to evaluate the model performance. The rsme was quite low, 0.00073, however, the R^2 was 0.3075, demonstrating the model has low performance.

The resulting feature importance rankings from the regression model was similar to the binary classification model. The top three features were nucleosome occupancy, % of CG methylation, and indel density (Figure 5).

4 Discussion

The binary classification and linear regression models ranked % CG site methylation as significant factors contributing to distinction between single CO regions and CO hotspots, suggesting that chromatin openness highly influences the CO process. Decreased methylation levels correspond with euchromatin regions, in which gene transcription levels are high and the DNA is considered "active"; within these regions, CO seems more likely to occur as it generates variation within coding regions that may contribute to evolutionary development (He et al., 2017). Both models also outputted indel density as another significant

factor, which may either be a consequence of increased CO events, as recombination has been demonstrated to generate mutations. On the other hand, increased CO events may influence where indel genesis occurs in the genome, as indels can also contribute to advantageous genetic variance (Liu et al., 2015).

However, both models performed at an average standard, so any conclusions drawn from their outputs are only tentative. Visualization demonstrated that the data clustered in an unexpected manner, with no clear distinction between CO hotspots and single CO events. These results were reflected by model performance, which also struggled to distinguish between these classes. A regression analysis did not improve the model significantly, although logistic and linear regression are not truly comparable due to differing means of evaluation. The relatively average model performance of both the binary classifier and regression analysis is likely due to bias introduced by the imbalanced dataset, as well as a lack of data points. The dataset is relatively small with a low number of features, low resolution (as indicated by the large interval sizes), and does not cover the entire genome. Future improvements may be made as more features that dictate CO formation are uncovered. The dataset can also be expanded with analysis of recombinationatorial history of the maize genome. Only two rounds of optimization were performed, so additional rounds of optimization may refine the models. Lastly, an unsupervised model can be developed on a different premise to determine what factors caused the clustering seen in the UMAP visualization.

Author Contributions

- Study design: R.E. and N.S.
- Coding: R.E. and N.S.
- Experiments: R.E. and N.S.
- Analyses: R.E. and N.S.
- Writing:
  - *Introduction:* R.E.
  - *Methods:* N.S.
  - *Results:* R.E.
  - *Discussion:* N.S.

References

Figures

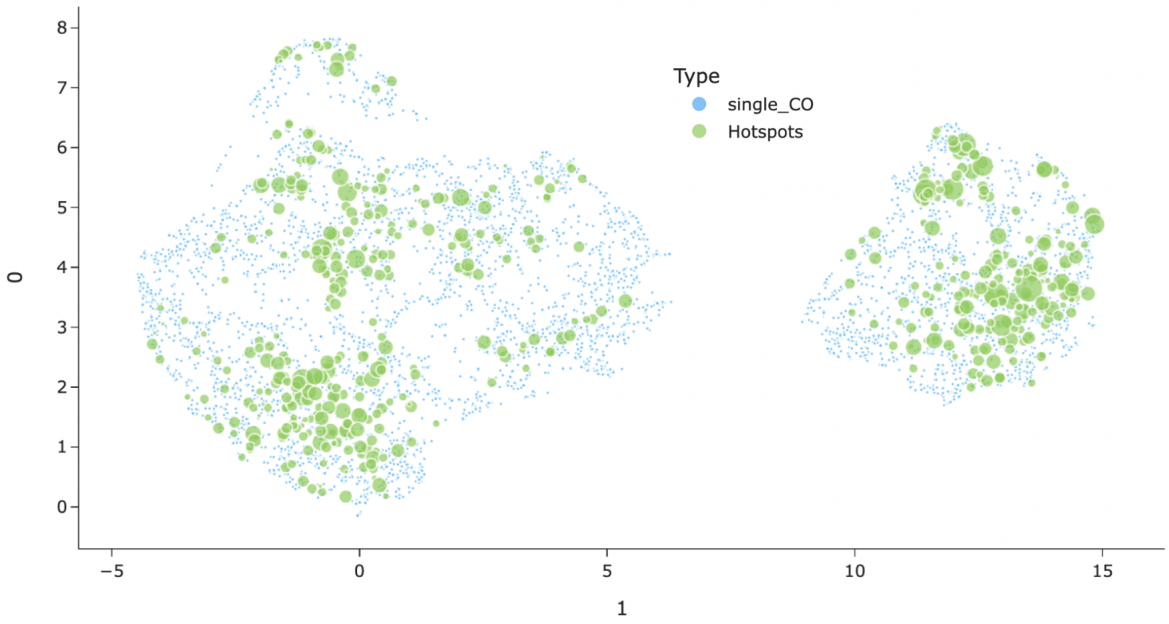

Figure 1. 2-dimensional reduction of all labeled binary classification data.

