Article

Methylation in the CHH Context Allows to Predict Recombination in Rice

Mauricio Peñuela [1] , **Jenny Johana Gallo-Franco** [1] , **Jorge Finke** [1] , **Camilo Rocha** [1,*] , **Anestis Gkanogiannis** [2] , **Thaura Ghneim-Herrera** [3] **and Mathias Lorieux** [4,*]

1   iÓMICAS, Facultad de Ingeniería y Ciencias, Pontificia Universidad Javeriana, Cali 760031, Colombia
2   The Alliance of Bioversity International and the International Center for Tropical Agriculture (CIAT), Palmira 763537, Colombia
3   Departamento de Ciencias Biológicas, Universidad ICESI, Cali 760031, Colombia
4   DIADE, University of Montpellier, CIRAD, IRD, 34394 Montpellier, France
*   Correspondence: camilo.rocha@javerianacali.edu.co (C.R.); mathias.lorieux@ird.fr (M.L.)

Abstract: DNA methylation is the most studied epigenetic trait. It is considered a key factor in regulating plant development and physiology, and has been associated with the regulation of several genomic features, including transposon silencing, regulation of gene expression, and recombination rates. Nonetheless, understanding the relation between DNA methylation and recombination rates remains a challenge. This work explores the association between recombination rates and DNA methylation for two commercial rice varieties. The results show negative correlations between recombination rates and methylated cytosine counts for all contexts tested at the same time, and for CG and CHG contexts independently. In contrast, a positive correlation between recombination rates and methylated cytosine count is reported in CHH contexts. Similar behavior is observed when considering only methylated cytosines within genes, transposons, and retrotransposons. Moreover, it is shown that the centromere region strongly affects the relationship between recombination rates and methylation. Finally, machine learning regression models are applied to predict recombination using the count of methylated cytosines in the CHH context as the entrance feature. These findings shed light on the understanding of the recombination landscape of rice and represent a reference framework for future studies in rice breeding, genetics, and epigenetics.

Keywords: epigenetic; DNA methylation; bisulfite sequencing; machine learning; modeling

1. Introduction

Meiotic recombination is recognized as a key process in genetics. During this process, maternally and paternally inherited homologous chromosomes exchange information by gene conversion or crossing over to create novel allelic combinations. Recombination is widely recognized for its role in promoting diversity to respond to continually shifting environments, in addition to preventing the build-up of genetic load by decoupling linked deleterious and beneficial variants [1]. However, meiotic recombination between homologous chromosomes is restricted by the number and location of crossover sites per chromosome. The crossover distribution and frequency along the genome are uneven, especially in plants [2]. Sites with high recombination rates have been linked to subtelomeric regions that are generally hypomethylated and have high gene and DNA transposon frequencies. In contrast, recombination is suppressed in the centromeric region, which is characterized by high frequencies of long terminal repeat retroelements and few genes [3].

The role of chromatin structure and DNA methylation in determining recombination rates has been previously reported. For example, high levels of histone H3 acetylation in *Arabidopsis* mutants were associated with changes in the crossover frequencies [4]. Likewise, studies using *met1* and *ddm1* mutants, which are globally hypomethylated, showed regional remodeling of crossover frequencies with increased recombination in chromosome

arms and decreased recombination in the pericentromeric region [5,6]. However, understanding how the DNA methylation patterns affect the recombination rates remains an open challenge.

In plants, DNA methylation occurs at cytosine nucleotides in all the sequence contexts CG, CHG, and CHH (H = C, T, or A). DNA methylation is a stable mark inherited from generation to generation and a crucial factor for plant development [7]. DNA methylation, in combination with histone and non-histone protein modifications, defines chromatin structure and accessibility, which helps to regulate gene expression, transposon silencing, chromosome interactions, and trait inheritance [8]. Several studies have shown that sexual reproduction in plants involves the reprogramming of DNA methylation patterns [9].

The methylation dynamics for each sequence context is determined by different mechanisms and related to specific biological functions [8]. The maintenance mechanism of plant DNA methylation depends on the context and is mediated by different enzymes. For example, in *Arabidopsis thaliana*, CG cytosine methylation is maintained by *MET1*, in a semiconservative manner in the DNA replication process, while CHG methylation is maintained by *CMT3* and *CMT2*, which enables the propagation of methylation through a positive feedback loop together with the *H3K9me2* in the cell division process. Meanwhile, CHH methylation is maintained by *DRM2* or *CMT2*, depending on the genomic region [8]. De novo methylation is carried out by *CMT2* for CHG and CHH context [9], and the *RdDM* pathway for all sequence contexts [9]. This process is not the same for all plants. In rice, CG cytosine methylation is carried out by two related genes *OsMET1-1* and *OsMET1-2*, with a possible redundant function, while *OsCMT3a* is the only functional ortholog of *CMT3* involved in CHG methylation during replication. For CHH methylation, no associated gene has yet been reported. There is some evidence that suggests that *OsCMT2* is closely related to *CMT2* and may play a role in CHH methylation [10]. More research on methylation and demethylation events and their precursors will be necessary to clarify these mechanisms.

Identifying factors influencing the meiotic recombination rates is important for breeders interested in transferring genes from one variety to another through crosses. Thus, developing new allelic combinations that allow breeders to meet the needs present in agricultural systems. Recently, a number of studies have addressed this issue and have developed different types of strategies to discover where crossovers occur most frequently and try to predict them. For example, Liu et al. [11] developed a predictor of recombination hot/cold spots in yeast using a machine learning approach combined with principal component analysis. Moreover, Demirci et al. [12] explored DNA sequence and shape features to train machine learning models for predicting crossover occurrence in *Arabidopsis*, maize, tomato, and rice. Moreover, Adrion et al. [13] used recurrent neural networks, a deep learning method for estimating genome-wide recombination in a natural population of African *Drosophila melanogaster.*

In recent years, rice has been a model monocotyledonous plant for understanding the methylation process because it is highly homozygous and self-pollinating. In addition, rice is of great importance in food security since half of the world's population depends on it as daily food [14]. However, few studies have analyzed methylation patterns in relation to recombination rates in rice. For instance, Habu et al. [15] developed an experiment crossing methylated and unmethylated rice varieties and concluded that the position and frequency of meiotic recombination in rice centromeric heterochromatin are regulated by the epigenetic state of the chromatin. Likewise, Choi et al. [16] explore how transposable elements interact with host plant epigenetics. They suggest that high levels of methylation at these elements have a role in suppressing deleterious ectopic recombination. Nevertheless, none of these studies have explored in detail how the methylation contexts are related with recombination rates.

In this work, the relationship between chromosomal recombination rates and different methylation contexts is explored by using *Oryza sativa* as a model. The focus is on the following objectives: (1) To estimate the correlation between recombination and methylation in all contexts, (2) to describe the effect of methylation within genes, transposons, and

retrotransposons with respect to recombination, and (3) to implement a machine learning model to predict recombination based on methylation data. The results provide evidence that recombination can be described by methylation in the context of CHH, regardless of whether it is outside or inside genes, transposons, and retrotransposons. Consequently, the use of machine learning models to predict chromosomal recombination rates in rice cultivars using CHH methylation is proposed.

