# Results

## Characterization of shared and species-specific chromatin states

We generated an universal chromatin states (CS) map annotation from ten common epigenomic marks using hiHMM software for the three widely-studied model plant species: *Arabidopsis thaliana*, *Oryza sativa* and *Zea mays*. We focused our analysis on a model with 16 states (**supplemental fig. S1**; see **Methods**). The states were divided into 5 functional groups (bivalent, active, divergent, repressive and quiescent/no-signal), with different levels of genome coverage, TE enrichment and overlap with other genomic features (**fig. 1**).

Co-occurrence of pairs of epigenetic marks exists between these species, but there are clearly specific patterns in both, CS and correlation analyses (**fig. 1**; **supplemental fig. S2**). Despite the diversity of the data, we found some conserved chromatin definitions such as Bivalent TSS/Promoter CS1, strongly linked to all active marks with very low enrichment in H3K27me3 and without clear presence of heavy repressive marks like 5mC and H3K9me2; and Active CS6, established in gene bodies and mainly constituted by H3K36me3, H3K4me2, H3K4me3 and H3K9ac in all the species. On the other hand, most of the states definitions strayed with some species-specific nuances at different levels, which could actually reflect our understanding of species-specific biology and how epigenomic complexity has evolved in plants. From less to more divergent: 1) States which shared genomic distribution and were constituted by marks with same roles, but covered with different marks like Heretochromatin 1 strong CS11 and Heterochromatin 2 weak CS12 (**fig. 1**). Repressive marks, also pinpointed in the correlation analysis with the highest inter-species variance (**supplemental fig. S2**), suggested two distinct types of heterochromatin across species requiring H3K27me3 for strong and H3K9me2 for weak definitions in *A. thaliana*,but not necessary in *O. sativa* and *Z. mays*. 2) Landscapes whose marks and genomic distribution gradually transitioned between species. A good case representing this could be Active weak TSS > TES CS8, mainly dominated by H3K36me3 deposition in gene bodies and TSS in *A. thaliana*,while in the two remaining species H3K4me2 is added and distribution changed towards the TES. 3) Ultimately, divergent region CS10 with a totally different mark and genomic distribution profile. CS10 corresponded to heterochromatic, bivalent and active states in *A. thaliana*, *O. sativa* and *Z. mays*, respectively.

We next performed additional annotation analyses based on non-common chromatin proteins and histone marks tracks to test our states definitions (**fig. 2, bottom panel**). There were evidence supporting our interpretation of the states for each species under study. For example: RNA polymerase II (Pol2) significantly located in all active and several bivalent states, and enrichment of the well-known H3K9-demethylase (IBM1) and transposon-methylase (CMT3) over heterochromatic states in *A. thaliana*. Essentially, all non-common active and repressive histone marks/variants evaluated were enriched in active/bivalent and heterochromatic states, respectively, with only two exceptions: H3K27me1 location in Bivalent Promoter CS2 in *A.thaliana*, which did not impact the state definition because this was already presented as bivalent due to the presence of H3K27me3; and H3K9me1/me3 in Active gradual bivalent flank > intergenic CS7 in *O. sativa*. Although the initial definition included gradual bivalent, this was only alluding to *Z. mays* as *O. sativa* CS7 was absent of any repressive mark, therefore, this would pontentially increase CS7 relation between both Poaceae-family members. We decided to stay conservative and keep our initial interpretation because H3K9me3 data is not available for all the species. It is worth mentioning that most of the transcription factors (TFs) observed in heterochromatin states were related with flowering, organ missed in our collection, and cell-cycle/division functions, previously described as present in chromatin barriers and strictly under control with low levels of expression (Feng & Michaels, 2015; Velay, Méteignier, & Laloi, 2022).

Taking advantage of the inter-species approach, we further evaluated if the states could involve evolutionary information. We observed a remarkable gradient across functional groups, excluding quiescent/no signal from the analysis due to the lack of epigenetic regulation (**fig. 3**; **supplemental fig. S3**). We found a decreasing trend in gene functional convergence (KO and GO), number of protein-coding genes and their corresponding proportion of orthologous relationships following active > bivalent > divergent > heterochromatin order (represented by CS6>CS1>CS10>CS11, respectively). Additionally, most of the PhastCons elements genomic overlaps were located in active and bivalent states (**fig. 2**). Conserved non-coding elements (CNEs) co-localization in the same states for *A. thaliana* and the greater number of CNEs enriched states when comparing both species of monocots, again showed how CS could reflect the closer distance between *O. sativa* and *Z. mays*. Even though the majority of the states enriched in Conserved TF binding-sites (BS) were active and bivalent in *A. thaliana* and *O. sativa*, we did not appreciate a constrained pattern for all the species in TF motifs and genetic variability annotation modules (**fig. 2**). On the opposite side to conservation, these results could indicate that state information is still useful, because significant overlaps were detected, but it would probably reflect species-specific features in genetic variability and TF motifs contexts.

