EPIPLANT 2022

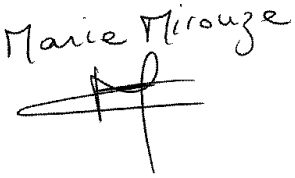
22-24 June 2022, Banyuls-sur-mer, France

I hereby certify that Mauricio Peñuela has participated in the above mentioned international conference.

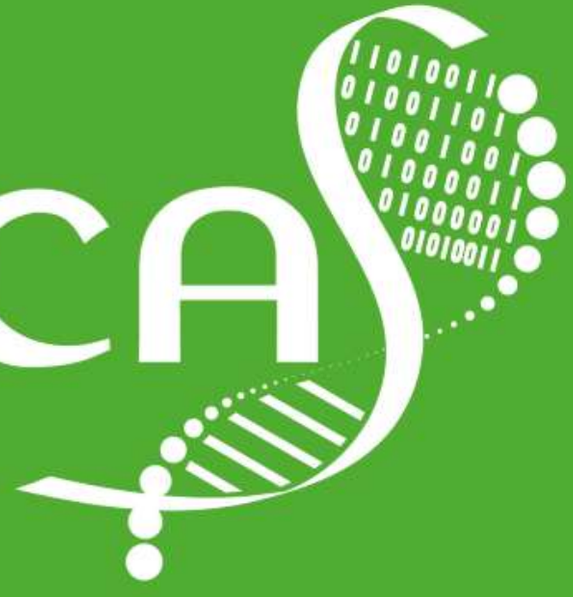
I also certify that Mauricio Peñuela has given a talk on Thursday, June 23rd.

Perpignan, July 4th, 2022

Marie Mirouze
For the Local Scientific Committee



ÓMICAS



El futuro es de todos
Gobierno de Colombia

COLOMBIA CIENTÍFICA

METHYLATION IN THE CHH CONTEXT ALLOWS TO PREDICT RECOMBINATION IN RICE

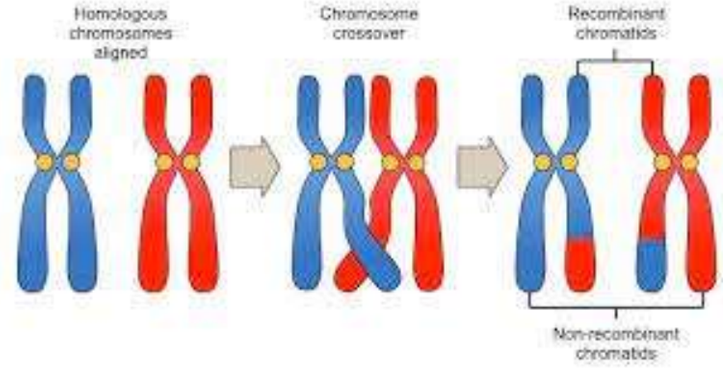
Mauricio Peñuela*, Jenny Johana Gallo-Franco, Jorge Finke, Camilo Rocha,
Anestis Gkanogiannis, Thaura Ghneim Herrera, Mathias Lorieux

mauricio.penuela@javerianacali.edu.co

In-silico

Introduction

- Crossover recombination
- Genetic breeders can use crossover recombination to develop new varieties
- Rice is of great importance in food security
- Rice has been a key crop to comprehension of the methylation process



Objectives

In this work we explore the relationship between chromosomal recombination rates and different methylation contexts, using *Oryza sativa* rice as a model

Focusing on the following objectives:

- Estimate the correlation between recombination and methylation in all contexts

- Implement a machine learning model to predict recombination based on methylation data

Materials and Methods

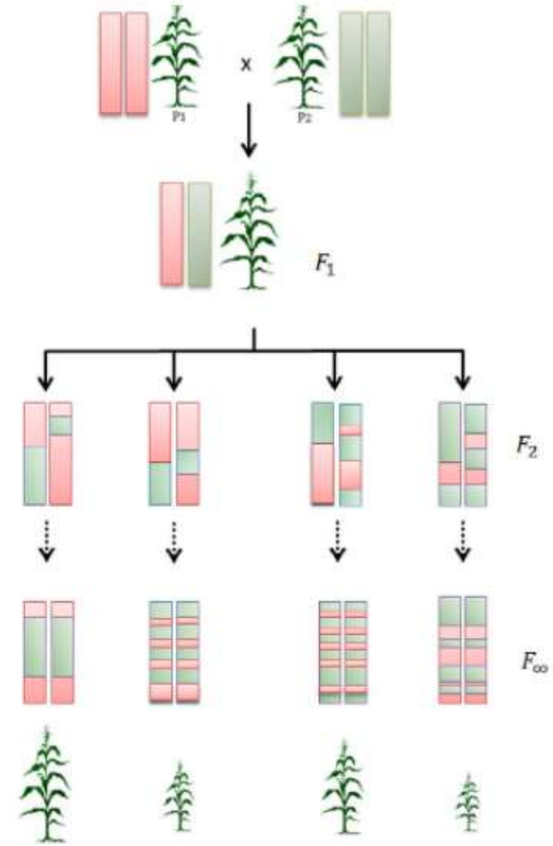
IR64 (Indica group) × **Azucena (Japonica group)**

F1

F2

F12 **212 RILs**

Recombinant Inbred Lines



(Behrouzi & Wit 2017)

In-silico

Materials and Methods

IR64 (Indica group)

Azucena (Japonica group)

F1

F2

F12 212 RILs

Recombinant Inbred Lines

1

PACBIO Whole Genome Sequencing

30x

→ GenBank: RWKJ00000000 PKQC000000000

2

3x PacBio → Recombination Map (100kb)

3

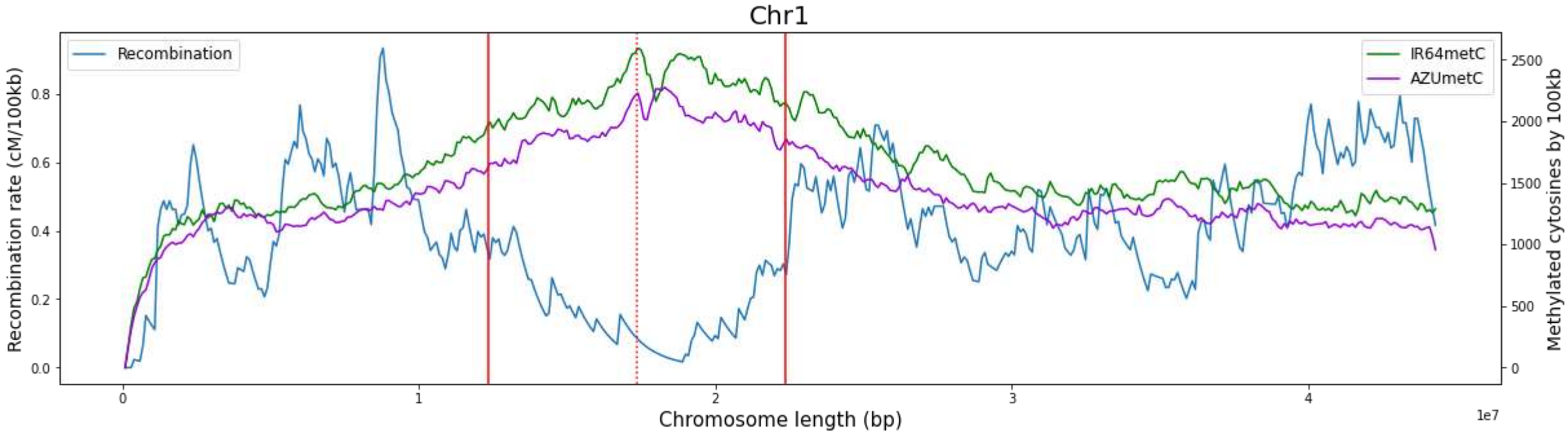
Illumina Whole Genome Bisulfite Sequencing 30x

Methylomes for IR64 and Azucena

CONTENT OF CG, CHG, CHH

In-silico

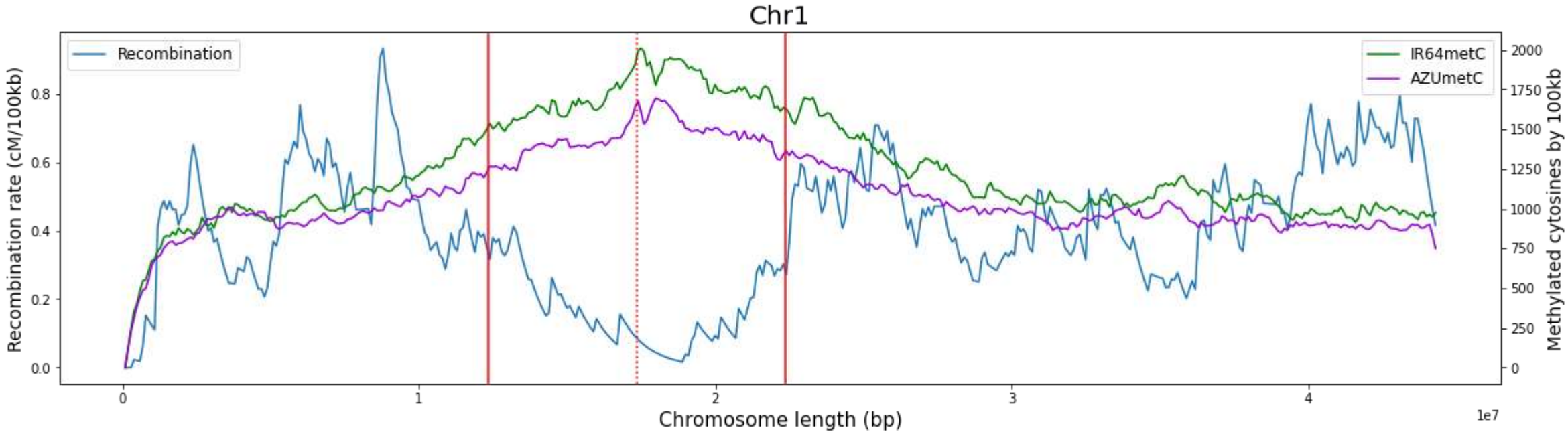
Recombination rates and abundance of methylated cytosines in all



Window Size: 100kb
Exponential Smothing: 0.9

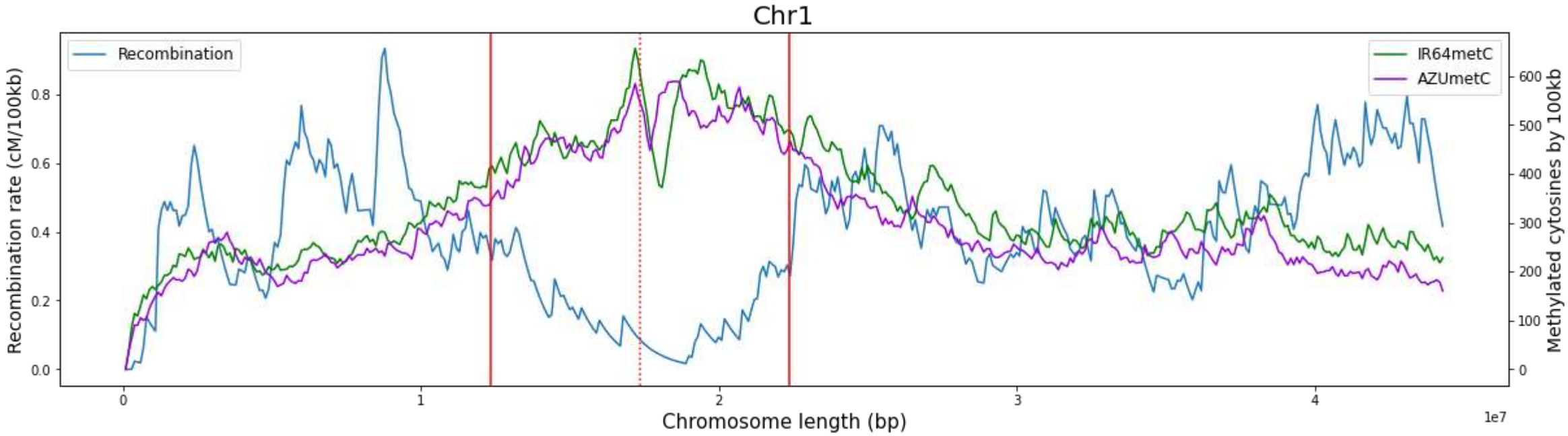
All Contexts

Recombination rates and abundance of methylated cytosines in the CG



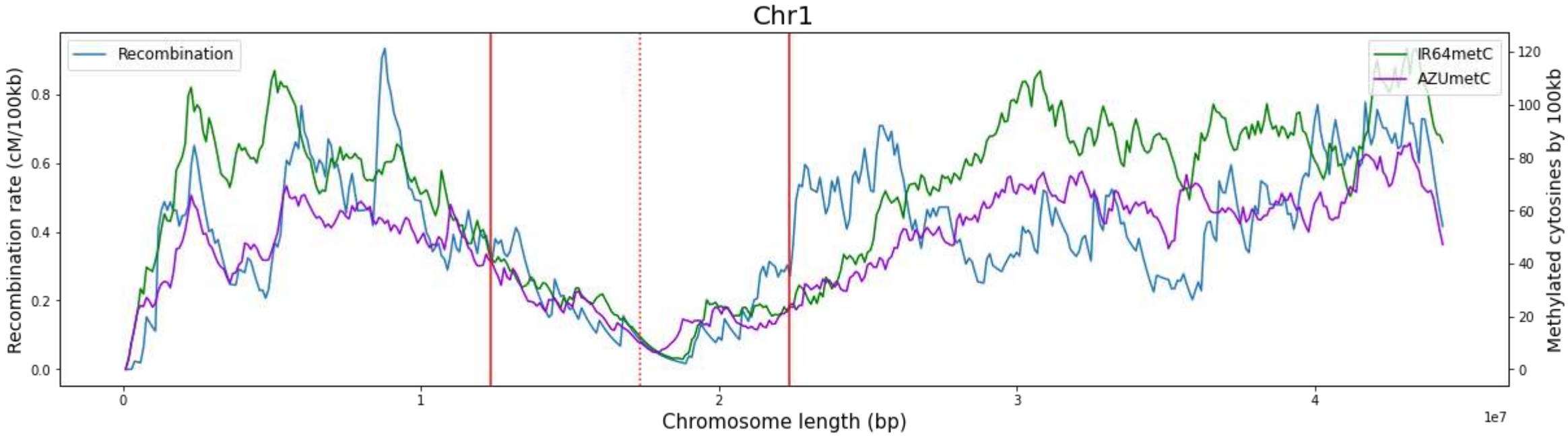
CG Context

Recombination rates and abundance of methylated cytosines in the CHG



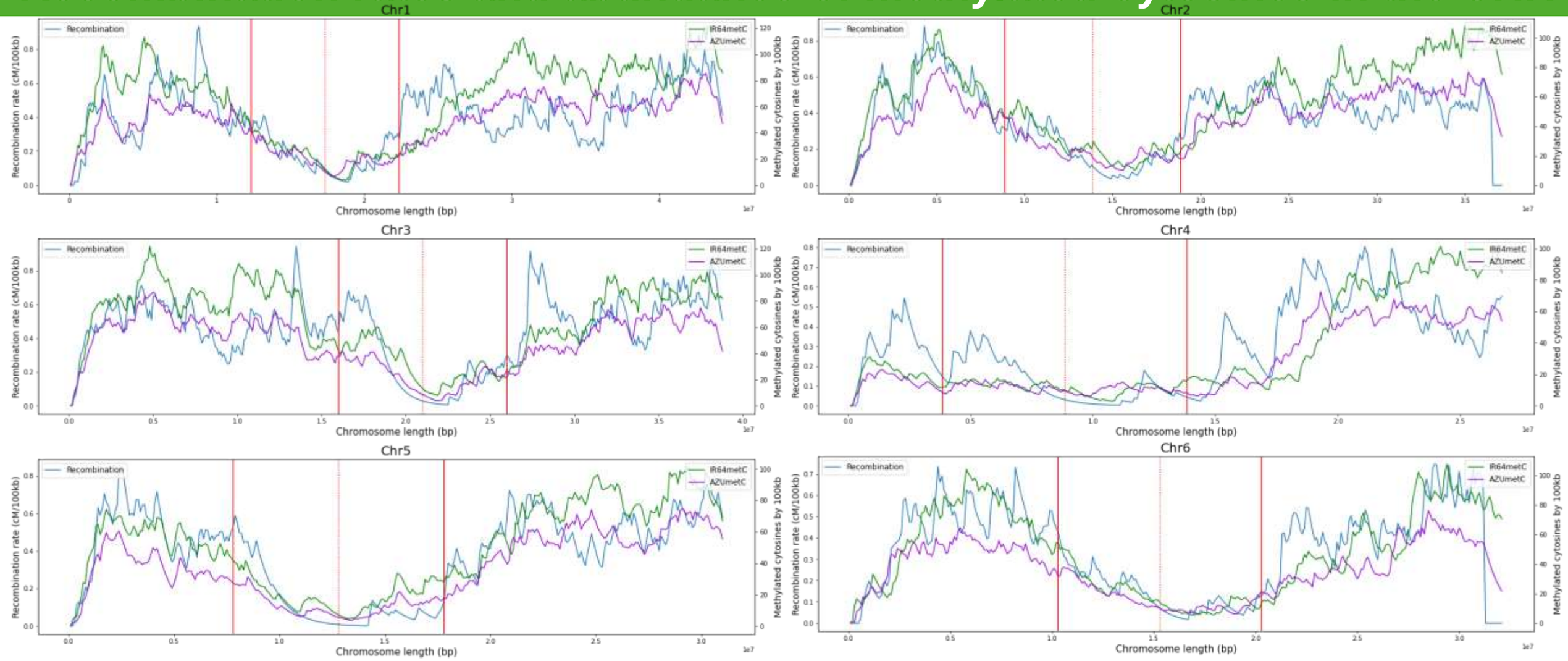
CHG Context

Recombination rates and abundance of methylated cytosines in the CHH



CHH Context

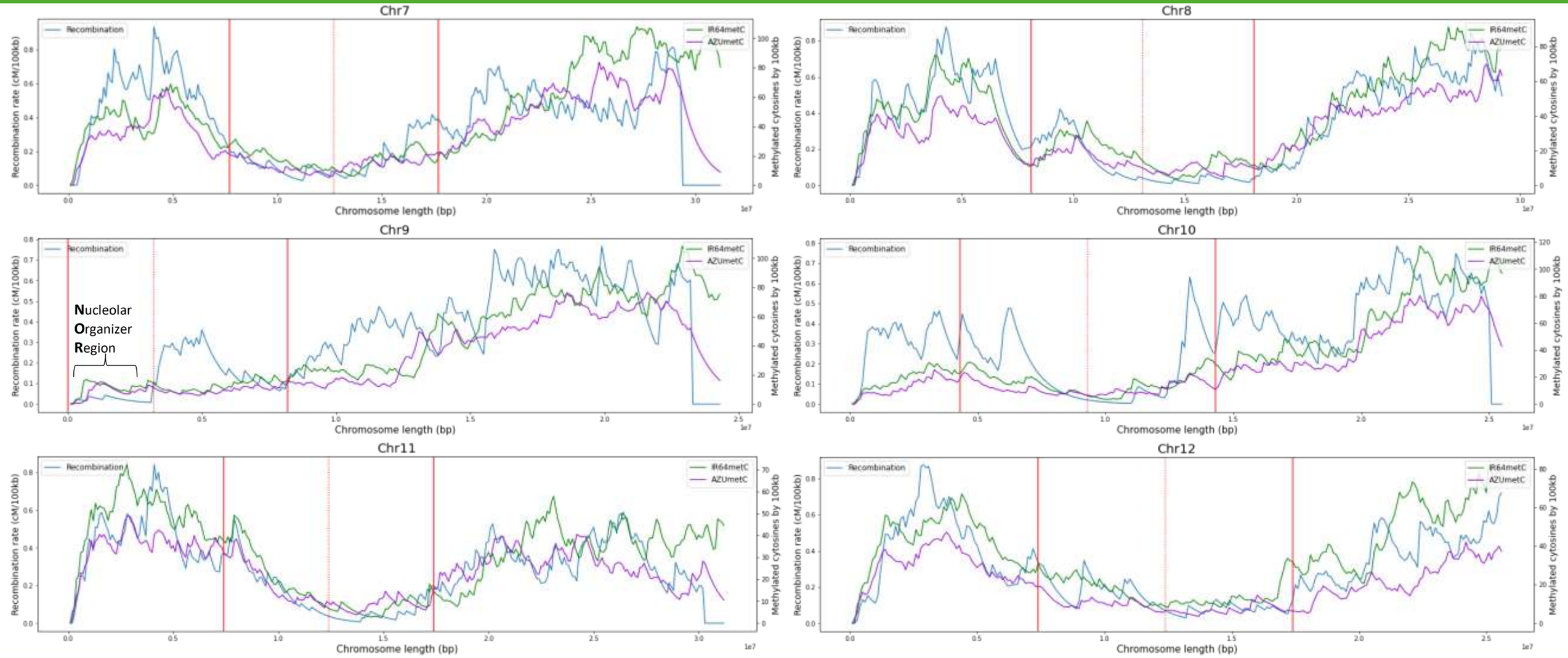
Recombination rates and abundance of methylated cytosines in the CHH