2. Results and Discussion

In this study, the correlation between recombination rates and the methylated cytosine counts for all chromosomes in two rice cultivars is evaluated (Figures 1 and 2). The correlation values are, on average, -0.44 ± 0.17 for all chromosomes of both varieties, with higher values in the centromere region. Similar results in rice were previously described by Yan et al. [17], revealing that DNA methylation patterns in the centromere are shaped by the DNA sequence and the centromeric domains. Habu et al. [15] described how artificial chromatin modification can vary the frequency of meiotic recombination. Overall, high levels of methylation in heterochromatin regions near the centromeres have been reported as a common pattern, where meiotic recombination is repressed. In the same way, recombination-free regions around centromeres are likely to be important for normal centromere function during meiosis [15,18].

By evaluating the CG and CHG methylation contexts independently, a decrease in recombination rates with increasing methylated cytosines is reported. On the contrary, methylated cytosines in the CHH context increase with recombination rates showing a positive correlation (Figure 3). The opposite relationship between the methylation contexts of CG and CHH has been reported in rice by Li et al. [19], who identified the tendency towards hypermethylation in CG context, but hypomethylation in CHH.

The positive correlation between methylated cytosine count and recombination rates observed in the context of CHH is not clear when all methylation contexts are assessed together because the total number of methylated cytosines in the CG and CHG contexts was higher. This trend is observed for both varieties, IR64 and Azucena, where the methylation data and the alignment process have been obtained independently. The positive relationship between the CHH methylated cytosine count and recombination rates has been reported by Rodgers-Melnick et al. [1], who include the CHH methylation as a feature of a linear model to predict recombination in maize. It is unclear what role methylated cytosines play in the CHH context with respect to recombination.

Variability in DNA methylation can be heritable or reversible, and this can allow for phenotypic variation and rapid response to environmental changes. Even the degree of intraspecies epigenomic diversity can be correlated with climate and geographic origin [10]. It has been reported that CHH methylation could be related to fruit size in apples [20] and silencing transposons in sugar beets [21]. A potential role in *A. thaliana* seed dormancy, with increases in CHH methylation in seeds during seed development and a decrease during germination, has also been reported in [8]. These observations suggest the multiple roles that CHH methylation can play in plant genomes. Recently, Wang et al. [22] reported that CHH methylation levels are higher in rice reproductive organs, such as panicles and pistils, than in seedlings, suggesting a positive feedback loop between DNA methylation and RNA-directed DNA methylation activity involved in sexual reproduction.

The functional analysis performed with annotation data of genes, transposons, and retrotransposons for each variety, shows that the increment in the number of genes per window is correlated with recombination rates in the chromosomes of both varieties (Figures 4 and 5). This positive trend has been previously evidenced in *Drosophila*, *A. thaliana*, yeast, finches, monkeyflowers, and dogs, with recombination hotspots typically located near the promoter regions of genes [23] and observed in the euchromatic regions of maize [24]. In contrast, a negative correlation between the number of transposons and retrotransposons has been found with respect to recombination rates across all chromosomes for both rice varieties. This can be explained by the abundance of such elements

near the centromere where recombination rates are low. Similar results have been found by Tian et al. [25], who suggested that the rice genome is organized along recombinational gradients due to the negative correlation of recombination with transposable elements and positive one with gene densities.
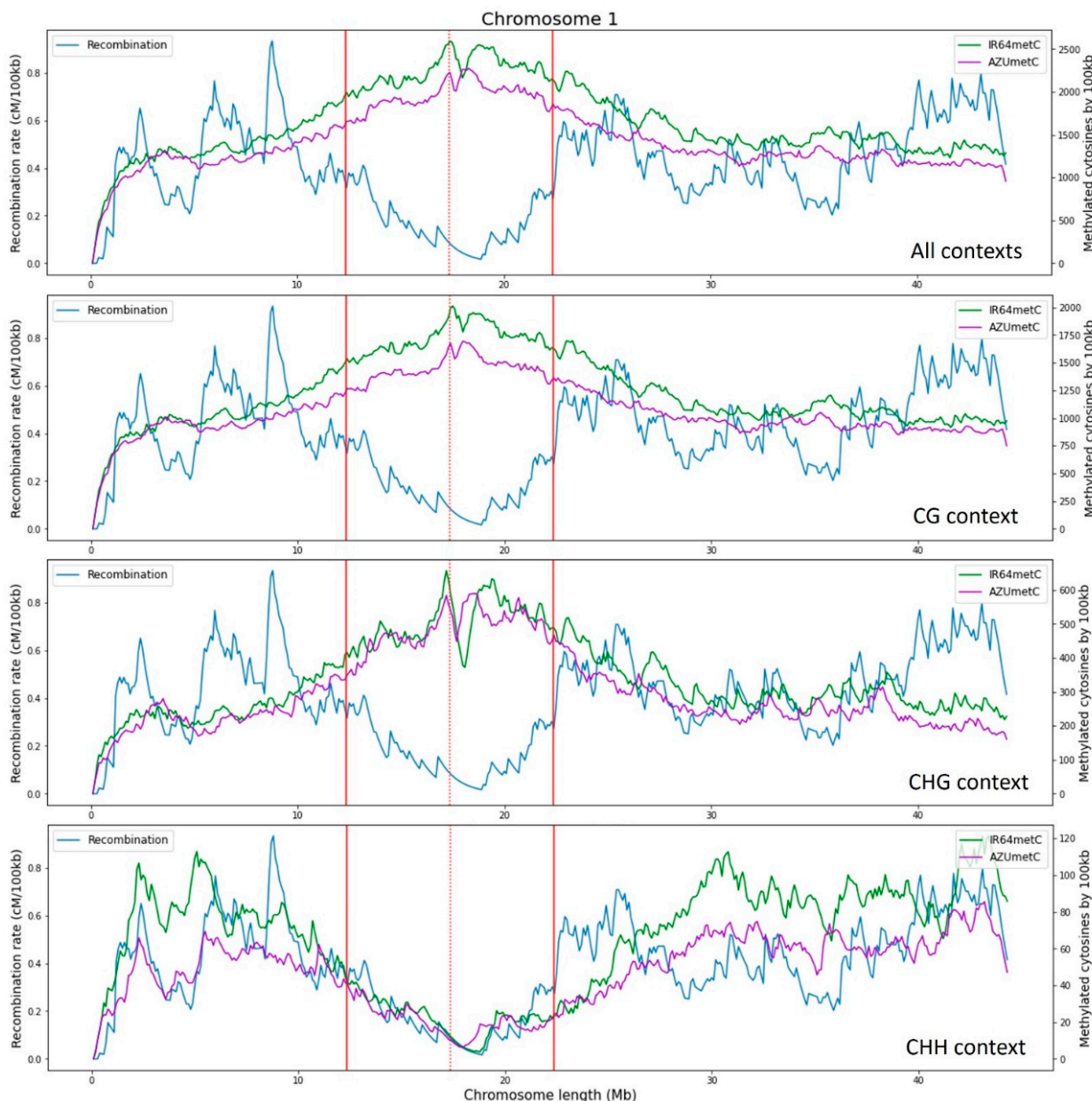

Figure 1. Recombination and methylated cytosines through Chromosome 1 for the rice varieties IR64 and Azucena. The centromere is represented by a red dotted line and the influence of the centromere region by solid red lines.