Taking together, these discoveries introduced a plant inter-species CS single annotation as a resource to provide conservation and diversity evolutionary epigenomic information for future research.

## Chromatin states features improve predictions of paralogs functional divergence

In order to exemplify an application of the resource generated, we reproduced two previously published models predicting *A. thaliana* genetic redundancy (Cusack et al., 2021; Ezoe, Shirai, & Hanada, 2021) including CS information. As far as we know, *A. thaliana* was the only organism under study with an experimentally validated set of mutants for paralogous gene pairs which allowed the development of these models. Under the initial hypothesis that two paralogs covered by different state profiles are more likely to have divergent functions, we computed simmilarity and distance metrics between both CS profiles and fed these data to the abovementioned models (**fig. 4, top panel**; see **Methods**).

For the models developed by Ezoe, Shirai, & Hanada, 2021 (**fig. 4a-d**), we first checked if the custom chromatin state metric (CCSM; see **Methods**) proposed could be a determinant of functional divergence using the same paralogous gene pairs as the original article (**fig. 4b**). High and low CCSM values were significantly associated to high and low diversified pairs, respectively (P-value = 3.4e-15, two tailed Wilcoxon rank sum test). In spite of the fact that epigenomic features tested in the reference did not pass this threshold, our CS metric even joined the two best explanatory variables Ka/Ks and Re/Ks (see **Methods**) in terms of relative importance (**fig. 4a**). These results pointed the need to use integrative metrics when predicting genome elements. Logistic regression models (see **Methods**) using different set of features were compared by calculating the area under the curve-receiver operating characteristic (AUC-ROC) and the area under-precision recall curve (AU-PRC) values (**fig. 4c**). Models including CS information had higher AUC-ROC and AU-PRC values and slightly improved the performance of the best final model reported in the original article (Ka/Ks+Re/Ks). This improvement was more obvious in the reduced formula (Ka/Ks+Re/Ks+CCSM) and the small range of improvement between full (Ka/Ks+Re/Ks+CCSM) and reduced formulas also agreed with the information reported by the main article. The degree of functional divergence (DFD) can be inferred from the best formula by logistic regression analysis. DFD values close to 0 and 1 reflected low (<0.5) and high (>0.5) functional divergence, respectively. To enable potential validation of paralogous pairs DFD in upcoming studies and to minimize the erronous assignment of high and low diversified duplicates, we calculated 5% FDR as a threshold. DFD thresholds were 0.93 and 0.46 for high and low diversified pairs, respectively (**fig. 4d**). A table containing labeled genome-wide predictions with additional filters to assist paralogs redundancy experimental verification (see **Methods**)is available at **supplementary table S3**.

In contrast, for the models developed by Cusack et al., 2021 (**fig. 4e-h**) redundancy was categorized into different definitions and a lot of features with distinct transformations were covered. Therefore, we decided to include all the CS metrics to model redundancy for each of the definitions resulting in four different sets: RD4 (extreme redundancy, single-mutants have no abnormal phenotype and the doublé-mutant is lethal; without CS information), RD4C (with CS information), RD9 (inclusive redundancy, general definition which also contained RD4 gene pairs; without CS information) and RD9C (with CS information). The number of variables and the relative importance of the six feature categories in the definitions without CS information mostly confirmed the discoveries in the reference (**fig. 4e**). Very briefly, the ranking from best to worst based on median importance ranks in those categories for RD4/RD9-based models was functional annotation (37/16) > network properties (57.5/64.5) > evolutionary properties (76/110) > gene expression (104/105) > protein properties (145/88) > epigenetic modifications (121/127), while gene expression was the category with the highest number of variables in both cases. These findings validated the reproducibility of the models and guaranteed a rigorous interpretation of the following results. Taking into account RD4C/RD9C-based models, chromatin state category was sixth/second in importance rankings and became the first in terms of number of variables for both cases. Thus, potentially indicating that CS information would be more useful when prediciting general rather than extreme redundancy. This idea was further verified when SVM models (see **Methods**) with different sets were compared using AUC-ROC and AU-PRC values (**fig. 4f-g**). While CS data clearly improved predictions for general redundancy, it also reduced the values for the extreme definition. Finally, we detected that the intersection with the highest number of features was common to all sets suggesting that the core predicting power remained constant for all the models and, again, ensuring accurate comparisons (**fig. 4h**).

