DNA sequence and shape are predictive for meiotic crossovers throughout the plant kingdom

Demirci, S., Peters, S. A., de Ridder, D., & van Dijk, A. D. J.

DNA sequence and shape are predictive for meiotic crossovers throughout the plant kingdom

Sevgin Demirci[1,2], Sander A. Peters[1], Dick de Ridder[2] and Aalt D.J. van Dijk[1,2,3*]

1 Business Unit Bioscience, Cluster Applied Bioinformatics, Wageningen University and Research, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands
2 Bioinformatics Group, Wageningen University and Research, Wageningen, The Netherlands
3 Biometris, Wageningen University and Research, Wageningen, The Netherlands
* To whom correspondence should be addressed: aaltjan.vandijk@wur.nl

Running head

Genomic features underlying meiotic crossovers

Keywords

Meiotic recombination; crossover; machine learning; prediction; genome accessibility; DNA shape; tomato; *Arabidopsis thaliana*; maize; rice

Summary

A better understanding of genomic features influencing the location of meiotic crossovers (COs) in plant species is both of fundamental importance and of practical relevance for plant breeding. Using CO positions with sufficiently high resolution from four plant species (*Arabidopsis thaliana*, tomato, maize, and rice) we have trained machine learning models to predict the susceptibility to CO formation. Our results show that CO occurrence within various plant genomes can be predicted by DNA sequence and shape features. Several features related to genome content (including LTR content) and to genomic accessibility were consistently positively or negatively related to COs in all four species. Other features were found as predictive only in specific species. Gene-annotation related features were especially predictive for maize, whereas in tomato and Arabidopsis propeller twist and helical twist (DNA shape features) and AT/TA dinucleotides were found as most important. In rice, high roll (another DNA shape feature) and low CA dinucleotide frequency in particular were found associated with CO occurrence. The accuracy of our models was sufficient for Arabidopsis and rice (AUROC > 0.5) and high for tomato and maize (AUROC>>0.5), demonstrating that DNA sequence and shape are predictive for meiotic crossovers throughout the plant kingdom.

Introduction

Meiosis is essential in most reproducing organisms to halve the number of chromosomes, in order to enable the restoration of ploidy levels during fertilization (Villeneuve and Hillers, 2001). At the first meiotic division, homologous chromosomes (homologs) are segregated. In most eukaryotes, accurate homolog segregation is ensured by the formation of at least one recombination event or crossover (CO) between the chromatids of homologs. COs represent a reciprocal exchange of genetic information between homologs (Mercier et al., 2015). In this way, meiotic crossover increases genetic diversity in a population of sexually reproducing eukaryotes. Understanding the genomic features influencing the location of COs is of fundamental importance for many areas of biology, ranging from chromosome evolution to population genetics. Knowledge of the location of COs is also key to plant breeding, as breeders are interested in manipulating COs either to introduce favorable genes from wild relatives to crops or to silence COs in order to generate stable genetic lines of successful crops (Wijnker and de Jong, 2008). However, there are still numerous gaps of knowledge with respect to meiotic CO and its genetic determinants in plants.

The mechanism leading to meiotic crossovers starts with the formation of double strand breaks (DSBs) at various chromosomal locations. The DSB distribution deviates from uniform in many species including mammals, birds, and plants (Lichten and Goldman, 1995; Kauppi *et al.*, 2004, Edlinger & Schlögelhofer 2011, He et al, 2017; Choi et al., 2018). If DSBs are not repaired immediately by DNA repair mechanisms, specific proteins (for example Rad51/Dmc1 in *A. thaliana*; Edlinger & Schlögelhofer 2011) guide one of the loose ends of the DSB to its homologous nonsister chromatid to form a double Holliday junction. Depending on how the junction is resolved, the resulting chromatids can have a non-crossover (for example, a gene conversion) or a crossover. In Arabidopsis, ~4% of the initial DSBs result in COs (Mercier *et al.*, 2015). COs are formed through two pathways, ZMM-dependent interfering (Class I) and ZMM-independent non-interfering (Class II) pathways. Class I COs are inhibited from occurring near other class I COs, while class II COs are unconstrained by the presence of adjacent class II COs; between class I and class II weak interference has been reported (Anderson et al., 2014; Mercier et al., 2015). In the current study we focus on the location of any resulting COs without discriminating between Class I or Class II COs.

It is an intriguing question how conserved or variable the mechanisms underlying CO formation be in various plant species. For example, variation exists in the mechanisms underlying DSB formation in different plant species (Lambing et al., 2017). Also, some proteins involved in CO formation have opposing roles in various species. One example is that down-regulation of ZYP1/ZEP1 leads to fewer COs in Arabidopsis yet to more COs in rice (Lambing et al., 2017). However, a general picture on conservation of determinants of CO formation in various plants is still lacking.

The location of COs is known to be correlated with several genomic features. In many plant species like tomato, maize, Arabidopsis and rice, COs are observed in euchromatic regions where genes are accumulated and are depleted in pericentromeric regions (Wu et al., 2003; Sato et al., 2012; Gao et al., 2013; Choi et al., 2013; Wijnker et al., 2013; Rodgers-Melnick et al., 2015). More specifically, COs occur preferentially upstream of transcription start sites (TSS) i.e. in gene promoters in tomato and Arabidopsis (Wijnker et al., 2013; Choi et al., 2013; Demirci et al., 2017; de Haas et al., 2017). In addition to their preferential occurrence in promoters, CO regions are also rich in particular sequence motifs, including for example, poly-A sequence motifs in Arabidopsis and tomato (Demirci et al., 2017; Choi et al., 2013; Wijnker et al., 2013). In maize, GC sequences are overrepresented in recombination regions (Rodgers-Melnick et al., 2015). Moreover, *Mu* retrotransposon insertion site frequencies are correlated with recombination in maize (Liu et al., 2009). Finally, DNA methylation was recently shown to be involved in CO silencing in Arabidopsis (Yelina et al., 2015). In this study, we will focus on genomic features rather than epigenetic factors.

To learn about genomic features correlated with CO formation in different plants, we take a predictive machine learning approach. There have been some previous attempts to predict recombination rate and CO position. In particular, Rodgers-Melnick et al. (2015) used several genomic and epigenetic features to construct a model to predict crossover density in maize at the megabase scale. Machine learning models were successfully used to predict meiotic recombination in yeast based on sequences only (Liu et al., 2012). A consistent, simultaneous analysis of multiple plant species in order to compare the genomic determinants of COs is however lacking. In this study, we apply machine learning to CO datasets from four different plant species in order to (i) develop predictive models for the occurrence of COs and (ii) to learn about relevant and important features in these species. This allows to gain insight into determinants of CO formation throughout the plant kingdom.

Results & Discussion

Comparison of the genomic features correlated with formation of COs requires a consistent analysis of multiple plant species. To this end, we pursued a machine learning approach, training computational models using available CO datasets obtained in populations derived from crossing parental lines. We specifically focused on high resolution (less than 2 kb long) crossover regions. The COs were identified from either recombinant inbred lines, tetrads or double haploid lines. Such data were available for tomato (cross between *S. lycopersicum* and *S. pimpinellifolium*), *Arabidopsis thaliana* (cross between Cvi x Ler and Col accessions, and between Col and Ler), maize (cross between SK and Zheng58 accessions) and rice (cross between PA64s and 93-11). Some general characteristics of the genomes of these four species are given in Table S1, and a more extensive description of the CO datasets is given in the Methods. Plots comparing transposable element density, gene density, SNP density and CO distribution for the four species are provided in Figure 1.

We first developed our predictive model on CO data obtained in tomato. Subsequently, we trained similar models for Arabidopsis, maize and rice. We used the crossover regions together with their flanking sequences, extending each region to a total length of 4 kb. In these regions, we analysed features based on sequence information, genome annotation and parental genome sequences. We used these features to construct classification (i.e. machine learning) models that predict the probability of meiotic recombination for a given sequence. After training such a model with a set of known CO regions, it can be applied to predict likely CO sites throughout the genome. More importantly, we can analyze how the model learned to perform these predictions: to what extent, and in what direction is the probability of CO occurrence influenced by the different features, according to the model? In other words, this allows to learn about genomic features related with CO frequency in different plants.

CO region prediction in tomato genome

As input for training a machine learning model both a positive set (regions containing COs) and a negative set (regions not containing COs) are needed. We prepared a positive set consisting of 4kb-long CO regions from tomato (n = 664) obtained in our previous study (Demirci et al., 2017). Because absence of a CO in a given region does not automatically imply that a CO could not occur, generating a negative set is not straightforward. Therefore we used a random set instead of a negative set. As a first strategy to generate a random set, we simply sampled the same number (n = 664) of 4kb-long regions randomly from the tomato genome, excluding the 664 CO regions.