Recombination tends to occur within and near genes and away from transposable elements. This may reflect the passive effects of recombination initiating in open chromatin [23]. Recent analyses of the localization of recombination at the fine scale tend to show negative correlations with local densities of repetitive elements. Actually, strong recombination suppression and a large accumulation of transposable elements are usual in pericentromeric regions [23]. For rice, this pattern is shared between *japonica* and *indica* groups [25]. There remains uncertainty about the directionality of cause and effect, the extent to which the correlation is driven by associations of both recombination and trans-

posable elements with other factors, or why patterns differ among species and types of repetitive elements [23].

Chromosome	All Contexts		CG Context		CHG Context		CHH Context	
	IR64	Azucena	IR64	Azucena	IR64	Azucena	IR64	Azucena
1	-0.17	-0.41	-0.43	-0.44	-0.45	-0.54	0.62	0.66
2	-0.34	-0.50	-0.42	-0.41	-0.52	-0.61	0.68	0.77
3	-0.54	-0.57	-0.56	-0.50	-0.66	-0.64	0.67	0.72
4	-0.55	-0.63	-0.55	-0.61	-0.68	-0.72	0.65	0.81
5	-0.50	-0.57	-0.59	-0.53	-0.70	-0.71	0.82	0.84
6	-0.35	-0.47	-0.44	-0.42	-0.55	-0.59	0.78	0.86
7	-0.47	-0.35	-0.46	-0.23	-0.53	-0.48	0.40	0.71
8	-0.42	-0.64	-0.62	-0.60	-0.67	-0.75	0.90	0.89
9	0.06	-0.11	-0.13	-0.16	-0.25	-0.34	0.65	0.78
10	-0.42	-0.55	-0.47	-0.45	-0.58	-0.65	0.62	0.71
11	-0.59	-0.64	-0.55	-0.54	-0.66	-0.69	0.73	0.82
12	-0.45	-0.60	-0.53	-0.44	-0.59	-0.67	0.81	0.87

Figure 2. Correlations between recombination rates and the count of methylated cytosines for rice varieties IR64 and Azucena. Blue and red colors correspond to positive and negative correlations, respectively. The higher the correlation value, the higher the color intensity.

The count of methylated cytosines is assessed within genes, transposons, and retrotransposons and compared to recombination rates (Supplementary Materials Figures S1 and S2). The analysis shows that methylated cytosine count in genes, transposons, and retrotransposons is negatively correlated with recombination rates when evaluated for all contexts together (Supplementary Materials Figure S3). This indicates that methylation inside these entities is higher when recombination is lower. The same negative trend is observed when methylated cytosines are analyzed in CG and CHG contexts. Methylation events in transposons and retrotransposons are associated with the prevention of their expression and movement in chromosomes, which can be damageable to the organism and even deleterious [23,26]. It should be noted that these methylation events can also affect surrounding genomic regions [26], potentially influencing the methylation status of nearby genes. In genes, methylation usually occurs at the promoters or within the body of the transcribed gene, inhibiting their expression [8]. However, the methylated cytosines in the CHH context are also positively correlated with the recombination rates. This is a consequence of low CHH methylation near the centromere region and greater presence in the chromosome arms. Gallo-Franco et al. [27] reported high CHH methylation levels of transposable elements close to genes in rice, which supports the conclusion of Martin et al. [28] for grass species that long genes and genes close to transposable elements tend to have CHH islands more frequently. It could be hypothesized that the presence of these CHH islands is promoting the positive correlation between methylation and recombination in gene-rich regions.

Chromosomal regions close to the centromere have a high incidence on methylation. When only the chromosome arms are evaluated, correlation trends change, from being high negative to being negative, for all contexts evaluated together and for the CG and CHG contexts evaluated independently (Figure 6). For CHH methylation, the markedly positive correlation also decreases but is still positive. In the context centromere regions are evaluated, negative correlations are evidenced in all contexts when they are evaluated together and for CG and CHG contexts independently. These results are in agreement with the reported importance of DNA methylation for plant chromosomal interactions in pericentromeric regions [9]. They also agree with the results obtained by Habu et al. [15], who indicate that the position and frequency of meiotic recombination in the centromeric heterochromatin of rice are regulated by the epigenetic state of the chromatin. With respect

to methylation in CHH contexts, the correlation of the centromere region is positive but weaker than that of the whole chromosome (Figure 6).
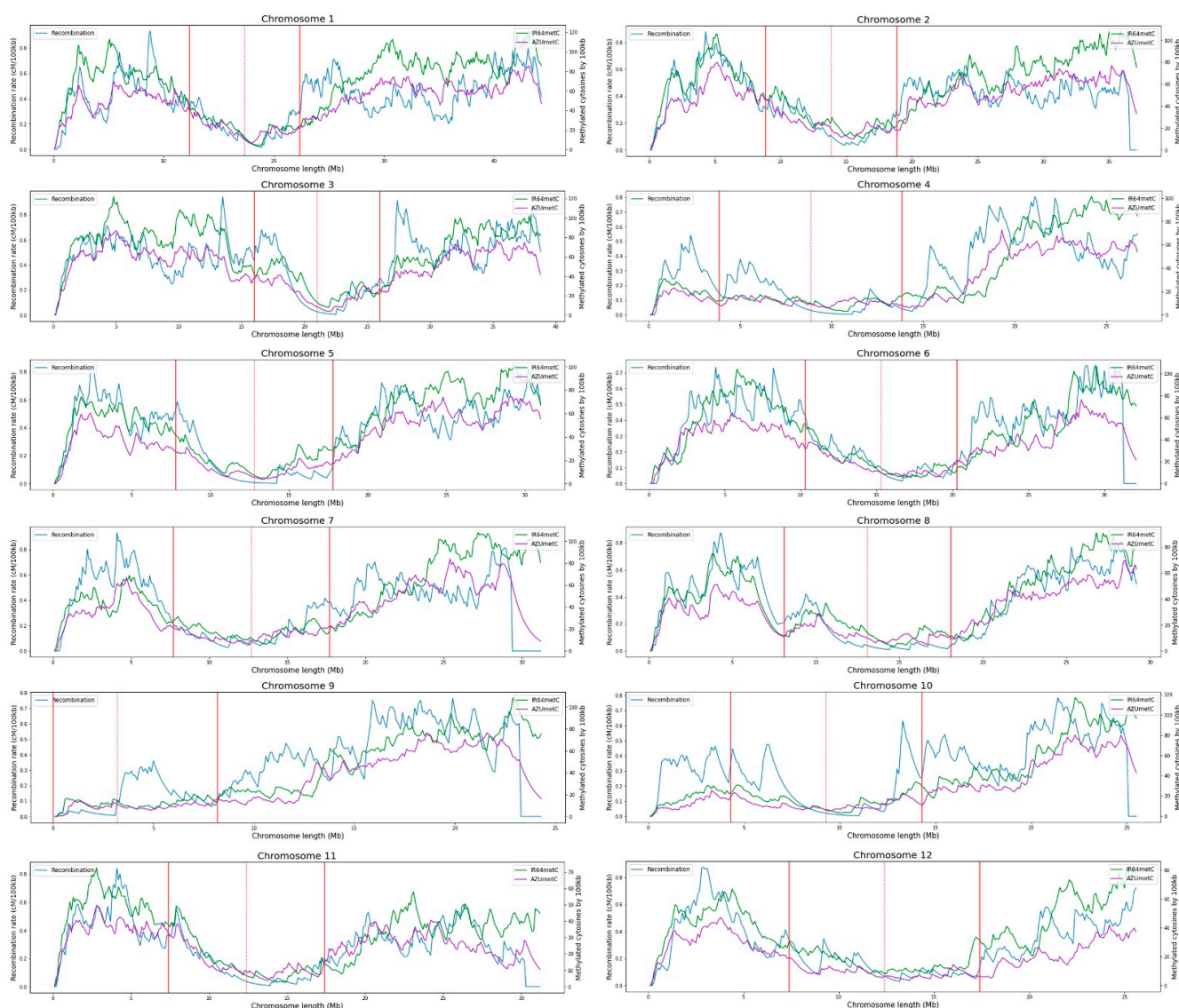

