# Figures

**Fig. 1. Inter-species chromatin states definition.** Top panel: From left to right chromatin state definitions, abbreviation, species relation, composition (emission probability) and genome coverage based on 10 common epigenomic marks. Chromatin states with “>” indicate definitions transitioning between species. Darkblue colors in relation heatmap higlight for which species the definition is similar and columns represent *A. thaliana (At)*, *O. sativa (Os)* and *Z. mays (Zm)*, respectively. Bottom panel: fold enrichments over different genomic features for each state and species.

**Fig. 2. Inter-species chromatin states annotation.** Heatmaps depicting significant (p < 0.05) genomic overlap-enrichment (odds ratio) of inter-species states with different annotation modules. From top to bottom: genetic variability represented by significant SNPs in GWAS, transcription factor (TF) motifs illustrated by TF binding sites (BS) accordint to PlantRegMap categories, conservation covered by PhastCons elements and pairwise conserved non-coding elements (CNEs) and non-common chromatin proteins and histone marks. Chromatin states with “>” indicate definitions transitioning between species. Darkblue colors in relation heatmap higlight for which species the definition is similar and rows represent *A. thaliana (At)*, *O. sativa (Os)* and *Z. mays (Zm)*, respectively.

**Fig. 3. Inter-species chromatin states description.** Each chromatin functional group is exemplified by a module with a single state (CS1 – bivalent; CS6 – active; CS10 – divergent; CS11 – heterochromatin). From left to right, each module is constituted by a dotplot showing significant KO enrichments for the genes covered by the CS and alluvial diagrams describing the distribution and correspondence between gene biotypes and orthologous for each species (*A. thaliana (At)*, *O. sativa (Os)* and *Z. mays (Zm)*). Colors denote species. Dot size indicates gene ratio. Bold KO terms highlight convergent terms for all the species. Minor gene biotypes are represented by different symbols.

**Fig. 4. Predictive models of paralogs degree of functional divergence including chromatin states metrics.** Chromatin states metrics were obtained dividing promoter and genes in a fixed number of windows, calculating frequency and presence vectors and computing several distance and simmilarity coefficients between genes from the same paralog pair comparing equivalent vector types (see **Methods**). **(a-d)** Results reproducing Ezoe et al., 2021 models including CS metrics. **(a)** Relative importance in explanatory variables. The relative importance was inferred based on the logistic regression algorithm. **(b)** Custom chromatin state metric (CCSM; see **Methods**) distribution of high and low diversified gene pairs. P-value, two-tailed Wilcoxon rank sum test. Numbers in parenthesis represent the number of duplicate pairs. **(c)** Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves in our prediction models. Colored lines indicate different generated models in six types of formula based on logistic regression algorithms using different sets of features. The are under the curve (AUC) values were calculated by the best prediction model in each formula. A perfect classification model would have AUC-ROC and AU-PRC score of 1.0; black dotted lines represent performance of random classification model, in which AUC-ROC and AU-PRC values would be 0.5. **(d)** Histogram of the inferred degree of functional divergence (DFD) in high and low duplicates of the training data. The inferred DFD was calculated for 463/111 high/low diversified pairs, respectively. The bottom 5% of the inferred high diversified DFD values were < 0.46 (i.e low DFD at 5% FDR). The top 5% of the inferred low diversified DFD values were > 0.93 (i.e high DFD at 5% FDR). **(e-h)** Results reproducing Cusack et al., 2021 models including CS metrics. **(e)** Top 200 final selected features distribution across groups of variables for extreme-inclusive redundancy definitions without (RD4-RD9, respectively) and with (RD4C-RD9C, respetively) CS information. Numbers in parenthesis denote the median importance ranks for all the features in that group. Feature importance was determined using SVM with a linear kernel and normalized features values. Colors represent distinct redundancy definitions and features sets. RD4 (light green): extreme redundancy definition without CS information; RD4C (dark green): extreme redundancy definition with CS information; RD9 (light purple): inclusive redundancy definition without CS information; RD9C (dark purple): inclusive redundancy definition with CS information. All gene pairs in RD4/RD4C are contained in RD9/RD9C. **(f)** ROC and PR curves of final SVM models for each redundancy definition/feature set. AUC values were calculated by the best prediction model in each formula. **(g)** AUC-ROC and AU-PRC for the heldout tests for models built with each redundancy definition/feature set. **(h)** Matrix layout for all intersections between top 200 variables in redundancy definition/feature sets, sorted by decreasing order. Dark circles in the matrix indicate sets that are part of the intersection.