Each positive and each random sample was represented by 62 features based on sequence, genome annotation and parental genome sequence variation. Sequence-based features included dinucleotide frequencies and DNA shape features: minor groove width, propeller twist, helical twist, and roll. Propeller twist describes how one base in a base-pair is rotated about the long axis of the base-pair relative to the other base; helix twist is the angle between two adjacent base-pairs as they twist in a

DNA helix structure; and roll is the angle between two consecutive base-pairs rolling over each other (Chiu et al., 2016). These DNA shape features were predicted using a model trained on experimental DNA structures (Methods). The values predicted for each nucleotide were averaged over a (positive or random) region to obtain a single value for each region. DNA shape features have recently been shown to be helpful for predicting e.g. binding of proteins to DNA, where these features showed improved performance compared to using more simple representations of the DNA sequence (Mathelier et al., 2016). Genome annotation features described repeat elements, gene elements and (eu)chromatin state. The latter was defined as described previously (Demirci et al., 2017) based on results from cytogenetic analyses of pachytene chromosomes, which display long continuous stretches of less condensed euchromatin in chromosome arms flanked by highly condensed heterochromatin at the telomere ends and centromeres. Finally, features based on parental genome information included SNPs and INDELs between the parental genomes. Additional information on the exact definition of these features is provided in the Experimental Procedures. To assess the discriminative power of features, we initially applied a *t*-test to compare the means of each individual feature in the positive and in the random set. This indicated that 43 out of 62 features were significantly discriminative (Table S2).

The *t*-test analyzed whether each single feature on its own displayed different values in the CO set compared to the random set. To investigate the discriminative power of features when they are combined, we constructed a classification model which uses all features together. Different types of classifiers were tested to find the best performing model. In particular, we trained a decision tree, a random forest and a logistic regression classifier. The performances of the prediction models are visualized using receiver operator characteristic (ROC) curves in Figure 2a. The random forest classifier was the best performing model, in terms of the area under the ROC curve (AUROC = 0.92). Note that performance is calculated using regions not used for training the model, in order to prevent over-optimistic performance estimates.

We subsequently analyzed the importance of each feature according to the random forest model (Figure 2b, Table S3). This revealed that whether a region is in euchromatin or not is the most contributing feature (with a positive association, i.e. a region in euchromatin is more likely to be a CO region); euchromatin is defined here as in (Demirci et al., 2017). Additional important features included DNA shape features (positive or negative association, depending on the feature), LTR repeat elements (negatively associated, i.e. a region containing LTR repeats is less likely to be a CO region) and the length of INDELs between the two parental genomes (positively associated). The strong contribution of euchromatin presence to CO prediction fits expectation, since CO regions are known to accumulate in euchromatic regions (Sherman and Stack, 1995; Demirci et al., 2017). However, the strong contribution of euchromatin in our model may overshadow the effect of other features: effectively, the model has learned to discriminate between euchromatin and heterochromatin. In order to find the most relevant features for CO prediction *within* euchromatin, we

subsequently followed a second strategy to generate a random set, focussing on the euchromatic regions of the tomato genome.

CO region prediction in euchromatic regions of tomato genome

To focus on the prediction of CO regions inside tomato euchromatin, we generated an alternative random set: instead of sampling from the whole genome, the regions were sampled randomly from euchromatic regions only. With this new random dataset and the same positive dataset as above, we again constructed three predictive models using a decision tree, a random forest and logistic regression. Similar to the results obtained with the first random set, the best performing classifier was the random forest classifier, although performance decreased slightly (AUROC = 0.86, Figure 2c) reflecting an increased difficulty of the prediction problem. As indicated by the AUROC, we could clearly discriminate CO regions from randomly chosen regions in euchromatin. Compared to the results obtained above, the order of most contributing features changed drastically (compare Figure 2b and 2d). Top features now are gene density-related features (gene, exon and CDS coverage), DNA shape, sequence-related features, and distance to transcription start site (TSS) (Figure 2d). This change in feature order, together with the high performance of the model inside euchromatic regions suggests, that not only the (eu)chromatin state but also local sequence properties influence the occurrence of crossovers. It is particularly revealing that features related to gene density (gene, exon, CDS) constitute the top 3 (Figure 2d). This is in line with existing knowledge on the preference of COs in tomato to be located near genes (Demirci et al., 2017). However, similar to the strong influence of euchromatin found above, we now have features describing high-level annotation which strongly influence the prediction model. In order to further reveal more local sequence properties that influence COs in gene-rich regions, we devised a third and final strategy to generate a random set.

CO region prediction in tomato gene-rich regions

Given the important role of gene annotation related features in the prediction model found above, we used a third sampling strategy which takes the gene distribution of the tomato genome into account. This new sampling strategy also largely distinguishes euchromatin versus heterochromatin, as euchromatin is more gene rich; moreover, genic regions in heterochromatin where CO potentially could occur are also taken into account. Briefly, this strategy involved construction of an estimate for whole genome gene density, followed by selection of random regions by sampling from this density. In doing so, the experimental COs were used to find the best value of the bandwidth parameter of the gene density estimation. This procedure ensures that, similar to the positive cases (experimental CO regions), the random cases will preferentially, but not exclusively, occur in gene-rich regions. Further details of this sampling strategy are described in the Experimental Procedures section.

We constructed three classification models using the same three classifiers with the new random set and the same positive set. Similar to previous trials, the best

performing classifier was the random forest classifier (Figure 2e), again with a slightly lower performance than before (AUROC = 0.79). In this model, the most relevant features related to DNA shape, sequence, LTR repeats, distance to TSS and parental sequence differences (Figure 2f). In particular, it revealed local DNA properties as predictive: the two most important features were the DNA shape features propeller twist and helical twist. Since the second (euchromatin-based) and third (gene-density based) sampling strategies both focus mostly on genic areas, we expect that feature importances for both strategies are correlated. To test this, we compared the feature orders obtained by the random forest classifier following the two sampling strategies by Spearman's rank correlation test. The test showed significant positive correlation between the importance scores for the features obtained with these two sampling strategies (Spearman's *rho* = 0.91, *p*-value << 0.001). Hence, as expected, out of all features, similar features were selected as important for predicting CO regions in euchromatin (second sampling strategy) and in gene-rich regions (final sampling strategy).

We were interested whether this robust behaviour of predictive features was also present between the three different classifiers (decision tree, random forest and logistic regression) trained using the sampling strategy based on gene-rich regions. Such robustness would give credibility to the obtained set of predictive features. To investigate this, we compared the feature importances between the three classifiers by Spearman's rank correlation test; we also included the significance order of features obtained from the *t*-test. As summarized in Table 1, even the lowest correlation was significant and positive (*rho* = 0.44 ; *p*-value < 0.001). Given that some of the features are related to each other, this correlation between feature importance scores might be an underestimate. It could be strongly influenced by the correlation between features: out of two features which are highly correlated, one may be ranked highly by one classifier and the other by another classifier. Note that the correlation between different features describes whether the feature values display similar trends in our dataset. Above we analyzed the correlation between feature importance scores, obtained for the same feature with different prediction models. The correlation between feature importance scores could be lowered by correlation between the feature values; to test this, we clustered all features, and labeled them with their cluster membership (Figure S1, Table S4). Subsequently we run the Spearman correlation test for feature importance on cluster ranks (where each cluster was ranked with the rank from its most important feature). As expected, the correlation between the cluster ranks of the features between the different classifiers increased and resulted in a minimum *rho* value of 0.56 ($p < 0.001$). The analysis of feature importance thus showed that the ranking of features is robust to the choice of sampling strategy and classifier.

Factors related to crossovers in tomato

As described above, we generated machine learning models predicting the likelihood of CO formation based on DNA sequence and shape features. In a next step, we aimed at obtaining insight into genomic determinants of CO formation by analyzing

how the models make these predictions. This is reflected in the feature importance scores (Figure 2); to interpret these, we also made use of the feature values in CO regions and random regions (Figure S2; Table S5). In particular, we observed that the most important features (Figure 2) could be grouped into those related to genomic content and those related to genome accessibility.

Two features related to genomic content are euchromatin (Figure 2b) and gene content of a region (Figure 2d) which are strongly positively correlated with the occurrence of CO regions in the first two predictive models. A third feature related to genome content is the presence of LTR repeat regions: according to the final model, the probability of a CO increases with decreasing occurrence of LTR repeats (Figure 2f). These three genomic features are related to each other, as LTR regions are preferentially positioned in the pericentromeric regions of the chromosomes where gene density is lower and the DNA is condensed into tightly packed heterochromatin (Sherman and Stack, 1995; Jouffroy et al., 2016).