Figure 3. Distribution of methylated cytosines in CHH context in the twelve rice chromosomes for the IR64 and Azucena varieties, in comparison with the chromosomal recombination between these two varieties. The centromere is represented by a red dotted line and the influence of the centromere region in recombination by solid red lines.

The contributions of methylation in CG, CHG, and CHH contexts to predict recombination as features of machine learning models are assessed using the Shapley package. The results show a great contribution of CHH for the prediction of recombination and a low contribution of CG and CHG for both varieties (Figure 7). This agrees with the fact that the CHH context has the highest correlation values with respect to chromosome recombination rates, while the CG and CHG contexts have lower correlations. The Shap summary plot also shows the same trend, evidencing the strongest effect on recombination when the CHH values are higher.
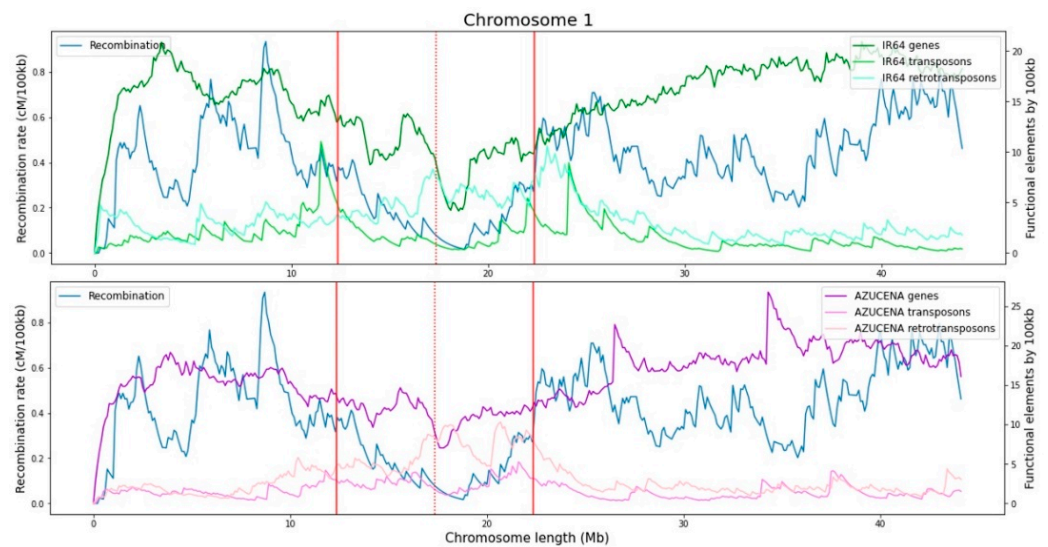
Figure 4. Genes, transposons, and retrotransposons compared to cross over recombination through chromosome 1 for rice varieties IR64 and Azucena. The centromere is represented by a red dotted line and the influence of the centromere region in recombination by solid red lines.

| Chromosome | Complete chromosome | | | | | | Chromosome arms | | | | | | Centromere region | | | | | |
| | Genes | | Transposons | | Retrotransposons | | Genes | | Transposons | | Retrotransposons | | Genes | | Transposons | | Retrotransposons | |
	IR64	Azucena	IR64	Azucena	IR64	Azucena	IR64	Azucena	IR64	Azucena	IR64	Azucena	IR64	Azucena	IR64	Azucena	IR64	Azucena
1	0.61	0.49	0.12	-0.34	-0.24	-0.52	0.25	0.13	0.06	-0.10	0.18	0.00	0.63	0.52	0.78	0.45	-0.44	-0.33
2	0.73	0.70	0.13	0.05	-0.39	-0.65	0.47	0.54	0.29	0.27	0.02	0.04	0.81	0.52	0.29	0.49	-0.70	-0.64
3	0.66	0.69	0.14	0.24	-0.57	-0.63	0.41	0.39	0.22	0.16	-0.04	0.07	0.72	0.76	0.03	0.42	-0.46	-0.88
4	0.55	0.63	0.10	-0.32	-0.60	-0.73	0.25	0.49	-0.07	-0.47	-0.27	-0.51	0.84	0.54	0.46	0.08	-0.71	-0.42
5	0.73	0.79	-0.29	-0.43	-0.75	-0.79	0.48	0.55	-0.13	0.06	-0.24	-0.33	0.60	0.70	0.28	-0.13	-0.65	-0.71
6	0.79	0.77	0.24	-0.26	-0.47	-0.62	0.71	0.69	0.05	-0.16	0.01	-0.12	0.75	0.71	0.36	-0.49	-0.52	-0.75
7	0.74	0.80	-0.09	-0.49	-0.68	-0.74	0.62	0.67	0.09	-0.22	-0.03	-0.19	-0.16	0.33	-0.35	0.11	-0.66	-0.74
8	0.84	0.84	-0.03	-0.05	-0.46	-0.61	0.83	0.75	-0.14	0.10	-0.26	-0.32	0.44	0.54	0.41	-0.07	-0.05	-0.46
9	0.79	0.80	-0.18	-0.27	-0.37	-0.73	0.65	0.67	-0.50	-0.36	-0.65	-0.78	0.12	0.22	-0.39	-0.09	0.73	0.51
10	0.74	0.64	-0.16	0.00	-0.65	-0.57	0.63	0.67	-0.19	-0.19	-0.01	-0.28	0.75	0.26	0.32	0.04	-0.76	-0.45
11	0.76	0.76	-0.10	-0.09	-0.67	-0.67	0.60	0.68	-0.30	0.03	-0.24	-0.04	0.80	0.86	0.06	0.36	-0.81	-0.80
12	0.88	0.80	-0.24	0.27	-0.43	-0.44	0.82	0.71	-0.38	0.19	0.12	-0.15	0.72	0.85	0.12	0.69	-0.56	-0.10

Figure 5. Correlations between recombination rates and the number of genes, transposons, and retrotransposons rates for rice varieties IR64 and Azucena. Blue and red colors correspond to positive and negative correlations, respectively. The higher the correlation value, the higher the color intensity.

Subsequently, the methylated cytosine count in the CHH context is used as a unique feature to evaluate regression algorithms of machine learning, because the performance of the model decreases when the other features are considered. The evaluation is carried out independently for each variety using the Lazy Predict package. The results show that the Extra Trees algorithm performed the best prediction (R^2 = 0.57, *RMSE* = 0.01 for IR64; R^2 = 0.69, *RMSE* = 0.01 for Azucena). Thus, this algorithm is used to develop the training and subsequent predictions.