In-silico

El futuro es de todos
Gobierno de Colombia
COLOMBIA CIENTÍFICA
ómica

Recombination rates and abundance of methylated cytosines in the CHH



In-silico

El futuro es de todos
Gobierno de Colombia
COLOMBIA CIENTÍFICA
ómica

Correlations between recombination rates and methylation contexts

	All Contexts		CG Context		CHG Context		CHH Context	
Chromosome	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

Model Training

Model Trainning

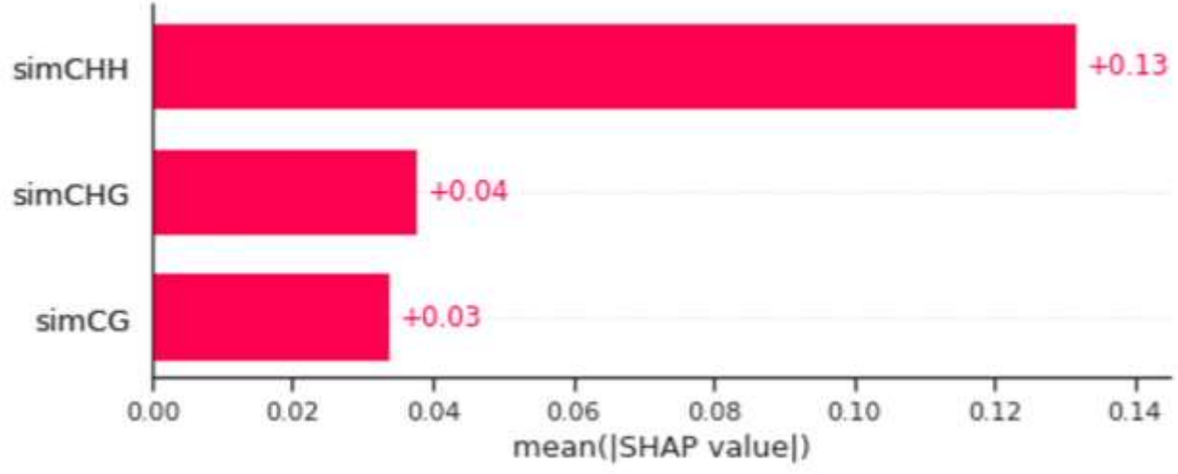
Model Output

CG
CHG
CHH

Model

Recombination rates

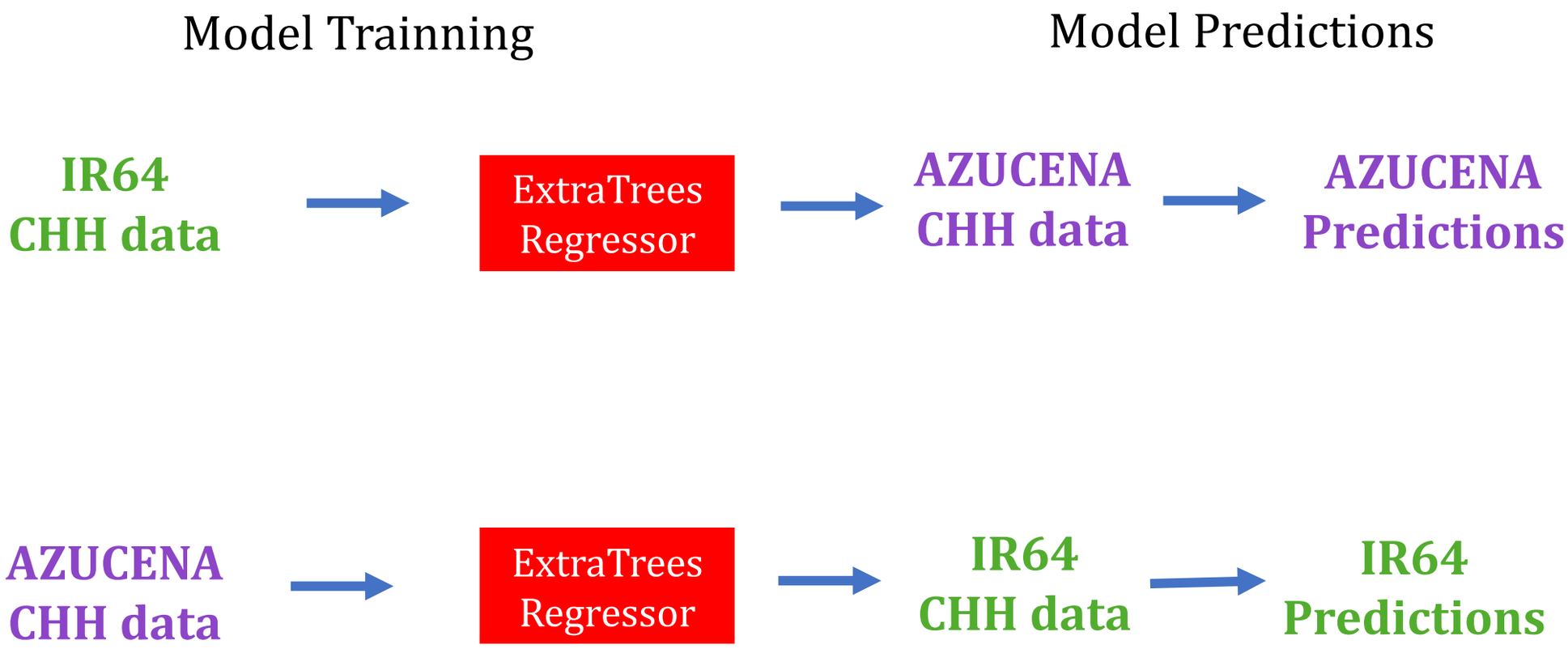
Contribution values



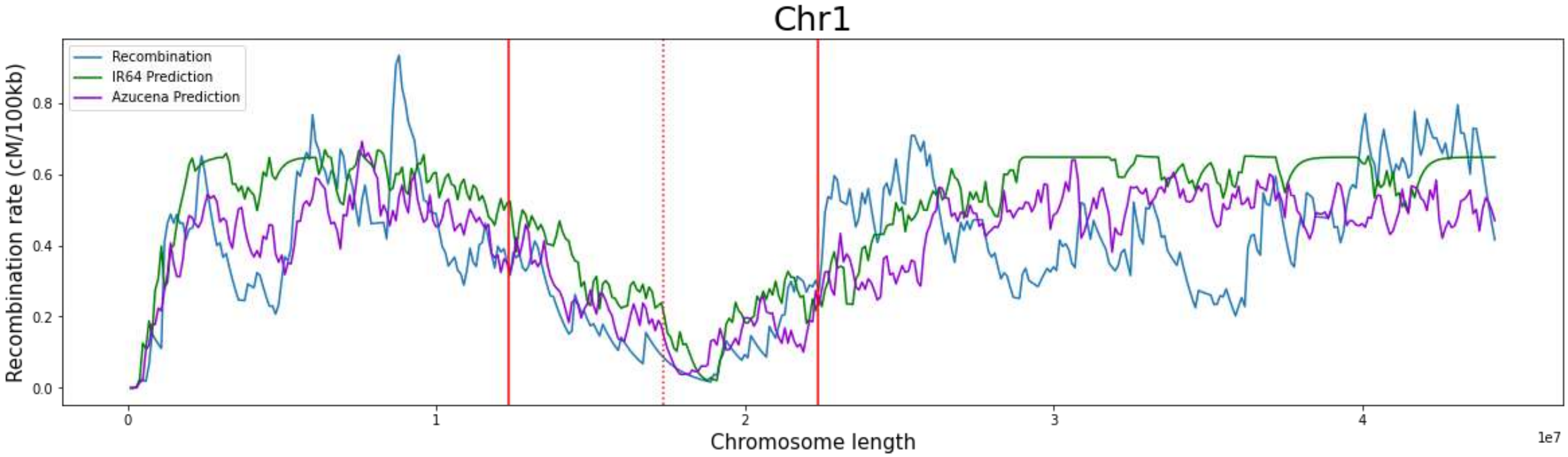
Contribution values of each feature (Shapley Package)/Evaluation of best ML model (LazyPredict

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
ExtraTreesRegressor	0.57	0.57	0.14	0.31
RandomForestRegressor	0.56	0.56	0.14	0.64
HistGradientBoostingRegressor	0.56	0.56	0.14	0.42
LGBMRegressor	0.56	0.56	0.15	0.11
GradientBoostingRegressor	0.56	0.56	0.15	0.27
SVR	0.55	0.55	0.15	0.23
NuSVR	0.55	0.55	0.15	0.48
KNeighborsRegressor	0.54	0.55	0.15	0.02
MLPRegressor	0.54	0.54	0.15	0.18
AdaBoostRegressor	0.53	0.53	0.15	0.04
XGBRegressor	0.52	0.52	0.15	0.35
BaggingRegressor	0.51	0.51	0.15	0.07
LassoLarsIC	0.47	0.48	0.16	0.01
OrthogonalMatchingPursuitCV	0.47	0.48	0.16	0.01
TransformedTargetRegressor	0.47	0.48	0.16	0.01
LinearRegression	0.47	0.48	0.16	0.01
Lars	0.47	0.48	0.16	0.02
LarsCV	0.47	0.48	0.16	0.03
LassoLarsCV	0.47	0.48	0.16	0.02
RidgeCV	0.47	0.48	0.16	0.02
Ridge	0.47	0.48	0.16	0.01
BayesianRidge	0.47	0.48	0.16	0.01
LassoCV	0.47	0.48	0.16	0.07
ElasticNetCV	0.47	0.48	0.16	0.05
SGDRegressor	0.47	0.47	0.16	0.03

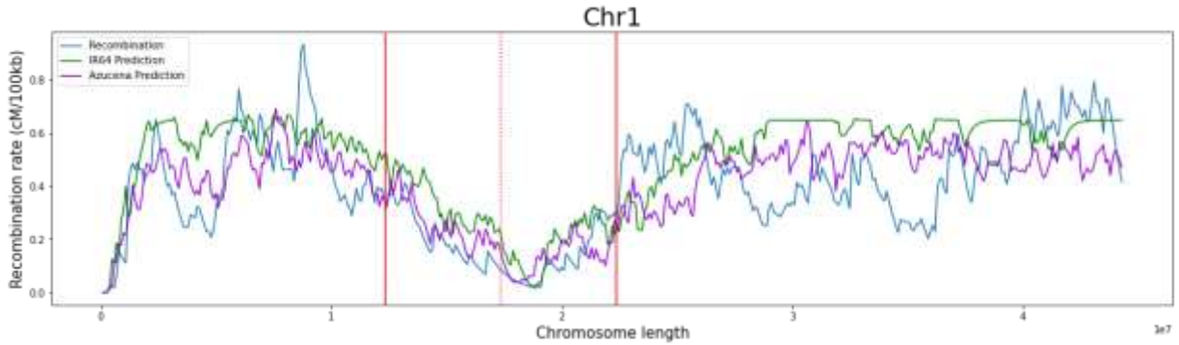
Model Training

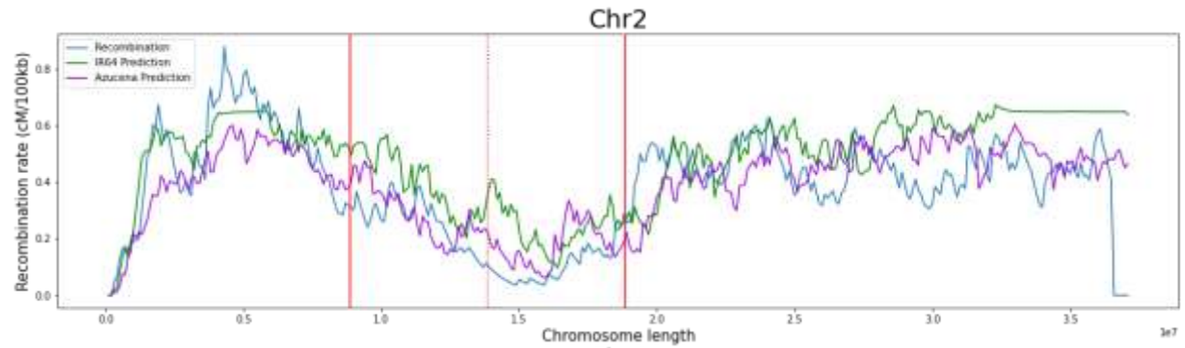


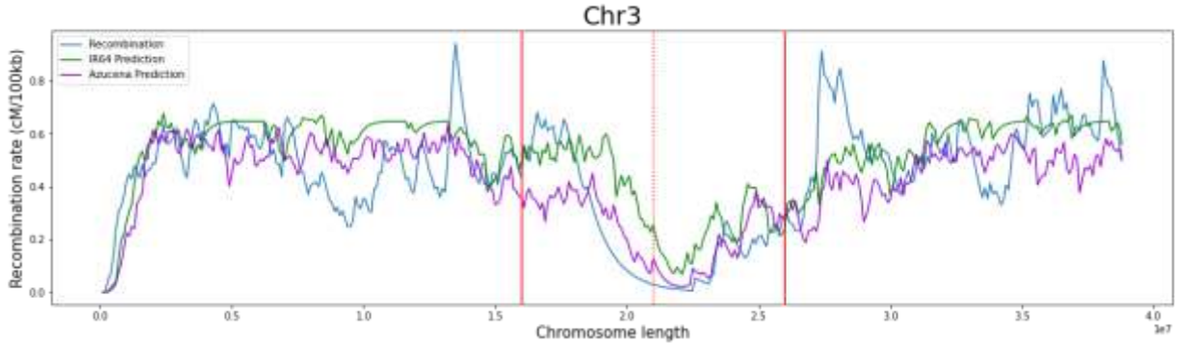
Chromosomal recombination predictions using the ExtraTrees model with CHH methylation as input

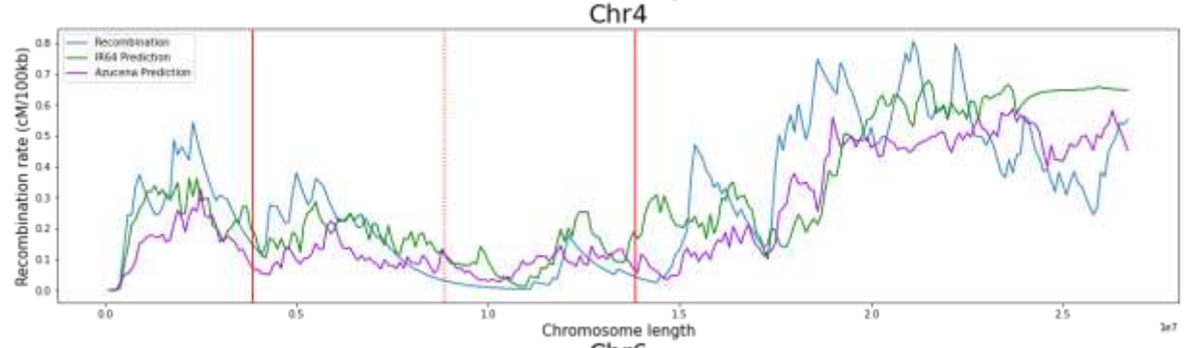


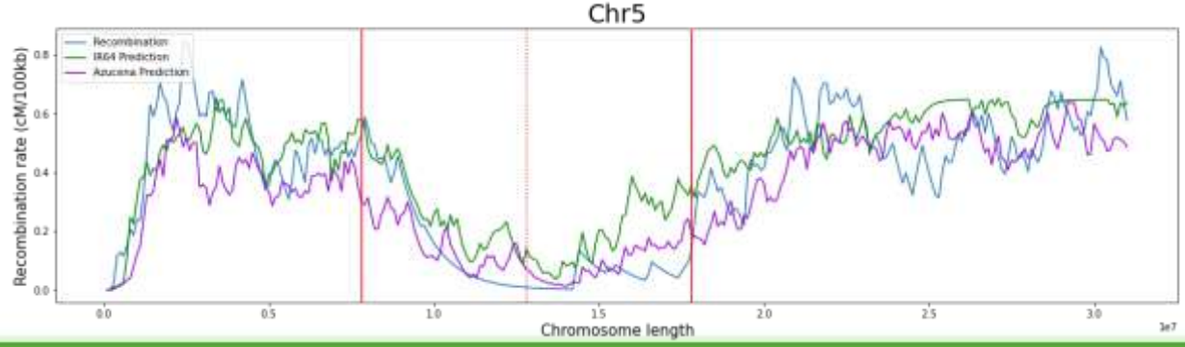
Chromosomal recombination predictions using the ExtraTrees model with CHH methylation as input feature

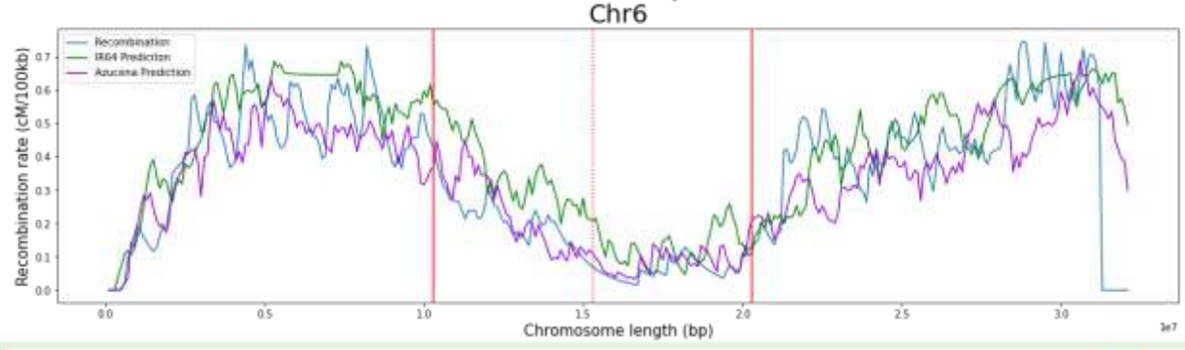








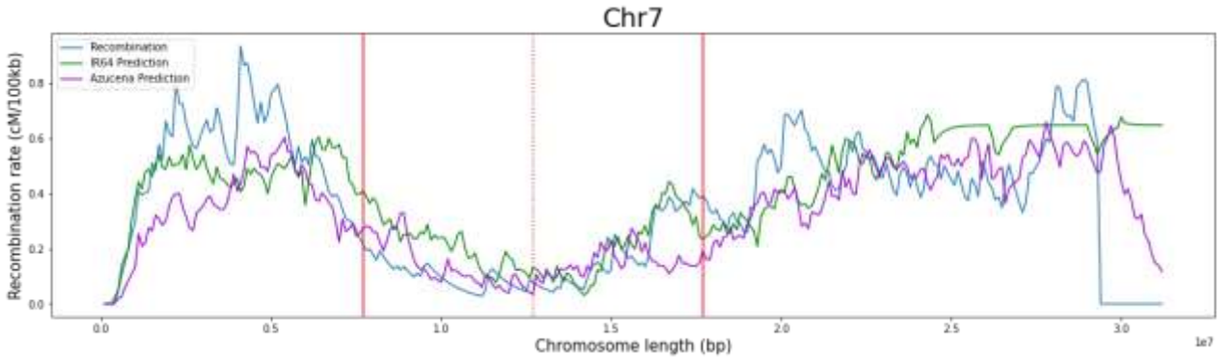


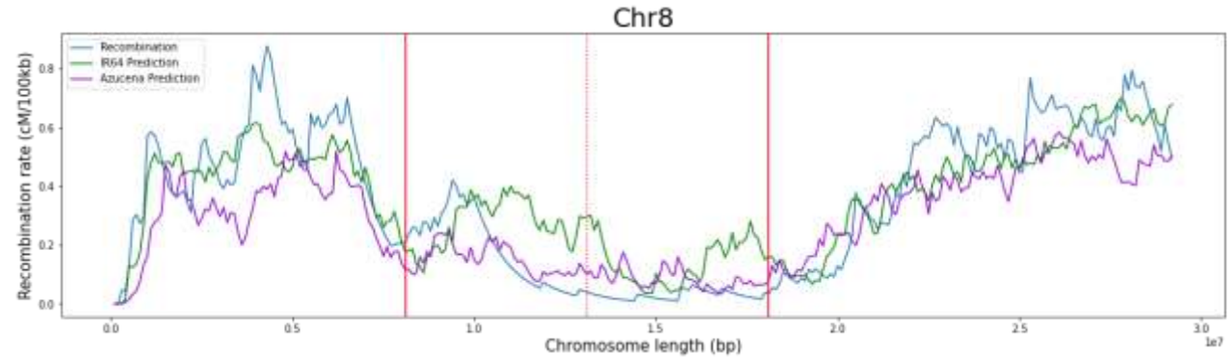


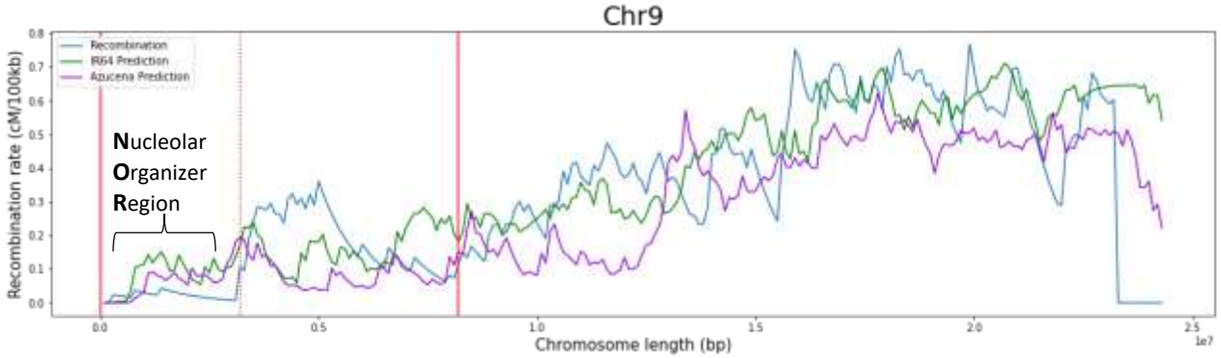
In-silico

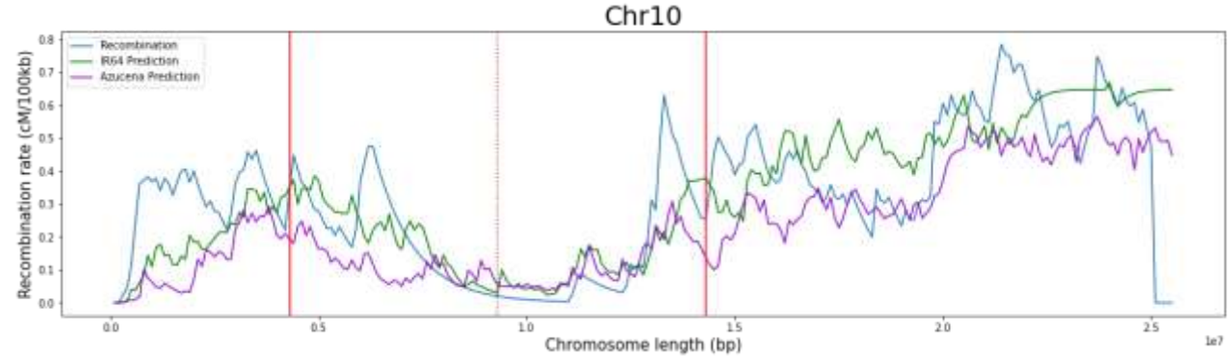
El futuro es de todos
Gobierno de Colombia
COLOMBIA CIENTÍFICA
ómicas

