Can Genome Similarity Predicts Recombination?

ABSTRACT

KEYWORDS

INTRODUCTION

MATERIALS AND METHODS

Plant Material

Rice varieties IR64 and Azucena from *Oryza sativa* L. were crossed to generate an F1 generation. From F2, five self-pollination cycles continued to generate 212 RILs. All crossing cycles were conducted under standard greenhouse conditions at International Center for Tropical Agriculture (CIAT) in Palmira-Colombia.

Whole Genome Sequencing

Leaf tissue from parent plants and RILs were collected and DNA was extracted following the protocol of NNN(XXX). For parent plants (IR64 and Azucena) a PacBio Whole Genome Sequencing with a deep of 120x was performed, while for RILs the sequencing deep was 1x. IR64 and Azucena genomes used are available on GenBank with the accession numbers XXXXX and XXXXX. For increasing sequencing deep in RILs and gain accuracy, an Illumina sequencing was performed in XXX RILs following the RADseq protocol from NNN(XXX) with the enzymes XXX.

Data imputation and recombination values

RILs genomes segments were categorized according to its parental origin using NOISYmputer Software (Lorieux *et al.* 2019). The genotyping data by chromosome obtained consisted in a matrix of markers (sequence position) versus individuals, filled with A or B depending on the parental origin of the corresponding sequence. To improve recombination estimation, we include a series of markers with the sequence positions for each 100 Kbp inside the matrix, and we imputed data from the nearest marker. These matrices were analyzed in MapDisto (Lorieux 2012) to extract positions of each marker in Centimorgans (cM). Only mapping position from 100Kbp series were retained and were used to estimate recombination in cM/Mbp along each chromosome.

Protocol development

With the hypothesis in mind that similar genome regions present more recombination events, we design a python script called *BasebyBase* which generates a data-frame that compare every nucleotide from parental chromosomes and identify base similarities. In detail, *BasebyBase* consists in an initial alignment for each pair of parental chromosomes using the `nucmer` command from MUMmer 3 (Kurtz *et al.* 2004) with default parameters. The resulting delta file is filtered using the command `delta-filter -r -q`. Then, resulting filter file is used to extract coordinates using the command `show-coords -r`. Sequence variants as SNPs and Indels are extracted from the initial delta file using the command `show-snps`. Following, *BasebyBase* develop a data-frame using the reference sequence as spine for which nucleotides are numbered consecutively and uses the coordinates file to match query bases and positions in neighboring columns, in the process the variants file is used to mark positions where a SNP or Indel occurs allowing a correct match based on MUMmer alignment. Other columns complement this dataframe by helping to indicate whether each base is part of an inversion, its possible variant type, and whether it is identical in both sequences. A basic scheme of the resulting dataframe generated by *BasebyBase* can be observed in Figure 1.

To analyze this information in a manageable way, we develop a second python script called *StatsbyWindow* which allows users to extract parameters as GC content from reference and query sequences, number of variants, number of bases in inversions, number of absent bases in the query sequence, percentage of identity among sequences, and the respective variance of each parameter inside each window. *StatsbyWindow* allows users to adjust the desired window size for extracting these parameters and compile information in a csv file. Both *BasebyBase* and *StatsbyWindow* scripts are available in supplementary material for its free use.

Model approach

For chromosome 1, correlations analyses were made between recombination values and the parameters extracted by *StatsbyWindow* for a window size of 100Kbp. The recombination values were also categorized into quartiles and analyses of variance were performed for their corresponding sequence parameters per window. The predictive recombination model was built based on this information, using a reward and penalty system according to the values of the sequence parameters for each window.

Recombination prediction

For the chromosome 1, correlation analyzes were performed between the model's predicted recombination and recombination values, using various levels of exponential smoothing to determine which might be the best alternative to choose. Chosen level of smoothing, the model was used to predict the recombination values per window in the remaining chromosomes, and its fit was evaluated with correlation values.