Predictions on Azucena's chromosomes, by training the Extra Trees algorithm with information from IR64, give an R^2 of 0.32 ± 0.13 and an *MSE* of 0.02 ± 0.00, on average. Meanwhile, predictions on IR64's chromosomes by training the Extra Trees algorithm with information from Azucena give an R^2 of 0.21 ± 0.21 and an *MSE* of 0.03 ± 0.00, on average. In both cases, the average correlation values between predictions and recombination rates are 0.67 ± 0.06 for Azucena and 0.65 ± 0.07 for IR64, evidencing a positive trend (Table 1, Figure 8).

Chromosome	All Contexts						CG Context					
	Complete Chromosome		Chromosome arms		Centromere region		Complete Chromosome		Chromosome arms		Centromere region	
	IR64	Azucena	IR64	Azucena	IR64	Azucena	IR64	Azucena	IR64	Azucena	IR64	Azucena
1	-0.17	-0.41	0.26	0.25	-0.22	-0.68	-0.43	-0.44	0.21	0.23	-0.78	-0.83
2	-0.34	-0.50	0.22	0.21	-0.86	-0.86	-0.42	-0.41	0.22	0.33	-0.84	-0.89
3	-0.54	-0.57	0.25	0.22	-0.77	-0.80	-0.56	-0.50	0.14	0.22	-0.85	-0.82
4	-0.55	-0.63	-0.23	-0.22	-0.71	-0.59	-0.55	-0.61	-0.25	-0.26	-0.19	-0.61
5	-0.50	-0.57	0.09	0.07	-0.78	-0.69	-0.59	-0.53	0.06	0.10	-0.67	-0.32
6	-0.35	-0.47	0.22	0.16	-0.79	-0.69	-0.44	-0.42	0.12	0.23	-0.83	-0.66
7	-0.47	-0.35	0.39	0.46	-0.22	-0.36	-0.46	-0.23	0.08	0.49	-0.37	-0.11
8	-0.42	-0.64	-0.25	-0.31	-0.59	-0.69	-0.62	-0.60	-0.31	-0.31	-0.70	-0.45
9	0.06	-0.11	0.01	0.04	0.62	0.67	-0.13	-0.16	-0.32	-0.13	0.67	0.63
10	-0.42	-0.55	0.15	0.15	-0.74	-0.82	-0.47	-0.45	0.03	0.05	-0.60	-0.60
11	-0.59	-0.64	-0.24	-0.24	-0.78	-0.87	-0.55	-0.54	-0.16	-0.15	-0.84	-0.91
12	-0.45	-0.60	-0.25	-0.28	-0.38	-0.69	-0.53	-0.44	-0.09	-0.12	-0.64	-0.78

Chromosome	CHG Context						CHH Context					
	Complete Chromosome		Chromosome arms		Centromere region		Complete Chromosome		Chromosome arms		Centromere region	
	IR64	Azucena	IR64	Azucena	IR64	Azucena	IR64	Azucena	IR64	Azucena	IR64	Azucena
1	-0.45	-0.54	0.17	0.08	-0.43	-0.74	0.62	0.66	0.25	0.35	0.77	0.72
2	-0.52	-0.61	0.10	0.02	-0.88	-0.87	0.68	0.77	0.36	0.55	0.80	0.74
3	-0.66	-0.64	0.00	0.04	-0.82	-0.84	0.67	0.72	0.38	0.43	0.79	0.88
4	-0.68	-0.72	-0.31	-0.45	-0.73	-0.61	0.65	0.81	0.48	0.74	0.64	0.63
5	-0.70	-0.71	-0.07	-0.13	-0.80	-0.75	0.82	0.84	0.59	0.63	0.68	0.80
6	-0.55	-0.59	-0.05	-0.06	-0.83	-0.75	0.78	0.86	0.60	0.73	0.86	0.88
7	-0.53	-0.48	-0.01	0.27	-0.28	-0.42	0.40	0.71	-0.03	0.52	0.60	0.62
8	-0.67	-0.75	-0.31	-0.48	-0.67	-0.75	0.90	0.89	0.85	0.81	0.69	0.82
9	-0.25	-0.34	-0.23	-0.19	0.63	0.68	0.65	0.78	0.41	0.66	-0.21	-0.10
10	-0.58	-0.65	0.00	-0.18	-0.76	-0.83	0.62	0.71	0.46	0.65	0.73	0.71
11	-0.66	-0.69	-0.28	-0.34	-0.84	-0.89	0.73	0.82	0.49	0.67	0.93	0.90
12	-0.59	-0.67	-0.15	-0.40	-0.49	-0.73	0.81	0.87	0.68	0.80	0.73	0.83

Figure 6. Correlations between recombination rates and the count of methylated cytosines for complete chromosomes, chromosome arms, and centromere region of rice varieties IR64 and Azucena. Blue and red colors correspond to positive and negative correlations, respectively. The higher the correlation value, the higher the color intensity.

Several studies have focused on predicting recombination using machine learning. For example, Liu et al. [11] combined support vector machines with consensus feature dinucleotide-based autocross covariance to predict the recombination of hot/cold spots in yeast. Demirci et al. [12] used features, such as gene annotation, propeller, and helical twist, AT/TA dinucleotides, and CA dinucleotides to train machine learning models for predicting crossover occurrences in *Arabidopsis,* maize, rice, and tomato. More recently, Adrion et al. [13] proposed an approach to predict the recombination landscape in African populations of *Drosophila melanogaster* using deep learning with recurrent neural networks. For all cases, the results have been satisfactory according to the specific objective of each study, which demonstrates the power of machine learning approaches to predict complex traits such as chromosomal recombination.

The Extra Trees regression model makes it possible to predict chromosomal recombination using a single feature: The CHH methylated cytosine count. It is possible due to the high correlation between this feature and the recombination rates, which behaved similarly in all chromosomes. The model was trained on a dataset of one variety and was tested it on the other, performing two independent tests and finding that results were consistent (Figure 8). This opens a door for future studies. The evidence suggests that these models can be used to predict chromosomal recombination rates in any variety of *Oryza sativa* rice.