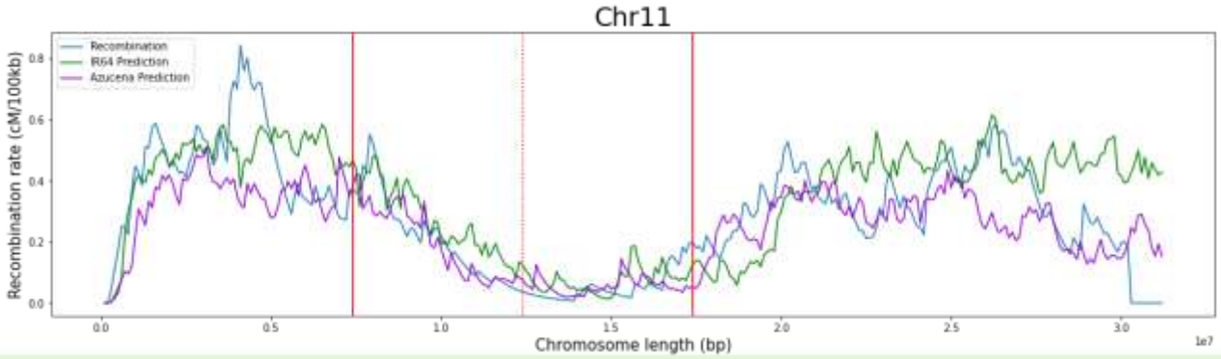
Chromosomal recombination predictions using the ExtraTrees model with CHH methylation as input feature

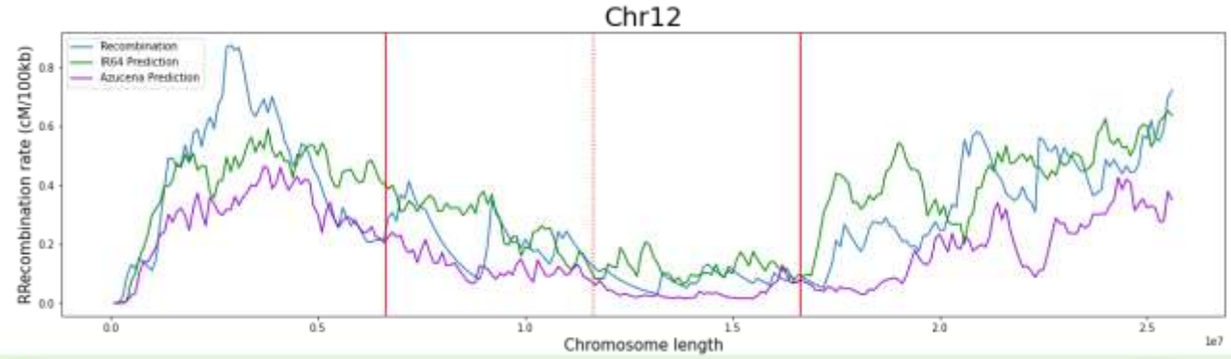












In-silico

Predicted vs Experimental values

IR64				Azucena			
Chromosome	*R2*	Correlation	MSE	Chromosome	*R2*	Correlation	MSE
1	0	0.63	0.03	1	0.44	0.67	0.02
2	0.04	0.66	0.03	2	0.53	0.73	0.01
3	0.37	0.7	0.02	3	0.49	0.72	0.02
4	0.44	0.72	0.02	4	0.6	0.81	0.01
5	0.59	0.81	0.02	5	0.67	0.84	0.01
6	0.44	0.78	0.02	6	0.68	0.82	0.01
7	0.16	0.53	0.04	7	0.5	0.73	0.02
8	0.71	0.85	0.01	8	0.67	0.88	0.02
9	0.32	0.65	0.03	9	0.5	0.75	0.02
10	0.41	0.7	0.02	10	0.28	0.69	0.03
11	0.3	0.7	0.02	11	0.52	0.77	0.01
12	0.54	0.77	0.01	12	0.35	0.85	0.02

Conclusions

- CHH context **positively correlates** with recombination rates along the twelve rice chromosomes
- CHH methylated cytosines **can be use to predict recombination** using machine learning models
- We invite colleagues to explore how the counting of CHH-methylated cytosines in other species behaves with respect to chromosomal recombination.

ómicas

Aliados

CIAT
International Center for Tropical Agriculture
Since 1967 Science to cultivate change

cenicaña

FEDEARROZ
FEDERACIÓN NACIONAL DE ARROCEROS

NIAB

intelecto
Soluciones y Tecnología

HITECH
AUTOMATIZACION S.A.

Caltech

GHENT UNIVERSITY

東京大学
THE UNIVERSITY OF TOKYO

ILLINOIS

Pontificia Universidad JAVERIANA Cali
IES Ancla

UNIVERSIDAD DEL QUINDÍO

Universidad de Ibagué

UNILLANOS
UNIVERSIDAD DE LOS LLANOS

Universidad de los Andes
Colombia

UNIVERSIDAD ICESI

Pontificia Universidad JAVERIANA
Bogotá

Apoyan

El futuro es de todos
Gobierno de Colombia

COLOMBIA CIENTÍFICA

THANKS!

mauricio.penuela@javerianacali.edu.co

METHYLATION IN THE CHH CONTEXT ALLOWS TO PREDICT RECOMBINATION IN RICE

Mauricio Peñuela, Jenny Johana Gallo-Franco, Jorge Finke, Camilo Rocha, Anestis Gkanogiannis, Thaura Ghneim Herrera, Mathias Lorieux

ABSTRACT

Variation of DNA methylation is the most studied epigenetic trait. It is considered a key factor in regulating plant development and physiology, and has been related to the regulation of several genomic features including transposon silencing, regulation of gene expression and recombination rates. In fact, several studies have reported increased methylation around centromere regions, generally recombination suppression regions. Nonetheless, characterizing relationships between DNA methylation and recombination rates remains a challenge. This work explores the relationship between recombination rates and DNA methylation data for two commercial rice varieties. Several correlation analyses were made between methylation levels, for each sequence context, CG, CHG, and CHG, and recombination rates. Our aim is to identify patterns that would help predict recombination behavior. Our results show negative correlations between recombination rates and methylated cytosines counts for all contexts tested at the same time and separately for CG and CHG contexts. A positive correlation between recombination rates and methylated cytosine count was reported in CHH contexts. A similar behavior is observed when considering only methylated cytosines within genes, transposons, and retrotransposons. Moreover, it was shown that the centromere region strongly affects the relationship between recombination rates and methylation, with the higher values inside the centromeric region. Finally, machine learning regression models are applied to predict recombination using the count of methylated cytosines in the CHH context as the entrance feature. Our findings shed light into 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*

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, and create novel allelic combinations. Recombination is widely recognized for its roles in promoting the diversity to respond to continually shifting environments, in addition to preventing the build-up of genetic load by decoupling linked deleterious and beneficial variants (Rodgers-Melnick et al., 2015). 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 (Lambing et al., 2017). 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 and characterized by high frequencies of long terminal repeat retroelements and few genes (Henderson, 2012).

The role of chromatin structure and DNA methylation in determining recombination rates has been widely reported. For example, high levels of histone H3 acetylation in *Arabidopsis* mutants were associated with changes in the crossover frequencies (Perrella et al., 2010). 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 (Melamed-Bessudo & Levy, 2012; Mirouze et al., 2012). 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 (Bräutigam & Cronk, 2018). Several studies have shown that sexual reproduction in plants involves the reprogramming of DNA methylation patterns (Kawashima & Berger, 2014). DNA methylation in combination with modifications of histones and non-histone proteins defines the structure and accessibility of chromatin, which helps to regulate gene expression, transposon silencing, chromosome interactions and trait inheritance (Zhang et al., 2018).

The methylation dynamics for each sequence context is determined by different mechanisms and related to specific biological functions (Zhang et al., 2018). 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 (Zhang et al., 2018). *De novo* methylation is carried out by CMT2 for CHG and CHH context (Kawashima & Berger, 2014) and the RdDM pathway for all sequence contexts (Zhang et al., 2018). 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 (Lanciano & Mirouze, 2017). More research on methylation and demethylation events and their precursors will be necessary to clarify these mechanisms.

Identifying factors influencing the meiotic recombination rates are important for breeders interested in transferring genes from one variety to another through crosses, thus developing new allelic combinations that allow them to meet the needs present in agricultural systems. Recently, several 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., 2016) developed a predictor of recombination hot/cold spots using a machine learning approach combined with principal component analysis in yeast. Otherwise, (Demirci et al., 2018) explored DNA sequence and shape features to train machine learning models for predicting crossover occurrence in *Arabidopisis,* maize, tomato and rice. Meanwhile, (Adrion et al., 2020) used recurrent neural networks, a deep learning method for estimating genome-wide recombination in a natural population of African *Drosophila melanogaster*. Finally, (Peñuela et al., 2022) proposed a mechanism-based model using sequence identity between two genomes to predict recombination along rice chromosomes. Recombination prediction has recently become important due to the possibility of extracting data from genome sequencing technologies and exploring how sequence features may affect it. Within these features, methylation has been reported as a prime factor in understanding recombination.

In recent years, rice has been a key crop to comprehension of the methylation process, because it is highly homozygous and self-pollinated, which is why it is known as a model to study methylation patterns in monocotyledonous plants. In addition, it is of great importance in food security, since half of the world's population depends on it as daily food (Cheng et al., 2001). However, few studies have analyzed methylation patterns in relation to recombination rates in rice. For instance, Habu et al., (2015) 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. In addition, Choi & Purugganan, (2018) 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 we explore the relationship between chromosomal recombination rates and different methylation contexts, using *Oryza sativa* rice as a model. Focusing on the following objectives 1) estimate the correlation between recombination and methylation in all contexts, 2) describe the effect of methylation within genes, transposons and retrotransposons with respect to recombination, and 3) implement a machine learning model to predict recombination based on methylation data. Our 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. Machine learning models helped predict chromosomal recombination along all twelve chromosomes of both rice varieties with mean $R^2 = 0.26$ and mean correlation values between predictions and recombination rates of 0.66.

MATERIALS AND METHODS

Recombination rates

The recombination rates were estimated from an inter-subspecific segregating population of 212 F11 recombinant inbred lines (RIL) 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 (~2x) followed by imputation with NOISYmputer (Lorieux et al. 2019) . Local recombination rates in cM/bp were calculated in sliding windows of 100 kbp using MapDisto v2 (Heffelfinger et al. 2017). The details of this process and the access to the data are available in Peñuela et al. (2022).

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°C. Total genomic DNA was extracted from frozen root tissue by CTAB 2X protocol with modifications (Maropola et al., 2015). Genomic DNA quality was evaluated on agarose gels and DNA quantity was measured using a Nanodrop spectrophotometer (Thermo Scientific).

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 was performed by CD Genomics (CD Genomics Inc., Shirley, New York, USA). Basic statistics on the quality of the raw reads was done using the FastQC tool. Sequencing adapters and low-quality data of the sequencing data were removed by Trimmomatic (version 0.36). Cleaned data were aligned to the reference genomes reported in the genebank repository for IR64 (Accession number: RWKJ00000000) and Azucena (Accession number: PKQC000000000) using Bismark v.0.16.3 (Krueger & Andrews, 2011) with default parameters. Only uniquely aligned reads were maintained. Methylation calling data obtained from Bismark were used for further analysis.

Comparison between recombination rates and methylation patterns

To compare the methylation patterns with the local recombination rates, the genomes were divided into 100 kbp 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.

Functional evaluation

Gene, transposon, and retrotransposon annotation information from both varieties were used (Supplementary data S1). Pearson correlation analyzes were carried between the number of genes, transposons and retrotransposons with respect to recombination along 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 centromere and non-centromere regions was also included.

Machine learning modeling

To assess the usefulness of methylation in predicting chromosome recombination, we explored different machine learning approaches. Counts of methylated cytosines belonging to the CG, CHG, and CHH contexts for each variety were evaluated as features for machine learning modeling using the Shapley package (https://github.com/slundberg/shap). Subsequently, the performance of different machine learning models was evaluated using the LazyPredict package (https://pypi.org/project/lazypredict/). An 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, meanwhile *MSE* was used for predictions. Pearson correlation analyzes 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 data set 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.

RESULTS AND DISCUSSION

Correlation analysis, performed between local recombination rate and the total count of methylated cytosines without differentiating their methylation context for IR64 and Azucena varieties, showed a negative trend in all chromosomes with higher levels of methylation in the centromere region (Table 1, Figure 1a). The correlation values were on average -0.44±0.17, for all chromosomes of both varieties. Similar results in rice have been previously described by (Yan et al., 2010) and (Habu et al., 2015), and this trend has been widely discussed by (Yan et al., 2005). High levels of methylation in heterochromatin regions near the centromeres have been reported as a common pattern, where meiotic recombination is also repressed. Likewise, recombination-free regions around centromeres are likely to be important for normal centromere function during meiosis (Habu et al., 2015; Yan et al., 2005). The correlation analyses for each methylation context individually, showed that the count of methylated cytosines of CG and CHG were similarly negatively correlated with recombination (Figure 1b,c). On the contrary, the CHH contexts showed an opposite trend. More specifically, CHH context methylated cytosines count was positively correlated with recombination rates in chromosomes, which is opposite to the behavior by the CG and CHG contexts (Table 1, Figure 1d and 2). This opposite relationship between the methylation contexts of CG and CHH has been reported by (Li et al., 2012).

The positive correlation between methylated cytosine count and recombination rates observed in the context of CHH, was not clear when all methylation contexts were assessed together, since the methylated cytosines count in the CG and CHG contexts was high. This trend was observed for both varieties, IR64 and Azucena, where the methylation data and the alignment process were obtained independently. To our knowledge, positive correlation between the CHH methylated cytosines count and recombination rates has not been reported before, representing a new finding for epigenetics. It is unclear what the role of methylated cytosines is in the CHH context with respect to recombination. For instance, it could be related to the biochemical signaling for the crossing-over events or could be also a consequence of these events.

It has been reported that CHH methylation could be related to fruit size in apples (Daccord et al., 2017), silencing transposons in sugar beets (Zakrzewski et al., 2017), and a potential role in *A. thaliana* seed dormancy, with increases in CHH methylation in seeds during seed development and a decrease during germination (Zhang et al., 2018). Demonstrating the multiple roles that CHH methylation can play in plant genomes. It must not be forgotten that DNA methylation variations can be hereditary or reversible, this ability can allow phenotypic variation and rapid response to environmental changes. Even the degree of intraspecies epigenomic diversity can be correlated with climate and geographic origin (Lanciano & Mirouze, 2017).

The functional analysis developed with annotation data of genes, transposons, and retrotransposons for each variety, evidenced a high positive correlation between the number of genes per window and the recombination rates along chromosomes for both varieties

(Table 2). This positive trend has been previously evidenced in *Drosophila*, *A. thaliana*, yeast, finches, monkeyflowers, and dogs, with clear hotspots typically located near promoter regions of genes (Kent et al., 2017) and also observed in the euchromatic regions of maize (Anderson et al., 2006). In contrast, we found a negative correlation between the number of transposons and retrotransposons with respect to recombination rates across all chromosomes for both rice varieties. This result can be explained by the abundance of these elements near the centromere (Table 2, Figure 3). Similar results have been found by (Tian et al., 2009) who suggested that the rice genome is organized along recombinational gradients, due to the negative correlation of recombination with transposable elements and the positive one with gene densities.

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 (Kent et al., 2017). Recent analyses of the localization of recombination at the fine scale, also tend to show negative correlations with local densities of repetitive elements. Strong recombination suppression and a large accumulation of transposable elements are usual in pericentromeric regions (Kent et al., 2017). For rice, this pattern is shared between *japonica* and *indica* groups (Tian et al., 2009). There remains uncertainty about the directionality of cause and effect, the extent to which the correlation is driven by associations of both recombination and transposable elements with other factors, or why patterns differ among species and types of repetitive elements. (Kent et al., 2017).

The count of methylated cytosines was assessed within genes, transposons, and retrotransposons and compared to recombination rates (Table 3, Figures 4 and 5). The analysis shows that the methylated cytosines count in genes, transposons, and retrotransposons are negatively correlated with recombination rates when evaluated for all contexts together. This indicates that methylation inside these entities was higher when recombination was lower. The same negative trends were observed when methylated cytosines are analyzed in CG and CHG contexts. Methylation events in transposons and retrotransposons are associated with prevention of their expression and movement along chromosomes, which can be damageable to the organism and even deleterious (Ahmed et al., 2011; Kent et al., 2017). It should be noted that these methylation events can also affect surrounding genomic regions (Ahmed et al., 2011), 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 (Zhang et al., 2018). However, when the methylated cytosines count was evaluated in the CHH context within these elements, correlation analyses showed a positive correlation with recombination rates (Table 3). This was a consequence of low CHH methylation near the centromere region (Figures 4 and 5).

Chromosomal regions close to the centromere have a high incidence of methylation. When these regions were removed from the correlation analyses, trends changed from being high negative to being lower, for all contexts evaluated together and for the CG and CHG contexts evaluated independently (Table 4). For the methylation in the CHH context, the markedly positive correlation also decreased but continued being positive. When only the

centromere regions were evaluated, negative correlations were evidenced in all contexts when they were evaluated together, and in the contexts of CG and CHG when they were evaluated independently. These results are in agreement with the reported importance of DNA methylation for plant chromosomal interactions in pericentromeric regions (Zhang et al., 2018) and with the results obtained by (Habu et al., 2015) 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 was positive but weaker than that of the whole chromosome (Table 4).

The contributions of methylation in CG, CHG, and CHH contexts to predict recombination as features of machine learning approaches were evaluated with the help of the Shapley package. The results showed a great contribution of CHH for the prediction of recombination, and a low contribution of CG and CHG for both varieties (Figure 6a,b). This was in agreement with previous results where the CHH context had the highest positive correlations with respect to chromosome recombination rates, while the CG and CHG contexts had negative correlations. The Shap summary plot also showed the same trend, evidencing the strongest effect on recombination when the CHH values were higher (Figure 6c,d).