Collectively, we revealed that CS information could give clues into duplicates general functional divergence corroborated by the replication of two independent previously published models.

## Defining functional genomics conservation score and database

Evolutionary functional-genomics/epigenomics is an emerging field of study with a growing body of literature reporting massive generation of functional genomics data, yet the determinants underlying these processes are still not well understood for a lack of a holistic point of view. To fill this gap, we adopted an integrative approach and expanded the resource generated with functional genomics conservation scores computed by LECIF algorithm (Kwon & Ernst, 2021). LECIF was applied integrating epigenomic, chromatin states, whole genome alignments and transcriptomic information for all pairwise comparisons. By querying LECIF-scores, we sought to identify regions with similar phenotypic properties (**fig. 5, topleft panel**).

To research elements highlighted by LECIF, we characterized genome distribution of the scores over genetic variability, chromatin states and conservation modules. In all the comparisons, LECIF-scores density decreased in centromeres due to the lower number of alignments in these regions (**fig. 5, middle panel**). As mentioned before, we did not find a constrained pattern in the genetic variability module. Whilst both *Z. mays* contrasts (**fig. 5, topright panel**) and *O. sativa* vs *Z. mays* (**fig. 5 bottomleft panel**) GWAS significant SNPs are enriched in regions with high functional conservation, both *A. thaliana* contrasts (**fig. 5, bottomright panel**) did not reflect any enrichment and *O. sativa* vs *A. thaliana* was even enriched in regions with low LECIF-scores. This could be explained by balanced significant-SNPs distribution through *A. thaliana* genome due to its architecture and higher number of GWA studies, more simmilarity in the traits studied between the monocots and/or *O. sativa* only being able to retain functional conservation information related to the closest species.

In the CS module, genome-wide distributions were shifted to the left because of the higher weights of negative (only aligned) vs positive (aligned and functionally conserved) samples to ensure that only regions with strong functional evidence were underlined (**fig. 5, bottomright-bottomleft-topright panel; histogram**). To validate that LECIF-score displays expected cross-species similarity in functional genomics features, we examined it in relation to CS annotation. In each of the six query vs target comparisons, CS linked to strong regulatory or transcription activity tended to have higher mean LECIF-score than the other states (**fig. 5, bottomright-bottomleft-topright panel; violinplots**). We investigated cross-species CS simmilarity for different ranges of the LECIF-score (**fig. 5, bottomright-bottomleft-topright panel; lineplots**). As LECIF-score increased, cross-species CS agreement was gradually higher in active, bivalent and heterochromatin functional groups. This pattern was not fulfilled for divergent and quies/no-signal states because simmilarity was not expected by definition and the absence of epigenetic regulation, respectively. To provide further proof, we analyzed CS annotations in regions where functiona genomics (LECIF) and comparative genomics (PhyloP) scores disagreed (**fig. 5, bottomright-bottomleft-topright panel; grouped barplots**). Specifically, for pairs of regions where the LECIF-score was high (percentile-rank>60) and PhyloP-score was low (percentile-rank<40), we computed CS simmilarity. We appreciated that such pairs were more likely to exhibit convergent states for all the groups and vice versa.

We next evaluated more deeply the relationships between functional/comparative-genomics scores and annotations (**fig. 5, bottomright-bottomleft-topright panel; boxplots**). It should be noted that as we are studying distant-related species, the scores of annotations with high coverage % in the aligning regions, like PhastCons/PhyloP (Tian, Yang, Meng, Jin, & Gao, 2020) sequence-based conservation, would be influenced by the high negative:positive weights ratio. We found that regions overlapping PhastCons elements did not have greater average LECIF-score compared to the genome-wide distribution and LECIF-score was not correlated with PhyloP-score (min-max range: 0.04-0.119 and 0.005-0.118 for PCC and SCC, respectively). Interestingly, CNEs followed the same trend as PhastCons elements except for Poaceae-members vs *A. thaliana* pairs, which had higher LECIF-scores. This is reasonable since CNEs preserved during longer timescales are more probable to be functionally conserved.

In summary, all these reports suggests that plants LECIF-score can capture functional conservation without being correlated with other comparative genomics and sequence-constraint scores. We expect the LECIF-score and inter-species CS would be useful tools to unify and extrapolate molecular mechanisms discoveries using different model systems, so we developed an integrated hub called PlantFUN(ctional)CO(nservation) to provide interactive user-friendly functionalities for further requests (**fig. 5, topleft panel**; see **Methods**). PlantFUNCO database is available at <https://rocesv.github.io/PlantFUNCO/>.