Among the features important for discriminating CO regions from non-CO regions, there were three features related to the accessibility of genomic regions. First, we found a negative correlation between distance to TSS and the occurrence of CO regions (Figure 2d and 2f). The distribution of TSS distances is shifted towards somewhat more negative values for CO regions compared to random regions. This implies that, compared to randomly chosen regions, CO regions on average are more often found upstream of the TSS, i.e. in promoter regions. Since promoters contain nucleosome depleted regions (Hartley and Madhani, 2009) and are accessible to transcription factor binding, it is likely that they are also accessible to the recombination machinery during the DSB formation stage, as was found in yeast (Pan *et al.*, 2011) and Arabidopsis (Choi et al., 2018). Moreover, AA/TT/TA/AT dinucleotide frequencies are positively correlated and predictive for CO regions (Figure 2f). This finding could be related to the enrichment of TATAT, poly-A and poly-T sequence motifs found in CO regions in tomato (Demirci et al., 2017) and in Arabidopsis (Wijnker et al., 2013; Choi et al., 2013). Similar to the role of promoters, it has been suggested that specific sequence motifs associated with CO occurrence indicate regions of open chromatin (Shilo et al., 2015) which might be explained by the exclusion of nucleosomes, leading to high double strand break levels (Choi et al., 2018). Thirdly, we found a relation between mean propeller twist angle (a DNA structural property) and CO regions (Figure 2f): a higher absolute value of propeller twist angle makes a region more likely to be a CO region. Importantly, in yeast a higher absolute propeller twist angle correlates with a lower nucleosome occupancy (Gan et al., 2012). A higher absolute propeller twist angle between particular base pairs could render the DNA more rigid, making the DNA harder to bend around e.g. histones (El Hassan and Calladine, 1996). Overall, our results indicate the relevance of genome accessibility for CO formation: nucleosome depletion could render genomic regions more accessible to the recombination machinery.

In addition to features related to genomic content and features related to genome accessibility, the genetic diversity between contributing parental sources is also suggested to be relevant by the model. In particular, the model showed a positive

relation for the number of homozygous SNPs and length of INDELs between parental genomes with CO region presence (Figure 2f). However, care should be taken when interpreting correlations between SNP rates and CO rates: since CO regions are defined by SNPs, it is likely that there is a bias in favor of positive correlation.

CO prediction in Arabidopsis, maize and rice

The results obtained for tomato indicate that it is possible to analyze genomic determinants of CO formation using the set of sequence- and annotation-based features. To investigate the role of these features in other plant species, we constructed prediction models for maize, Arabidopsis and rice. For these three species, we obtained CO regions with sufficient resolution needed for training the models (Wijnker *et al.*, 2013; Li *et al.*, 2015; Si *et al.*, 2015). We prepared positive sets as 4kb-long regions around CO positions from rice (n = 468), maize (n = 63) and Arabidopsis (n = 159), respectively. We sampled the same number of 4kb-long regions as in the positive set for each species, using the gene-density based sampling strategy as described above. We prepared the same features as for tomato, except for the parental sequence based features. In addition, there are small differences in feature sets between the species as different genomes have different repeat content.

We initially tested the individual discriminative power of features by *t*-test. This yielded 15 significant features among 59 features for Arabidopsis, 13 significant features among 64 features for rice, 7 significant features among 55 features for maize and 28 significant features among 56 features for tomato with p-values < 0.05 (Table S6). For tomato, the number of significant features was lower than what was found above when using a random set from the whole genome. This is caused by the fact that it is more difficult to discriminate between CO regions and random regions which are both sampled from gene-rich areas in the genome. Given the smaller number of COs available for Arabidopsis, rice and maize, it is also not surprising that fewer features were found significant in these species compared to tomato. Subsequently, we trained a random forest classifier for each of the three species separately. To compare these three models in a fair way with the tomato model, we also trained a model for tomato without the parental sequence based features. According to the performance results given in Table 2, CO sites are well predictable for both models of tomato and maize (AUROC >> 0.5) and reasonably predictable for Arabidopsis and rice (AUROC > 0.5). The difference in predictive power is not dependent on the number of COs in our training set: tomato has the most data and maize the least, while in both CO is easier to predict than in Arabidopsis and rice.

To obtain additional validation for the models, we followed two strategies. One was to obtain a set of true negative cases from pericentromeric regions. Reassuringly, as shown in Table 2, accuracy obtained by applying the models to these regions was again quite decent for Arabidopsis and rice (66%-76% correct) and in particular high for tomato and maize (>90% correct). The second strategy was specific for Arabidopsis, for which we used a genome-wide set of recombination rates (Choi *et al.*, 2013). As expected, CO regions in our dataset showed clearly higher rates compared to random regions (Figure S3A; p-value based on *t*-test: $10^{-9)}$. The recombination rate

for CO regions correctly predicted by the model was similar to the rate for CO regions not correctly predicted by the model (Figure S3B). Strikingly however, recombination rates for random regions predicted by our model to be CO regions were clearly higher than rates for random regions predicted to be random regions (Figure S3B; *p*-value based on *t*-test: 10^{-7}). This provides clear validation for our model, because it demonstrates that for a set of randomly chosen genome regions, the model discriminates between regions with low and with high recombination rate.

Factors related to crossovers in Arabidopsis, maize and rice

We further investigated whether similar features are important for CO prediction in the four different species (a complete overview of feature importance values is given in Figure S4). We compared the order of feature importances between species with Spearman's correlation test, as shown in Figure 3. On the one hand this revealed that tomato and maize displayed only a modest non-significant correlation, whereas on the other hand, all other pairs of species displayed positive significant correlations. The highest correlation was observed between tomato and Arabidopsis, for which very similar features were important to predict CO regions.

To identify common and species-specific features we selected the top ten most contributing features of each species' CO prediction model. Features contributing to the top ten in at least one species are displayed in Figure 4, showing their importance and their influence on the likelihood of COs. Note that features reported in Figure 4 are not necessarily the same as those reported as the result of the *t*-test in Table S6. This is because the *t*-test considers each feature separately, whereas the random forest uses combinations of features, and then orders the features individually based on their contribution to the model. In addition, for tomato there are small differences between the features shown in Figure 2f and those shown in Figure 4, as the latter includes only features relevant for all four species.

Interestingly, Figure 4 shows there is a large group consisting of the DNA shape feature helix twist, AT, TA, AA and TT dinucleotide frequencies, that are predicted to have a positive effect in all four studied species, with higher feature values indicating a higher likelihood to be a CO region. Similarly, another group consisting of the DNA shape feature propeller twist, and GG, GA, TC, CC and AG dinucleotide frequencies has a negative effect in all four studied species. In addition, the LTR/Gypsy feature has a negative relation to COs in three of the four species: CO regions are not favored near LTR repeats in maize, tomato and rice. For Arabidopsis the LTR/Gypsy feature is not relevant, since LTR repeats to a large extent are absent from the Arabidopsis genome (The Arabidopsis Genome Initiative, 2000).

These two groups of conserved features, which are consistently positively or negatively related to CO in all four species, can be broadly related to genome content and genome accessibility as found above for tomato. In particular, the importance of genomic content is reflected in the negative correlation of CO regions with the occurrence of LTR/Gypsy repeats. The negative correlation between recombination and transposon occurrence along chromosome arms has recently been reviewed (Lambing et al., 2017); transposon content increases towards the centromere while

the recombination rate decreases towards the centromere. Several other features conserved between species are related to genome accessibility. CO regions are positively correlated with AT/TA/AA/TT dinucleotide frequencies and propeller twist absolute angles. As discussed above, the nucleosome occupancy of these regions is expected to be low. This suggests that COs tend to localize in regions of open chromatin that are accessible for the recombination machinery.

In addition to these features that are invariant between species, a more species-specific role was observed for other features. The most important features found in maize were gene-annotation related features like exon, CDS and 3'UTR, whereas in tomato and Arabidopsis propeller twist, helical twist and AT/TA dinucleotides were most important. This difference could partially relate to the observation of CO regions in maize preferentially in 5'UTRs and 3'UTRs (Li et al., 2015), and in tomato and Arabidopsis primarily in promoters (Wijnker et al., 2013; Choi et al., 2013; Demirci et al., 2017). Furthermore, in rice, high roll (a DNA shape feature) and low CA dinucleotide frequency in particular favored the occurrence of COs. Two additional features with a species-specific role were Minor Groove Width (MGW) and the distance to TSS. MGW has a negative relation to COs in Arabidopsis and tomato and a positive (albeit non-significant) relation in maize and rice. MGW can strongly influence the binding of proteins to DNA (Rohs *et al.*, 2009). As described in the Introduction, some knowledge exists on different effects of CO regulators on CO formation in different plant species. The potential influence of MGW on binding of such CO regulator suggests a possible explanation for why the relation between MGW and CO formation is positive in some species and negative in others: higher MGW would have the same effect on binding of the protein in all species, which subsequently would have a differential effect on CO formation. As for distance to TSS, this feature again hints at the importance of genome accessibility. CO regions are localized upstream of the TSS (i.e. in promoter regions) in tomato, rice and Arabidopsis, while they are located downstream of TSS (i.e. at 3' UTR ends of genes and gene bodies) in maize. Even though CO regions localize at different ends of genes, apparently these positions are associated with nucleosome depleted regions (Bell et al., 2011) rendering them accessible to the recombination machinery.

Conclusions

We present the first comprehensive application of machine learning to predict CO regions throughout the plant kingdom. CO regions are reasonably predictable in Arabidopsis and rice and can be predicted with high accuracy in tomato and maize. A few different factors might influence the predictive power. One is that we focus on prediction of COs in gene-rich regions to be able to find local features, which inevitably means losing predictive power as the difference between random and CO regions gets smaller. The second reason is that there is no proper negative dataset to compare; irrespective of the way we sample, some regions in the random dataset may actually be prone to CO formation.