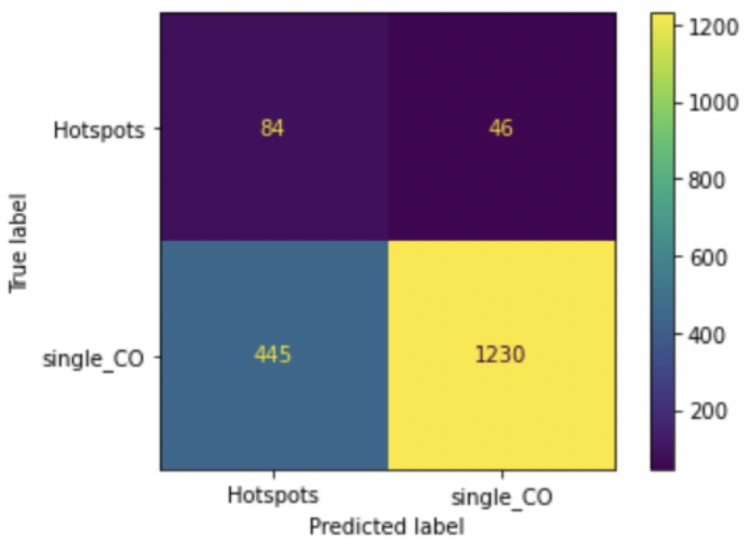

Figure 2. Confusion matrix output from binary classification model

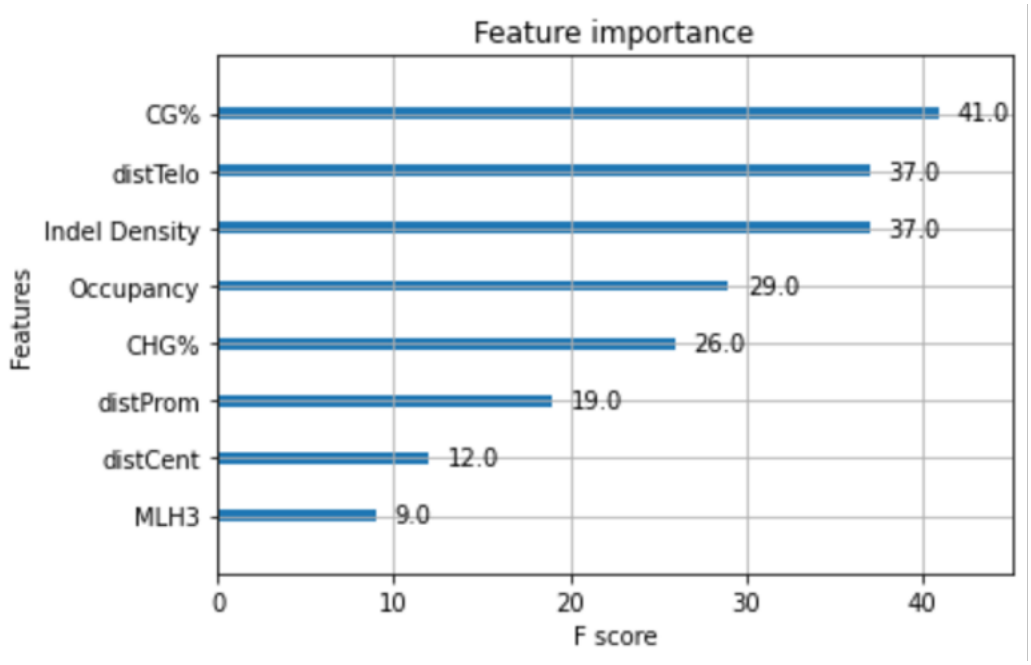

Figure 3. Feature importance ranking output from binary classification model.

