Rice Recombination Predictor

Camila Riccio, Mauricio Peñuela, Camilo Rocha, and Jorge Finke

Pontificia Universidad Javeriana, Cali, Colombia
{camila.riccio,mauricio.penuela,camilo.rocha,jfinke}@javerianacali.edu.co

Abstract. 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. The sequence identity between the genomes of two rice varieties 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 rice chromosomes. The model consists of different steps that fit the original identity values using a series of parameters in 100 kbp windows. The model can be adjusted for any of the twelve chromosomes and obtain similar performance 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.

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* x *qry* pair, the model performs a comparison via an alignment process. Additionally, the CentO sequence is aligned with each of *ref* and *qry* to approximate the location of their centromeres. 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 location of centromeres (Fig 1).

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.

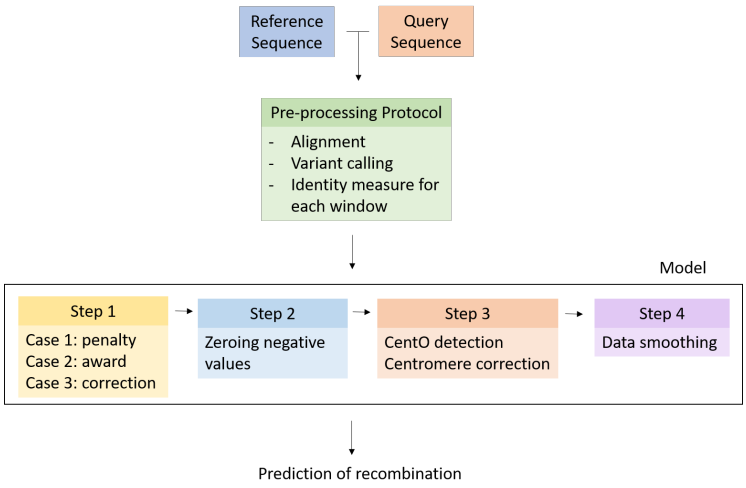

Fig. 1: **Model workflow.** Schematic representation of data preprocessing and model steps to predict recombination.

Let $W = \{1, 2, ..., n\}$ be the set representing the n windows partitioning a chromosome. The identity of each window w, denoted by $Id_0 : W \rightarrow [0, 1]$, 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 $V : W \rightarrow [0, 1]$, the number of bases in inversions $I : W \rightarrow [0, 1]$, and the absent bases $A : W \rightarrow [0, 1]$ are the features that can modify the identity criteria directly, and if these features are not present in a certain window, its identity value is set at its maximum value:

$$Id_0(w) = window\_size - V(w) - I(w) - A(w) \tag{1}$$

The identity is taken as a starting point to predict recombination. The model adapts those 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 involves a total of 7 parameters ($p_i \in [0, 1] \quad \forall i \in \{1, 2, ..., 7\}$) throughout these three cases, 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 cases two and three variants must be below the same threshold (p_7). Thus, an updated identity function $Id_1 : W \rightarrow \mathbb{R}$ is defined for each window $w \in W$ as:

$$Id_1(w) = \begin{cases} Id_0(w) - p_1 \ , \ Id_0(w) < p_2 \wedge V(w) > p_7 \\ Id_0(w) + p_3 \ , \ Id_0(w) < p_4 \wedge V(w) < p_7 \\ Id_0(w) - p_5 \ , \ A(w) > p_6 \wedge V(w) < p_7 \\ Id_0(w) \quad , \quad \text{otherwise} \end{cases} \tag{2}$$