This is because the two varieties used in this study, IR64 and Azucena, are highly distant genetically, belonging to the *indica* and *japonica* groups, respectively.
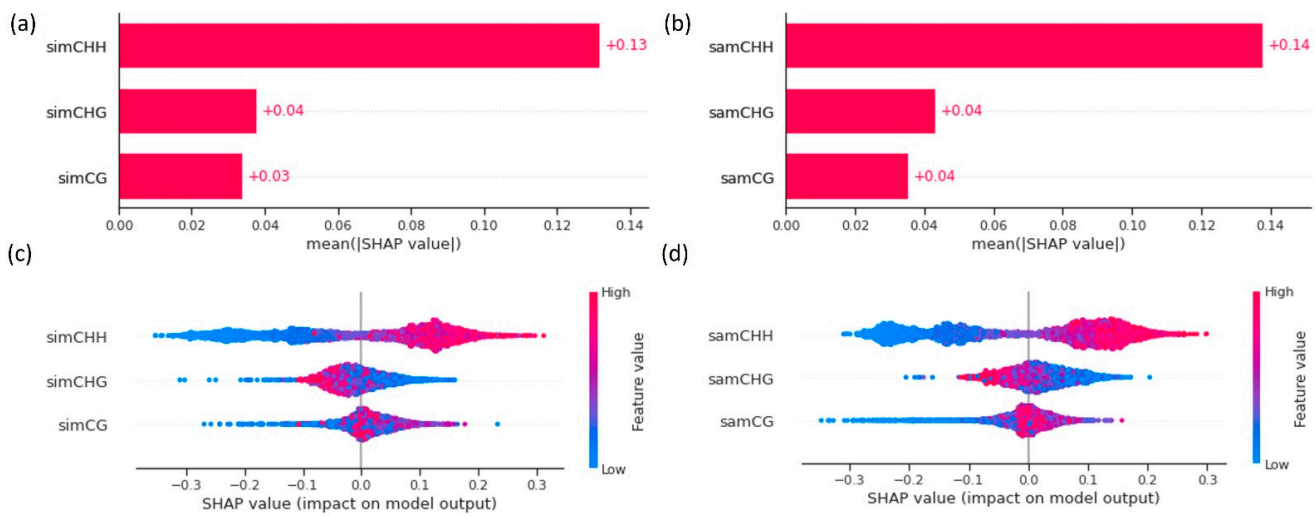

Figure 7. Shap values and contributions of features CG, CHG, and CHH to the prediction of recombination rates, using IR64 and Azucena data: (**a**) Shap summary plot of features for the IR64 variety; (**b**) Shap summary plot of features for the Azucena variety; (**c**) Shap values for the IR64 variety; and (**d**) Shap values for the Azucena variety.

Table 1. Performance of chromosome recombination rates predictions of IR64 and Azucena rice varieties using the Extra Trees model trained with CHH methylation data.

Chromosome	IR64			Azucena		
	R^2	Correlation	MSE	R^2	Correlation	MSE
1	0.00	0.63	0.03	0.44	0.67	0.02
2	0.04	0.66	0.03	0.53	0.73	0.01
3	0.37	0.70	0.02	0.49	0.72	0.02
4	0.44	0.72	0.02	0.60	0.81	0.01
5	0.59	0.81	0.02	0.67	0.84	0.01
6	0.44	0.78	0.02	0.68	0.82	0.01
7	0.16	0.53	0.04	0.50	0.73	0.02
8	0.71	0.85	0.01	0.67	0.88	0.02
9	0.32	0.65	0.03	0.50	0.75	0.02
10	0.41	0.70	0.02	0.28	0.69	0.03
11	0.30	0.70	0.02	0.52	0.77	0.01
12	0.54	0.77	0.01	0.35	0.85	0.02
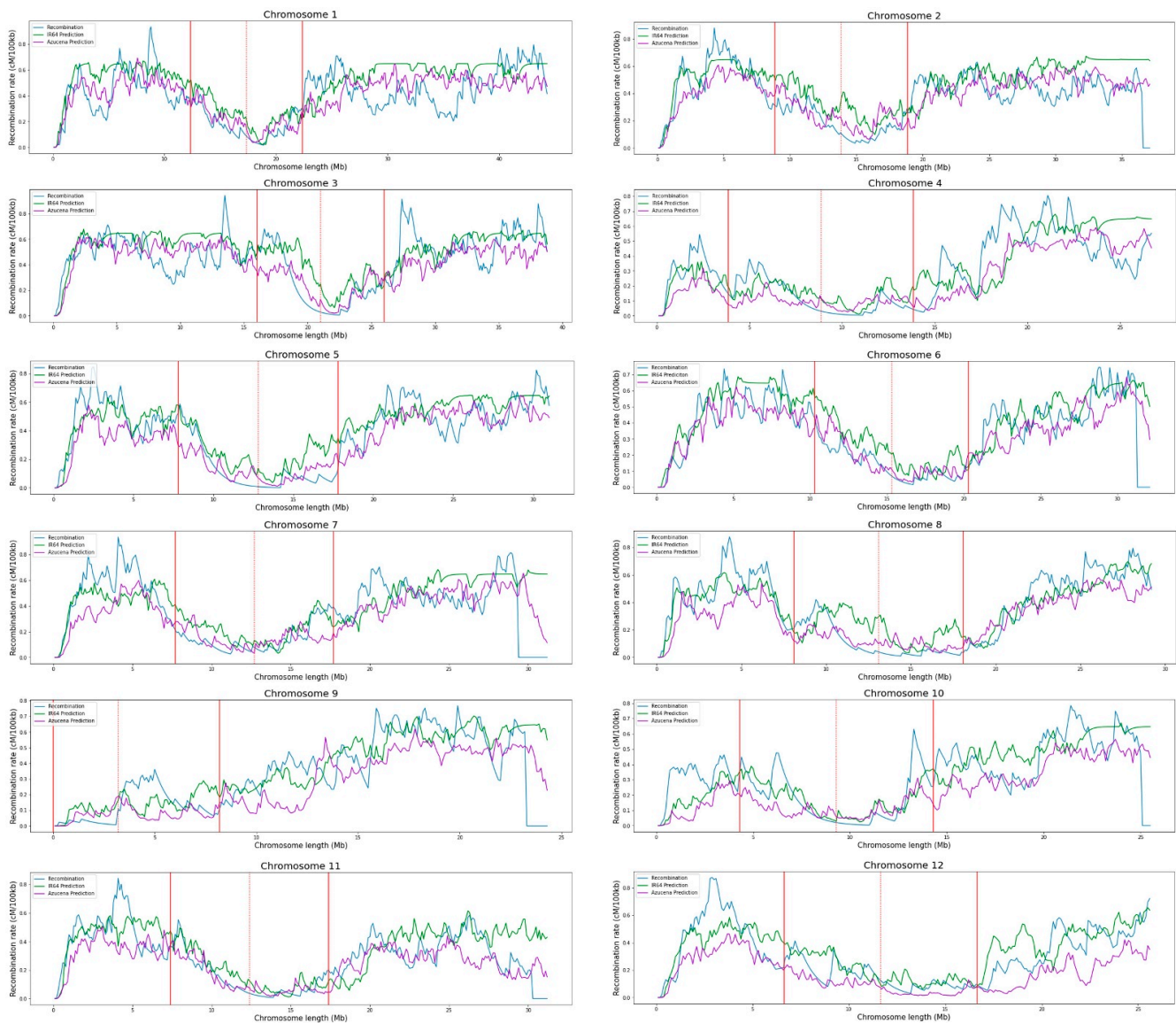
Figure 8. Recombination predictions between IR64 and Azucena varieties by the Extra Trees machine learning model using the count of methylated cytosines in the CHH context as a feature. Predictions on the IR64 manifold are made using Azucena methylation as the training dataset, and predictions on the Azucena manifold are made using IR64 methylation as the training dataset. The centromere is represented by a red dotted line and the influence of the centromere region in recombination by solid red lines.

3. Materials and Methods

3.1. Recombination Rates

The recombination rates are estimated from an inter-subspecific segregating population of 212 F11 recombinant inbred lines (RIL). They are obtained by single seed descent, derived from the cross between the rice varieties IR64 (*indica* group) and Azucena (tropical *japonica* group), and genotyped using shallow Illumina sequencing (~2×) followed by imputation with NOISYmputer. Local recombination rates in cM/bp are calculated in sliding windows of 100 kb using MapDisto.