**Fig. 5. Functional genomics conservation (LECIF) score overview and downstream analyses.** This figure is constituted by 5 panels (top-left, top-right, bottom-left, bottom-right and middle). Top-left panel: Overview of the LECIF-score. Very briefly, LECIF algorithm was applied integrating epigenomic, chromatin states, whole genome alignments and transcriptomic information to obtain functional genomics conservation scores for all pairwise comparisons. These scores, together with previosuly generated resources, are stored in PlantFUNCO database to allow future applications and further hypothesis testing such as paralog functional evolution. Bottom-left, top-right and bottom-right panels illustrate LECIF-score downstream analyses for *O. sativa (Os)*, *Z. mays (Zm)* and *A. thaliana (At)*, respectively. Each of this panels are divided into left and right sides according to the two target remaining species and three description modules: 1) Genetic variability as genomic overlap-enrichment of GWAS significant SNPs over regions divided into five bins based on LECIF scores. Black bars indicate significance (p < 0.05). 2) Comparative genomics represented by boxplots showing the distribution of LECIF scores against PhatCons elements/CNEs and correlation values for LECIF versus PhyloP scores (PCC = Pearson correlation coefficient; SCC = Spearman correlation coefficient). Gray lines in boxplots denote genome-wide median and mean. 3) Chromatin states module with genome-wide (histogram) and state-specific (violinplots) LECIF scores distribution. Additionally, this module is covered by CS simmilarity between high/low (percentile rank > 60 / < 40; dark colors) and low/high (light colors) functional (LECIF) /comparative (PhyloP) genomics score regions, respectively (horizontal barplot); and between regions with low, medium and high LECIF score (lineplot). CS simmilarity was computed using the Dice coefficient. Lastly, middle panel depicted by a circos to visualize gene density (first track), scores (second to fourth track) and CS (inner track; colors indicate chromatin functional groups) across nuclear chromosomes and species. *A. thaliana* and *O. sativa* chromosomes are zoomed in to reach *Z. mays* scale. PlantFUNCO DB is available at <https://rocesv.github.io/PlantFUNCO/>.

**Fig. 6. Experimental validation of potential high diversified AOX.** From left to right degree of functional divergence (DFD) values, genic models, chromatin states and LECIF scores, when applicable, for each of the AOX paralogs evaluated. Rows represent genotypes and columns indicate distinct conditions. For each column representative images of 7 days seedlings and cotyledons after 3,3-Diaminobenzidine (DAB) staining are displayed. The white bar represents 1 cm. Furthermore, root phenotype boxplots of root length, hypocotyl length and root length: hypocotyl length ratio are presented in the bottom panel projection of the column. After two paired conditions (Control vs PEG x Heat; Mock vs Antimycin A) an additional column is added to illustrate DAB quantification intra-genotype results. The staining intensity was quantified after 32-bit gray scale transformation as: integrated density – (area selected \* mean intensity of background readings). Phenotypic differences were determined based on at least twelve biological replicates for root phenotypes and at least three biological replicates for DAB staining. A difference is considered significant with p < 0.05.

# Supplementary Material

**Supplementary fig. S1. Overview of the methods workflow.**

**Supplementary fig. S2. Genome-wide intra and inter-species correlation between epigenetic marks showing inter-species variance.**

**Supplementary fig. S3. Inter-species chromatin states description based on GO treemaps highlighting convergent terms.**

**Supplementary table S1. Epigenomic data collection.**

**Supplementary table S2. Transcriptomic data collection.**

**Supplementary table S3. Degree of functional divergence genome-wide predictions and mutants information.**

**Supplementary table S4. LECIF training and tunning.**