## Experimental validation of potential divergent duplicates

To illustrate that functional uses of the database could be translated into solutions for complex biological problems, we focused in the experimental validation of mitochondrial alternative oxidases (AOX) redundancy in *A. thaliana*. Despite these pairs do not pass the stringent threshold (**fig. 4d**), they presented high enough DFD values to be considered high divergent paralogs (**fig. 6**). We decided to assess AOX redundancy by monitoring root phenotypes under two different stresses, considering previously described roles of these genes in response and retrograde-signalling (Fuchs et al., 2022); 2/5 paralogs are not root expressed (Papatheodorou et al., 2020), simplifying the system and evaluation in seedling stages. The DFD of duplicates can be inferred based on the phenotypes of knockout plants. When single knockout exhibit abnormal phenotypes related to the wild-type (WT, Col-0) under a specific condition, the duplicates are not compensated by the other gene copies so are assumed to be functional divergent and conversely (Ezoe, Shirai, & Hanada, 2021).

Root phenotypes followed the same pattern for control and mock conditions, there were significant differences for all AOX genotypes in root length (WT>*aox1c*>*aox1a*>*aox1d*), hypocotyl length (*aox1c*>*aox1d*>*aox1a*>WT) and root:hypocotyl ratio (WT>*aox1a/aox1c*>*aox1d*) (**fig. 6**). In PEGxheat stress, significant differences were appreciated too but with two exceptions: *aox1c* root length and *aox1a* hypoctyl length. We decided to establish an additional stress assay using Antimycin A (AA), a complex III inhibitor that can be tolerated in plants due to electon bypass via AOX, but not when the activity of these genes is supressed/dimished (Strodtkotter et al., 2009). Because of the small size of *aox1a* seedlings only root length was monitorized. Again, significant changes were found for all AOX genotypes measured in root length and root:hypoctyl ratio. Hypocotyl length greater p-values in PEGxheat and no significance in AA suggested a general-stress hyopoctyl elogation mechanism in these mutants. In view of AOX genes roles in redox state, DAB staining quantification was performed to have a raw hydrogen peroxide measure. Although both stresses agreed in WT, *aox1d* relevantincrease and *aox1c* no significance, *aox1a* trends were not congruent. Hydrogen peroxide *aox1a* change was nonmeaningul for PEGxheat while a significant increase was detected during AA. Finally, in terms of functional genomics the dominant isoform *AOX1A* seems to be the most crucial because was covered by active CS and was marked with high LECIF-scores when compared to *O. sativa*.

In brief, these findings validated our high divergence predictions and setted a sceneario where *AOX1A* appeared to retain ancestral function allowing the understanding of the remaining AOX genes redundancy relations related to this reference.

# Methods

An overview of the methods workflow used in this study is shown in **supplementary fig. S1**.

## Data collection

We collected epigenomic (ChIP-, MeDIP-, ATAC- and DNase-seq) and transcriptomic (RNA-seq) data from three plant model species: *Arabidopsis thaliana*, *Oryza sativa* and *Zea mays*.

For the epigenomic data we used the previously published collection from the PCSD (Y. Liu et al., 2018) to ensure high-quality data. Then, we expanded the abovementioned list to include new common epigenetic marks published in the last years (**supplementary table S1**).

For the transcriptomic data we used the baseline collection of the manually curated database EBI-ATLAS (Papatheodorou et al., 2020). We filtered this list to only include studies that covered multiple tissues/organs (**supplementary table S2**).

## Epigenomic data processing

Raw reads were trimmed and adapters were removed using trim\_galore v.0.6.6 as interface to CutAdapt (Martin, 2011). The remaining reads were aligned to the reference genome (*A. thaliana*: TAIR10, *O. sativa*: IRGSP-1.0, *Z. mays*: RefGen v4) using bowtie2 algorithm (Langmead & Salzberg, 2012). Mapped reads with MAPQ > 30 were used to secure optimal quality of the data. Aligned reads were sorted using SAMtools v.1.9 and duplicate reads were removed using Picard v.2.26 (<https://github.com/broadinstitute/picard>). For all the subsequent analysis we performed peak calling (narrow and broad), signal tracks building, correlation and formatting with MACS2 and deepTools (Ram et al., 2016; Zhang et al., 2008). Very briefly, the *–g* argument was changed for each species (*A. thaliana*: 91254070, *O. sativa*: 215463918, *Z. mays*: 1975365725), FDR < 0.1 was used for broad peaks calling and the arguments *--nomodel --shift 75 --extsize 150* were added for ATAC- and DNase-seq files processing. To guarantee the reproducibility of the analysis a docker was created and it is available at <https://hub.docker.com/r/rocesv/plantina-chiplike>.