Our results indicate conservation and variation of genomic features influencing CO formation throughout the plant kingdom. We found two main groups of conserved features important for predicting CO regions in all four species: genome content and genome accessibility. CO regions are more likely to lie in euchromatic, gene-rich chromosomal regions, be A/T rich, have high absolute propeller twist angles and be depleted of LTR repeats. This could well relate to nucleosome depletion, leading to accessibility by the recombination machinery. In addition to these general rules, we observed that in Arabidopsis, rice and tomato, CO regions are often found in 5' UTR ends of genes while in maize CO regions are more prevalent in 3' UTR ends of genes. Yet, in general, in Arabidopsis, rice, tomato and maize, CO regions localize around the UTR ends of genes which suggests that gene regulatory regions are involved in the crossover mechanism.

In addition to these gross similarities between species, our results also indicate the importance of species-specific aspects of CO formation. One example is that minor groove width is negatively related to CO formation in tomato and Arabidopsis and positively related in rice and maize. Our findings that both conserved and species specific genomic features are correlated with COs might be related to the differential effect that proteins have on CO formation. For example, PRDM9 has a specific role in CO formation in human and mouse (Myers *et al.*, 2010; Edlinger and Schlögelhofer, 2011). Similarly, PCH2/CRC1 and ZYP1/ZEP1 seem to have a differential effect on CO formation in Arabidopsis and rice (Lambing et al., 2017). The finding that DNA shape features are important according to our prediction models could be related to interactions of such proteins with DNA, given that DNA shape is known to be relevant for protein-DNA interactions (Mathelier et al., 2016). The characteristics of the (spatial) interaction between such proteins and their DNA targets is relatively unknown and in our opinion calls for more detailed studies, involving for example ChIPseq technology.

Generally speaking, our results indicate the importance of both conservation and variation of features influencing COs in various plant species. Our work lays the ground for a comprehensive analysis of features underlying crossover formation in plants. Using additional high resolution datasets, as well as additional relevant features such as epigenetic modifications, will be the next step in order to understand CO regions better. This will be of fundamental biological relevance and will provide further opportunities for application in plant breeding.

Experimental Procedures

Dataset preparation

Sequences for positive (CO regions) and negative cases were prepared for tomato, rice, thale cress (Arabidopsis thaliana) and maize by using the corresponding genome information.

For tomato, 1015 CO positions were obtained from Demirci et al. (2017). CO events were detected in an F6 generation of interspecies recombinant inbred lines (RILs). Parental lines of the RILs were *S. lycopersicum* Moneymaker and *S. pimpinellifolium*. The reference genome *Solanum lycopersicum* Heinz version SL2.50 was used. The genome sequence and gene annotation files (ITAG2.4 gene models and ITAG2.4 repeats aggressive files in gff3 format) were obtained from https://solgenomics.net.

For rice, 1287 CO positions were obtained from Si et al. (2015). CO events were detected in F2 lines grown in different environmental conditions; the parental lines were *PA64s* (a hybrid between *O. sativa* indica and javanica) and *93-11* (*O. sativa* indica group). The reference genome *Oryza sativa* Nipponbare version IRGSP-1.0 was used. The genome sequence and gene annotation files were obtained from http://rapdb.dna.affrc.go.jp/download/irgsp1.html.

For Arabidopsis, 191 CO positions in total were obtained from tetrads and double haploids of *Arabidopsis thaliana* (Wijnker et al. 2013). The parental lines of tetrads were Cvi X Ler and Col accessions of *A*. *thaliana*. The parental lines of double haploids were Col and Ler accessions. The reference genome version TAIR 10 genome sequence and gene annotation (gff3) were obtained from arabidopsis.org.

For maize, 924 CO positions from tetrads were obtained from Li et al. (2015). The parental lines of tetrads were SK and Zheng58 accessions of *Zea mays*. The reference genome B73 RefGen v3 (aka AGPv3) genome sequences and the gene annotation file were downloaded from Ensembl Genomes release 21 (ftp://ftp.ensemblgenomes.org/pub/plants/release-21/fasta/zea_mays/).

Repeats for rice, Arabidopsis and maize genomes were inferred using RepeatMasker (Smit et al., 2013-2015) together with its dependencies Tandem Repeat Finder (Benson, 1999) and NCBI blastn programs. As repeat database, Genetic Information Research Institute Repbase Update database (Bao et al., 2015) was used.

For positive data set preparation, CO sites smaller than 2 kb were selected and extended to 4kb from their midpoint. After this step, the number of CO regions was 749 for tomato, 485 for rice, 69 for maize and 161 for Arabidopsis. For cases where CO regions overlapped, one of the two overlapping regions was randomly removed when the overlap was more than 25%, i.e. more than 1 kb. Moreover, CO regions were filtered if they overlapped with gaps in the reference genome. After filtering, the number of CO regions was 664 for tomato, 468 for rice, 63 for maize and 159 for Arabidopsis.

Sampling random cases from euchromatin or whole genome in tomato

We randomly selected 664 non-overlapping regions from tomato euchromatin excluding CO regions and assembly gaps (i.e. N bases). Euchromatic region positions were previously calculated in Demirci et al. (2017). To sample these random regions, the bedtools version 2.25.0 (Quinlan and Hall, 2010) shuffle function was used with the 'chrom' option, which protects the distribution of sequences among chromosomes. For example, if 10 sequences were present in chromosome 1 in the positive set, 10 sequences will be randomly selected on that chromosome for the random set. The same procedure was used to sample from the whole genome.

Sampling random cases from gene-dense regions

First, we generated a whole genome gene density estimate using a kernel density procedure (scikit-learn version 0.18 (Pedregosa et al., 2011), Python 3.5.2 (Python Software Foundation, https://www.python.org/)). We used the center position of every gene from the corresponding species annotation as a representation of the genes. The value of the kernel bandwidth was chosen such that the density would optimize the probability of the experimental CO distribution: the maximum log likelihood of the experimental CO distribution was found using a grid of 1000 different bandwidths, ranging from 1,000 to 1,000,000 with increments of 1000. The optimum bandwidths obtained were 36,000, 7,000, 171,000 and 54,000 for tomato, maize, rice and Arabidopsis, respectively. Then, to generate the negative set, for each chromosome, *n* regions were randomly sampled, where *n* is the number of CO regions in that chromosome in the positive set. Then, the candidate regions were filtered for the presence of gaps (N's), overlaps between each other and overlaps with any region in the positive set. If any of the initial candidates failed to pass the filtering, a new candidate was sampled from the distribution and the same filtering was applied. This process was repeated until *n* candidate negative regions passed all the filtering steps.

Feature preparation

For the positive and negative cases, the following features were calculated:

(i) Features derived from sequence information:

Dinucleotide frequencies: for each of the 16 possible dinucleotides, the following calculation was performed:

$$F_{AA} = n_{AA} / (l\text{-}1)$$

where F_{AA} indicates the frequency of dinucleotide AA, n_{AA} is the number of occurrences of AA in the given sequence, and *l* is the length of the sequence.

CTT and CCN motifs: as motifs, we used TCTTCTTC (Wijnker et al., 2013) and CCNCCNCCN (Shilo et al., 2015). Motif absence or presence in a region was described with a binary feature (motif presence), and the number of times a motif

occurred in a region was described in the feature motif occurrence. Finally, motif search scores were obtained with FIMO (Grant et al., 2011); in case of multiple occurrences of a given motif in a region, the following score was used to represent repetitive motifs:

score = (“motif score” / “motif length”) * “total length”

where “total length” means the total length of sequences covered by the motif.

The *DNA structural features*: helix twist angle, propeller twist angle, minor groove width (MGW) and Roll were estimated for each nucleotide position in each region using the DNAshapeR algorithm (Chiu et al., 2016). This approach predicts these structural properties for a given sequence using a model trained on experimental DNA structures (Zhou et al., 2013): (i) propeller twist angle is a negative value which measures the perpendicular twist between two paired bases from different strands; (ii) helix twist angle is a positive angle between two adjacent base-pairs as they twist in a DNA helix structure; (iii) minor Groove Width (MGW) is the width of the DNA minor groove in armstrong (Å); (iv) roll angle is the angle between two consecutive base-pairs rolling over each other, which can be positive or negative. The values predicted for each nucleotide were averaged over a (positive or random) region to obtain a single value for each region. In addition, we calculated the minimum and maximum values estimated for each DNA structural feature for each region.

(ii) Features derived from genome annotation information:
The *distance from the centre of sequences to the nearest transcription start site (TSS)* was calculated as described in Demirci et al. (2017). Briefly, the directed distance from the closest TSS position was calculated with the bedtools version 2.25.0 closest function; a negative value means that the midpoint of a sequence lies upstream of the TSS. Since the 5’ UTR regions were incomplete in the tomato genome annotation, we used mRNA start positions as TSS. For rice, maize and Arabidopsis 5’ UTR regions were used.

The *coding region fraction* was calculated for each region. The gene elements which overlap with the regions were extracted by the bedtools version 2.25.0 intersect function from gene annotation files (ITAG 2.4 gene models file for tomato, IRGSP-1.0 representative locus and transcripts exon files for rice, TAIR 10 genes for arabidopsis, AGPv3.21 annotation file for maize). Subsequently, for each region, the total length of exonic regions was divided by the length of the region and reported as the coding region fraction of that region.