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:

$$Id_2(w) = \max(0, Id_1(w)) \tag{3}$$

Step 3: Centromere correction

The alignments of the CentO sequence help approximating the location of chromosomes 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:

$$c_0 = \min(wcentO(ref) \cup wcentO(qry)) \tag{4}$$

$$c_1 = \max(wcentO(ref) \cup wcentO(qry)) \tag{5}$$

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:

$$f(w) = \begin{cases} 1 & , \ 0 \leq Id_2(w) \leq c_0 - 50 \\ \frac{-1}{50}(w - c_0) & , \ c_0 - 50 < Id_2(w) \leq c_0 \\ 0 & , \quad c_0 < Id_2(w) \leq c_1 \\ \frac{1}{50}(w - c_0) & , \ c_1 < Id_2(w) \leq c_1 + 50 \\ 1 & , \ c_1 + 50 < Id_2(w) < n \end{cases} \tag{6}$$

$$g(w) = \begin{cases} 0 \ 0 \leq Id_2(w) < c_1 \\ 1 \ c_1 \leq Id_2(w) \leq n \end{cases} \tag{7}$$

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

$$Id_3(w) = \begin{cases} Id_2(w) \cdot f(w) \ c_1 > n/4 \\ Id_2(w) \cdot g(w) \ \text{otherwise} \end{cases} \tag{8}$$

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:

$$Id_4(w) = \begin{cases} 0 & w = 0 \\ \alpha Id_3(w) + (1-\alpha) Id_4(w) & w > 0 \end{cases} \tag{9}$$

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:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{10}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{11}$$

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 12) of a single chromosome. The parameter optimization is carried out through a gradient-based algorithm called 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.

$$X_s(w) = \begin{cases} X(w) & w = 0 \\ \alpha X(w) + (1-\alpha) X_s(w) & w > 0 \end{cases} \tag{12}$$

Case Study: IR64 X Azucena

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 [1].

Whole Genome Sequencing

Leaf tissue from parent plants and F12 lines were collected, and DNA was extracted following a protocol similar to [1]. Platinum-grade PacBio assemblies of the parental genomes were obtained at the Arizona Genomics Institute (AGI, Tucson, Arizona) [8]. 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 [6]. 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 [5,2], using the Kosambi mapping function to convert recombination fractions into centimorgans (cM) [3].

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

The alignment process is performed using MUMmer3 [4]. MUMmer3 is an open source software package for the rapid alignment of very large DNA and amino acid sequences. The latest version, release 3.0, includes a new suffix tree algorithm that has further improved the efficiency of the package and has been integral to making MUMmer an open source product. The initial alignment for each pair of parental chromosomes is done with the `nucmer` command with the 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`, and is also used to extract the variants 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. These measurements are used to calculate the identity and finally apply the predictive model.

0.1 Prediction

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 2 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.

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

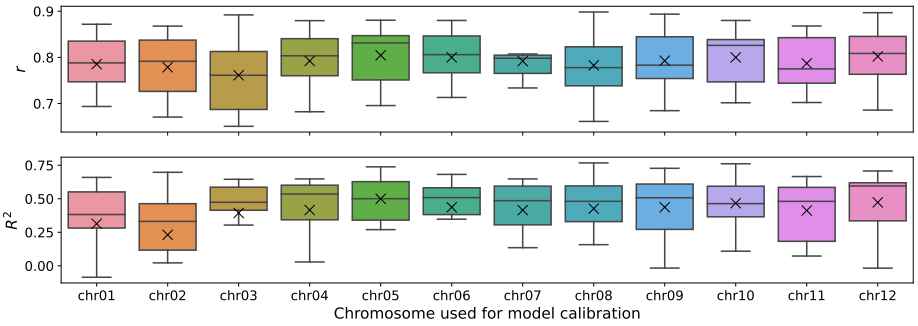

Fig. 2: **Boxplot distributions of model performance.**
r and R^2 distributions for each model calibration evaluated in the 12 chromosomes.

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 3 and Fig 4 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 3 and Fig 4 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. 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 5. 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.

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 [7], 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.

A model implementation is publicly available:

- Project name: Rice recombination predictor
- Project home page: `https://github.com/criccio35/Rice-recombination-predictor`
- Operating system(s): platform independent.
- Programming language: Python 3.
- Other requirements: ncbi-blast+.
- License: GNU GPL v3.