3.2. Plant Material and Growth Conditions for Methylation Experiment

Seeds of rice varieties IR64 and Azucena were germinated and grown in a growth chamber at 30 °C and 12:12 dark/light conditions for 10 days. Seedlings were transferred

to a hydroponic medium with a Kimura B solution (pH 7) and Arnon micronutrients. Roots from three weeks-old seedlings were collected and stored at $-80\ ^\circ$C. Total genomic DNA was extracted from frozen root tissue by CTAB 2X protocol with modifications [29]. Genomic DNA quality was evaluated on agarose gels, and DNA quantity was measured using a Nanodrop spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA).

3.3. Whole-Genome Bisulfite Sequencing and Data Analysis

Bisulfite-seq (BS-seq) libraries were made from genomic DNA isolated from IR64 and Azucena seedling roots. DNA from three independent seedlings for each genotype was pooled as one sample and sequenced. Bisulfite conversion of DNA, library construction, and sequencing were performed by CD Genomics (CD Genomics Inc., Shirley, New York, NY, USA). Raw data are available in the GenBank repository for IR64 (Accession number: SRR20325840) and Azucena (Accession number: SRR20325842). Basic statistics on the quality of the raw reads was done with the FastQC tool (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 5 September 2021)). Sequencing adapters and low-quality data of the sequencing data were removed by Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic (accessed 21 November 2021)). Cleaned data were aligned to the reference genomes reported in the GenBank repository for IR64 (Accession number: RWKJ00000000) and Azucena (Accession number: PKQC000000000) using Bismark v.0.16.3 [30] with default parameters. Only uniquely aligned reads were maintained. Methylation calling data obtained from Bismark were used for further analysis.

3.4. Comparison between Recombination Rates and Methylation Patterns

To compare the methylation patterns with the local recombination rates, the genomes were divided into 100 kb windows, in which the number of cytosines with a methylation level greater than 75% was calculated for each of the CG, CHG, and CHH contexts. Exponential smoothing with $\alpha = 0.1$ was applied to the recombination and methylation data to remove noise associated with the abrupt change in the count of methylated cytosines in adjacent windows. Subsequently, a Pearson correlation analysis per chromosome was developed to evaluate the linear relationships between the recombination rates and the methylation patterns of both varieties.

3.5. Functional Evaluation

Gene, transposon, and retrotransposon annotation information from both varieties was used (Supplementary Materials Tables S1 and S2). Pearson correlation analyses were carried out between the number of genes, transposons, and retrotransposons with respect to recombination of the chromosome to investigate their relationship with the recombination landscape. Later, the start and end coordinates of these elements were used to extract the count of methylated cytosines inside them. New correlation analyses were performed to learn the trends between methylated cytosines for each context within these functional elements with respect to recombination. A differentiation between the centromere and non-centromere regions was also included.

3.6. Machine Learning Modeling

To assess the usefulness of methylation in predicting chromosome recombination, different machine learning approaches were explored. The total counts of methylated cytosines in windows of 100 kb belonging to the CG, CHG, and CHH contexts for each variety were evaluated as features for machine learning modeling using the Shapley package (https://shap.readthedocs.io/en/latest/index.html (accessed on 2 February 2022)). Subsequently, the performance of different machine learning models was evaluated using the LazyPredict package (https://pypi.org/project/lazypredict/ (accessed on 2 February 2022)). Exponential smoothing with $\alpha = 0.1$ was applied to the data input before training the model and another one to the model output with $\alpha = 0.3$. The coefficient of determination R^2 and the root of the mean square error *RMSE* were used to evaluate the performance

of the models. *MSE* was used for predictions. Pearson correlation analyses were also performed to discover general linear trends between the predictions and the experimental data. The resulting best model was fitted, and the information from the twelve chromosomes of one variety was used as a training dataset to predict the recombination rates in each of the twelve chromosomes of the other variety. All these analyses and the previous ones were run in Python.

4. Conclusions

This study reported on how methylated cytosines in the CHH context positively correlate with recombination rates in the twelve rice chromosomes of two genetically distant rice varieties: IR64 and Azucena. However, a negative correlation was found between methylation and recombination rates when only CG and CHG contexts were tested, as well as in the three methylation contexts together. For this case, the positive correlation of CHH was hidden due to the high number of methylated cytosines from the CG and CHG contexts. In addition, functional analysis showed that genes were positively correlated with recombination rates, unlike transposons and retrotransposons, which showed a negative correlation. The correlation between methylation and recombination suggests the same trends for the entire genome with respect to only methylation in genes, transposons, and retrotransposons. The influence of the centromere on methylation patterns and its correlation with recombination rates was evident, supporting the hypothesis that the position and frequency of meiotic recombination in rice centromeric heterochromatin are regulated by the epigenetic state of the chromatin. Finally, a machine learning model was proposed and trained using the CHH methylated cytosine count to predict recombination rates, which obtained consistent results in two independent data sets. This suggests that the extraction of methylation data and the use of machine learning models in future studies is a promising path to focus on predicting recombination rates using the count of CHH-methylated cytosines in rice as a feature. Colleagues are invited to explore how the counting of CHH-methylated cytosines in other species behaves with respect to chromosomal recombination.