For each region, the *transposon family fractions* were calculated in a similar way as coding region fractions. Repeats which overlap with the regions were extracted by the bedtools version 2.25.0 intersect function from the repeat annotation files (ITAG 2.4 annotation repeat file ITAG2.4_repeats_aggressive.gff3 for tomato and repeat

annotation files generated by RepeatMasker (see above) for other species). Then, for each region, the overlap fractions were calculated for all defined repeat families: the total length of the annotated repeats was divided by the length of the region. Repeat families were excluded as features if they were not present in any region in the dataset of each species. For tomato, in addition, eu(chromatin) state was used as a feature in the first model (sampling from the whole genome); it was assigned as described in Demirci et al. (2017).

(iii) Features derived from parental genome information:

S*equence divergence* between parental genomes for a given region was calculated from VCF files of tomato parental genomes (*S. lycopersicum* Moneymaker and *S. pimpinellifolium*). The fastq files were downloaded from European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ena) for *S. pimpinellifolium* (SAMEA2625653) under project number PRJEB6659 (Aflitos et al., 2015) and for *S. lycopersicum* Moneymaker (SAMEA2340764) under the project number PRJEB5235 (Aflitos et al., 2014). These were mapped to the *Solanum lycopersicum* Heinz version SL2.50 reference genome and variants were called with the same settings as described in Aflitos et al. (2014). From the resulting variant VCF files for each parental genome, containing SNPs w.r.t. the reference genome, SNPs were compared to each other and homozygous SNPs having the same alternative alleles in the two parents, i.e. identical variants w.r.t. the reference, were removed. The remaining SNPs from the two genomes were combined to obtain parental SNPs and analysed to calculate the total number of SNPs, heterozygous SNPs and homozygous SNPs present in the regions as three separate features. In a similar way, INDELs with different lengths in the parental genomes were analysed to calculate the number of INDEL positions and the total length of differential INDEL lengths for each region. All five features from SNPs and INDELs were reported as a fraction of each analysed region.

Features were scaled individually by subtracting the mean and dividing by the standard deviation. The scaled features were used in later steps unless otherwise stated. To cluster features, the absolute value of Pearson correlation between features was converted to a dissimilarity matrix using the equation:

$D = 1 - \text{abs}(rho)$

where *D* is the distance and *rho* is Pearson correlation coefficient.

Based on the dissimilarity matrix, we performed hierarchical clustering with the hclust function in R using complete linkage. After manual inspection a threshold of 0.4 was applied to define clusters.

To inspect the role of individual features, we performed a *t*-test on non-scaled feature data using scipy 0.17.0 (Jones et al., 2001). *p*-values were Benjamini-Hochberg corrected using the multiple test function in statsmodels version 0.8.0 (Seabold and

Perktold, 2010). To visualize and detect the most significant features, the *p*-values were log-transformed.

Comparative genomic analysis

For each species, we used the above mentioned genome annotation files for transposable element (TE) density and gene density graphs. For CO density, we used the filtered set of CO regions which were used as positive set to build the models. For SNP density, we used the parental marker set if provided by the original study (for tomato, Arabidopsis and rice); if not provided (in the case of maize), we identified the differential SNPs between parental genomes. To do so, raw sequence datasets of parental genomes Zheng58 (accession no SRR449340, SRR449342 and SRR449343) and SK (accession no SRR1585475) were downloaded from the European Nucleotide Archive (https://www.ebi.ac.uk/ena). After trimming with Trimmomatic v0.36 (Bolger et al., 2014), reads were mapped to the reference genome AGPv3 by bowtie2 version 2.2.6 (Langmead and Salzberg, 2012) with fast mapping option, PCR duplicates were removed, and SNPs for each parent were called by samtools version 0.1.19 (Li et al., 2009) and bcftools version 0.1.19 (Li, 2011). SNPs having coverage less than 4 or more than 100 were filtered by bcftools. Finally, we reported the homozygous SNPs between parental genomes. Centromere information was obtained as follows: for Arabidopsis, we used Table S26 from Ziolkowski et al. (2017); for maize, we used 1 Mb flanking region of CRM repeats as identified by Repeatmasker; for tomato, we used Data S1 in Demirci et al. 2017; for rice, we inferred the approximate locations from Si et al.'s study (2015), Figure 3. Counts for different elements (COs, TEs, genes, SNPs) were obtained in 1Mb bins across all chromosomes for a given species.

Classifiers

Decision tree classifier: We used the decision tree classifier algorithm implemented in scikit-learn v0.18 with the Gini impurity criterion to split the nodes. To prevent overfitting, the minimum number of samples on each leaf was set to 5 and the rest of the settings was left as default.

Random forest classifier: the random forest algorithm implemented in scikit-learn was used with 1000 trees in the forest. The remaining settings were kept at their defaults, with the number of features used at each split in each tree equal to the square root of the number of features, and the Gini criterion for splitting nodes.

Logistic Regression: the logistic regression algorithm implemented in scikit-learn was applied. To optimize the regularization factor *C,* necessary to prevent overfitting, we used cross-validation over 10 different values in the range of $1x10^{-4}$ to $1x10^{4}$. After the prediction model was built, we used the absolute values of the coefficients to determine the feature importances.

Comparison of feature importances

Spearman rank correlation was calculated between feature importances from different classifiers and different species. The resulting *rho* value per pair of feature importances and the corresponding *p*-value were reported to assess the similarity of the order of two feature importances.

(Cor)relation of the features to CO prediction

To determine if the predictive features have a positive or negative relation on the CO prediction, the mean value of a feature in the random set was subtracted from the mean value of a feature in the positive set. A positive sign means that higher values of that feature favor CO regions, and vice versa.

Evaluation of the performance of classifiers

The regions which include COs were defined as positive cases, whereas negative cases are the randomly selected regions. By comparing the prediction for a given case with its real label (CO or random), the following four values can be obtained: FP, the number of false positives (random cases predicted as CO); TP, the number of true positives (CO cases predicted as CO); FN, the number of false negatives (CO cases predicted as random); and TN, the number of true negatives (random cases predicted as random). To evaluate the performance of each predictor, we used the following evaluation metrics based on the values of FP, TP, FN and TN:

(i) The AUROC is the area under the Receiver Operator Characteristic (ROC) curve, which visualizes the True Positive Rate (TPR) versus the False Positive Rate (FPR). Here,

TPR = TP / (TP + FN), probability of detection of COs;

FPR = FP / (TN + FP), probability of wrongly predicting a random case as CO.

(ii) Precision measures how many of the CO regions were correct among the cases predicted to be CO: Precision = TP / (TP + FP)

(iii) Recall measures how many of the experimental CO regions were correctly predicted to be CO: Recall = TP / (TP + FN), which is identical to the TPR.

(iv) Accuracy measures how many of the instances are correctly predicted.

Validation of prediction models

We used 10-fold cross-validation to validate the prediction model. The dataset was randomly split into 10 parts, which in 10 iterations each serve as a test set for a model trained on the remaining 9 parts. The performance evaluation metrics are reported as average and standard deviation over the 10 test sets.

To obtain additional validation on independent data, for the prediction models trained on CO regions and random regions obtained from gene-rich areas in the four species, a negative set was generated by sampling from pericentromeric regions. The same number of regions as in the positive set (CO regions) was sampled from pericentromeric regions (excluding assembly gaps) with the same method as above (bedtools shuffle algorithm). The pericentromeric region locations were obtained as follows: for Arabidopsis, we used Table S26 from Ziolkowski et al. (2017); for maize, we used 20 Mb flanking regions of CRM repeats as identified by Repeatmasker

(excluding the CRM repeats); for tomato, we used heterochromatin regions defined in Data S1 in Demirci et al. (2017); for rice, we used cold spot regions defined in Si et al.'s study (2015), Table S4. Features were constructed for these regions in the same way as described above. To estimate the accuracy of the models, it was assessed for how many of the pericentromeric regions the models predicted that these regions would not be CO regions.
In addition, for Arabidopsis, we used a genome-wide set of recombination rates (Choi *et al.*, 2013) for validation. For each genome region used in our Arabidopsis model, a single recombination rate was obtained by averaging the values provided by Choi *et al*. The distributions of these values were obtained separately for CO regions vs. random regions, and for both types of regions separately based on whether the model predicted a region to be a CO region or a random region.

The scripts used for the analyses are available on https://github.com/sdemirci/predCO.

Acknowledgments

The work presented here is supported by the EU FP7 COMREC Marie Curie Initial Training Networks Programme project number 606956.

The authors declare no conflicts of interest.

Supporting Information

Figure S1. Dissimilarity tree of features in tomato.

Figure S2. Histograms of features.

Figure S3. Arabidopsis recombination rate for CO regions and random regions.

Figure S4. Features used in the random forest model in four species.

Table S1: Genome information of the studied species.

Table S2: t-test results of tomato features.

Table S3: Feature importances based on random forest in tomato.

Table S4: Cluster membership of the features in tomato.

Table S5: Feature means in random and positive regions.

Table S6: *p*-values of *t*-test on the features in maize, rice and Arabidopsis and tomato.