References

1. Fragoso, C.A., Moreno, M., Wang, Z., Heffelfinger, C., Arbelaez, L.J., Aguirre, J.A., Franco, N., Romero, L.E., Labadie, K., Zhao, H., et al.: Genetic architecture of a rice nested association mapping population. G3: Genes, Genomes, Genetics **7**(6), 1913–1926 (2017)
2. Heffelfinger, C., Fragoso, C.A., Lorieux, M.: Constructing linkage maps in the genomics era with mapdisto 2.0. Bioinformatics **33**(14), 2224–2225 (2017)
3. Kosambi, D.D.: The estimation of map distances from recombination values. In: DD Kosambi, pp. 125–130. Springer (2016)

4. Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S.L.: Versatile and open software for comparing large genomes. Genome biology **5**(2), 1–9 (2004)
5. Lorieux, M.: Mapdisto: fast and efficient computation of genetic linkage maps. Molecular Breeding **30**(2), 1231–1235 (2012)
6. Lorieux, M., Gkanogiannis, A., Fragoso, C., Rami, J.F.: Noisymputer: genotype imputation in bi-parental populations for noisy low-coverage next-generation sequencing data. bioRxiv p. 658237 (2019)
7. Wu, J., Mizuno, H., Hayashi-Tsugane, M., Ito, Y., Chiden, Y., Fujisawa, M., Katagiri, S., Saji, S., Yoshiki, S., Karasawa, W., et al.: Physical maps and recombination frequency of six rice chromosomes. The Plant Journal **36**(5), 720–730 (2003)
8. Zhou, Y., Chebotarov, D., Kudrna, D., Llaca, V., Lee, S., Rajasekar, S., Mohammed, N., Al-Bader, N., Sobel-Sorenson, C., Parakkal, P., et al.: A platinum standard pan-genome resource that represents the population structure of asian rice. Scientific data **7**(1), 1–11 (2020)