RESULTS

DISCUSSION

REFERENCES

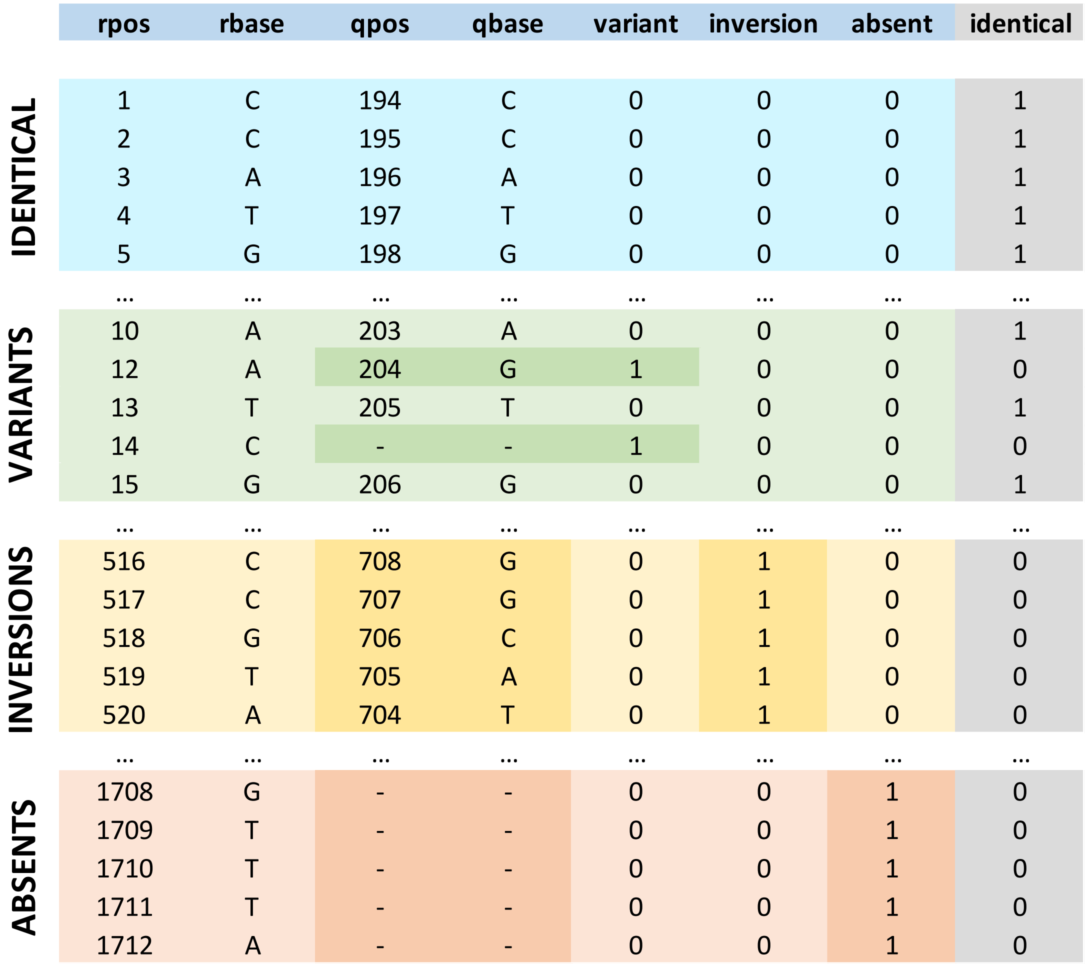
Lorieux, M. 2012. MapDisto: fast and efficient computation of genetic linkage maps. Mol Breeding 30: 1231-1235.

Lorieux, M. Gkanogiannis, A. Fragoso, C. Rami, J-F. 2019. NOISYmputer: genome imputation in bi-parental populations for noisy low-coverage next-generation sequencing data. BioRxiv. doi.org/10.1101/658237

Kurtz, S. Phillippy, A. Delcher, A. L. Smoot, M. Shumway, M. Antonescu, C. Salzberg, S. L. 2004. Versatile and open software for comparing large genomes. Genome Biology, 5: R12.

ACKNOWLEDGMENTS

TABLES AND FIGURES



	rpos	rbase	qpos	qbase	variant	inversion	absent	identical
IDENTICAL	1	C	194	C	0	0	0	1
	2	C	195	C	0	0	0	1
	3	A	196	A	0	0	0	1
	4	T	197	T	0	0	0	1
	5	G	198	G	0	0	0	1
	...	...	...	...	...	...	...	...
VARIANTS	10	A	203	A	0	0	0	1
	12	A	204	G	1	0	0	0
	13	T	205	T	0	0	0	1
	14	C	-	-	1	0	0	0
	15	G	206	G	0	0	0	1
	...	...	...	...	...	...	...	...
INVERSIONS	516	C	708	G	0	1	0	0
	517	C	707	G	0	1	0	0
	518	G	706	C	0	1	0	0
	519	T	705	A	0	1	0	0
	520	A	704	T	0	1	0	0
	...	...	...	...	...	...	...	...
ABSENTS	1708	G	-	-	0	0	1	0
	1709	T	-	-	0	0	1	0
	1710	T	-	-	0	0	1	0
	1711	T	-	-	0	0	1	0
	1712	A	-	-	0	0	1	0

Figure 1. Basic schema of the resulting data-frame generated by *BasebyBase* showing the four possible alignment cases (horizontal color blocks) and how information is registered across columns according to the comparison among **r**eference and **q**uery bases. Intense colors signalize differences among reference and query sequences.