References

Aflitos, S.A., Sanchez-Perez, G., Ridder, D. de, Fransz, P., Schranz, M.E., Jong, H. de and Peters, S.A. (2015) Introgression browser: high-throughput whole-genome SNP visualization. *Plant J.*, **82**, 174–182.

Aflitos, S., Schijlen, E., Jong, H. de, et al. (2014) Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. *Plant J.*, **80**, 136–148.

Anderson, L.K., Lohmiller, L.D., Tang, X., *et al.* (2014) Combined fluorescent and electron microscopic imaging unveils the specific properties of two classes of meiotic crossovers. *Proc. Natl. Acad. Sci.*, **111**, 13415–13420.

Bao, W., Kojima, K.K. and Kohany, O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*, **6**, 11.

Bell, O., Tiwari, V.K., Thomä, N.H. and Schübeler, D. (2011) Determinants and dynamics of genome accessibility. *Nat. Rev. Genet.*, **12**, 554–564.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–80.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114.

Chiu, T.-P., Comoglio, F., Zhou, T., Yang, L., Paro, R. and Rohs, R. (2016) DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics*, **32**, 1211–3.

Choi, K., Zhao, X., Kelly, K. a, et al. (2013) Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. *Nat. Genet.*, **45**, 1327–36.

Choi, K., Zhao, X., Tock, A.J., et al. (2018) Nucleosomes and DNA methylation shape meiotic DSB frequency in Arabidopsis transposons and gene regulatory regions. *Genome Res.*, **28**, 532-546.

Demirci, S., Dijk, A.D.J. van, Sanchez Perez, G., Aflitos, S.A., Ridder, D. de and Peters, S.A. (2017) Distribution, position and genomic characteristics of crossovers in tomato recombinant inbred lines derived from an interspecific cross between Solanum lycopersicum and Solanum pimpinellifolium. *Plant J.*, **89**, 554–564.

Edlinger, B. and Schlögelhofer, P. (2011) Have a break: Determinants of meiotic DNA double strand break (DSB) formation and processing in plants. *J. Exp. Bot.*, **62**, 1545–1563.

Gan, Y., Guan, J., Zhou, S. and Zhang, W. (2012) Structural features based genome-wide characterization and prediction of nucleosome organization. *BMC Bioinformatics*, **13**, 49.

Gao, Z.-Y., Zhao, S.-C., He, W.-M., et al. (2013) Dissecting yield-associated loci in super hybrid rice by resequencing recombinant inbred lines and improving parental genome sequences. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 14492–7.

Grant, C.E., Bailey, T.L. and Noble, W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.

Haas, L.S. de, Koopmans, R., Lelivelt, C.L.C., Ursem, R., Dirks, R. and Velikkakam James, G. (2017) Low-coverage resequencing detects meiotic recombination pattern and features in tomato RILs. *DNA Res.*, **3**, 1213–6.

Hartley, P.D. and Madhani, H.D. (2009) Mechanisms that Specify Promoter Nucleosome Location and Identity. *Cell*, **137**, 445–458.

Hassan, M.A. El and Calladine, C.R. (1996) Propeller-Twisting of Base-pairs and the Conformational Mobility of Dinucleotide Steps in DNA. *J. Mol. Biol.*, **259**, 95–103.

He, Y., Wang, M., Dukowic-Schulze, S., et al. (2017) Genomic features shaping the landscape of meiotic double-strand-break hotspots in maize. *Proc. Natl. Acad. Sci.*, **114**, 12231–12236.

Jones, E., Oliphant, T., Peterson, P., et al. (2001-) SciPy: Open Source Scientific Tools for Python. Available at http://www.scipy.org.

Jouffroy, O., Saha, S., Mueller, L., et al. (2016) Comprehensive repeatome annotation reveals strong potential impact of repetitive elements on tomato ripening. *BMC Genomics*, **17**, 624.

Kauppi, L., Jeffreys, A.J. and Keeney, S. (2004) Where the crossovers are: recombination distributions in mammals. *Nat. Rev. Genet.*, **5**, 413–424.

Lambing, C., Franklin, F.C.H. and Wang, C.-J.R. (2017) Understanding and Manipulating Meiotic Recombination in Plants. *Plant Physiol.*, **173**, 1530–1542.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li, X., Li, L. and Yan, J. (2015) Dissecting meiotic recombination based on tetrad analysis by single-microspore sequencing in maize. *Nat. Commun.*, **6**, 6648.

Lichten, M. and Goldman, A.S. (1995) Meiotic recombination hotspots. *Annu. Rev. Genet.*, **29**, 423–44.

Liu, G., Liu, J., Cui, X. and Cai, L. (2012) Sequence-dependent prediction of recombination hotspots in Saccharomyces cerevisiae. *J. Theor. Biol.*, **293**, 49–54.

Liu, S., Yeh, C.-T., Ji, T., Ying, K., Wu, H., Tang, H.M., Fu, Y., Nettleton, D. and Schnable, P.S. (2009) Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. *PLoS Genet.*, **5**, e1000733.

Mathelier, A., Xin, B., Chiu, T.P., Yang, L., Rohs, R., Wasserman, W.W. (2016) DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell Syst.*, **3**, 278-286.

Mercier, R., Mézard, C., Jenczewski, E., Macaisne, N. and Grelon, M. (2015) The Molecular Biology of Meiosis in Plants. *Annu. Rev. Plant Biol*, **66**, 297–327.
Myers, S., Bowden, R., Tumian, A., Bontrop, R.E., Freeman, C., MacFie, T.S., McVean, G. and Donnelly, P. (2010) Drive Against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination. *Science*, **327**, 876–879.
Pan, J., Sasaki, M., Kniewel, R., et al. (2011) A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell*, **144**, 719–731.
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
Python Software Foundation. The Python Language Reference, version 3.5.2. Available at https://docs.python.org/3.5/reference/index.html.
Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–2.
Rodgers-Melnick, E., Bradbury, P.J., Elshire, R.J., Glaubitz, J.C., Acharya, C.B., Mitchell, S.E., Li, C., Li, Y. and Buckler, E.S. (2015) Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proc. Natl. Acad. Sci. U. S. A.*, **112**, 3823–8.
Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S. and Honig, B. (2009) The role of DNA shape in protein–DNA recognition. *Nature*, **461**, 1248–1253.
Sato, S., Tabata, S., Hirakawa, H., et al. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635–641.
Seabold, S., and Perktold, J. (2010) Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference.* 57-61.
Sherman, J.D. and Stack, S.M. (1995) Two-Dimensional Spreads of Synaptonemal Complexes from Solanaceous Plants. VI. High-Resolution Recombination Nodule Map for Tomato (Lycopersicon esculentum). *Genetics*, **141**, 683–708.
Shilo, S., Melamed-Bessudo, C., Dorone, Y., Barkai, N. and Levy, A.A. (2015) DNA Crossover Motifs Associated with Epigenetic Modifications Delineate Open Chromatin Regions in Arabidopsis. *Plant Cell*, **27**, 2427–36.
Si, W., Yuan, Y., Huang, J., et al. (2015) Widely distributed hot and cold spots in meiotic recombination as shown by the sequencing of rice F 2 plants. *New Phytol.*, **206**, 1491–1502.
Smit, AFA, Hubley, R & Green, P. (2013-2015) *RepeatMasker Open-4.0*. <http://www.repeatmasker.org>.
The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
Villeneuve, A.M. and Hillers, K.J. (2001) Whence meiosis? *Cell,* **106**, 647-650.
Wijnker, E. and Jong, H. de (2008) Managing meiotic recombination in plant breeding. *Trends Plant Sci.*, **13**, 640–6.
Wijnker, E., Velikkakam James, G., Ding, J., et al. (2013) The genomic landscape of meiotic crossovers and gene conversions in Arabidopsis thaliana. *Elife*, **2**, e01426.

Wu, J., Mizuno, H., Hayashi-Tsugane, M., et al. (2003) Physical maps and recombination frequency of six rice chromosomes. *Plant J.*, **36**, 720–730.

Yelina, N.E., Lambing, C., Hardcastle, T.J., Zhao, X., Santos, B. and Henderson, I.R. (2015) DNA methylation epigenetically silences crossover hot spots and controls chromosomal domains of meiotic recombination in Arabidopsis. *Genes Dev.*, **29**, 2183–202.

Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A.C., Ghane, T., Felice, R. Di and Rohs, R. (2013) DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res.*, **41**, W56–W62.

Ziolkowski, P.A., Underwood, C.J., Lambing, C., et al. (2017) Natural variation and dosage of the HEI10 meiotic E3 ligase control Arabidopsis crossover recombination. *Genes Dev.*, **31**, 306–317.

Legends

Figure 1. Distribution of genomic elements (genes, repeats, COs and SNPs) for (a) Arabidopsis, (b) tomato, (c) rice and (d) maize. The number of nucleotides covered per kb by the different genomic elements is given in 1Mb bins for the different chromosomes (indicated with numbers on the horizontal axes). Dashed lines indicate centromere locations.