References

1. Rodgers-Melnick, E.; Bradbury, P.J.; Elshire, R.J.; Glaubitz, J.C.; Acharya, C.B.; Mitchell, S.E.; Li, C.; Li, Y.; Buckler, E.S. Recombination in Diverse Maize Is Stable, Predictable, and Associated with Genetic Load. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 3823–3828. [CrossRef]
2. Lambing, C.; Franklin, F.C.H.; Wang, C.-J.R. Understanding and Manipulating Meiotic Recombination in Plants. *Plant Physiol.* **2017**, *173*, 1530–1542. [CrossRef]
3. Henderson, I.R. Control of Meiotic Recombination Frequency in Plant Genomes. *Curr. Opin. Plant Biol.* **2012**, *15*, 556–561. [CrossRef] [PubMed]
4. Perrella, G.; Consiglio, M.F.; Aiese-Cigliano, R.; Cremona, G.; Sanchez-Moran, E.; Barra, L.; Errico, A.; Bressan, R.A.; Franklin, F.C.H.; Conicella, C. Histone Hyperacetylation Affects Meiotic Recombination and Chromosome Segregation in Arabidopsis: Histone Acetylation in At Meiosis. *Plant J.* **2010**, *62*, 796–806. [CrossRef]
5. Melamed-Bessudo, C.; Levy, A.A. Deficiency in DNA Methylation Increases Meiotic Crossover Rates in Euchromatic but Not in Heterochromatic Regions in Arabidopsis. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, E981–E988. [CrossRef]
6. Mirouze, M.; Lieberman-Lazarovich, M.; Aversano, R.; Bucher, E.; Nicolet, J.; Reinders, J.; Paszkowski, J. Loss of DNA Methylation Affects the Recombination Landscape in Arabidopsis. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 5880–5885. [CrossRef] [PubMed]
7. Bräutigam, K.; Cronk, Q. DNA Methylation and the Evolution of Developmental Complexity in Plants. *Front. Plant Sci.* **2018**, *9*, 1447. [CrossRef] [PubMed]
8. Zhang, H.; Lang, Z.; Zhu, J.-K. Dynamics and Function of DNA Methylation in Plants. *Nat. Rev. Mol. Cell Biol.* **2018**, *19*, 489–506. [CrossRef]
9. Kawashima, T.; Berger, F. Epigenetic Reprogramming in Plant Sexual Reproduction. *Nat. Rev. Genet.* **2014**, *15*, 613–624. [CrossRef]
10. Lanciano, S.; Mirouze, M. DNA Methylation in Rice and Relevance for Breeding. *Epigenomes* **2017**, *1*, 10. [CrossRef]
11. Liu, B.; Liu, Y.; Jin, X.; Wang, X.; Liu, B. IRSpot-DACC: A Computational Predictor for Recombination Hot/Cold Spots Identification Based on Dinucleotide-Based Auto-Cross Covariance. *Sci. Rep.* **2016**, *6*, 33483. [CrossRef] [PubMed]
12. Demirci, S.; Peters, S.A.; de Ridder, D.; van Dijk, A.D.J. DNA Sequence and Shape Are Predictive for Meiotic Crossovers throughout the Plant Kingdom. *Plant J.* **2018**, *95*, 686–699. [CrossRef]
13. Adrion, J.R.; Galloway, J.G.; Kern, A.D. Predicting the Landscape of Recombination Using Deep Learning. *Mol. Biol. Evol.* **2020**, *37*, 1790–1808. [CrossRef] [PubMed]
14. Cheng, Z.; Buell, C.R.; Wing, R.A.; Gu, M.; Jiang, J. Toward a Cytological Characterization of the Rice Genome. *Genome Res.* **2001**, *11*, 2133–2141. [CrossRef]
15. Habu, Y.; Ando, T.; Ito, S.; Nagaki, K.; Kishimoto, N.; Taguchi-Shiobara, F.; Numa, H.; Yamaguchi, K.; Shigenobu, S.; Murata, M.; et al. Epigenomic Modification in Rice Controls Meiotic Recombination and Segregation Distortion. *Mol. Breed.* **2015**, *35*, 103. [CrossRef]
16. Choi, J.Y.; Purugganan, M.D. Evolutionary Epigenomics of Retrotransposon-Mediated Methylation Spreading in Rice. *Mol. Biol. Evol.* **2018**, *35*, 365–382. [CrossRef]
17. Yan, H.; Kikuchi, S.; Neumann, P.; Zhang, W.; Wu, Y.; Chen, F.; Jiang, J. Genome-Wide Mapping of Cytosine Methylation Revealed Dynamic DNA Methylation Patterns Associated with Genes and Centromeres in Rice: Genome-Wide Methylation of Rice. *Plant J.* **2010**, *63*, 353–365. [CrossRef]
18. Yan, H.; Jin, W.; Nagaki, K.; Tian, S.; Ouyang, S.; Buell, C.R.; Talbert, P.B.; Henikoff, S.; Jiang, J. Transcription and Histone Modifications in the Recombination-Free Region Spanning a Rice Centromere. *Plant Cell* **2005**, *17*, 3227–3238. [CrossRef]
19. Li, X.; Zhu, J.; Hu, F.; Ge, S.; Ye, M.; Xiang, H.; Zhang, G.; Zheng, X.; Zhang, H.; Zhang, S.; et al. Single-Base Resolution Maps of Cultivated and Wild Rice Methylomes and Regulatory Roles of DNA Methylation in Plant Gene Expression. *BMC Genom.* **2012**, *13*, 300. [CrossRef]
20. Daccord, N.; Celton, J.-M.; Linsmith, G.; Becker, C.; Choisne, N.; Schijlen, E.; van de Geest, H.; Bianco, L.; Micheletti, D.; Velasco, R.; et al. High-Quality de Novo Assembly of the Apple Genome and Methylome Dynamics of Early Fruit Development. *Nat. Genet.* **2017**, *49*, 1099–1106. [CrossRef]
21. Zakrzewski, F.; Schmidt, M.; Van Lijsebettens, M.; Schmidt, T. DNA Methylation of Retrotransposons, DNA Transposons and Genes in Sugar Beet (*Beta vulgaris* L.). *Plant J.* **2017**, *90*, 1156–1175. [CrossRef]
22. Wang, L.; Zheng, K.; Zeng, L.; Xu, D.; Zhu, T.; Yin, Y.; Zhan, H.; Wu, Y.; Yang, D.-L. Reinforcement of CHH Methylation through RNA-Directed DNA Methylation Ensures Sexual Reproduction in Rice. *Plant Physiol.* **2022**, *188*, 1189–1209. [CrossRef] [PubMed]
23. Kent, T.V.; Uzunović, J.; Wright, S.I. Coevolution between Transposable Elements and Recombination. *Philos. Trans. R. Soc. B Biol. Sci.* **2017**, *372*, 20160458. [CrossRef]
24. Anderson, L.K.; Lai, A.; Stack, S.M.; Rizzon, C.; Gaut, B.S. Uneven Distribution of Expressed Sequence Tag Loci on Maize Pachytene Chromosomes. *Genome Res.* **2006**, *16*, 115–122. [CrossRef] [PubMed]
25. Tian, Z.; Rizzon, C.; Du, J.; Zhu, L.; Bennetzen, J.L.; Jackson, S.A.; Gaut, B.S.; Ma, J. Do Genetic Recombination and Gene Density Shape the Pattern of DNA Elimination in Rice Long Terminal Repeat Retrotransposons? *Genome Res.* **2009**, *19*, 2221–2230. [CrossRef]
26. Ahmed, I.; Sarazin, A.; Bowler, C.; Colot, V.; Quesneville, H. Genome-Wide Evidence for Local DNA Methylation Spreading from Small RNA-Targeted Sequences in Arabidopsis. *Nucleic Acids Res.* **2011**, *39*, 6919–6931. [CrossRef] [PubMed]

27. Gallo-Franco, J.J.; Ghneim-Herrera, T.; Tobar-Tosse, F.; Romero, M.; Chaura, J.; Quimbaya, M. Whole-genome DNA Methylation Patterns of *Oryza sativa* (L.) and *Oryza glumaepatula* (Steud) Genotypes Associated with Aluminum Response. *Plant Direct* **2022**, *6*, e430. [CrossRef] [PubMed]

28. Martin, G.T.; Seymour, D.K.; Gaut, B.S. CHH Methylation Islands: A Nonconserved Feature of Grass Genomes That Is Positively Associated with Transposable Elements but Negatively Associated with Gene-Body Methylation. *Genome Biol. Evol.* **2021**, *13*, evab144. [CrossRef]

29. Maropola, M.K.A.; Ramond, J.-B.; Trindade, M. Impact of Metagenomic DNA Extraction Procedures on the Identifiable Endophytic Bacterial Diversity in Sorghum Bicolor (L. Moench). *J. Microbiol. Methods* **2015**, *112*, 104–117. [CrossRef]

30. Krueger, F.; Andrews, S.R. Bismark: A Flexible Aligner and Methylation Caller for Bisulfite-Seq Applications. *Bioinformatics* **2011**, *27*, 1571–1572. [CrossRef]