Subsequently, the methylated cytosines count in the CHH context was used as a unique feature to evaluate regression algorithms of machine learning. This evaluation was carried out independently for each variety using the Lazy Predict package. The results showed that the Extra Trees algorithm performed best (R^2= 0.57, *RMSE* = 0.01 for IR64; R^2= 0.69, *RMSE* = 0.01 for Azucena). We thus chose this algorithm to develop the training and subsequent predictions.

Predictions on Azucena's chromosomes, by training the Extra Trees algorithm with information from IR64, gave an R^2 of 0.32±0.13 and a *MSE* of 0.02±0.00 on average. Meanwhile, predictions on IR64's chromosomes by training the Extra Trees algorithm with information from Azucena, gave an R^2 of 0.21±0.21 and an *MSE* of 0.03±0.00 on average. For both cases, the average correlation values between predictions and recombination rates were 0.67±0.06 for Azucena and 0.65±0.07 for IR64, evidencing a positive trend (Table 5, Figure 7).

Several studies have focused on predicting recombination using machine learning. For example, Liu et al., (2016) combines a support vector machine with a consensus feature (called dinucleotide-based autocross covariance) to predict recombination of hot/cold spots in yeast. Authors such as Demirci et al., (2018) have used features as gene annotation, propeller and helical twist, AT/TA dinucleotides, and CA dinucleotides to train machine learning models to predict crossover occurrences in *Arabidopsis,* maize, rice, and tomato. More recently, Adrion et al., (2020) predicted 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.

In the case of this paper, ExtraTrees made it possible to predict chromosomal recombination using a single feature: the CHH methylated cytosines count. It was possible due the high correlation between this feature and the recombination rates, which behaved similarly along all chromosomes (Figure 7). We trained the model on a dataset of one variety and tested it on the other, performing two independent tests and finding that results were consistent. This opens a door for future studies we anticipate that these trained models can be used to predict chromosomal recombination rates in any variety of *Oryza sativa* rice, since the two varieties used in this study, IR64 and Azucena, are highly distant genetically, belonging to the *indica* and *japonica* groups, respectively.

Compared to the previous model described by Peñuela et al. (2022), which uses a measure of genomic identity between parental genomes to predict recombination rates, the approximation presented here has some advantages. First, methylation data is only required for one of the varieties involved in the cross; this information is sufficient to predict recombination rates using the CHH methylated cytosines count as a feature in a machine learning regression model. Additionally, no assumptions are required to apply the model, and thresholds and penalty values to increase the prediction values are not needed. Most importantly, there is no need for centromere correction because the CHH methylated cytosines count decreases naturally around the centromere region and increases in telomeric regions following the recombination landscape. This is the most remarkable discovery of this work. However, it must be noted that these two models work with different data types: the identity model of (Peñuela et al 2022) uses DNA sequences of parental genomes; meanwhile, the one presented here uses methylation data obtained by bisulfite sequencing experiments. The choice of one model over the other will depend on the availability of the types of data or their possibility of extraction.

CONCLUSION

In this study, we reported how methylated cytosines in the CHH context positively correlate with recombination rates along the twelve rice chromosomes for two genetically distant rice varieties evaluated, IR64 and Azucena. However, a negative correlation was obtained 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 higher 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 showed 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, we trained a machine learning model using the CHH methylated cytosines count to predict recombination rates, which obtained consistent results in two independent data sets. We recommend the extraction of methylation data and the application of machine learning models to future studies interested in predicting recombination rates using as feature the count of CHH methylated cytosines in rice. We invite colleagues to explore how the counting of CHH-methylated cytosines in other species behaves with respect to chromosomal recombination.

REFERENCES

Adrion, J. R., Galloway, J. G., & Kern, A. D. (2020). Predicting the Landscape of Recombination Using Deep Learning. *Molecular Biology and Evolution*, *37*(6), 1790–1808. https://doi.org/10.1093/molbev/msaa038

Ahmed, I., Sarazin, A., Bowler, C., Colot, V., & Quesneville, H. (2011). Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis. *Nucleic Acids Research*, *39*(16), 6919–6931. https://doi.org/10.1093/nar/gkr324

Anderson, L. K., Lai, A., Stack, S. M., Rizzon, C., & Gaut, B. S. (2006). Uneven distribution of expressed sequence tag loci on maize pachytene chromosomes. *Genome Research*, *16*(1), 115–122. https://doi.org/10.1101/gr.4249906

Bräutigam, K., & Cronk, Q. (2018). DNA Methylation and the Evolution of Developmental Complexity in Plants. *Frontiers in Plant Science*, *9*, 1447. https://doi.org/10.3389/fpls.2018.01447

Cheng, Z., Buell, C. R., Wing, R. A., Gu, M., & Jiang, J. (2001). Toward a Cytological Characterization of the Rice Genome. *Genome Research*, *11*(12), 2133–2141. https://doi.org/10.1101/gr.194601

Choi, J. Y., & Purugganan, M. D. (2018). Evolutionary Epigenomics of Retrotransposon-Mediated Methylation Spreading in Rice. *Molecular Biology and Evolution*, *35*(2), 365–382. https://doi.org/10.1093/molbev/msx284

Colome-Tatche, M., Cortijo, S., Wardenaar, R., Morgado, L., Lahouze, B., Sarazin, A., Etcheverry, M., Martin, A., Feng, S., Duvernois-Berthet, E., Labadie, K., Wincker, P., Jacobsen, S. E., Jansen, R. C., Colot, V., & Johannes, F. (2012). Features of the Arabidopsis recombination landscape resulting from the combined loss of sequence variation and DNA methylation. *Proceedings of the National Academy of Sciences*, *109*(40), 16240–16245. https://doi.org/10.1073/pnas.1212955109

Daccord, N., Celton, J.-M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E., van de Geest, H., Bianco, L., Micheletti, D., Velasco, R., Di Pierro, E. A., Gouzy, J., Rees, D. J. G., Guérif, P., Muranty, H., Durel, C.-E., Laurens, F., Lespinasse, Y., Gaillard, S., … Bucher, E. (2017). High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nature Genetics*, *49*(7), 1099–1106. https://doi.org/10.1038/ng.3886

Demirci, S., Peters, S. A., de Ridder, D., & van Dijk, A. D. J. (2018). DNA sequence and shape are predictive for meiotic crossovers throughout the plant kingdom. *The Plant Journal*, *95*(4), 686–699. https://doi.org/10.1111/tpj.13979

Habu, Y., Ando, T., Ito, S., Nagaki, K., Kishimoto, N., Taguchi-Shiobara, F., Numa, H., Yamaguchi, K., Shigenobu, S., Murata, M., Meshi, T., & Yano, M. (2015). Epigenomic modification in rice controls meiotic recombination and segregation distortion. *Molecular Breeding*, *35*(4), 103. https://doi.org/10.1007/s11032-015-0299-0

Henderson, I. R. (2012). Control of meiotic recombination frequency in plant genomes. *Current Opinion in Plant Biology*, *15*(5), 556–561. https://doi.org/10.1016/j.pbi.2012.09.002

Kawashima, T., & Berger, F. (2014). Epigenetic reprogramming in plant sexual reproduction. *Nature Reviews Genetics*, *15*(9), 613–624. https://doi.org/10.1038/nrg3685

Kent, T. V., Uzunović, J., & Wright, S. I. (2017). Coevolution between transposable elements and recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1736), 20160458. https://doi.org/10.1098/rstb.2016.0458

Krueger, F., & Andrews, S. R. (2011). Bismark: A flexible aligner and methylation caller for

Bisulfite-Seq applications. *Bioinformatics*, *27*(11), 1571–1572. https://doi.org/10.1093/bioinformatics/btr167

Lambing, C., Franklin, F. C. H., & Wang, C.-J. R. (2017). Understanding and Manipulating Meiotic Recombination in Plants. *Plant Physiology*, *173*(3), 1530–1542. https://doi.org/10.1104/pp.16.01530

Lanciano, S., & Mirouze, M. (2017). DNA Methylation in Rice and Relevance for Breeding. *Epigenomes*, *1*(2), 10. https://doi.org/10.3390/epigenomes1020010

Li, X., Zhu, J., Hu, F., Ge, S., Ye, M., Xiang, H., Zhang, G., Zheng, X., Zhang, H., Zhang, S., Li, Q., Luo, R., Yu, C., Yu, J., Sun, J., Zou, X., Cao, X., Xie, X., Wang, J., & Wang, W. (2012). Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression. *BMC Genomics*, *13*(1), 300. https://doi.org/10.1186/1471-2164-13-300

Liu, B., Liu, Y., Jin, X., Wang, X., & Liu, B. (2016). iRSpot-DACC: A computational predictor for recombination hot/cold spots identification based on dinucleotide-based auto-cross covariance. *Scientific Reports*, *6*(1), 33483. https://doi.org/10.1038/srep33483

Maropola, M. K. A., Ramond, J.-B., & Trindade, M. (2015). Impact of metagenomic DNA extraction procedures on the identifiable endophytic bacterial diversity in Sorghum bicolor (L. Moench). *Journal of Microbiological Methods*, *112*, 104–117. https://doi.org/10.1016/j.mimet.2015.03.012

Melamed-Bessudo, C., & Levy, A. A. (2012). Deficiency in DNA methylation increases meiotic crossover rates in euchromatic but not in heterochromatic regions in Arabidopsis. *Proceedings of the National Academy of Sciences*, *109*(16), E981–E988. https://doi.org/10.1073/pnas.1120742109

Mirouze, M., Lieberman-Lazarovich, M., Aversano, R., Bucher, E., Nicolet, J., Reinders, J., & Paszkowski, J. (2012). Loss of DNA methylation affects the recombination landscape in Arabidopsis. *Proceedings of the National Academy of Sciences*, *109*(15), 5880–5885. https://doi.org/10.1073/pnas.1120841109

Monroe, J. G., Srikant, T., Carbonell-Bejerano, P., Becker, C., Lensink, M., Exposito-Alonso, M., Klein, M., Hildebrandt, J., Neumann, M., Kliebenstein, D., Weng, M.-L., Imbert, E., Ågren, J., Rutter, M. T., Fenster, C. B., & Weigel, D. (2022). Mutation bias reflects natural selection in Arabidopsis thaliana. *Nature*. https://doi.org/10.1038/s41586-021-04269-6

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. (2010). Histone hyperacetylation affects meiotic recombination and chromosome segregation in Arabidopsis: Histone acetylation in At meiosis. *The Plant Journal*, *62*(5), 796–806. https://doi.org/10.1111/j.1365-313X.2010.04191.x

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. (2015). Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proceedings of the National Academy of Sciences*, *112*(12), 3823–3828. https://doi.org/10.1073/pnas.1413864112

Tian, Z., Rizzon, C., Du, J., Zhu, L., Bennetzen, J. L., Jackson, S. A., Gaut, B. S., & Ma, J. (2009). Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? *Genome Research*, *19*(12), 2221–2230. https://doi.org/10.1101/gr.083899.108

Yan, H., Jin, W., Nagaki, K., Tian, S., Ouyang, S., Buell, C. R., Talbert, P. B., Henikoff, S., & Jiang, J. (2005). Transcription and Histone Modifications in the Recombination-Free Region Spanning a Rice Centromere[W]. *The Plant Cell*, *17*(12), 3227–3238. https://doi.org/10.1105/tpc.105.037945

Yan, H., Kikuchi, S., Neumann, P., Zhang, W., Wu, Y., Chen, F., & Jiang, J. (2010). Genome-wide mapping of cytosine methylation revealed dynamic DNA methylation patterns associated with genes and centromeres in rice: Genome-wide methylation of rice. *The Plant Journal*, *63*(3), 353–365. https://doi.org/10.1111/j.1365-313X.2010.04246.x

Zakrzewski, F., Schmidt, M., Van Lijsebettens, M., & Schmidt, T. (2017). DNA methylation of

retrotransposons, DNA transposons and genes in sugar beet (*Beta vulgaris* L.). *The Plant Journal*, *90*(6), 1156–1175. https://doi.org/10.1111/tpj.13526

Zhang, H., Lang, Z., & Zhu, J.-K. (2018). Dynamics and function of DNA methylation in plants. *Nature Reviews Molecular Cell Biology*, *19*(8), 489–506. https://doi.org/10.1038/s41580-018-0016-z

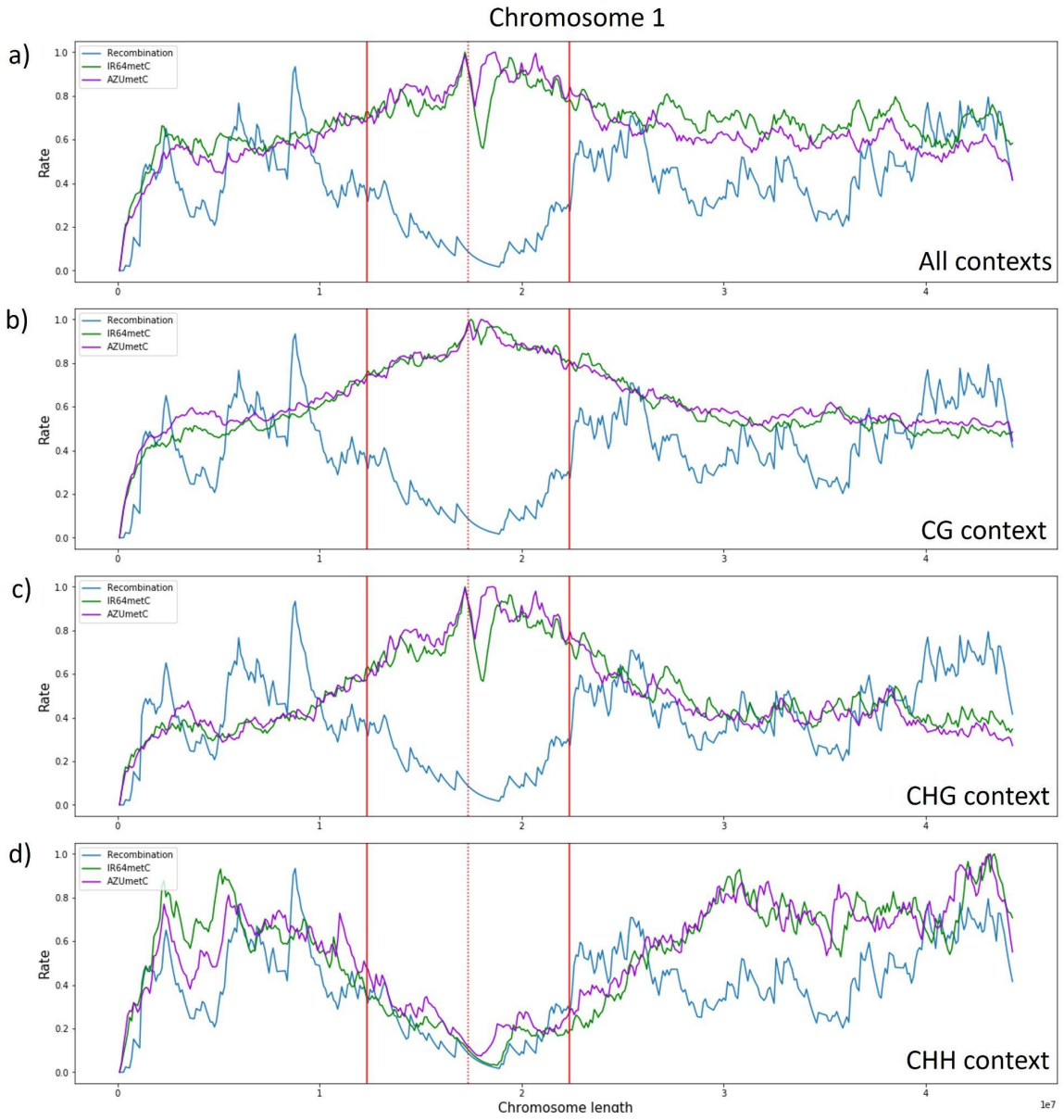


Figure 1. Recombination and methylated cytosines through chromosome 1 for the rice varieties IR64 and Azucena. The centromere is represented by a red dotted line and the influence of the centromere region by red solid lines. a) methylated cytosines of all contexts. b) Methylated cytosines of CG context. c) Methylated cytosines of CHG context. d) Methylated cytosines of CHH context.

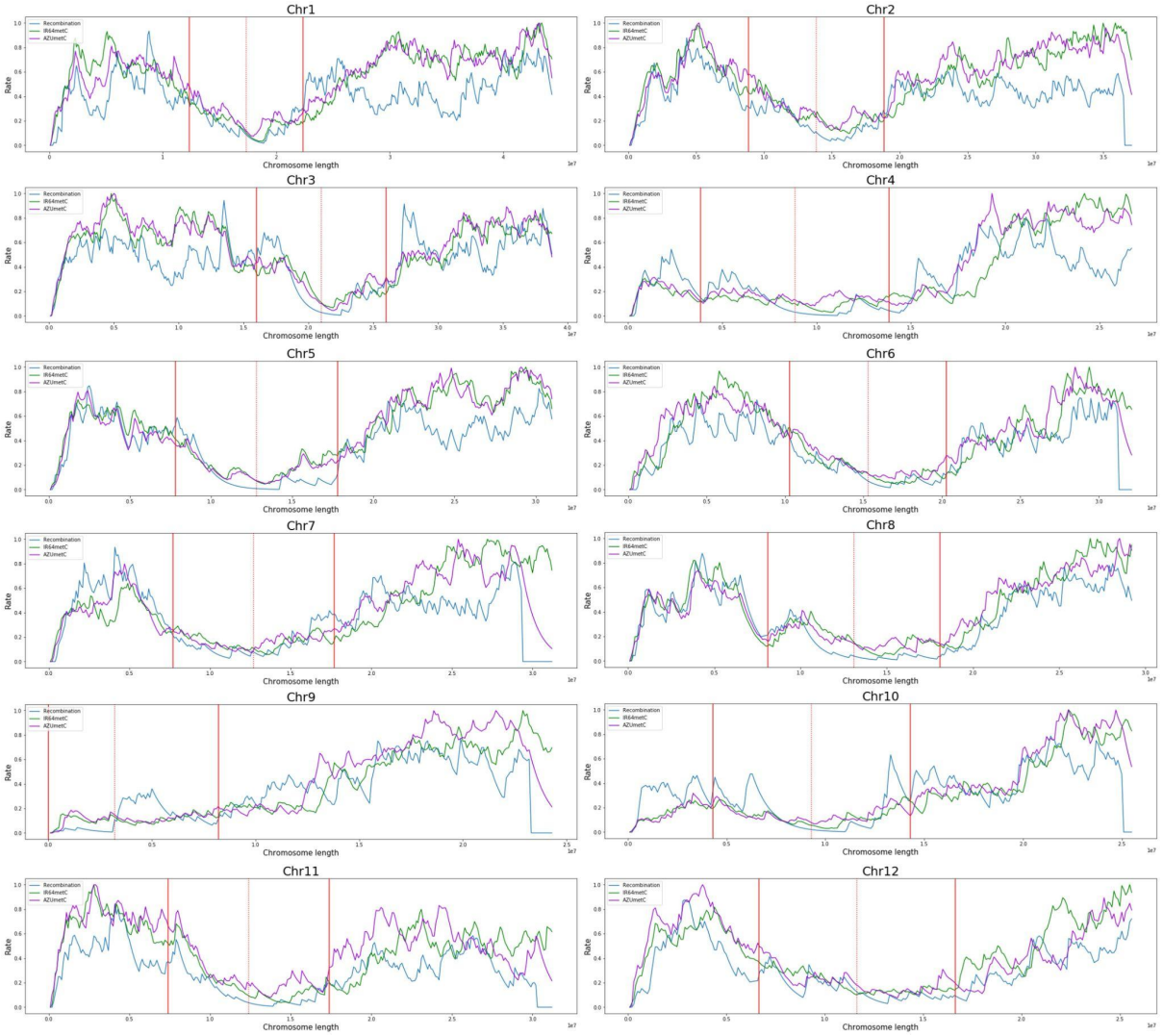


Figure 2. Distribution of methylated cytosines in CHH context along the twelve rice chromosomes for the IR64 and Azucena varieties, in comparison with the chromosomal recombination between these two varieties. The centromere is represented by a red dotted line and the influence of the centromere region in recombination by red solid lines.