## Inter-species chromatin states definition and annotation

We applied hiHMM (Sohn et al., 2015) to jointly infer multiple species chromatin states (CS) using commons marks signal tracks from several tissues as input. Signal tracks consisted in scaled log2 (fold change + 0.5) values averaged in 200 bp bins in all three species as described in the original application (Ho et al., 2014). The analysis was restricted to nuclear chromosomes. hiHMM can handle an unbounded number of hidden states so the number of states is learned from the training data instead of a pre-specified value by the user. The model inferred a total of 15 chromatin states with unmappable regions added a posteriori as the sixteenth state to avoid any bias in the segmentation. We defined the chromatin states based on the co-localization of marks and overlap enrichments of different genomic features using ChromHMM (Ernst & Kellis, 2017).

To further improve the interpretability of the states, additional annotation and description was performed. The annotation was based on significant overlap enrichments using the LOLA package (Sheffield & Bock, 2016) and was divided in: 1) Genetic variability represented by significant SNPs compiled in GWAS-ATLAS and AraGWAS (X. Liu et al., 2023; Togninalli et al., 2020). 2) Transcription factor binding motifs collected in PlantRegMap (Tian, Yang, Meng, Jin, & Gao, 2020). 3) Conservation covered by PhastCons elements in PlantRegMap and pairwise CNEs. 4) Other epigenomic features employing non-common liftovered information in PCSD. The description involved KEGG-Orthology/Gene-Ontology enrichments using clusterProfiler/REVIGO, respectively, and gene biotype-orthology correspondence using inParanoid information stored in Phytozome (Goodstein et al., 2012).

## Modelling paralogs degree of functional divergence

We reproduced two published models that predict genetic redundancy in *A. thaliana* paralogs (Cusack et al., 2021; Ezoe, Shirai, & Hanada, 2021) including our inter-species chromatin states distance metrics. To define state distance metrics, we first binned different genomic features (promoters and genes) into a fixed number of windows and computed both, presence (1 = present; 0 = absent) and frequency (% of bp covered in a window) vectors for each state and gene. Additionally, we also included a third type of vector being each element the frequency of a particular state over a non-binned genomic feature. Lastly, distinct distance metrics were calculated between genes of the same paralog pair comparing equivalent vectors using philentropy package (Drost, 2018).

To reproduce both studies we followed the workflow originally stablished for the best performing model. In brief, for the model described by Ezoe, Shirai & Hanada, 2021 feature selection was executed by two-tailed Wilcoxon rank sum test p-values between pairs labeled as redundant or divergent followed by logistic regression relative importance to examine the explanatory weights of the best variables. Due to the fact that this model is designed to perform genome-wide predictions and that only some of the distance state metrics could be informative, a small number of features is desirable. We combined the information of the best scored features into a single metric defined as custom chromatin state metric (CCSM) (**supplementary table S3**). To compare the performance of logistic regression models using different set of features we calculated the AUC-ROC and AU-PRC values. All the analysis were conducted in R software environment ([Team R Development Core 2013](javascript:;)).

On the other hand, in the model developed by Cusack et al., 2021 multiple transformations and interpretations of the same feature were included so all the distance state metrics and CCSM were considered. Only the available extreme (RD4) and inclusive (RD9) redundancy gene pair sets were analyzed deleting variables identified as mispredictors in the main article. Non-redundant gene pairs were randomly downsampled to generate balanced cross-validation sets. Feature selection was executed by random forest top 200 best transformed variables (determined by the feature importance) for sets without (RD4-RD9) and with (RD4C-RD9C) chromatin information. The C value for SVM algorithm was set as hyperparamenter during the tunning. To measure SVM performance using different feature sets we calculated AUC-ROC and AU-PRC values. All the analyses were conducted using the pipeline implemented and developed by the authors (<https://github.com/ShiuLab/ML-Pipeline>).

## Redundancy genome-wide predictions

To generate genome-wide predictions we used the best performing model from the first pipeline described above. The stringent threshold for identifying high and low diversified pairs with the logistic regression formula (DFD=degree of functional divergence) was defined by 100 cross-validation test where the FDR was under 5 %. As a result, high/low divergent pairs have >0.5/<0.5 and >0.93/<0.46 DFD values with relaxed and stringent thresholds, respectively. *A. thaliana* genes (longest sequence) were used as queries to search for self-match homologous with DIAMOND v2 (E-value=1e-04) (Buchfink, Reuter, & Drost, 2021). We only focused on pairs with the best hits, > 30 % identity and > 50 % coverage. We identified 7852 pairs of which 1444/6898 were predicted as high and 723/954 as low diversified duplicates with strict/relaxed thresholds, respectively. Ka/Ks (number of nonsynonymous/synonymous substitutions per nonsynonymous/synonymous site) and the similarity of expression patterns (Re) were calculated as described by Ezoe, Shirai & Hanada, 2021. An additional table is provided with filters such as same second closest paralog and expression in stress and seedling stages to assist experimental validation in future studies (**supplementary table S3**).