Figure 2. Crossover (CO) prediction in tomato. (a, c, e) Assessment of prediction performance with receiver operator characteristic (ROC) curves for models trained on (a) whole genome, (c) euchromatin and (e) gene-rich regions. FPR: False Positive Rate, TPR: True Positive Rate, DT: Decision tree, LR: Logistic regression, RF: Random forest. Values between brackets indicate AUROC. Dashed line indicates performance of a random predictor, with AUROC equal to 0.5; the higher the AUROC, the better the predictor. (b, d, f) The top 10 most important features (ordered from left to right) according to the random forest classifier using (b) whole genome, (d) euchromatin and (f) gene-rich regions. The higher the Gini index, the more important the feature, i.e. the bigger its role in determining the CO prediction. The color of the bar for each feature indicates its positive (red) or negative (blue) relation to the occurrence of CO regions. The DNA shape features given here are mean angular values.

Figure 3. Correlation (Spearman's correlation coefficient *rho*) between the order of feature importances for crossover prediction in tomato, Arabidopsis, rice and maize. Significance is given as *** p <0.001; ** p <0.01.

Figure 4. Features (vertical axis) contributing to the top ten most important features in at least one of the species (horizontal axis). Color represents the relation of features to CO prediction (red: positive, blue: negative); intensity represents feature importance. The color-coded full set of features for each species is given in Figure S4.

Table 1. Spearman correlation coefficients (*rho*) between the feature importances of classifiers in tomato gene-rich regions.

Table 2. Performance statistics of random forest model for tomato, rice, maize and Arabidopsis.

Tables

Table 1. Spearman correlation coefficients (*rho*) between the feature importances of classifiers in tomato gene-rich regions.

	t-test	Decision Tree	Logistic Regression	Random Forest
t-test	-	0.44***	0.46***	0.70***
Decision Tree	*0.57***	-	0.46***	0.75***
Logistic Regression	*0.56***	*0.59***	-	0.50***
Random Forest	*0.62***	*0.84***	*0.59***	-

P-values: *** $p < 0.001$. Underlined values (above diagonal) are between the order of individual feature importances; *italic* values (below diagonal) are between the order of importances of feature clusters.

Table 2. Performance statistics of random forest model for tomato, rice, maize and Arabidopsis.

	tomato	tomato[a]	Arabidopsis	rice	maize
AUROC[b]	0.79 (s= 0.04)	0.77 (s= 0.03)	0.63 (s= 0.08)	0.67 (s= 0.05)	0.72 (s= 0.14)
recall[b]	0.82 (s= 0.03)	0.82 (s= 0.03)	0.64 (s= 0.09)	0.68 (s= 0.08)	0.76 (s= 0.10)
precision[b]	0.69 (s= 0.04)	0.67 (s= 0.03)	0.58 (s= 0.06)	0.60 (s= 0.05)	0.70 (s= 0.12)
accuracy[c]	0.95	0.94	0.66	0.76	0.92

[a.] Tomato dataset without the features from parental genome sequence.
[b.] AUROC, recall and precision are calculated with ten-fold cross-validation using the positive set consisting of experimental CO regions and the random set obtained by sampling from gene-rich regions. Values are mean values obtained with ten-fold cross-validation; s = standard deviation.
[c.] Accuracy values are calculated on the pericentromeric regions dataset after training with the positive set and the random set.

Figures

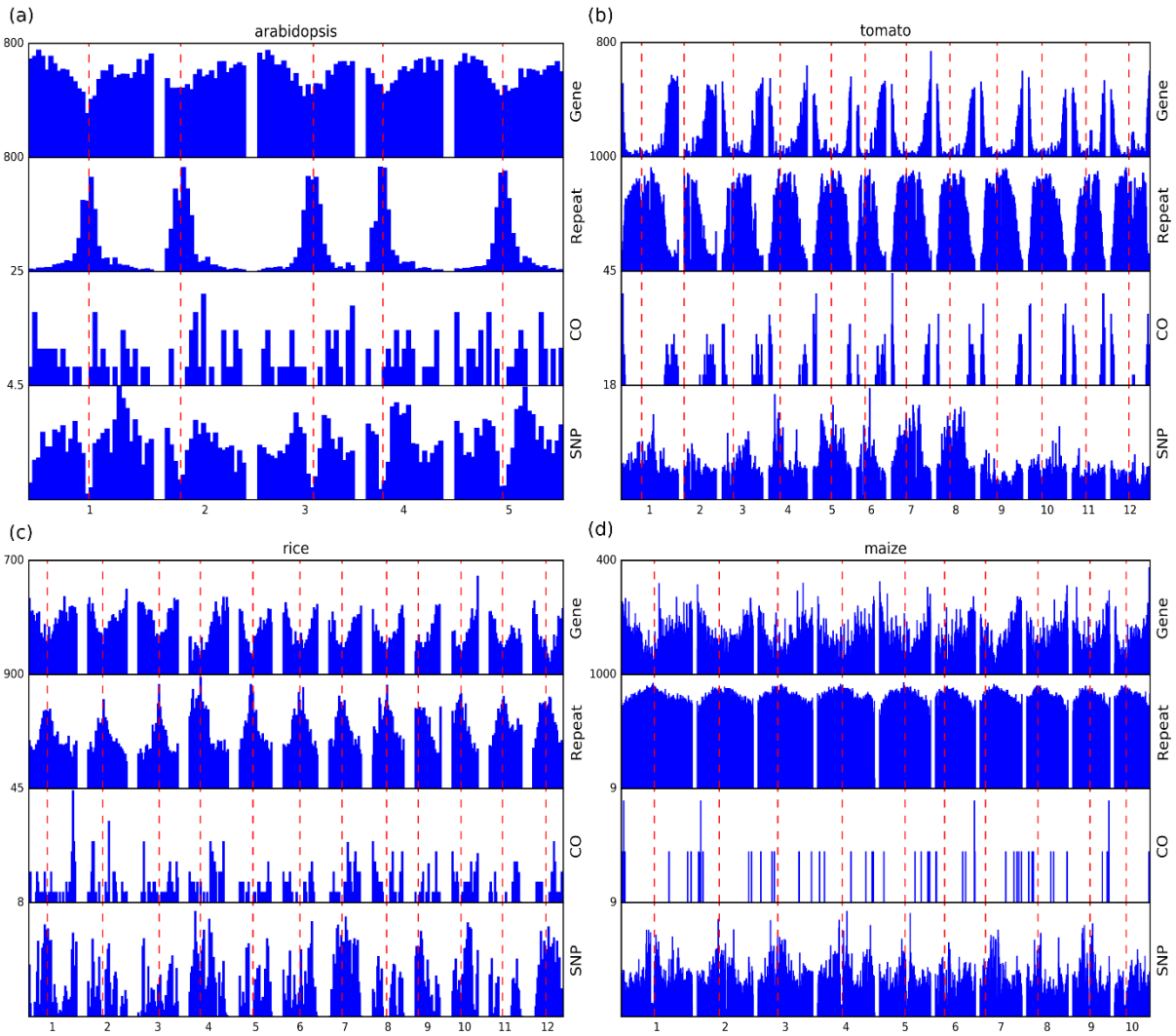

Figure 1. Distribution of genomic elements (genes, repeats, COs and SNPs) for (a) Arabidopsis, (b) tomato, (c) rice and (d) maize. The number of nucleotides covered per kb by the different genomic elements is given in 1Mb bins for the different chromosomes (indicated with numbers on the horizontal axes). Dashed lines indicate centromere locations.