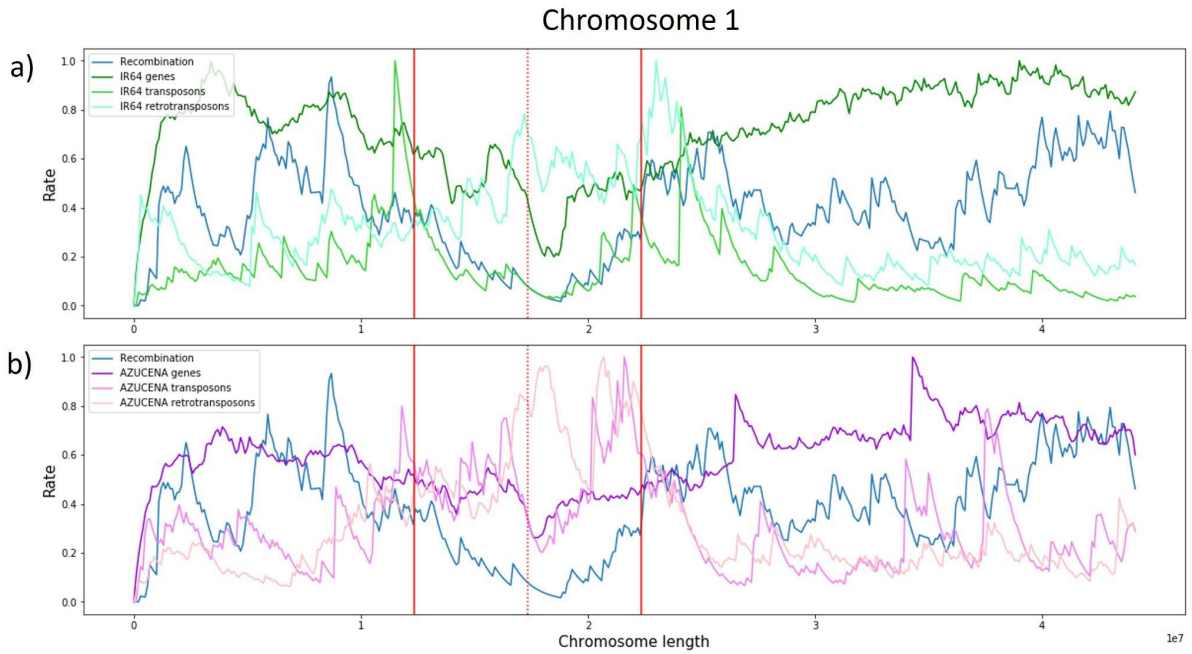


Figure 3. Genes, transposons, retrotransposons, compared to cross over recombination through chromosome 1 for rice varieties a) IR64 and b) Azucena. The centromere is represented by a red dotted line and the influence of the centromere region in recombination by red solid lines.

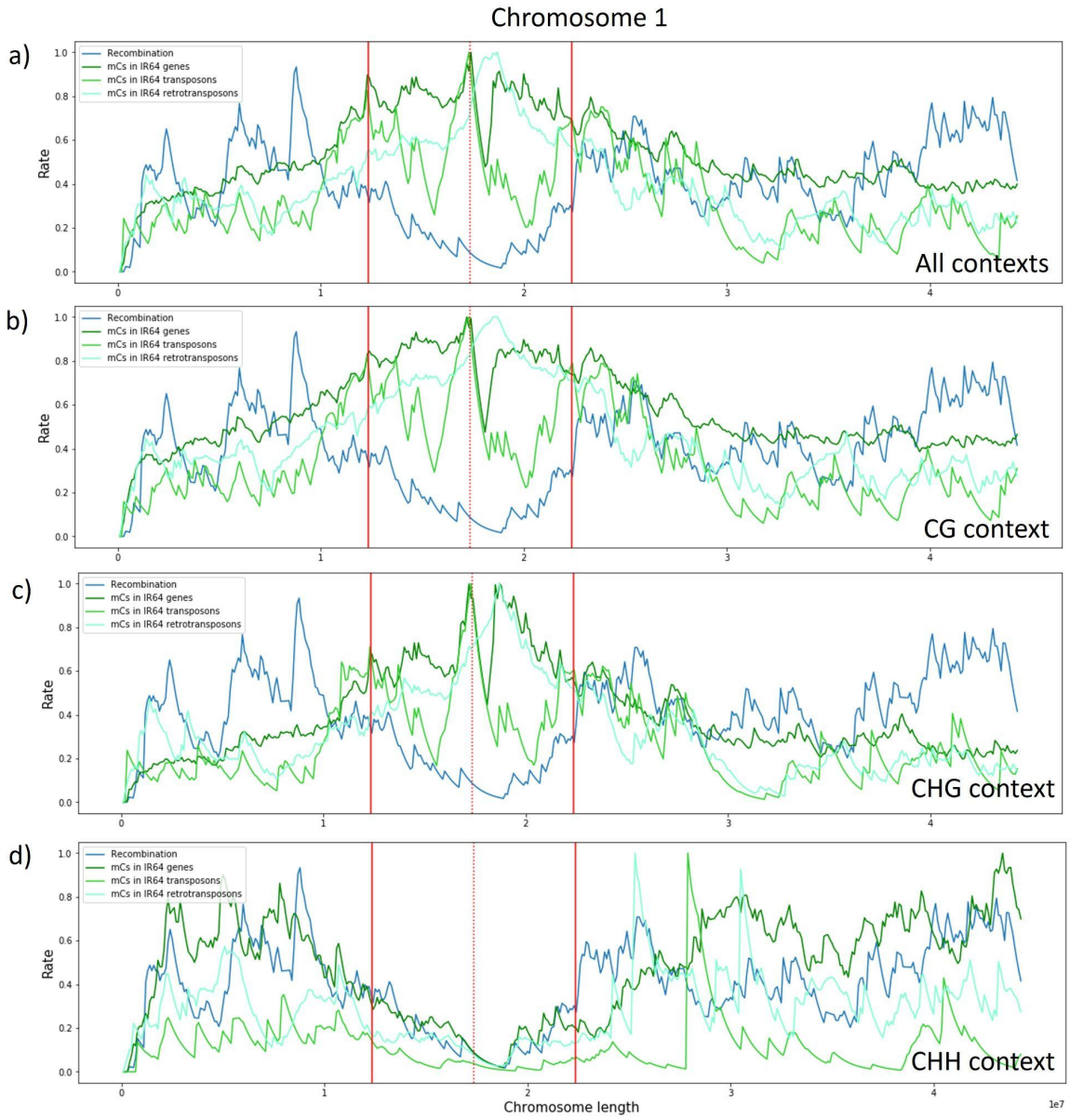


Figure 4. Recombination and methylated cytosines inside genes, transposons, and retrotransposons through chromosome 1 for the rice variety IR64. The centromere is represented by a red dotted line and the influence of the centromere region in recombination by red solid lines.

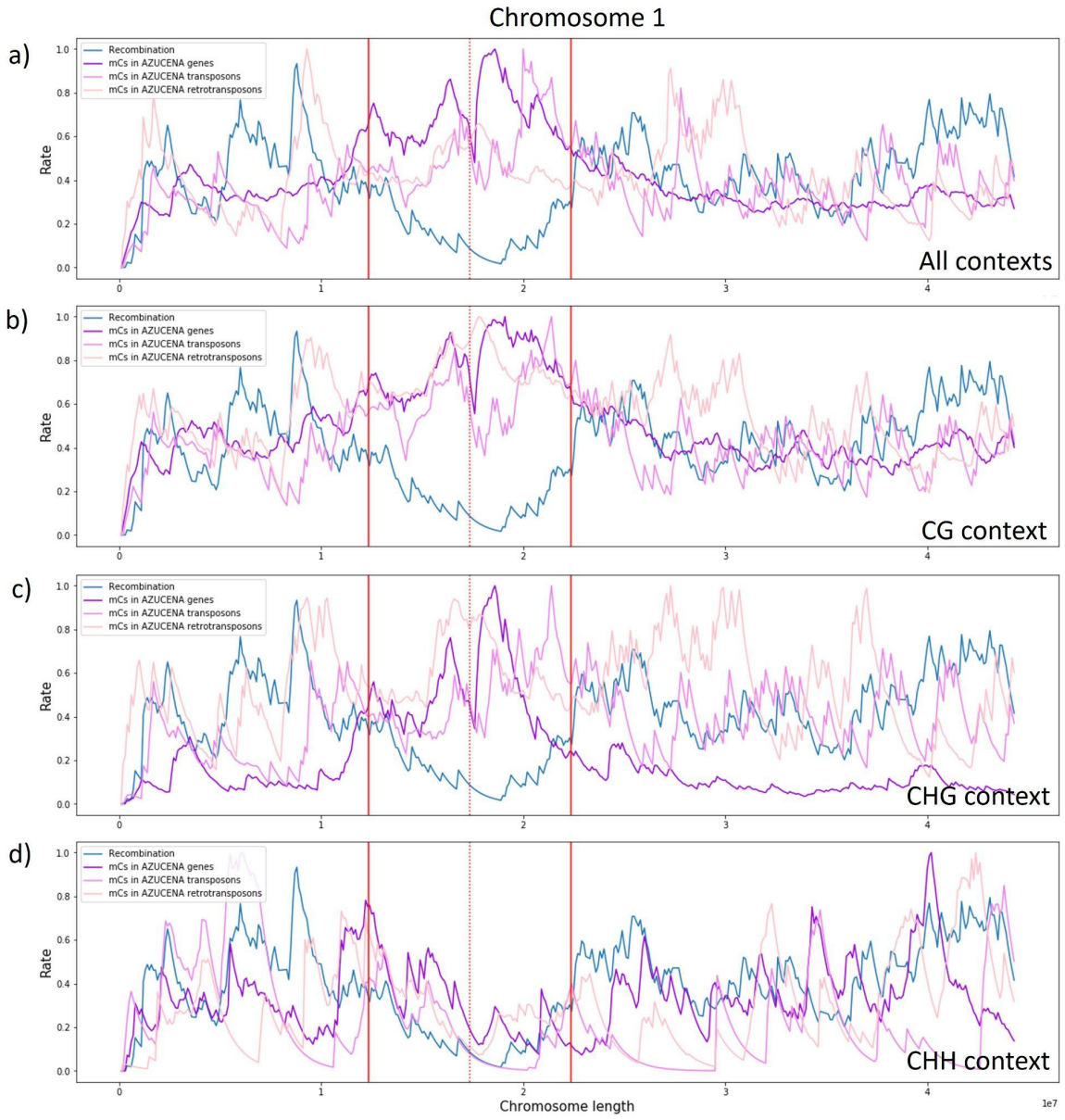


Figure 5. Recombination and methylated cytosines inside genes, transposons, and retrotransposons through chromosome 1 for the rice variety Azucena. The centromere is represented by a red dotted line and the influence of the centromere regionin recombination by red solid lines.

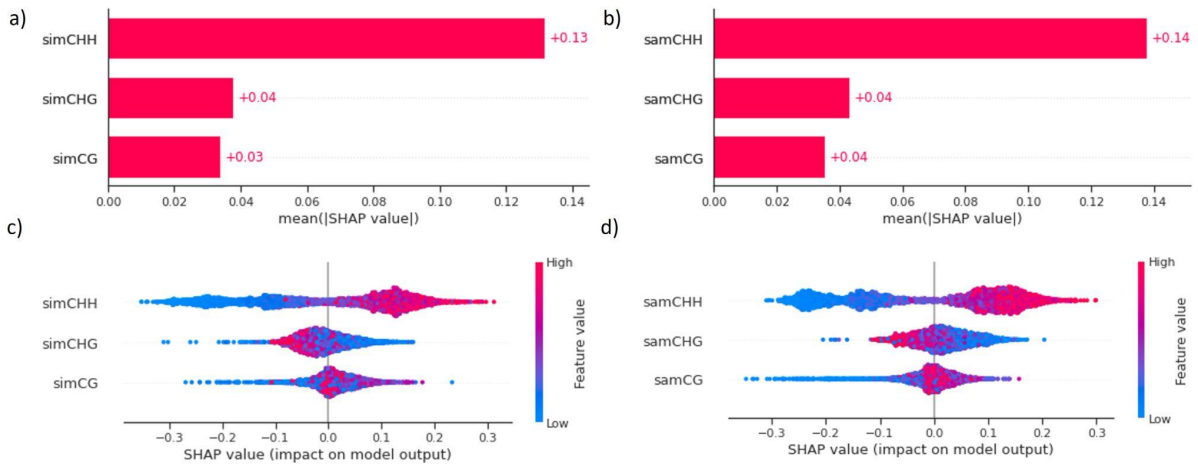


Figure 6. Shapley values and contributions of features CG, CHG and CHH to the prediction of recombination rates, using IR64 and Azucena data. a) Contribution values of features for the IR64 variety. b) Contribution values of features for the Azucena variety. c) Shapley values for the IR64 variety. d) Shapley values for the Azucena variety.

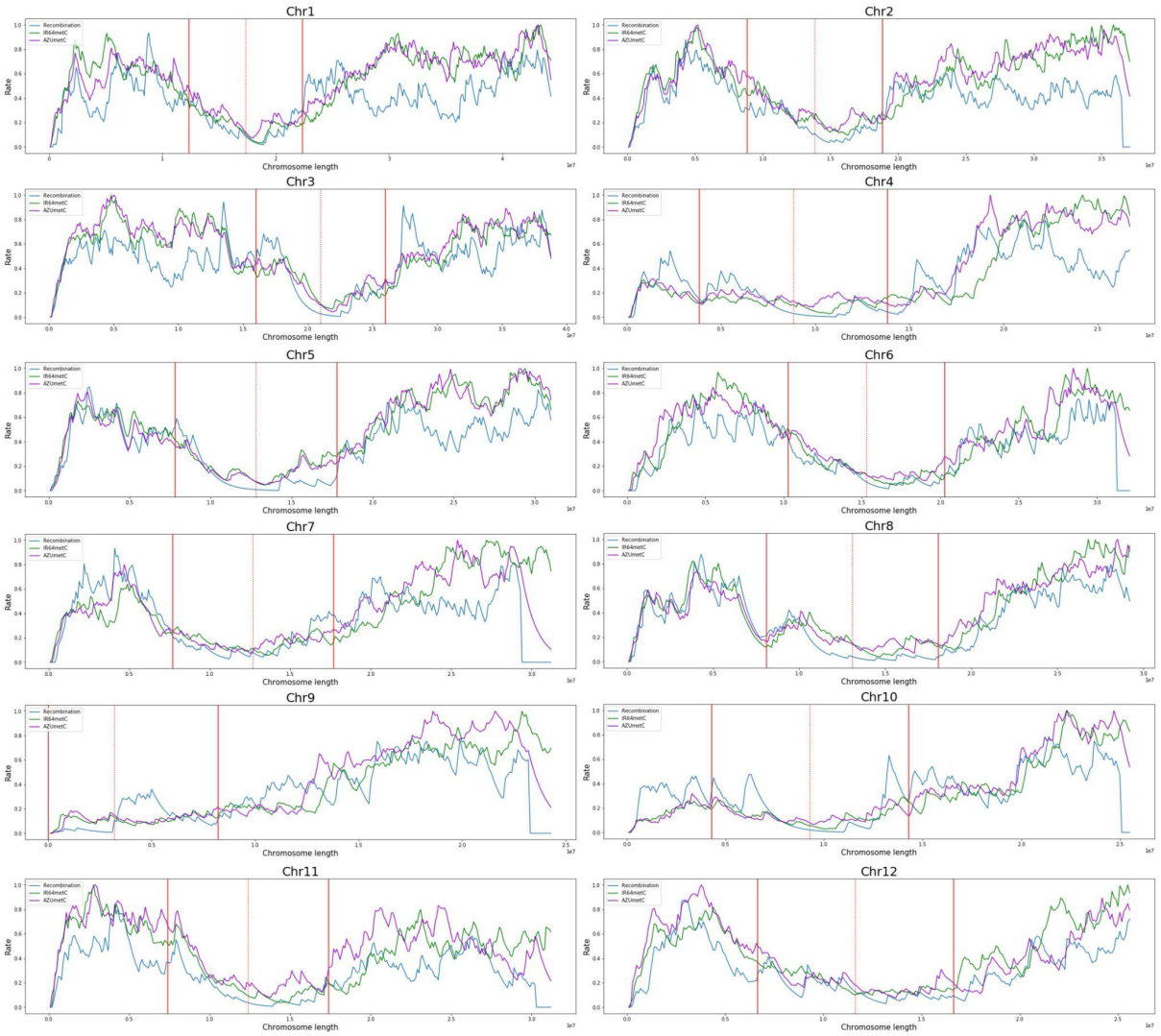


Figure 7. Cross-recombination predictions between IR64 and Azucena varieties, through the Extratrees machine learning model using the count of methylated cytosines in the CHH context as a feature. Predictions on the IR64 manifold were made using Azucena methylation as the training dataset, and predictions on the Azucena manifold were made using IR64 methylation as the training dataset.

PLOS ONE

Prediction of Crossover Recombination using Parental Genomes

--Manuscript Draft--

Manuscript Number:	PONE-D-22-01079
Article Type:	Research Article
Full Title:	Prediction of Crossover Recombination using Parental Genomes
Short Title:	Prediction of Crossover Recombination
Corresponding Author:	Mauricio Peñuela Pontificia Universidad Javeriana - Cali Cali, COLOMBIA
Keywords:	rice, mathematical model, recombination landscape, RILs, Whole genome sequencing
Abstract:	Meiotic recombination is a crucial cellular process, being one of the major drivers of evolution and adaptation of species. In plant breeding, crossing is used to introduce genetic variation among individuals and populations. A better characterization of the variation of the recombination rates along the chromosomes would enable breeding programmes to increase the chances of creating novel allele combinations, and more generally, to introduce new varieties with a collection of desirable traits. While different approaches to predict recombination rates for different species have been developed, they fail to estimate the outcome of a crossing between two specific accessions. This is missing in the panel of tools that breeders can use to reduce costs and execution times of crossing experiments. This papers builds on the hypothesis that chromosomal recombination correlates positively to a measure of sequence identity. In particular, we develop a model that uses sequence identity, combined with other features derived from genome alignment (including the number of variants, inversions, absent bases, and CentO sequences) to predict local chromosomal recombination in rice. Model performance is validated in an inter-subspecific indica x japonica cross, using 212 recombinant inbred lines. Across all 12 chromosomes, an average correlation of about 0.8 between experimental and prediction rates is achieved.
Order of Authors:	Mauricio Peñuela
	Camila Riccio-Rengifo
	Jorge Finke
	Camilo Rocha
	Anestis Gkanogiannis
	Rod A. Wing
	Mathias Lorieux
Opposed Reviewers:	
Additional Information:	
Question	**Response**
Financial Disclosure Enter a financial disclosure statement that describes the sources of funding for the work included in this submission. Review the submission guidelines for detailed requirements. View published research articles from *PLOS ONE* for specific examples. This statement is required for submission	This work was funded by the OMICAS program: Optimización Multiescala In-silico de Cultivos Agrícolas Sostenibles (Infraestructura y Validación en Arroz y Caña de Azúcar), anchored at the Pontifcia Universidad Javeriana in Cali and funded within the Colombian Scientifc Ecosystem by The World Bank, the Colombian Ministry of Science, Technology and Innovation, the Colombian Ministry of Education and the Colombian Ministry of Industry and Turism, and ICETEX, under GRANT ID: FP44842-217-2018. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

and **will appear in the published article** if the submission is accepted. Please make sure it is accurate. **Unfunded studies** Enter: *The author(s) received no specific funding for this work.* **Funded studies** Enter a statement with the following details: • Initials of the authors who received each award • Grant numbers awarded to each author • The full name of each funder • URL of each funder website • Did the sponsors or funders play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript? • **NO** - Include this sentence at the end of your statement: *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.* • **YES** - Specify the role(s) played. * typeset	
Competing Interests Use the instructions below to enter a competing interest statement for this submission. On behalf of all authors, disclose any competing interests that could be perceived to bias this work—acknowledging all financial support and any other relevant financial or non-financial competing interests. This statement is required for submission and **will appear in the published article** if the submission is accepted. Please make sure it is accurate and that any funding sources listed in your Funding Information later in the submission form are also declared in your Financial Disclosure statement. View published research articles from *PLOS ONE* for specific examples.	The authors have declared that no competing interests exist.