## Experimental validation of potential divergent paralogs

The *A. thaliana* T-DNA insertion line *aox1a* (SALK\_084897) was previously described as knockout and was validated by genotyping before using (Fuchs et al., 2022). We characterized *aox1c* (Sail\_420\_A04) and *aox1d* (SM\_3\_24421) insertion lines as homozygous and knockout by genotyping and RT-PCR analysis, respectively. Briefly, RNA was extracted as described by Valledor et al, 2014 and quantified by a Navi UV/Vis Nano Spectrophotometer, integrity was evaluated by agarose gel electrophoresis. cDNA was obtained from 500 ng of RNA using the RevertAid kit (ThermoFisherScientific), where random hexamers were used as primers following the manufacturer's instructions. RT-PCR analysis reported these lines as knockouts because no amplification was detected in the mutants (all primers available in **supplementary table S3**).

For stress evaluation, *aox1a*, *aox1c* and *aox1d* seeds were surface-sterilized in a 2.8 % hypochlorite solution and washed several times with sterile water; they were stratified for 3 days at 4 ºC in darkness. The in vitro culture of seeds was carried out in 12x12 plates (Greiner) containing 50 ml of MS medium, 5.8 pH, 1 % (w/v) sucrose and 0.8 % (w/v) agar and they were vertically placed under long-day photoperiod (16 h light 21 ºC, 8 h dark 18 ºC) for control conditions. To avoid a position effect, the four genotypes (Col-0 as wildtype, *aox1a*, *aox1c* and *aox1d*) were located in every plate position by rotating sectors in different plates. For the combined drought-heat stress, 2.5 % PEG8000 (ThermoFisherScientific) was added to the initial plates and seedlings were subjected to 1 h 37 ºC stress every day at the same hour, gradually increasing and decreasing temperature. For the antimycin A (AA) treatment, 50 μM AA (Sigma-Aldrich) was added to the initial plates; control conditions were setted as a mock due to AA being dissolved in ethanol. Phenotypic monitoring was conducted 5 days after germination by scanning culture plates with high-resolution scans (EpsonPerfectionV600); hypocotyl and root lengths were measured with ImageJ software (Schneider, Rasband, & Eliceiri, 2012)for at least twelve biological replicates. Furthermore 3,3-Diaminobenzidine (DAB) staining (Sigma-Aldrich) was performed 5 days after germination for at least 3 biological replicates per treatment, following the protocol described by Daudi & A. O’Brien, 2012; DAB quantification was carried out by ImageJ.

## RNA-seq data processing

Sequence quality of RNA-seq libraries was evaluated by FastQC and multiQC (Andrews, 2013; Ewels, Lundin, & Max, 2016). Raw reads were trimmed and adapters were removed using trim\_galore v.0.6.6. Cleaned reads were mapped using STAR v.2.7.10 (Dobin et al., 2013) changing reference genome and minimum/maximum intron size accordingly to species. Bigwig files were obtained using *bamCoverage* command from deepTools (Ram et al., 2016).

## Whole genome alignments and identification of conserved non-coding elements

Whole genome alignments (WGA) were computed for each pairwise comparison. In summary, *lastz* alignments with far(vs *A. thaliana*; >100 MYA according to TimeTree (Kumar et al., 2022)) and medium(*O. sativa* vs *Z. mays*; >15 & <100 MYA) *distance* arguments were performed using CNEr package interface (Tan, Polychronopoulos, & Lenhard, 2019). This was followed by format conversion, chains building and processing using lavToPsl, maf-convert, axtChain and chainMergeSort. RepeatFiller (Osipova, Hecker, & Hiller, 2019) was applied to the chains in order to improve the identification of conserved non-coding elements (CNEs). After RepeatFiller, we executed ChainCleaner (Suarez, Langer, Ladde, & Hiller, 2017) to improve alignment specificity and chains were then converted into aligments nets using Hillerlab chainNet and netToAxt. Finally, Axt files were used as input to the pairwise identification of CNEs using the CNEr package with 45-identities/50-length windows while taking into account the difference in whole genome duplications history between these species as decribed in Ren et al., 2018.

In order to take advantage of previously processed epigenetic tracks in PCSD that are not included in our initial collection (not common for all the species), we executed another WGA pipeline to liftover these files to the new reference assemblies. In summary, we used near as *distance* argument, skipped the RepeatFiller-ChainCleaner step because we aligned the same species, and the liftover was carried out using CrossMap v.0.6.2 (Hao Zhao et al., 2014). To guarantee the reproducibility of the analysis a docker was created and it is available at <https://hub.docker.com/r/rocesv/compcnes>.