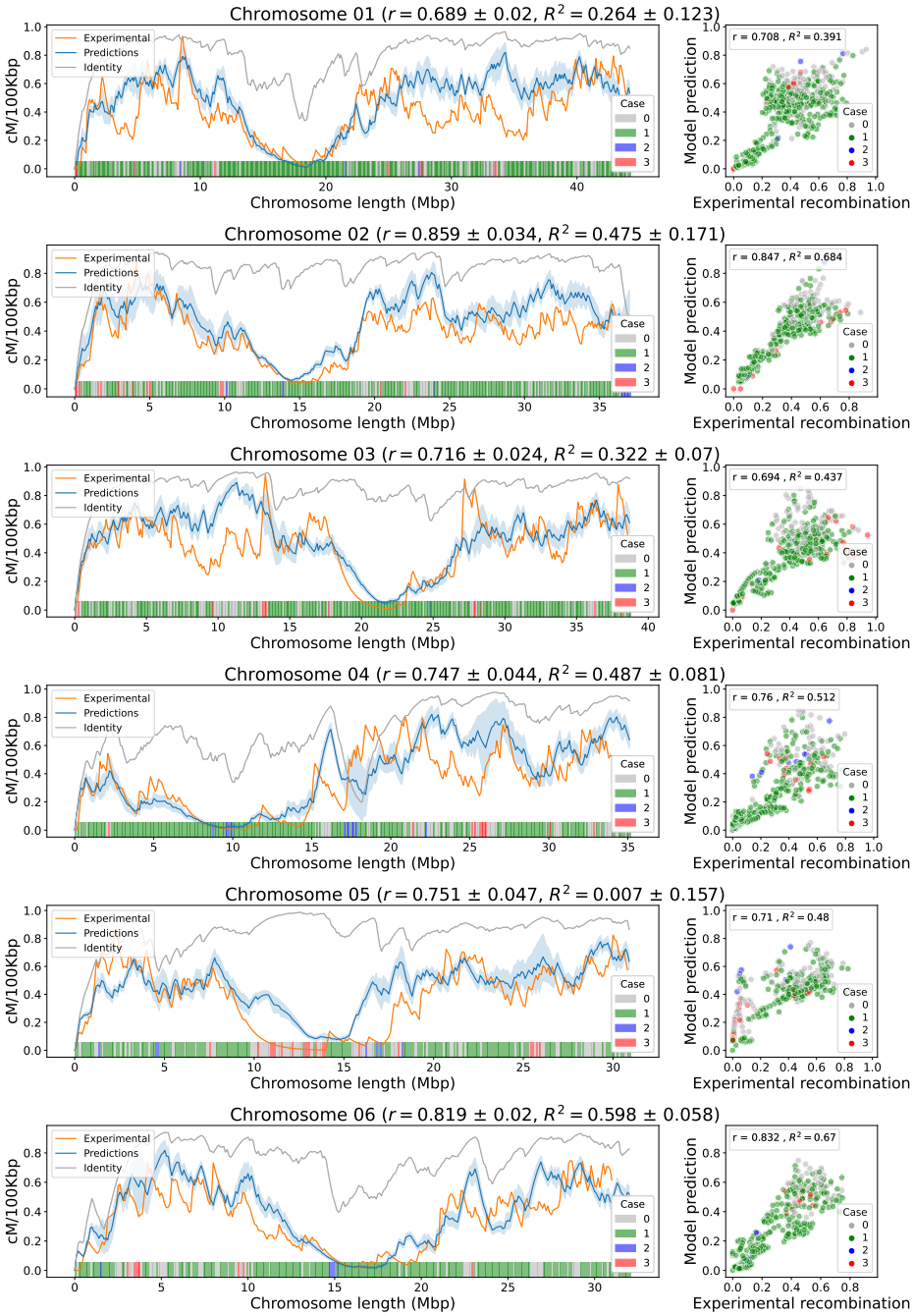

Fig. 3: **Model correlation analysis in chromosomes 1 to 6.** Landscape and correlation between chromosomal recombination and model prediction for rice chromosomes 1 to 6 (cross IR64 x Azucena).The identity criteria is included for comparative purposes only. The colored bars at the bottom of the landscapes indicate which case from the first step of the model is applied in each window.