NO authors have competing interests Enter: *The authors have declared that no competing interests exist.* **Authors with competing interests** Enter competing interest details beginning with this statement: *I have read the journal's policy and the authors of this manuscript have the following competing interests: [insert competing interests here]* * typeset	
Ethics Statement Enter an ethics statement for this submission. This statement is required if the study involved: • Human participants • Human specimens or tissue • Vertebrate animals or cephalopods • Vertebrate embryos or tissues • Field research Write "N/A" if the submission does not require an ethics statement. General guidance is provided below. Consult the submission guidelines for detailed instructions. **Make sure that all information entered here is included in the Methods section of the manuscript.**	N/A

Format for specific study types **Human Subject Research (involving human participants and/or tissue)** • Give the name of the institutional review board or ethics committee that approved the study • Include the approval number and/or a statement indicating approval of this research • Indicate the form of consent obtained (written/oral) or the reason that consent was not obtained (e.g. the data were analyzed anonymously) **Animal Research (involving vertebrate animals, embryos or tissues)** • Provide the name of the Institutional Animal Care and Use Committee (IACUC) or other relevant ethics board that reviewed the study protocol, and indicate whether they approved this research or granted a formal waiver of ethical approval • Include an approval number if one was obtained • If the study involved *non-human primates*, add *additional details* about animal welfare and steps taken to ameliorate suffering • If anesthesia, euthanasia, or any kind of animal sacrifice is part of the study, include briefly which substances and/or methods were applied **Field Research** Include the following details if this study involves the collection of plant, animal, or other materials from a natural setting: • Field permit number • Name of the institution or relevant body that granted permission	
Data Availability Authors are required to make all data underlying the findings described fully available, without restriction, and from the time of publication. PLOS allows rare exceptions to address legal and ethical concerns. See the PLOS Data Policy and FAQ for detailed information.	Yes - all data are fully available without restriction

A Data Availability Statement describing where the data can be found is required at submission. Your answers to this question constitute the Data Availability Statement and **will be published in the article**, if accepted. **Important:** Stating 'data available on request from the author' is not sufficient. If your data are only available upon request, select 'No' for the first question and explain your exceptional situation in the text box. Do the authors confirm that all data underlying the findings described in their manuscript are fully available without restriction?	
Describe where the data may be found in full sentences. If you are copying our sample text, replace any instances of XXX with the appropriate details. • If the data are **held or will be held in a public repository**, include URLs, accession numbers or DOIs. If this information will only be available after acceptance, indicate this by ticking the box below. For example: *All XXX files are available from the XXX database (accession number(s) XXX, XXX.*). • If the data are all contained **within the manuscript and/or Supporting Information files**, enter the following: *All relevant data are within the manuscript and its Supporting Information files.* • If neither of these applies but you are able to provide **details of access elsewhere**, with or without limitations, please do so. For example: *Data cannot be shared publicly because of [XXX]. Data are available from the XXX Institutional Data Access / Ethics Committee (contact via XXX) for researchers who meet the criteria for access to confidential data.* *The data underlying the results presented in the study are available from (include the name of the third party*	All relevant data are within the manuscript and its Supporting Information files.

and contact information or URL). • This text is appropriate if the data are owned by a third party and authors do not have permission to share the data. * typeset	
Additional data availability information:	

Dear editor,

We are submitting the manuscript “Prediction of Crossover Recombination using Parental Genomes” for review for publication in your journal. Cross recombination is a natural event that occurs in gametogenesis and is responsible for generating diversity in sexual organisms, generating DNA exchange between homologous chromosomes. Knowing where these cross events occur is essential for breeders as it allows them to plan crosses between varieties to introduce new gene combinations. Here we present a model that predicts recombination along chromosomes using identity between two genomes using rice as a model. Since the objective of this work is a novel model to predict recombination, it is important to locate the "Materials and methods" section before the results, because this is where the model is presented. We hope that the manuscript falls within the scope of the journal and is of great interest to its audience.

Thank you for your consideration.

The authors

Prediction of Crossover Recombination using Parental Genomes

Mauricio Peñuela[1☯*], Camila Riccio-Rengifo[1☯], Jorge Finke[1], Camilo Rocha[1], Anestis Gkanogiannis[2], Rod A. Wing[3], Mathias Lorieux[2,4*],

1 Facultad de Ingeniería y Ciencias, Pontificia Universidad Javeriana. Cali, Colombia.
2 AgroBiotechnology Unit, Alliance Bioversity-CIAT. Cali, Colombia.
3 Arizona Genomics Institute, University of Arizona. Tucson, AZ, US
4 DIADE, University of Montpellier, CIRAD, IRD. Montpellier, France.

☯These authors contributed equally to this work.
* mauricio.penuela@javerianacali.edu.co, mathias.lorieux@ird.fr

Abstract

Meiotic recombination is a crucial cellular process, being one of the major drivers of evolution and adaptation of species. In plant breeding, crossing is used to introduce genetic variation among individuals and populations. A better characterization of the variation of the recombination rates along the chromosomes would enable breeding programmes to increase the chances of creating novel allele combinations, and more generally, to introduce new varieties with a collection of desirable traits. While different approaches to predict recombination rates for different species have been developed, they fail to estimate the outcome of a crossing between two specific accessions. This is missing in the panel of tools that breeders can use to reduce costs and execution times of crossing experiments. This papers builds on the hypothesis that chromosomal recombination correlates positively to a measure of sequence identity. In particular, we develop a model that uses sequence identity, combined with other features derived from genome alignment (including the number of variants, inversions, absent bases, and CentO sequences) to predict local chromosomal recombination in rice. Model performance is validated in an inter-subspecific *indica* x *japonica* cross, using 212 recombinant inbred lines. Across all 12 chromosomes, an average correlation of about 0.8 between experimental and prediction rates is achieved.

Author summary

Crossover recombination is the event by which large portions of DNA are exchanged between homologous chromosomes during meiosis. For genetic breeders, it is of great interest to know where these exchange events occur. They can use the highest recombination regions to introduce, through genetic crosses, relevant genes from one variety into another that lacks them. In this paper, we demonstrate how the sequence identity between the genomes of two rice varieties (IR64 and Azucena) is positively correlated with chromosomal recombination. On this basis, we build a model that uses information from the alignment between the two genomes, such as variants, inversion bases, absent bases, and CentO sequences, to predict recombination along the chromosomes. The model consists of different steps that fit the original identity values using a series of parameters in 100 kbp windows. We verify that the model can be

adjusted for any of the twelve chromosomes and obtain similar predictions in all cases. We expect this model will help breeders to predict high and low recombination regions, facilitating the genetic improvement of rice varieties without the need to incur in the expense of time, effort, and money involved in calculating experimental recombination.

Introduction

Crossover recombination refers to the exchange of genetic material across homologous chromosomes. It is an important process during meiosis in the production of gametes and contributes to the creation of novel allele combinations [1–3]. Both biological and biochemical factors influence the recombination rates along each chromosome. In rice, for example, it has been shown that recombination rates play a key role for adaptive evolution in rapidly changing environments and vary with exposure to different stresses [4]. Furthermore, a number of studies have shown that recombination rates across different regions along a chromosome (i.e., for windows of a certain size along a chromosome) are not uniformly distributed [5,6]. Instead, there exist the so-called hot and cold spots, which represent regions that, when compared to regular regions, exhibit relatively high and low rates of recombination. According to [4,7,8], the location of such regions varies between populations, primarily as a result of population history.

Over generations, recombination has played an important role in the evolution of the genome in plants [6]. Evidence suggests that recombination responds not only to direct selection but also to the effects of indirect selection over different traits [7]. From the perspective of agricultural growth and development, understanding recombination rates enables plant breeders to develop better criteria for determining (i) which varieties represent the most suitable parents for crosses and (ii) which progeny make the selection process highly effective [9]. More specifically, estimating the recombination rates along the chromosomes accelerates the fine mapping of genetic traits [10], which lies at the heart of efforts to design better crops [2].

The design and development of experiments to measure recombination rates between varieties is a demanding task, both in terms of costs and time. Such efforts require, first, a large number of recombinant descendants and second, a large number of markers from high throughput next generation sequencing. Not surprisingly, several studies have introduced different strategies to characterize recombination rates along the chromosomal arms [2,3,8,11–15]. These studies generally evaluate several varieties to construct a genomic recombination landscape for a species as a whole. They tend to follow one of two general approaches. The first approach seeks to discover and understand which factors explain recombination. The second approach aims to predict either the location of hot and cold spot, or to estimate the recombination rates along the chromosome using different types of genome sequence information.

Following the first approach, the work by Rodgers-Melnick et al. [11] identifies recombination breakpoints in populations of U.S. and Chinese maize. The authors show that the distribution of gene density and CpG methylation explains, on a broad scale, cross-overs. In another closely-related study, Colomé-Tatché et al. [12] evaluate the combined effect of removing sequence polymorphisms and repeat-associated DNA methylation on the meiotic recombination landscape of an Arabidopsis mapping population. Similarly, Horton et al. [13] test 1,307 worldwide Arabidopsis accessions to characterize the pattern of recombination history. The authors observe an enrichment of hot spots in regions of intergenic space and repetitive DNA. Finally, Haas et al. [2] identify AT-rich DNA motifs associated with recombination breakpoints in 60 recombinant inbred lines of tomato.

One of the first studies to take the second approach is the work by Liu et al. [8]. Based on sequence k-mer frequencies, the authors predict hot and cold spots in yeast

using a machine learning method known as increment of diversity combined with quadratic discriminant analysis. The work is extended in [14], by introducing an algorithm to predict hot and cold spots in yeast. Unlike [14], the work by Demirci et al. [15] applies features related to genome content and genomic accessibility, such as gene annotation, propeller twist and helical twist, and AT/TA dinucleotides to train different machine learning models (specifically, decision trees, logistic regression, and random forest models). The work predicts hot and cold spots in maize, rice, tomato, and Arabidopsis. The more recent work by Adrion et al. [3] proposes a method to predict the recombination landscape based on deep learning algorithms; they evaluate model predictions in African populations of *Drosophila melanogaster*.

A number of studies that follow the second approach characterize broad-scale recombination rates for windows of certain size along a chromosome. They tend to focus on a given population or species. However, little attention has been paid to developing analytical frameworks that help explain recombination rates for a specific crossing between two particular varieties. The lack of such models limits the applicability of the outcome of studies that follow the second approach for breeding programmes. To overcome this limitation, the validation of such models is required. The lofty aim of the mechanism-based models is that the principles for prediction are generalizable and applicable to other varieties or species.

A large number of studies that aim to estimate recombination rates focus on rice for several reasons. Among them, rice (*O. sativa* L.) is highly homozygous, which makes haplotype reconstruction easy and also eliminates the need of phasing. Moreover, rice provides food for more than half the world's population [16]. This paper focuses on predicting specific recombination rates that result as the product of a crossing between the rice varieties of IR64 (indica) and Azucena (japonica). In particular, this work explores the hypothesis that an identity measure between genome sequences of the parents is correlated with chromosomal recombination. The analysis is performed based on whole genome sequencing of both rice varieties and their recombinant inbred lines.

The main result suggests that the sequence identity is positively correlated with chromosomal recombination. Model is proposed to predict recombination using parental sequences as its input. Unlike the previously models based on machine learning or deep learning methods, this model is a mechanism-based model, whose outcome is the result of a series of steps applied to specific features measured after the alignment process between parental sequences. The model is calibrated on chromosome 1 and tested on the remaining 11 chromosomes. The validation of the model shows that the prediction for the 12 rice chromosomes has an average correlation of 80% with the recombination rates. The model offers a tool to help improve the plant breeding programs in rice cultivars.

Materials and methods

The IR64 (indica cluster) and Azucena (tropical japonical cluster) varieties were crossed to generate a F1 generation. A total of 212 F8 recombinant inbred lines (RIL) were generated in the greenhouse at IRD, France by single-seed descent (SSD) from the F2. Then, the lines were advanced in the field to the F12 generation at the International Center for Tropical Agriculture (CIAT, now "Alliance Bioversity-CIAT") in Palmira, Colombia. This population is also part of a Nested association Mapping design [17].

Whole Genome Sequencing

Leaf tissue from parent plants and F12 lines were collected, and DNA was extracted following a protocol similar to [17]. Platinum-grade PacBio assemblies of the parental genomes were obtained at the Arizona Genomics Institute (AGI, Tucson, Arizona) [18].

The IR64 and Azucena genomes that were used are available in the GenBank repository with the accession numbers RWKJ00000000 and PKQC000000000, respectively. The F12 RIL genomes were sequenced using paired-end Illumina with a depth of approximately 1x.

Data imputation and recombination values

SNP features for the F12 genomes were extracted using a standard bioinformatics pipeline. Briefly, Illumina reads were mapped on the IR64 RefSeq, and SNP features were extracted with the GATK package. Genotypes and recombination breakpoints (that is, meiotic crossovers) were imputed and corrected using the NOISYmputer algorithm introduced in [19]. The resulting genotypes data for each chromosome consist of a matrix of genetic markers (arranged by sequence position) versus individuals. An entry is encoded as A or B depending on the parental origin of the corresponding sequence. Genetic recombination maps were calculated with MapDisto v2 [20, 21], using the Kosambi mapping function to convert recombination fractions into centimorgans (cM) [22].

Recombination measurement

Cublic spline smoothing of local recombination rates, expressed as cM/bp, were calculated in sliding windows of 100 Kbp in MapDisto v2.

Data pre-processing protocol

Since we wished to test the hypothesis that crossover frequency is a function of genome similarity, sequence features were extracted and a measure of "identity" was introduced. In particular, the script consisted of an initial alignment for each pair of parental chromosomes using the *nucmer* command from MUMmer3 [23] with default parameters. The outcome is a delta file which is filtered using the command `delta-filter -r -q`. The filtered file is used to extract coordinates using the command `show-coords -r`. Sequence variants are extracted from the initial delta file using the command `show-snps`. Subsequently, the sequence is divided into windows of 100 Kbp of size. Each window is built and associated with parameters such as mapped and absent bases, number of variants (bases corresponding to SNPs or deletion polymorphism), and bases in inversions. The identity of each window w, denoted by $Id(w)$, is constructed to measure how similar the two sequences *ref* and *qry* are in equivalent regions depending on how many nucleotides they share. The number of variants V, the number of bases in inversions I, and the absent bases A are the features that can modify the identity criteria directly. If these features are not present in a certain window, its identity value is set at its maximum value:

$$Id(w) = window\_size - V(w) - I(w) - A(w). \quad (1)$$

Testing hypothesis

Under the hypothesis that similar genomic regions recombine more frequently, a correlation analysis was developed between the identity criteria and the local recombination values for the twelve rice chromosomes. The Pearson's correlation coefficient was used as the measure of correlation r. The identity and the recombination were exponentially smoothed to reduce noise and find the best fit with the trend of the data. For example, the actual recombination measured on a sequence of windows, denoted by $X(w)$, was exponentially smoothed as follows:

$$X_s(w) = \begin{cases} X(w) & w = 0 \\ \alpha X(w) + (1-\alpha) X_s(w) & w > 0, \end{cases} \tag{2}$$

where $\alpha \in (0, 1)$ is the smoothing factor. For the correlation analysis, both identity and experimental recombination were smoothed with the same factor. Various exponential smoothing factors were evaluated to try to reduce noise and find the best fit with the data trend (Fig 1 and Fig 2), being $\alpha = 0.1$ the one that gives the best fit.

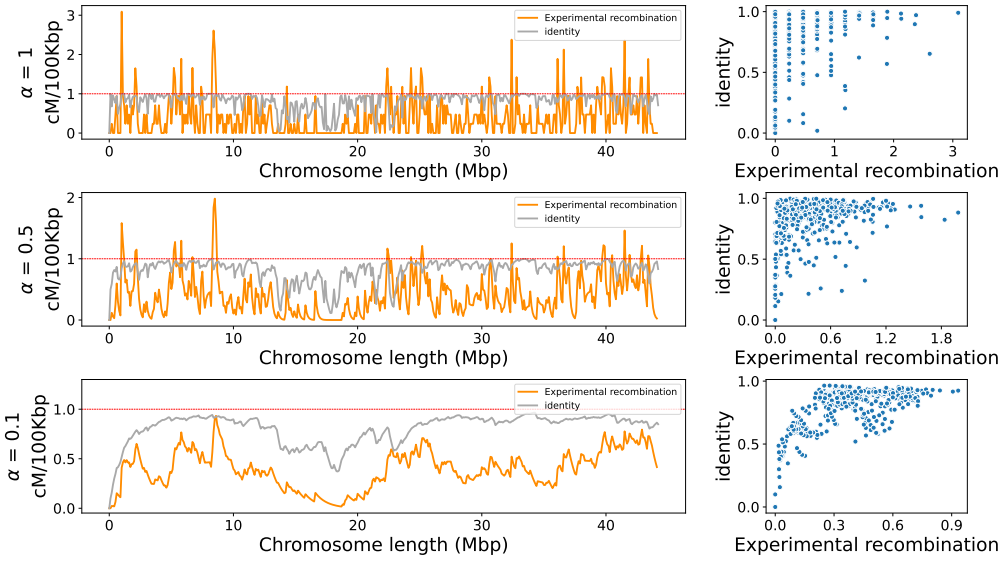


Fig 1. Effect of smoothing on recombination landscape. Landscape of identity and recombination, with its corresponding correlation, at different levels of smoothing for rice chromosome 1 in the IR64 x Azucena cross.

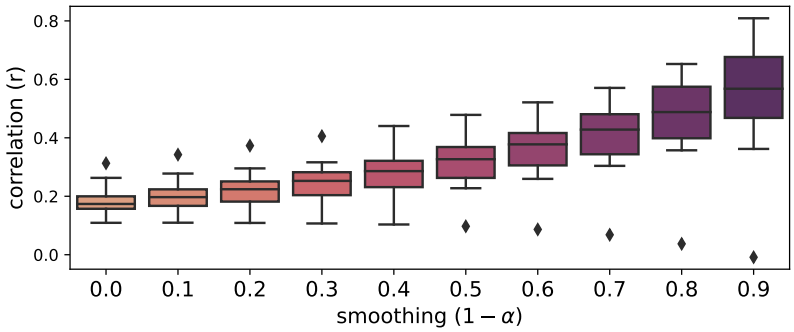


Fig 2. Effect of smoothing on correlation distribution. Boxplots of correlations between identity and recombination for 12 rice chromosomes (cross IR64 x Azucena) at different smoothing factor.

Model description

The proposed model predicts recombination for each pair of homologous chromosomes from two parental organisms. Arbitrarily, one of the parental organisms is taken as reference. Each pair of homologous chromosomes is identified by a reference chromosome (*ref*) and a query chromosome (*qry*). For each (*ref*, *qry*) pair, the model compares using

an alignment process. Additionally, the CentO sequence is aligned with each of *ref* and *qry* to approximate the location of each centromere. Moreover, the reference chromosome is subdivided into $n \in \mathbb{N} > 0$ windows of length 100Kbp each. The model then assigns a recombination value to each window, depending on a set of features from the *ref-qry* alignment and the approximate location of the centromeres (Fig 3).

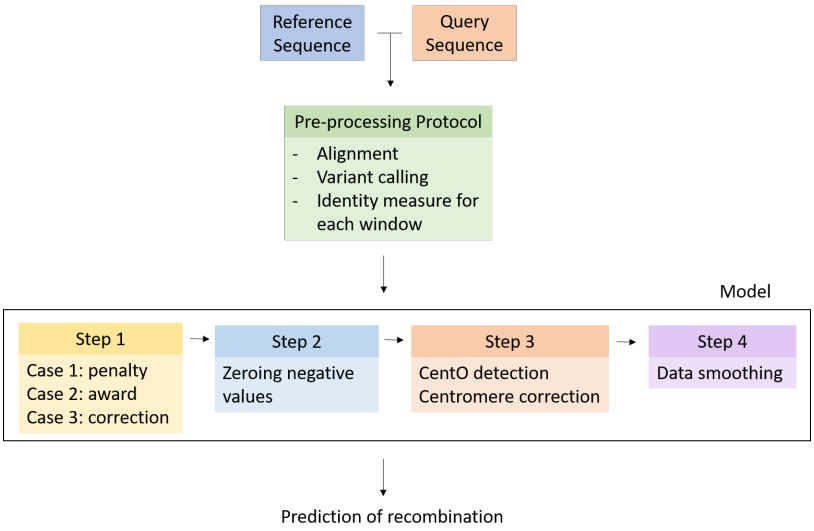


Fig 3. Model workflow. Schematic representation of data preprocessing and model steps to predict recombination.

Three features from the *ref-qry* alignment are considered for each window:

- Identity: proportion of identical base pairs.
- Variants: proportion of SNPs and deletion polymorphisms.
- Absent bases: proportion of query bases that are not mapped in the reference chromosome.

Let $W = \{1, 2, ..., n\}$ be the set representing the n windows partitioning a given chromosome, and $Id_0 : W \to [0, 1]$, $V : W \to [0, 1]$, and $A : W \to [0, 1]$ functions representing the identity, the variants, and the absent bases respectively. The identity is taken as a starting point to predict recombination. The model adapts the identity values in four sequential steps.