## Functional genomics conservation score

LECIF algorithm (Kwon & Ernst, 2021) was applied to obtain functional genomics conservation score between all the possible pairwise comparisons integrating whole genome alignments, epigenomic, chromatin states, and transcriptomic information. The negative to positive sample weight ratio was setted to 10 because species under study are distantly related, with lower number of samples aligning but more likely to be functional conserved. For the training and evalutation we adopted the same approach as the authors based in odd and even chromosomes (**supplementary table S4**). LECIF downstream analyses were performed in R software environment ([Team R Development Core 2013](javascript:;)).

## Database resource

We developed PlantFUN(ctional)CO(nservation) database to provide public availability of the functional integrative tracks generated in this work and to facilitate future research in evolutionary functional genomics. PlantFUNCO contains three main tools: 1) Search section with interactive tables to retrieve gene- or superenhancer-level (Hainan Zhao et al., 2022) functional and comparative genomics information. 2) Shiny-application to compute LOLA genomic overlap enrichments of user query bed files over chromatin states and LECIF/PhyloP binned scores. 3) JBrowse2 genome browser (Diesh et al., 2023). PlantFUNCO is available at <https://rocesv.github.io/PlantFUNCO>.

# Data availability

All the data generated in this study is available at PlantFUNCO database <https://rocesv.github.io/PlantFUNCO> and <https://zenodo.org/record/7852329>. All the code used in this work is available at <https://github.com/RocesV/PlantFUNCO_manuscript>.

# Acknowledgments

This work is generously financed by the Spanish Ministry of Science, Innovation and Universities (PID2020-113896GB-I00). VR and AA are supported by a fellowship from Spanish Ministry of Universities (FPU18/02953 and FPU19/01142, respectively). SG is supported by Severo Ochoa Predoctoral Program (BP19-145). JP is supported by Juan de la Cierva Incoporación Programme (IJC-2019-040330-I).

# Conflict of interest

The authors declare there is no conflict of interest.

# Author’s contributions

VR and MM conceived the study. VR designed the research. VR and AA collected the data and built the figures. SG performed all mutant generation, validation and stress experiments. VR performed computational analyses, analyzed-interpreted the data and wrote the manuscript. JP and MM supervised the study. All authors revised, read, and approved the final manuscript.

# References

Andrews, S. (2013). Babraham Bioinformatics -FastQC A Quality Control tool for High Throughput Sequence Data.

Buchfink, B., Reuter, K., & Drost, H. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, *18*(April). https://doi.org/10.1038/s41592-021-01101-x

Cusack, S. A., Wang, P., Lotreck, S. G., Moore, B. M., Meng, F., Conner, J. K., … Shiu, S. (2021). Predictive Models of Genetic Redundancy in Arabidopsis thaliana. *Molecular Biology and Evolution*, *38*(8), 3397–3414. https://doi.org/10.1093/molbev/msab111

Daudi, A., & A. O’Brien, J. (2012). Detection of Hydrogen Peroxide by DAB Staining in Arabidopsis Leaves. *Bio Protoc.*, *2*(18), 4–7.

Diesh, C., Stevens, G. J., Xie, P., Martinez, T. D. J., Hershberg, E. A., Leung, A., … Holmes, I. H. (2023). JBrowse 2 : a modular genome browser with views of synteny and structural variation. *Genome Biology*, 1–21. https://doi.org/10.1186/s13059-023-02914-z

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Drost, H. (2018). Philentropy : Information Theory and Distance Quantification with R. *The Journal of Open Source Software*, *1*, 1–4. https://doi.org/10.21105/joss.00765

Ernst, J., & Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. *Nature Publishing Group*, *12*(12), 2478–2492. https://doi.org/10.1038/nprot.2017.124

Ewels, P., Lundin, S., & Max, K. (2016). MultiQC : summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(June), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Ezoe, A., Shirai, K., & Hanada, K. (2020). Degree of Functional Divergence in Duplicates Is Associated with Distinct Roles in Plant Evolution. *Molecular Biology and Evolution*, *38*(4), 1447–1459. <https://doi.org/10.1093/molbev/msaa302>

Feng, W., & Michaels, S. D. (2015). Accessing the Inaccessible : The Organization , Transcription , Replication , and Repair of Heterochromatin in Plants. *Annual Review of Genetics*, *49*, 439–459. https://doi.org/10.1146/annurev-genet-112414-055048