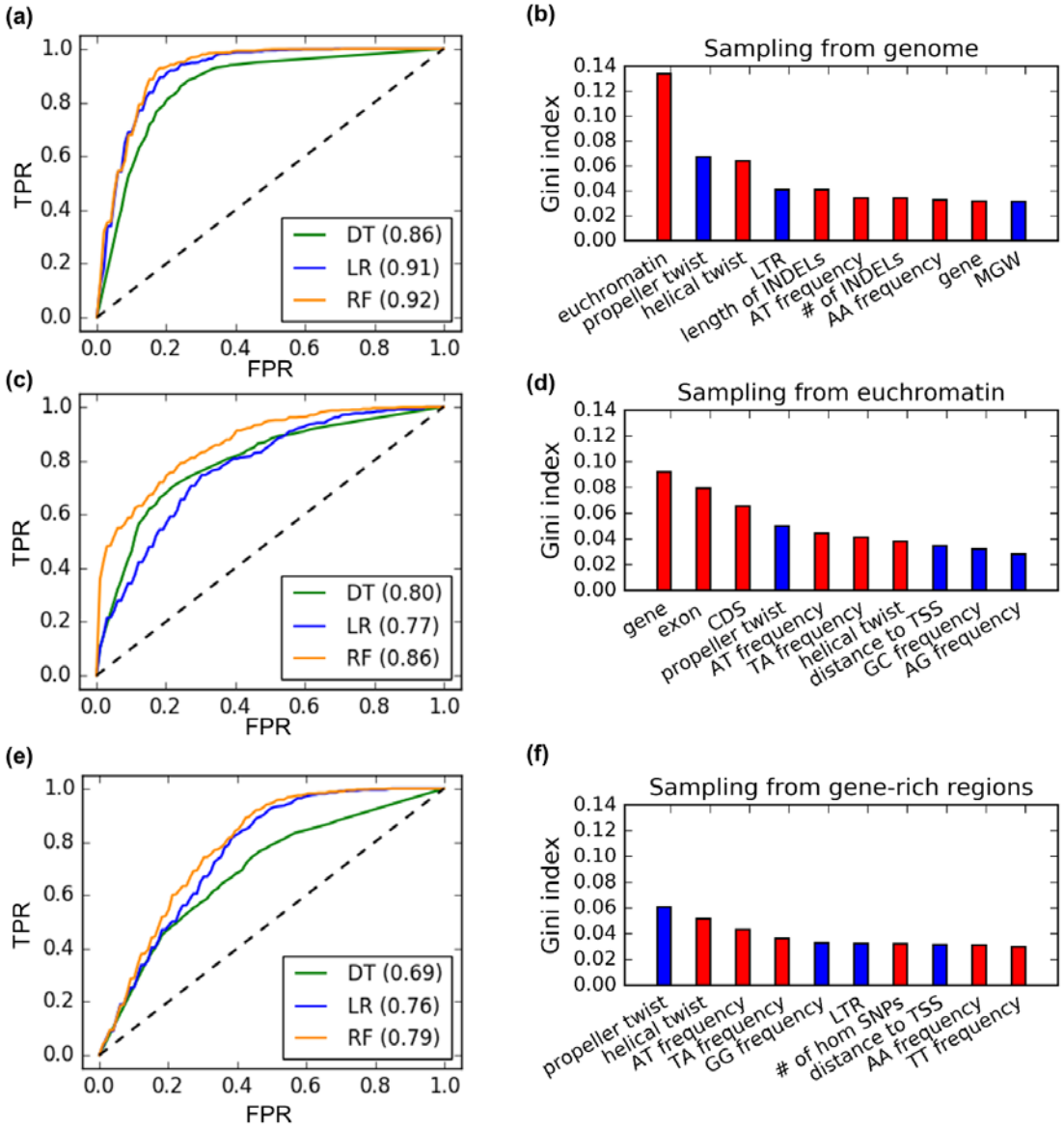

Figure 2. Crossover (CO) prediction in tomato. (a, c, e) Assessment of prediction performance with receiver operator characteristic (ROC) curves for models trained on (a) whole genome, (c) euchromatin and (e) gene-rich regions. FPR: False Positive Rate, TPR: True Positive Rate, DT: Decision tree, LR: Logistic regression, RF: Random forest. Values between brackets indicate AUROC. Dashed line indicates performance of a random predictor, with AUROC equal to 0.5; the higher the AUROC, the better the predictor. (b, d, f) The top 10 most important features (ordered from left to right) according to the random forest classifier using (b) whole genome, (d) euchromatin and (f) gene-rich regions. The higher the Gini index, the more important the feature, i.e. the bigger its role in determining the CO prediction. The color of the bar for each feature indicates its positive (red) or negative (blue) relation to the occurrence of CO regions. The DNA shape features given here are mean angular values.

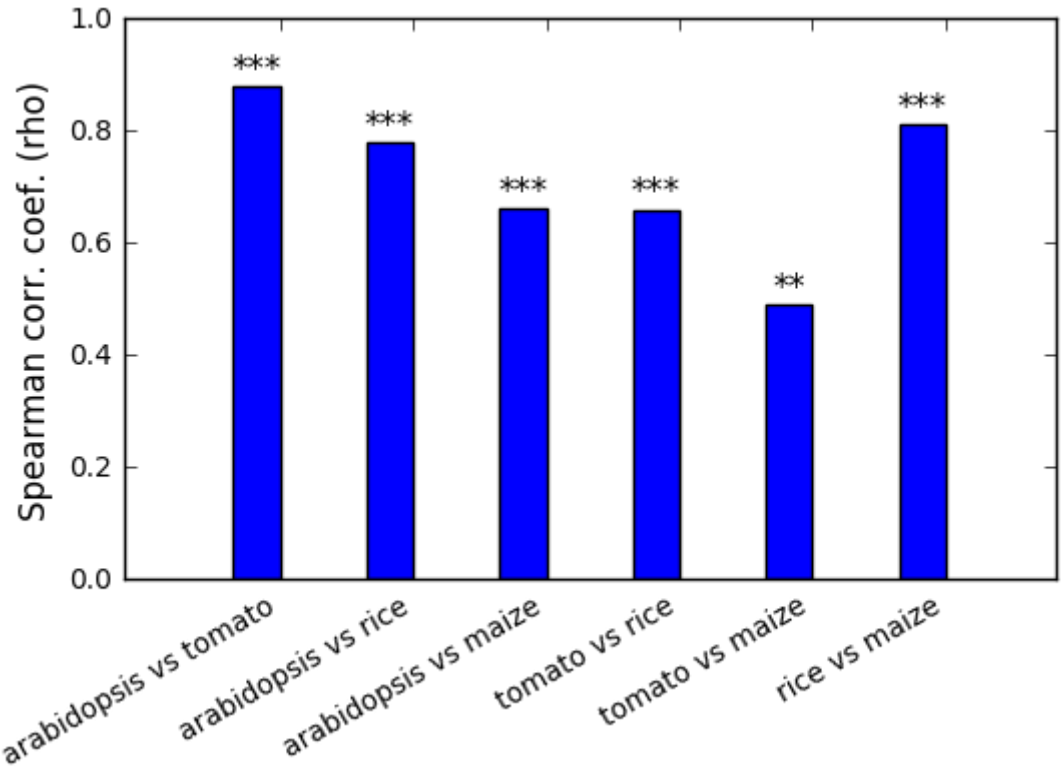

Figure 3. Correlation (Spearman's correlation coefficient *rho*) between the order of feature importances for crossover prediction in tomato, Arabidopsis, rice and maize. Significance is given as *** p <0.001; ** p <0.01.

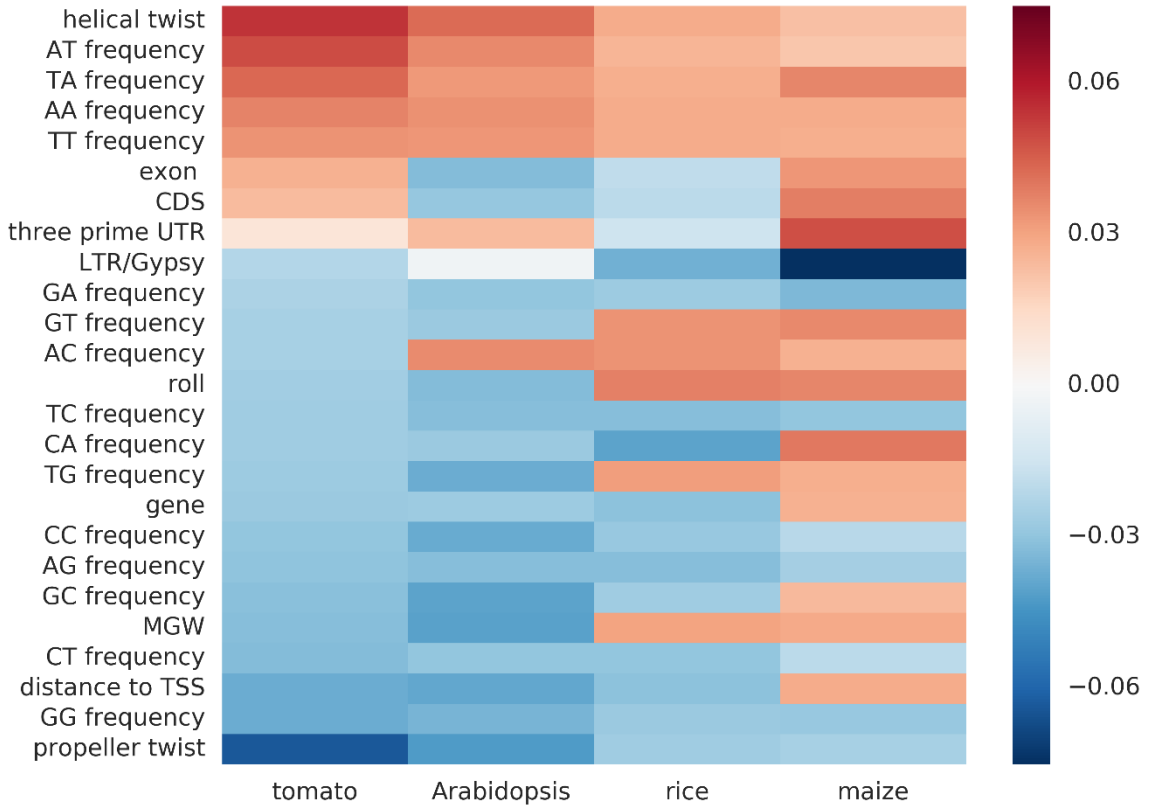

Figure 4. Features (vertical axis) contributing to the top ten most important features in at least one of the species (horizontal axis). Color represents the relation of features to CO prediction (red: positive, blue: negative); intensity represents feature importance. The color-coded full set of features for each species is given in Figure S4.