Step 1: Cases

Three mutually exclusive cases are considered starting from the identity values mapped by Id_0. The model has a total of 7 parameters ($p_i \in [0, 1], \quad \forall i \in \{1, 2, ..., 7\}$), which transform the identity values as follows. The first case penalizes with p_1 those windows with identity values inferior to p_2. The second case rewards with p_3 those windows with identity values inferior to p_4. The third case penalizes with p_5 those windows with absent bases greather than p_6. An additional constraint to apply Case one is that the variants must be above p_7, while for the cases two and three variants must be below the same threshold (p_7). Thus, an updated identity function $Id_1 : W \to \mathbb{R}$ is defined for each window $w \in W$ as:

$$Id_1(w) = \begin{cases} Id_0(w) - p_1 & , \quad Id_0(w) < p_2 \wedge V(w) > p_7 \\ Id_0(w) + p_3 & , \quad Id_0(w) < p_4 \wedge V(w) < p_7 \\ Id_0(w) - p_5 & , \quad A(w) > p_6 \wedge V(w) < p_7 \\ Id_0(w) & , \quad \text{otherwise} \end{cases} \tag{3}$$

Step 2: Negative values

Negative recombination values do not make biological sense. Therefore, only non-negative values are considered by correcting negative values to be zero. Mathematically, this step produces a function $Id_2 : W \to \mathbb{R} \geq 0$, defined for each $w \in W$ as:

$$Id_2(w) = \max(0, Id_1(w)) \tag{4}$$

Step 3: Centromere correction

The alignments of the CentO sequence helps in approximating the location of a chromosome centromeres. Let $wcentO$ be a function that maps each of the reference and query chromosomes to the set of windows having the greatest number of alignments with the CentO sequence. Note that $wcentO(ref) \subseteq W$, $wcentO(qry) \subseteq W$, and both sets are non-empty. Then, the centromere boundaries can be approximated by the interval $[c_0, c_1]$ defined by:

$$c_0 = \min(wcentO(ref) \cup wcentO(qry)) \tag{5}$$

$$c_1 = \max(wcentO(ref) \cup wcentO(qry)) \tag{6}$$

That is, c_0 and c_1 are the left- and right-most windows with the greatest number of alignments with the CentO sequence, between the two chromosomes input to the model.

Next, the weight functions f for centromeric chromosomes and g for the telomeric chromosomes are defined:

$$f(w) = \begin{cases} 1 & , \quad 0 \leq Id_2(w) \leq c_0 - 50 \\ \frac{-1}{50}(w - c_0) & , \quad c_0 - 50 < Id_2(w) \leq c_0 \\ 0 & , \quad c_0 < Id_2(w) \leq c_1 \\ \frac{1}{50}(w - c_0) & , \quad c_1 < Id_2(w) \leq c_1 + 50 \\ 1 & , \quad c_1 + 50 < Id_2(w) < n \end{cases} \tag{7}$$

$$g(w) = \begin{cases} 0 & 0 \leq Id_2(w) < c_1 \\ 1 & c_1 \leq Id_2(w) \leq n \end{cases} \tag{8}$$

Finally, the identity values are corrected by the function $Id_3 : W \to \mathbb{R}$, using the weight functions as follows:

$$Id_3(w) = \begin{cases} Id_2(w) \cdot f(w) & c_1 > n/4 \\ Id_2(w) \cdot g(w) & \text{otherwise} \end{cases} \tag{9}$$

Step 4: Smoothing

The final part of the model is to smooth the data to reduce noise. Here, an adaptation of the exponential smoothing beginning at zero is used with a smooth factor $\alpha = 0.1$. Thus, the final prediction of recombination is given by the function $Id_4 : W \to \mathbb{R} \geq 0$ defined by:

$$Id_4(w) = \begin{cases} 0 & w = 0 \\ \alpha Id_3(w) + (1 - \alpha) Id_4(w) & w > 0. \end{cases} \tag{10}$$

Parameter optimization and model evaluation

The two metrics involved in the evaluation and calibration of the model are the Pearson correlation r and the coefficient of determination R^2. Given paired data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ consisting of n pairs, these two metrics are defined as follows:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{11}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{12}$$

where $\bar{x}$ is the sample mean and $\hat{y}$ is the fitted linear regression between x and y.

The 7 model parameters ($p_i \in [0, 1] \quad \forall i \in \{1, 2, ..., 7\}$) are adjusted by maximizing the coefficient of determination R^2 between the final prediction of the model Id_4 and the experimental recombination X_s (see Eq 2) of a single chromosome. The parameter optimization was done by the Sequential Least Squares Programming (SLSQP) minimizing $(1 - R^2)$. The model is adjusted from information on one chromosome and the adjusted model is used to predict recombination on the remaining 11 chromosomes. The prediction performance for each chromosome is evaluated based on Pearson correlation r and coefficient of determination R^2 between its output and the experimental recombination.

Results and Discussion

Sequence identity versus recombination

Our identity criteria values between parental chromosome sequences correlates positively with their progeny experimental recombination rates, as shown in Fig 4 and 5. This supports the hypothesis that similar genome regions recombine more frequently than regions with higher structural difference, a relationship that could explain several evolutionary mechanisms. Regarding plant crossing, this is coherent with the observation that recombination rates are higher in related varieties than in genetically distant ones. The identity *sensus stricto* measures the ratio of identical bases between two sequences and can accurately represent the structural variability because every base that is not equal between sequences is marked as a variant, inversion, or absent base, this even eliminates a common problem such as repetitive sequences because they are absorbed by the identity measure. The identity is in great proportion conditioned to the alignment process. However a good alignment process by itself is not sufficient for a proper identity estimation, because contigs do not follow a strict pattern due to structural rearrangements. As a result the resulting alignment is filled with paired and unpaired regions, and in many cases with inversion events or overlapping, without counting on the abundant variants such as SNPs and indels polymorphisms. Therefore, we develop a protocol which allows to quantify the identity and other variables using a windows-based approach.

The mean correlation between recombination rates and sequence identity evaluated for the 12 rice chromosomes in the IR64 x Azucena cross is $r = 0.53 \pm 0.21$. This positive correlation is important because a single variable is supporting a considerable magnitude of the explanation. However, identity is a condensed variable that implicitly carries the information of other structural variables. More specifically, identity is the ratio of bases that do not correspond to variants, inversions, or absent bases within a genome interval.

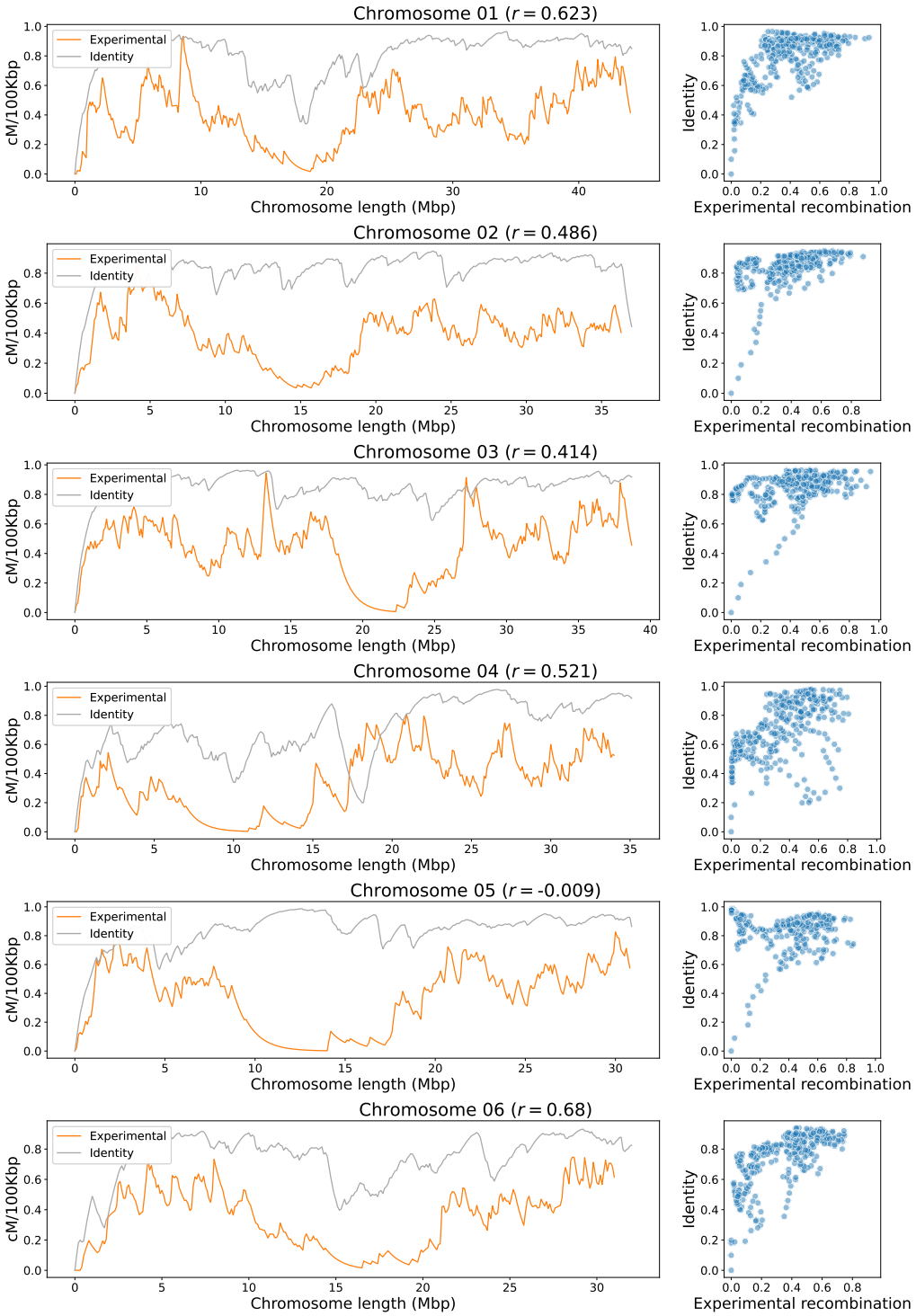


Fig 4. Identity correlation analysis for chromosomes 1 to 6. Landscape and correlation between chromosomal recombination and sequence identity for rice chromosomes 1 to 6 (cross IR64 x Azucena).

The higher correlations are found on chromosomes 9 and 10 with 0.809 and 0.705 respectively, meanwhile, lower correlations are found on chromosomes 5 and 12 with

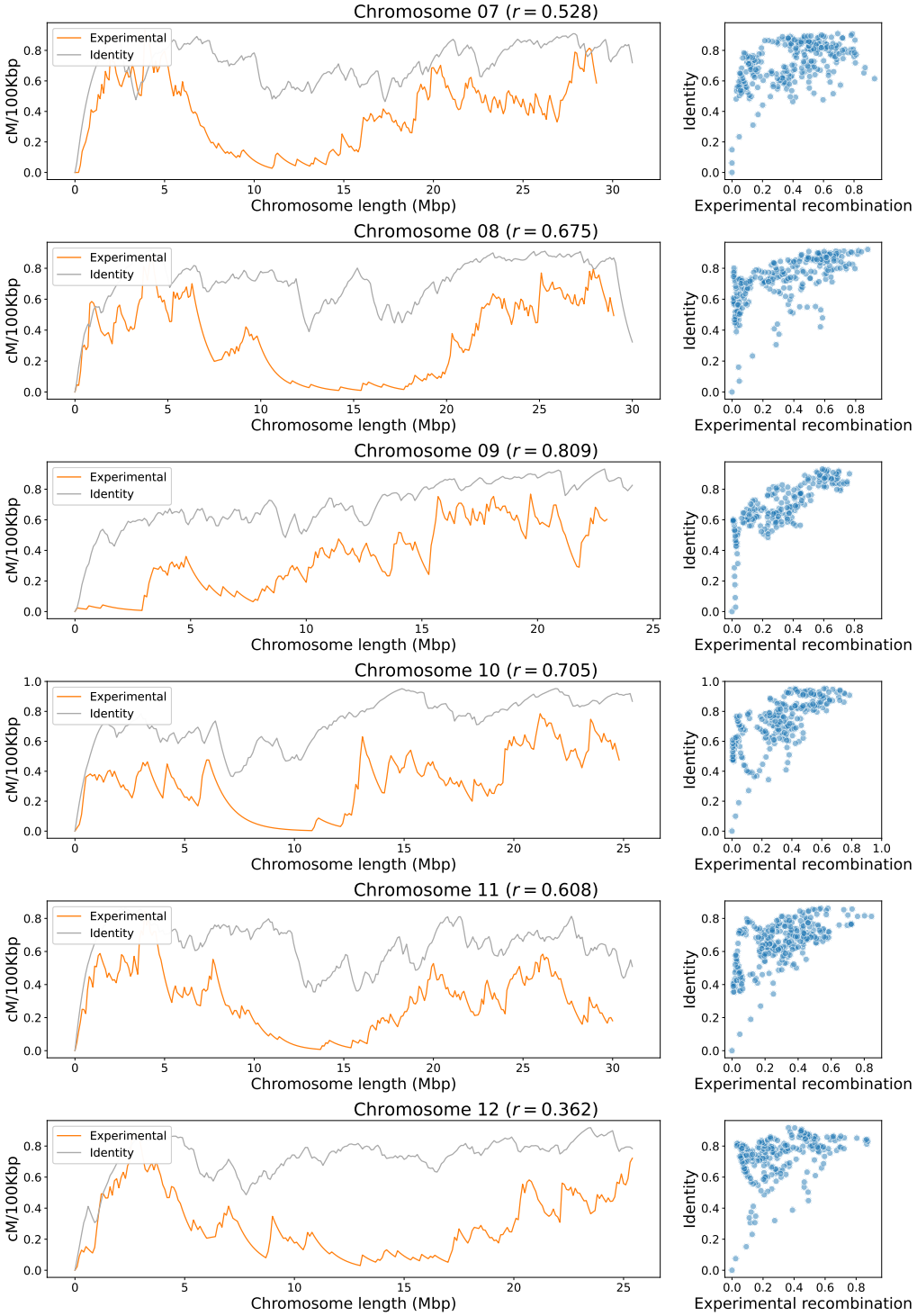


Fig 5. Identity correlation analysis for chromosomes 7 to 12. Landscape and correlation between chromosomal recombination and sequence identity for rice chromosomes 7 to 12 (cross IR64 x Azucena).

−0.009 and 0.362, being chromosome 5 the unique with near zero, negative correlation. This can be explained because the alignment of chromosome 5 between these two

varieties has a high identity in the centromere region, originating a trend opposite to that observed in other chromosomes, which usually report low identity values in centromeric regions.

Sequence identity by itself can reproduce some peaks and valleys of the recombination landscape, indicating that recombination is greatest in regions where identity between genomes is greatest and least where it is not. Thus, if genomic identity is highly correlated with chromosomal recombination, it can be used as a starting point for the construction of a model that aims to predict recombination. We thus developed a model based on sequence identity.

Rationale behind the model

In the first step of the model, three cases are defined to alter the identity of some windows, and to better fit valleys and peaks of real recombination using sequence information (recall Eq 3). The first case, the penalty stage, is compatible with the idea that regions with low identity recombine less. Therefore, a window with low identity value should be penalized, in contrast to a window with high identity values that should remain intact. This stage causes regions with predominant low identity values to form valleys, thus increasing the correlation with chromosomal recombination rates. Biologically, these adjustments model the fact that few recombination events are expected if there is no high genomic identity between parental chromosomal regions. This observation goes in accordance with the initial hypothesis of this study.

The second case, the reward stage, consists of rescuing windows with low identity values. The reason for doing this is that there could be alignment fragments with high (almost perfect) identity values, and with size smaller than the 100 Kpb window and having low variants proportion. Therefore, this case is useful to predict recombination peaks in regions with low or average identity.

The third case, the correction stage, is included in order to deal with windows with an over-adjustment in the alignment process; mainly, windows with high identity values that are not dealt with by the previous two cases. If there are absent bases in a window, it means that the data in the window is constructed from more than one contig. Furthermore, such a window contains few variants, probably because the information depends on multiple contigs that do not accurately represent the structure of the corresponding chromosomal region. For windows in which none of the three previous cases are applied, the initial identity values are assigned.

The second step of the model consists of zeroing the negative values resulting from the first step. This is necessary because, biologically, recombination rates are always positive.

The third step of the model tries to predict the boundaries of the centromeric region. It applies a weight function to correct the predictive values close to the centromere where recombination is expected to be lower than in the rest of the chromosome. CentO(AA) sequence reported by Lee et al. [24] was mapped on the reference and query chromosomes to predict their centromere positions. The weight function was applied to the data obtained from step two with the following aim: the predictive values furthest from the centromere are unchanged, while those close to the centromere are multiplied by values that increase linearly from zero (at the maximum CentO density point, Cent0max) to one at the bounds of a defined window around the Cent0max. The best range, according to the in-silico experimentation, for decreasing and increasing linear functions near the centromere is defined in 50 windows. In the case of Chromosome 9, a special centromere correction was proposed since its telomeric centromere has a Nucleolar Organizer Region on the short arm, which is known to block recombination [25,26]. This special case should be applied, in general, when the region

of the centromere is predicted to be within the first quarter of the chromosome (Figure 6).

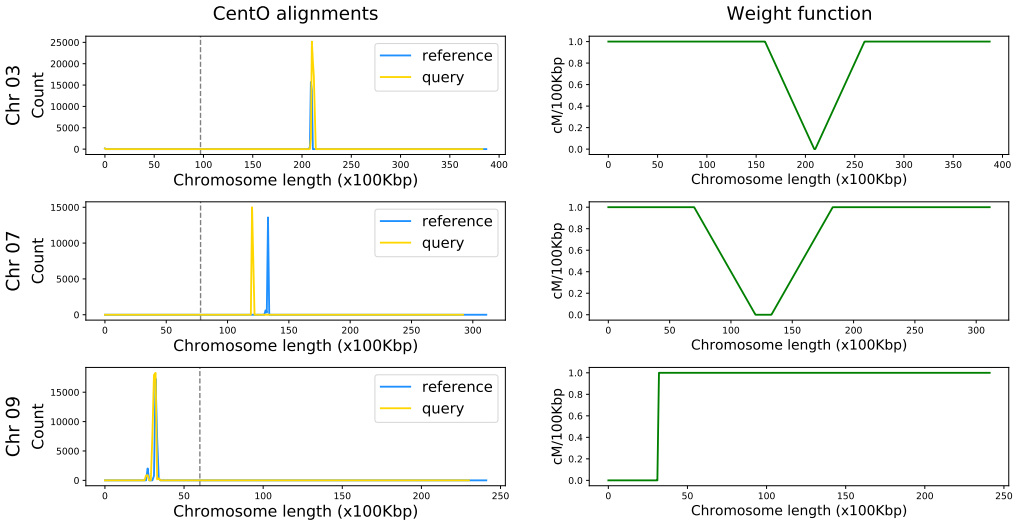


Fig 6. Centromere detection Centromere detection using CentO sequences and centromere correction distribution based on CentO. The vertical dotted line indicates a quarter of the length of the chromosome, the limit on which depends the weight function chosen for the correction of the centromere.

Finally, the fourth step, consisting of applying a special adaptation of exponential smoothing that replaces the value of the first window (i.e. telomeric regions) with zero, allows the prediction of the recombination rate to start at zero as actually occurs in the experimental data. Several smoothing factors were evaluated, with $\alpha = 0.1$ the one that mostly increases the prediction rate in the in-silico experimentation with the model.

Parameter optimization and model evaluation

The model was calibrated on each of the twelve chromosomes. Each calibration resulted in a different set of optimal parameters shown in Table 1.

Table 1. Parameters for each model calibration.

parameter	chr01	chr02	chr03	chr04	chr05	chr06	chr07	chr08	chr09	chr10	chr11	chr12
p1	0.529	0.578	0.568	0.563	0.470	0.469	0.508	0.488	0.476	0.467	0.380	0.504
p2	0.970	0.960	0.950	0.940	0.940	0.970	0.930	0.940	0.920	0.960	0.920	0.940
p3	1.000	0.000	0.102	1.000	0.000	0.998	1.000	0.135	0.000	0.000	0.000	1.000
p4	0.900	0.300	1.000	0.100	0.600	0.900	0.600	1.000	0.700	0.300	0.700	0.900
p5	1.000	1.000	0.500	0.665	1.000	1.000	0.700	0.100	0.500	1.000	1.000	0.536
p6	0.000	0.000	0.100	0.000	0.000	0.000	0.100	0.100	0.100	0.000	0.000	0.000
p7	0.002	0.002	0.001	0.004	0.002	0.002	0.005	0.004	0.001	0.002	0.005	0.003

The columns indicate the chromosome on which the model was calibrated and its corresponding set of optimum parameters.

The 12 model calibrations were used to test the prediction on the remaining eleven chromosomes. Fig 7 shows the distribution of the values r and R^2 obtained when evaluating the twelve predictions of each model calibration. The results look similar in all cases for both r and R^2. Furthermore, a two-sample Kolmogorov-Smirnov test, was performed between the evaluations of each pair of model calibrations. The test output indicated that the difference between the R^2 distributions is not statistically significant

(all p-values > 0.05). The same happens with the distributions of r (all p-values > 0.05). Therefore, the 12 distributions of R^2 can be considered equal to each other, as can the 12 distributions of r. This means that using the model calibrated on any arbitrarily chosen chromosome does not generate significant changes in the prediction performance. With this in mind and for practical reasons, some results discussed below are focused on the prediction obtained with the model calibrated on chromosome 1, which turns out to be the longest and therefore the one that provides the greatest amount of data for calibration.

