# Methods

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

## Data collection

We initially collected epigenomic (ChIP-, MeDIP-, ATAC- and DNase-seq) and transcriptomic (RNA-seq) data for three model plant 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. The epigenomic final collection included 10 common marks represented by 130, 63 and 96 files for *A.thaliana*, *O.sativa* and *Z.mays*, respectively (**supplementary table S1, Supplementary Material online**).

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. The transcriptomic final collection contained 118, 43 and 411 files for *A.thaliana*, *O.sativa* and *Z.mays*, respectively (**supplementary table S2, Supplementary Material online**).

## Epigenomic data processing

Raw reads were trimmed and adapters 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 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. While this flexible framework provides a consistent definition of chromatin states across multiple genomes, thus making easier direct comparison between them, the “full-stack” approach allows the understanding of the potential epigenomic regulation over several tissues/conditions such as differentiating constitutively active regions (Vu & Ernst, 2022). Therefore, we adopted this holistic approach simplifying genome annotations across tissues and species through a single segmentation annotation to allow future evolutionary epigenomics applications.

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 *OverlapEnrichment* in 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 thirdth 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, Supplementary Material online)**. 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 RD4 and RD9 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 in the trained models) 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 focused only in 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 (number of nonsynonymous substitutions per nonsynonymous site), Ks values (number of synonymous substitutions per 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 other future studies (**supplementary table S3, Supplementary Material online)**.

## Experimental validation of potential divergent paralogs

To illustrate that potential functional applications of the resources generated could reveal clues into complex biological problems, we focused in the experimental validation of mitochondrial alternative oxidases (AOX). Despite these pairs do not pass the stringent threshold, they presented high enough DFD values to be considered high divergent paralogs. Furthermore, we decided to asses AOX redundancy in roots phenotypes (because 2/5 paralogs are not expressed simplying the system and it is easier to monitorize in seedling stages) under two different stresses considering previously described roles of these genes in response and retrograde-signalling (cita).

The *A.thaliana* T-DNA insertion line *AOX1A* (SALK\_08489) was previously described as knockout and was validated by genotyping before using. We characterized *AOX1C* (Sail\_420\_A04) and *AOX1D* (SM\_3\_24421) insertion lines as homozygous by genotyping. Then, RNA was extracted as described by Valledor et al, 2014, quantified in a Navi UV/Vis Nano Spectrophotometer, integrity was evaluated by agarose gel electrophoresis and DNA contamination was checked by PCR employing *GLYCERALDEHYDE-3-212 PHOSPHATE DEHYDROGENASE* (*GADPH*) primer pair (all primers available in **supplementary table S3, Supplementary Material online**). cDNA was obtained from 500 ng of RNA using the RevertAid kit (ThermoFisher Scientific), 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.

Surface-sterilized, stratificated seeds were grown under long-day conditions (16 h light 22ºC, 8 h dark 18 ºC) on vertical MS medium plates in control conditions. For the combined drought x heat stress, 2.5 % PEG was added to the initial plates and seedlings were subjected to 1 h 37 ºC stress every day at the same hour. For the antimycin A (AA) treatment, mock was setted as control because AA was dissolved in ethanol and 50 μM AA was added to the initial plates. After 7 days, image phenotypic analyses were conducted, measuring hypocotyl and root length for at least twelve biological replicates and 3,3-Diaminobenzidine (DAB) staining (Daudi & A. O’Brien, 2012) for at least 3 biological replicates with ImageJ software (Schneider, Rasband, & Eliceiri, 2012). All the plates contained the four genotypes (*Col-0* as *WT*, *AOX1A*, *AOX1C* and *AOX1D*) and each genotype was represented in each sector of the plate twice.

## 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 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 aligmnets (WGA) were computed for each pairwise comparison of these three species. 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 chainNet (modified version that computes real score of partial nets) and netToAxt. Finally, Axt files were used as input to the pairwise identification of CNEs using the CNEr package with 45 identities over 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 and 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 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, Supplementary Material online**).

LECIF downstream analyses were performed in R software environment ([Team R Development Core 2013](javascript:;)) and were divided intro three modules: 1) Genetic variability as LOLA genomic enrichments of GWAS significant SNPs over regions divided into five bins based on LECIF scores. 2) Comparative genomics represented by distribution and correlation of LECIF scores against PhatCons elements and PhyloP scores (Tian et al., 2020). 3) Chromatin states with LECIF scores distribution for each state and CS simmilarity between high/low (percentile rank > 60 / < 40) and low/high functional (LECIF) /comparative (PhyloP) genomics scores regions, respectively, and between regions with low, medium and high LECIF score. CS simmilarity was computed using the Dice coefficient. We used circlize (Gu, Gu, Eils, Schlesner, & Brors, 2014) to visualize LECIF score distribution across genomes.

## Database resource

We developed PlantFUN(ctional)CO(nservation) database to provide public availability to 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>.

# 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

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

Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). circlize implements and enhances circular visualization in R. *Bioinformatics*, *30*(19), 2811–2812. https://doi.org/10.1093/bioinformatics/btu393

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

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

Tan, G., Polychronopoulos, D., & Lenhard, B. (2019). CNEr : A toolkit for exploring extreme noncoding conservation. *PLoS Computational Biology*, *15*((8)), 1–16.

Tian, F., Yang, D., Meng, Y., Jin, J., & Gao, G. (2020). PlantRegMap : charting functional regulatory maps in plants. *Nucleic Acids Research*, *48*(November 2019), 1104–1113. https://doi.org/10.1093/nar/gkz1020

Togninalli, M., Freudenthal, J. A., Monroe, J. G., Meng, D., Nordborg, M., Weigel, D., … Grimm, G. (2020). AraPheno and the AraGWAS Catalog 2020 : a major database update including RNA-Seq and knockout mutation data for Arabidopsis thaliana. *Nucleic Acids Research*, *48*(October 2019), 1063–1068. https://doi.org/10.1093/nar/gkz925

Valledor, L., Escandón, M., Meijón, M., Nukarinen, E., Cañal, M. J., & Weckwerth, W. (2014). A universal protocol for the combined isolation of metabolites, DNA, long RNAs, small RNAs, and proteins from plants and microorganisms. *Plant Journal*, *79*(1), 173–180. https://doi.org/10.1111/tpj.12546

Vu, H., & Ernst, J. (2022). Universal annotation of the human genome through integration of over a thousand epigenomic datasets. *Genome Biology*, *23*(9), 1–37.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., … Liu, X. S. (2008). Open Access Model-based Analysis of ChIP-Seq ( MACS ). *Genome Biology*, *R137*(9). https://doi.org/10.1186/gb-2008-9-9-r137

Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J., & Wang, L. (2014). CrossMap : a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, *30*(7), 1006–1007. https://doi.org/10.1093/bioinformatics/btt730

Zhao, H., Yang, M., Bishop, J., Teng, Y., Cao, Y., Beall, B. D., … Fang, C. (2022). Identification and functional validation of super-enhancers in Arabidopsis thaliana. *PNAS*, *119*(48), 1–11. https://doi.org/10.1073/pnas.