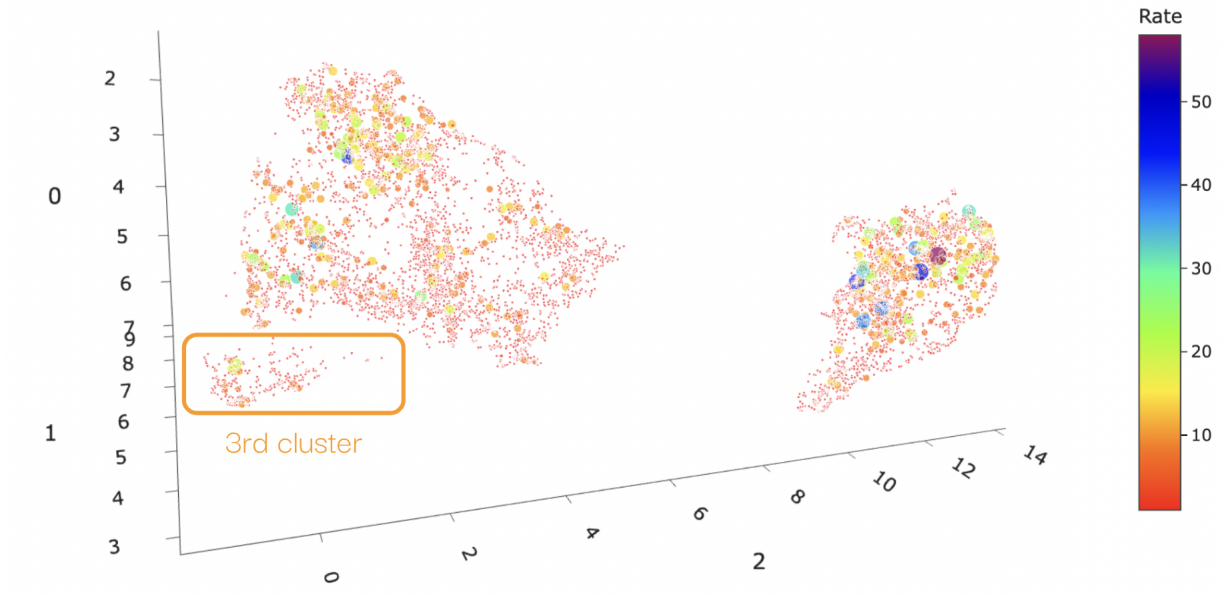

Figure 4. 3-dimensional reduction of labeled CO rate data.

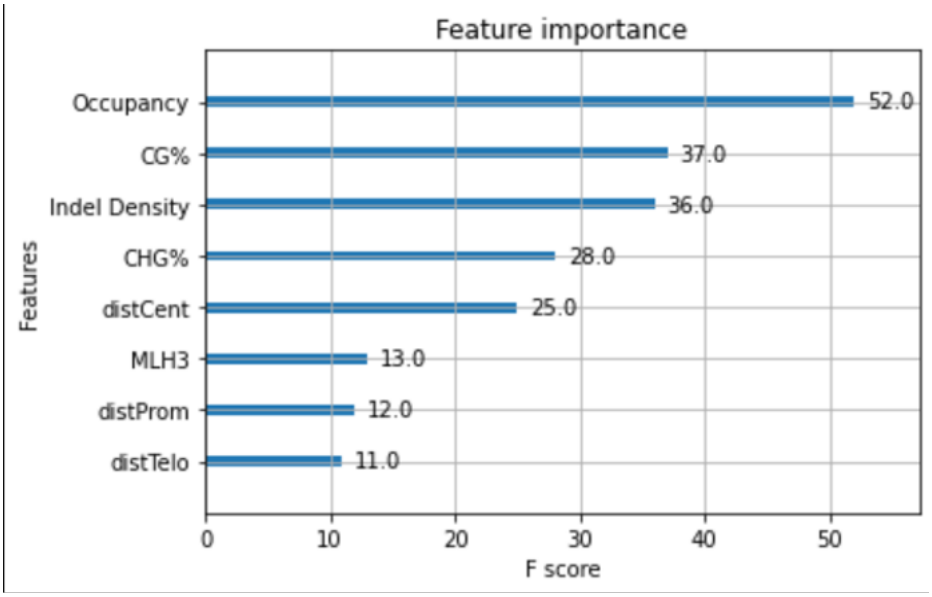

Figure 5. Feature importance ranking output from regression analysis.