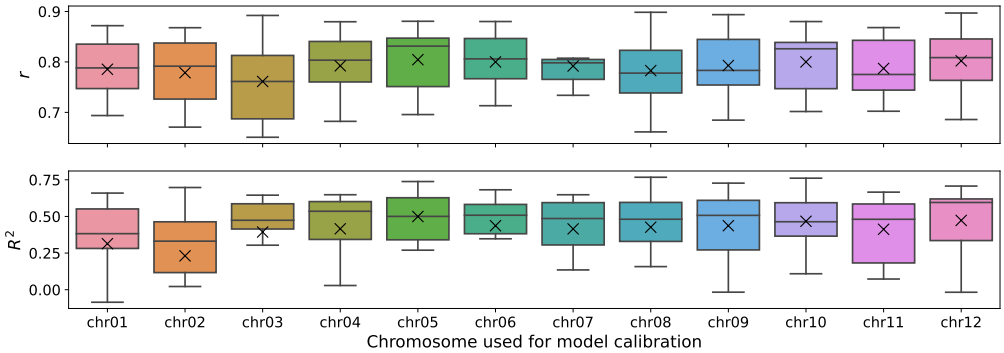


Fig 7. Boxplot distributions of model performance.
r and R^2 distributions for each model calibration evaluated in the 12 chromosomes.

Overall, for all 12 calibrations of the model, the predicted recombination have a correlation of $r = 0.8 \pm 0.012$ and a coefficient of determination $R^2 = 0.41 \pm 0.073$, which shows the power of the model to reproduce recombination trends along chromosomes. In terms of correlation, the lowest average value belongs to the model calibrated with chromosome 3 ($r = 0.761 \pm 0.081$). The lowest average coefficient of determination belongs to the model calibrated with chromosome 2 ($R^2 = 0.231 \pm 0.482$). While, the model calibrated with chromosome 5 has the highest average performance for both evaluation metrics: $r = 0.804 \pm 0.062$ and $R^2 = 0.5 \pm 0.157$.

In particular, the predictions of the model calibrated with chromosome 1 yields on $r = 0.785 \pm 0.06$ and $R^2 = 0.314 \pm 0.406$. It should be noted that the correlation on the calibrated chromosome ($r = 0.708$) is lowest than the correlations of the remaining predictions on the other 11 chromosomes ($r = 0.792 \pm 0.057$). The latter indicates that this model is not overfitted to the observed data, and is capable of predicting recombination rates of independent datasets, even achieving better performance.

Fig 8 and Fig 9 depicts, on the left, the landscape for the experimental recombination, identity, and model predictions. The shaded blue band on each chromosome represents the standard deviation of the predictions made with the 12 calibrated models. The width of these bands indicates that the predictions from any of the model calibrations are consistent across all chromosomes. Fig 8 and Fig 9 also depicts, on the right, the linear relationship between the experimental recombination and the prediction of the model calibrated with chromosome 1. The marker color in the scatter plot, and the bar color at the bottom of the line plots, represents the case of the model that was applied in a specific window.

It is important to analyze the incidence of the cases, from step 1 of the model, in the prediction of recombination. For all chromosomes, regardless of model calibration, the first case is the most applied in 67.2% of the chromosome windows on average, followed by the non-application of any case 26.2%. Meanwhile, the cases two and three are the least applied, with an average of 4.2% and 2.4% respectively. This indicates that the first case of step 1 is the one that contributes the most to the prediction of the model for all chromosomes, allowing the formation of medium and low recombination regions.

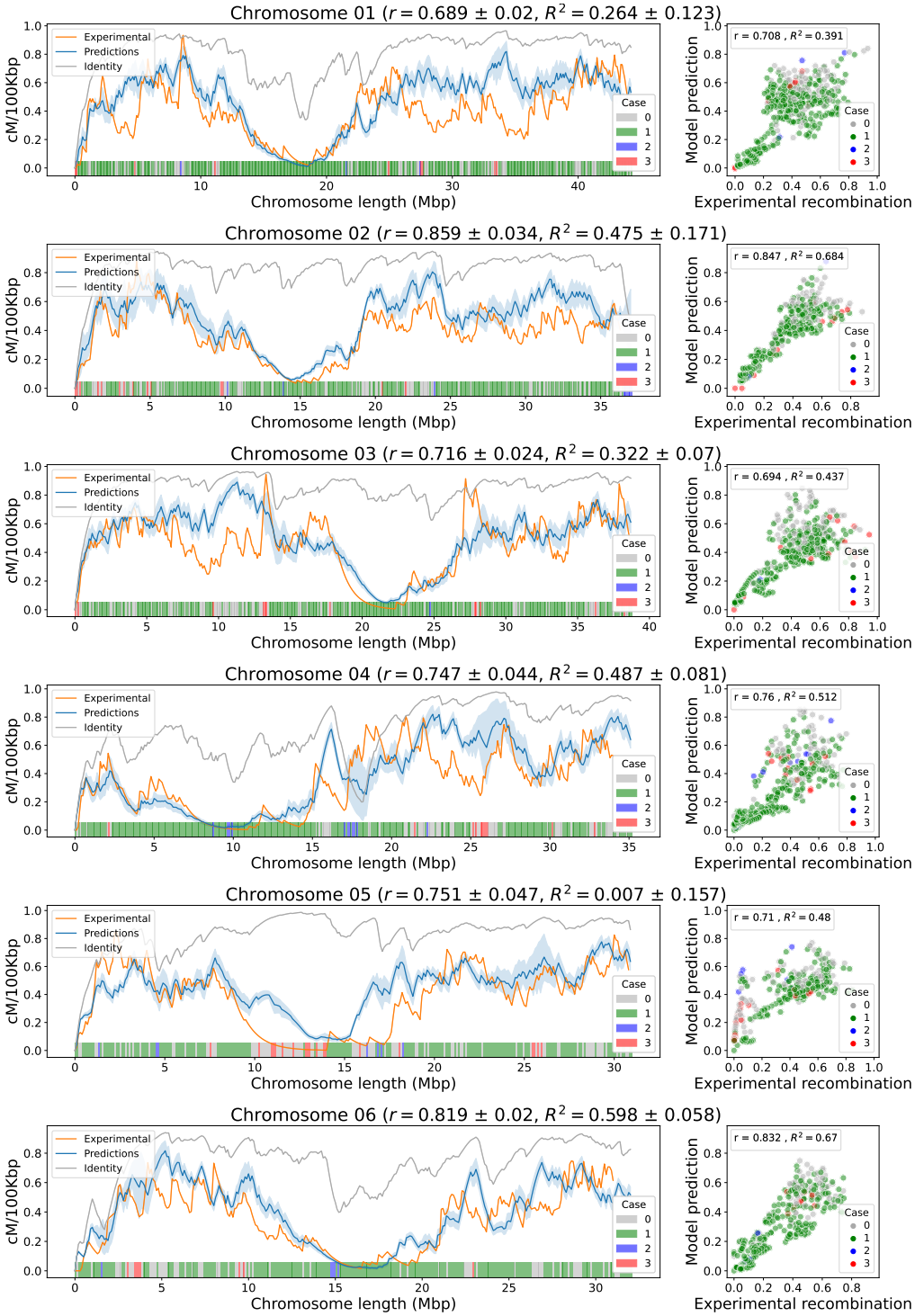


Fig 8. Model correlation analysis in chromosomes 1 to 6.
Landscape and correlation between chromosomal recombination and model prediction for rice chromosomes 1 to 6 (cross IR64 x Azucena).The identity criteria is included for comparative purposes only. The colored bars at the bottom of the landscapes indicate which case from the first step of the model is applied in each window.

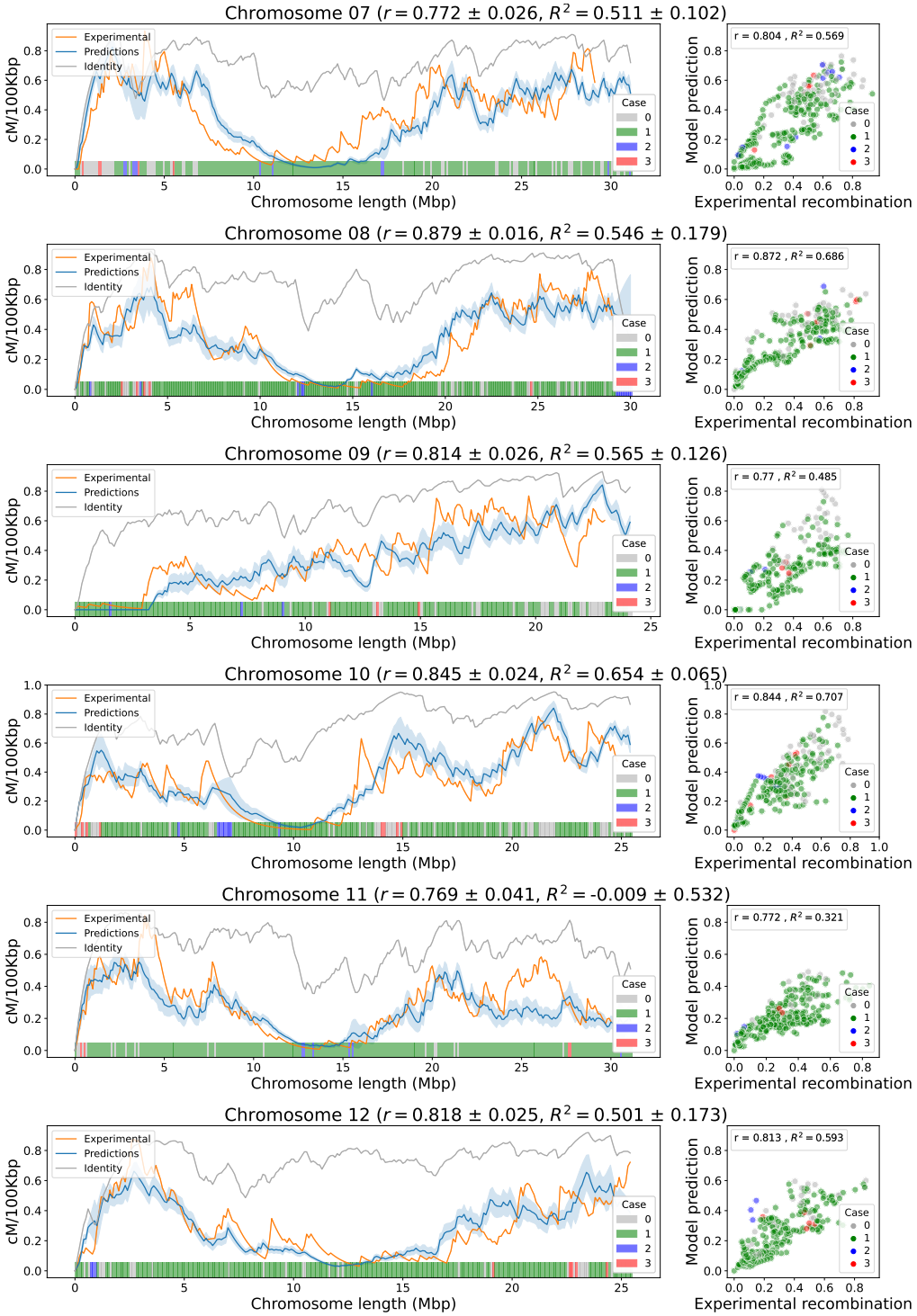


Fig 9. Model correlation analysis in chromosomes 7 to 12.
Landscape and correlation between chromosomal recombination and model prediction for rice chromosomes 7 to 12 (cross IR64 x Azucena).The identity criteria is included for comparative purposes only. The colored bars at the bottom of the landscapes indicate which case from the first step of the model is applied in each window.

Despite the fact that cases two and three have a low incidence in the chromosomal windows, they help to define particular areas that escape the action of the first case.

Note that, with respect to identity, the proposed model markedly increased the correlation and the coefficient of determination, as shown in Fig 10. The average increase in correlation, across all calibrations and tested chromosomes, is 0.256 ± 0.202, meanwhile the increase in the coefficient of determination is 8.98 ± 4.741, being the gain of prediction different for each chromosome. This gain is obtained because the different steps of the model transform the identity values of each 100 Kbp window, which helps to better represent peaks and valleys in the chromosomal arms and, in general, to define the centromeric regions. The chromosomes with the highest prediction gains are those whose identity in the centromeric region is greatest, with chromosome 5 being the most extreme case, gaining 0.760 correlation points with respect to identity. Other chromosomes such as 2, 3, and 12 gain approximately 0.37 correlation points, mainly because the model help define the low recombination rates around the centromere. The opposite case is observed in chromosome 9, where the average correlation gain is only 0.005. For this chromosome, the sequence identity is sufficient to describe recombination rates, even approaching the mean correlation achieved by the model.

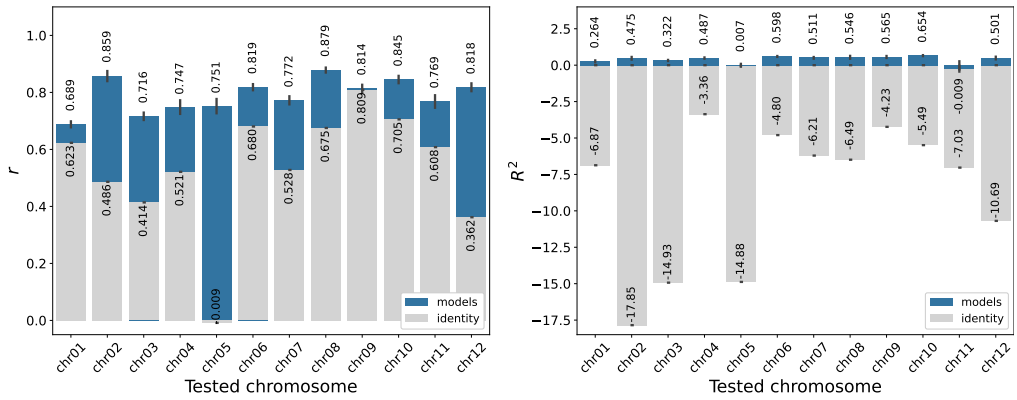


Fig 10. Gains in model performance versus identity. Correlation and coefficient of determination of identity criteria and model prediction with respect to recombination rates from 12 rice chromosomes (IR64 x Azucena cross).

Chromosome 9 is unique with its telomeric centromere in rice and is treated differently in the third step of the model, avoiding the centromere correction applied to the other chromosomes. This special treatment is due to the existence of the Nucleolar Organizer Region (NOR) in the short arm of the chromosome. The NOR of chromosome 9 is widely known to be a region where recombination is suppressed in rice [25], hence the special centromere correction. However, the effect of this correction on the chromosome 9 prediction is focused on the short arm only, and the prediction on the long arm is completely determined by the other steps of the model. Although sequence identity by itself can generate a high correlation with the recombination rate for this cross (IR64 x Azucena) on chromosome 9, the predictive values of the model continue to be preferred since the magnitude of the values is closer to those of recombination.

Finally, it should be noted that the model predictions reach a high correlation rate for all the chromosomes evaluated, being able to reproduce the recombination landscape of the crossing of the rice varieties IR64 and Azucena.

Conclusion

The results presented in this paper showed that the proposed criteria for sequence identity is strongly correlated with chromosomal recombination. The strength of this correlation allowed us to propose a model based on window "identities", which accurately predicts recombination rates along the length of the chromosome. The model is developed using data on the first chromosome of rice (accessions IR64 and Azucena). It is cross-validated using the remaining eleven chromosomes. Across all 12 chromosomes, an average correlation of about 80% between experimental and prediction rates is achieved. Similar results are found when model training is performed on other chromosomes, being of great importance the gain in the determination coefficient.

Application of this model could allow predicting chromosome recombination landscapes among rice varieties using only the parental genomes as a source. Such an approach is particularly useful for breeding purposes, for it offers the potential to optimize crossing experiments. In particular, model prediction could allow to identify varieties that should better recombine than others with recipient genomes, and to uncover recombination hot spots of vertical gene transfer. We hope that the proposed model will help breeders to reduce costs and execution times of crossing experiments. Finally, we hope to see other research studies extending the proposed methodology to other rice varieties, to other cereal species and even other plant and animal organisms.

Supporting information

S1 File. Experimental recombination. Experimental recombination values for the 12 rice chromosomes, Azucena x IR64 cross, in 100 kbp windows.

Acknowledgments

The authors thank to Nicolás López-Rozo for comments that greatly improved the manuscript.

References

1. Nicklas RB. Chromosome segregation mechanisms. Genetics. 1974;78(1):205–213.
2. de Haas LS, Koopmans R, Lelivelt CL, Ursem R, Dirks R, Velikkakam James G. Low-coverage resequencing detects meiotic recombination pattern and features in tomato RILs. DNA Research. 2017;24(6):549–558.
3. Adrion JR, Galloway JG, Kern AD. Predicting the landscape of recombination using deep learning. Molecular biology and evolution. 2020;37(6):1790–1808.
4. Si W, Yuan Y, Huang J, Zhang X, Zhang Y, Zhang Y, et al. Widely distributed hot and cold spots in meiotic recombination as shown by the sequencing of rice F2 plants. New Phytologist. 2015;206(4):1491–1502.
5. Choi K. Advances towards controlling meiotic recombination for plant breeding. Molecules and cells. 2017;40(11):814.
6. Gaut BS, Wright SI, Rizzon C, Dvorak J, Anderson LK. Recombination: an underappreciated factor in the evolution of plant genomes. Nature Reviews Genetics. 2007;8(1):77–84.

7. Butlin RK. Recombination and speciation. Molecular Ecology. 2005;14(9):2621–2635.

8. Liu G, Liu J, Cui X, Cai L. Sequence-dependent prediction of recombination hotspots in Saccharomyces cerevisiae. Journal of theoretical biology. 2012;293:49–54.

9. Brandariz SP, Bernardo R. Predicted genetic gains from targeted recombination in elite biparental maize populations. The plant genome. 2019;12(1):180062.

10. Wijnker E, de Jong H. Managing meiotic recombination in plant breeding. Trends in plant science. 2008;13(12):640–646.

11. Rodgers-Melnick E, Bradbury PJ, Elshire RJ, Glaubitz JC, Acharya CB, Mitchell SE, et al. Recombination in diverse maize is stable, predictable, and associated with genetic load. Proceedings of the National Academy of Sciences. 2015;112(12):3823–3828.

12. Colomé-Tatché M, Cortijo S, Wardenaar R, Morgado L, Lahouze B, Sarazin A, et al. Features of the Arabidopsis recombination landscape resulting from the combined loss of sequence variation and DNA methylation. Proceedings of the National Academy of Sciences. 2012;109(40):16240–16245.

13. Horton MW, Hancock AM, Huang YS, Toomajian C, Atwell S, Auton A, et al. Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. Nature genetics. 2012;44(2):212–216.

14. 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. Scientific reports. 2016;6(1):1–9.

15. Demirci S, Peters SA, de Ridder D, van Dijk AD. DNA sequence and shape are predictive for meiotic crossovers throughout the plant kingdom. The Plant Journal. 2018;95(4):686–699.

16. Cheng Z, Buell CR, Wing RA, Gu M, Jiang J. Toward a cytological characterization of the rice genome. Genome research. 2001;11(12):2133–2141.

17. Fragoso CA, Moreno M, Wang Z, Heffelfinger C, Arbelaez LJ, Aguirre JA, et al. Genetic architecture of a rice nested association mapping population. G3: Genes, Genomes, Genetics. 2017;7(6):1913–1926.

18. Zhou Y, Chebotarov D, Kudrna D, Llaca V, Lee S, Rajasekar S, et al. A platinum standard pan-genome resource that represents the population structure of Asian rice. Scientific data. 2020;7(1):1–11.

19. Lorieux M, Gkanogiannis A, Fragoso C, Rami JF. NOISYmputer: genotype imputation in bi-parental populations for noisy low-coverage next-generation sequencing data. bioRxiv. 2019; p. 658237.

20. Lorieux M. MapDisto: fast and efficient computation of genetic linkage maps. Molecular Breeding. 2012;30(2):1231–1235.

21. Heffelfinger C, Fragoso CA, Lorieux M. Constructing linkage maps in the genomics era with MapDisto 2.0. Bioinformatics. 2017;33(14):2224–2225.

22. Kosambi DD. The estimation of map distances from recombination values. In: DD Kosambi. Springer; 2016. p. 125–130.

23. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome biology. 2004;5(2):1–9.

24. Lee HR, Zhang W, Langdon T, Jin W, Yan H, Cheng Z, et al. Chromatin immunoprecipitation cloning reveals rapid evolutionary patterns of centromeric DNA in Oryza species. Proceedings of the National Academy of Sciences. 2005;102(33):11793–11798.

25. Wu J, Mizuno H, Hayashi-Tsugane M, Ito Y, Chiden Y, Fujisawa M, et al. Physical maps and recombination frequency of six rice chromosomes. The Plant Journal. 2003;36(5):720–730.

26. Mizuno H, Sasaki T, Matsumoto T. Characterization of internal structure of the nucleolar organizing region in rice (Oryza sativa L.). Cytogenetic and genome research. 2008;121(3-4):282–285.