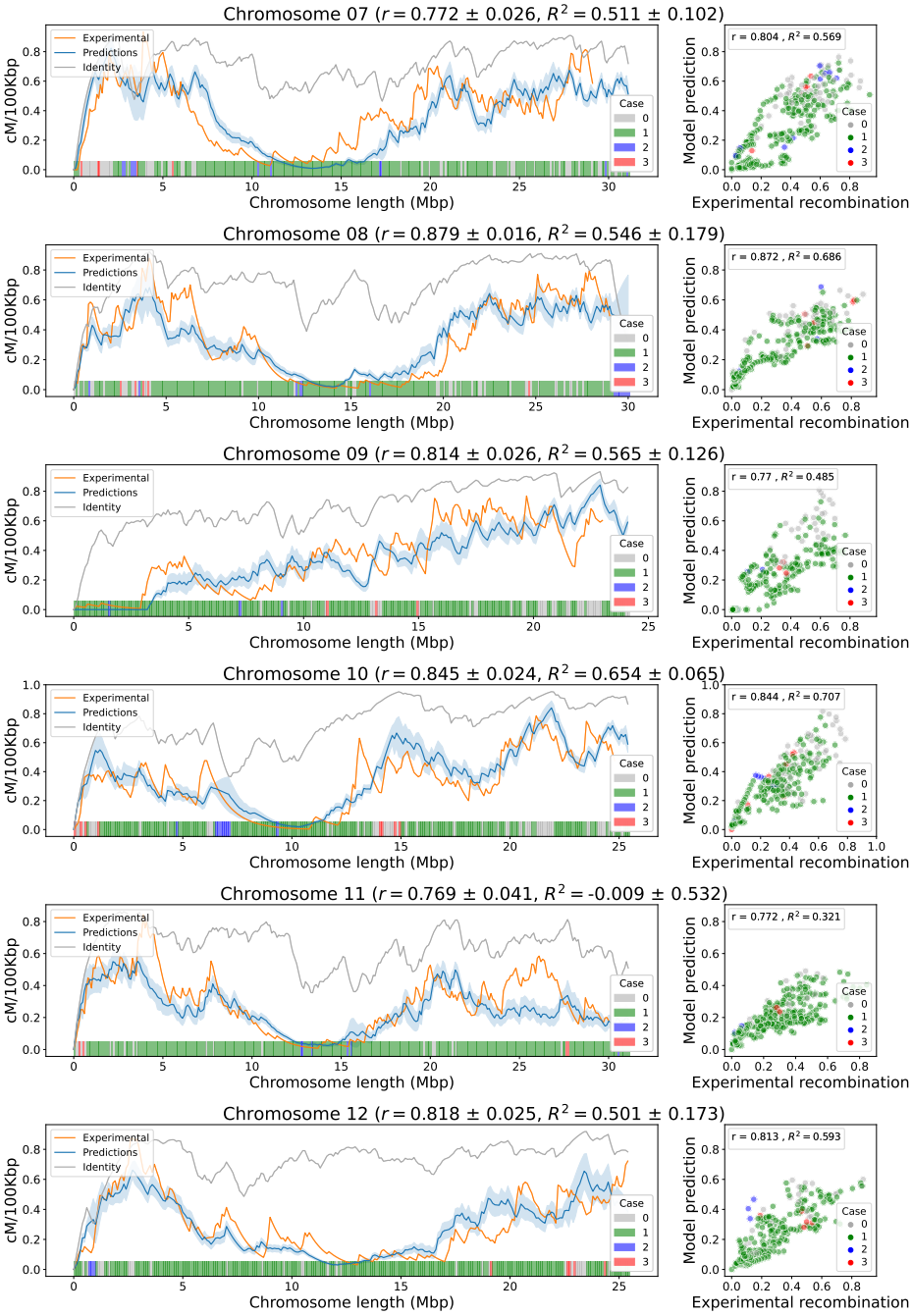

Fig. 4: **Model correlation analysis in chromosomes 7 to 12.** Landscape and correlation between chromosomal recombination and model prediction for rice chromosomes 7 to 12 (cross IR64 x Azucena).The identity criteria is included for comparative purposes only. The colored bars at the bottom of the landscapes indicate which case from the first step of the model is applied in each window.

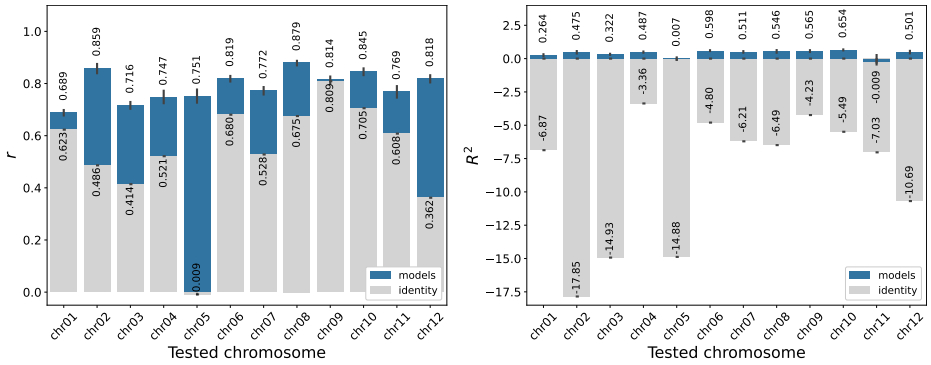

Fig. 5: **Gains in model performance versus identity.** Correlation and coefficient of determination of identity criteria and model prediction with respect to recombination rates from 12 rice chromosomes (IR64 x Azucena cross).