Fuchs, P., Bohle, F., Lichtenauer, S., Ugalde, M., Araujo, E. F., Mansuroglu, B., … Schwarzla, M. (2022). Reductive stress triggers ANAC017-mediated retrograde signaling to safeguard the endoplasmic reticulum by boosting mitochondrial respiratory capacity. *The Plant Cell*, *34*, 1375–1395.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., … Rokhsar, D. S. (2012). Phytozome : a comparative platform for green plant genomics. *Nucleic Acids Research*, *40*(November 2011), 1178–1186. https://doi.org/10.1093/nar/gkr944

Ho, J. W. K., Jung, Y. L., Liu, T., Alver, B. H., Lee, S., Ikegami, K., … Park, P. J. (2014). Comparative analysis of metazoan chromatin organization. *Nature*, *512*(7515), 449–452. https://doi.org/10.1038/nature13415

Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., … Hedges, S. B. (2022). TimeTree 5 : An Expanded Resource for Species Divergence Times. *Molecular Biology and Evolution*, *39*(8), 1–6. https://doi.org/10.1093/molbev/msac174

Kwon, S. Bin, & Ernst, J. (2021). Learning a genome-wide score of human–mouse conservation at the functional genomics level. *Nature Communications*, *12*, 2495. https://doi.org/10.1038/s41467-021-22653-8

Liu, X., Tian, D., Li, C., Tang, B., Wang, Z., Zhang, R., … Song, S. (2023). GWAS Atlas : an updated knowledgebase integrating more curated associations in plants and animals. *Nucleic Acids Research*, *51*(October 2022), 969–976.

Liu, Y., Tian, T., Zhang, K., You, Q., Yan, H., Zhao, N., … Xu, W. (2018). PCSD : a plant chromatin state database. *Nucleic Acids Research*, *46*(October 2017), 1157–1167. https://doi.org/10.1093/nar/gkx919

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10–12.

Osipova, E., Hecker, N., & Hiller, M. (2019). RepeatFiller newly identifies megabases of aligning repetitive sequences and improves annotations of conserved non-exonic elements. *GigaScience*, *8*, 1–10. https://doi.org/10.1093/gigascience/giz132

Papatheodorou, I., Moreno, P., Manning, J., George, N., Fexova, S., Fonseca, N. A., … Brazma, A. (2020). Expression Atlas update : from tissues to single cells Anja F ullgrabe. *Nucleic Acids Research*, *48*(October 2019), 77–83. https://doi.org/10.1093/nar/gkz947

Ram, F., Ryan, D. P., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., … Manke, T. (2016). deepTools2 : a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, *44*(April), 160–165. https://doi.org/10.1093/nar/gkw257

Ren, R., Wang, H., Guo, C., Zhang, N., Zeng, L., Chen, Y., … Qi, J. (2018). Widespread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms. *Molecular Plant*, *11*, 414–428. https://doi.org/10.1016/j.molp.2018.01.002

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ : 25 years of Image Analysis. *Nature Methods*, *9*(7), 671–675.

Sheffield, N. C., & Bock, C. (2016). LOLA : enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics*, *32*(October 2015), 587–589. https://doi.org/10.1093/bioinformatics/btv612

Sohn, K. A., Ho, J. W. K., Djordjevic, D., Jeong, H. H., Park, P. J., & Kim, J. H. (2015). HiHMM: Bayesian non-parametric joint inference of chromatin state maps. *Bioinformatics*, *31*(13), 2066–2074. https://doi.org/10.1093/bioinformatics/btv117

Strodtkotter, I., Padmasreea, K., Dinakara, C., Spetha, B., Niazi, P. S., Wojtera, J., … Scheibe, R. (2009). Induction of the AOX1D Isoform of Alternative Oxidase in A . thaliana T-DNA Insertion Lines Lacking Isoform AOX1A Is Insufficient to Optimize Photosynthesis when Treated with Antimycin A. *Molecular Plant*, *2*(2). https://doi.org/10.1093/mp/ssn089

Suarez, H. G., Langer, B. E., Ladde, P., & Hiller, M. (2017). ChainCleaner improves genome alignment specificity and sensitivity. *Bioinformatics*, *33*(January), 1596–1603. https://doi.org/10.1093/bioinformatics/btx024

Velay, F., Méteignier, L.-V., & Laloi, C. (2022). You shall not pass ! A Chromatin barrier story in plants. *Frontiers in Plant Science*, *13*(September), 1–9. <https://doi.org/10.3389/fpls.2022.888102>

# Figure legends

**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, Shirai, & Hanada, 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 (violinplot) 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 grouped 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. Coverage (%) referes to the aligning regions overlap. 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 5 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: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.**