**PlantFUNCO: integrative functional genomics database reveals clues into duplicate divergence evolution**

Víctor Roces1, Sara Guerrero1, Ana Álvarez1, Jesús Pascual1\*, Mónica Meijón1\*

1 Plant Physiology, Department of Organisms and Systems Biology, Faculty of Biology and Biotechnology Institute of Asturias, University of Oviedo, Asturias, Spain

\* Co-correspondence: [meijonmonica@uniovi.es](mailto:meijonmonica@uniovi.es), [pascualjesus@uniovi.es](mailto:pascualjesus@uniovi.es)

# Abstract

Evolutionary epigenomics and, more generally, evolutionary functional genomics are emerging fields that study how non-DNA-encoded alterations in gene expression regulation are an important form of plasticity and adaptation. Previous evidence analysing plants’ comparative functional genomics has mostly focused on comparing same assay-matched experiments, missing the power of heterogeneous datasets for conservation inference. To fill this gap, we developed the PlantFUN(ctional)CO(nservation) database, which is constituted by several tools and two main resources: interspecies chromatin states and functional genomics conservation scores. In this work, we presented and analysed three well-established plant models (*Arabidopsis thaliana*, *Oryza sativa* and *Zea mays*). Overall, PlantFUNCO elucidated evolutionary information in terms of cross-species functional agreement. Therefore, this study provides a new complementary comparative genomics source for assessing evolutionary studies. To illustrate the potential applications of this database, we replicated two previously published models predicting genetic redundancy in *A. thaliana* and found that chromatin states are a determinant of paralogs’ degree of functional divergence. These predictions were validated based on the phenotypes of mitochondrial alternative oxidase knockout mutants under two different stressors. Considering these results, PlantFUNCO utilises data diversity to extrapolate molecular mechanism findings from different model organisms and determine the extent of functional conservation, thus deepening our understanding of how plant epigenomics and functional genomics have evolved. PlantFUNCO is available at <https://rocesv.github.io/PlantFUNCO>.

**Keywords**: evolutionary epigenomics, functional genomics, integrative approach, database, paralogs.

## Introduction

A fundamental question in biology is how complex patterns of gene expression are determined to explain different phenotypes (Schmitz, Grotewold, and Stam, 2022; Marand et al., 2023). Today, it is largely known that genome function is dynamically regulated in part by chromatin organisation, which consists of histones, non-histone proteins and RNA molecules that package DNA (Ho et al., 2014). In this sense, the generation of comprehensive chromatin state (CS) maps, defined as the homogeneous coexistence of multiple chromatin modifications at the whole genome level, provides valuable information for annotating coding and non-coding genome features, including the identification of various types of regulatory elements. CSs can facilitate our understanding of regulatory elements and variants associated with core life processes, such as development and disease and stress responses (Liu et al., 2018). Great efforts have been made by the plant research community to contribute to the comprehension of chromatin mechanisms using different models (Zhao et al., 2020; Jamge et al., 2023); nevertheless, universal annotation allowing the extrapolation and unification of earlier conclusions across species/conditions still needs to be addressed.

Evolutionary theory has been dominated by the idea that selection proceeds by changes in allele frequencies within and between populations and mutations that occur randomly with respect to their consequences. The last theoretical and experimental advances in the field point to phenotypic plasticity as an adaptative trait subjected to natural selection; therefore, similar genotypes that differently develop appropriate phenotypes without sequence changes are equally responsible for evolutionary changes (Ashe, Colot, and Oldroyd, 2021; Monroe et al., 2022). This brings us to evolutionary epigenomics and, more generally, evolutionary functional genomics, which are emerging fields evaluating how alterations in the conservation of epigenome regulators and cytosine methylation over multiple generations represent a crucial form of plasticity and epigenetic adaptation. Regulatory element states have begun to be regarded as major targets of evolution, given that their diversity plays a critical role in phenotypic variance across all organisms, enabling them to adapt to various environmental niches (Yocca and Edger, 2022). Although relevant research in plants has lagged behind animal species (Schmitz et al., 2022), some of the most controversial findings in evolutionary biology use plants as model species; for example, mutations occur less often in functionally constrained regions, and epimutations are located in hotspots with specific chromatin features (Hazarika et al., 2022; Monroe et al., 2022). These findings support the clear importance of the plant kingdom in evolutionary functional genomics. Plants present a series of interesting molecular features that allow same sequence, different function scenarios; for instance, cytosine methylation is more easily transgenerationally transmitted due to soft epigenetic reset during meiosis and early development. Epialleles are quite common and have a relatively high rate of duplication events; thus, multiple original exact gene copies with distinct selection pressures in response to the environment may exist (Ashe et al., 2021; Cusack et al., 2021). Many comparative genomics studies have identified sequence-conserved loci of interest across a wide range of species, and their functions are determined by evaluating their homologues in a single model organism. In this context, a maze of opportunities and challenges appears to systematically and confidently determine the extent of conservation at the functional genomics level between model species (Kwon and Ernst, 2021).

Previous evidence analysing comparative functional genomics has mostly focused on comparing same assay-matched experiments (Maher et al., 2018; Lu et al., 2019). These works have been crucial for the in-depth study of molecular machinery but lack the power of diverse datasets for conservation inference. In contrast to this narrow but deep knowledge bottleneck, we adopted a broad but shallow approach using heterogeneous functional genomics to search directly for simple large-scale answers that we would never have contemplated asking based on our understanding of single assay/species information (Kliebenstein, 2019). In the current Earth Biogenome era, an increasing number of genomes and functional tracks are becoming available (Expósito-Alonso et al., 2020), thus highlighting the urge to use integrative tools that consider the vast diversity of biological strategies and enable wide genomic element characterisation. Considering the abovementioned knowledge trade-off, in the present study, we introduced PlantFUN(ctional)CO(nservation), an integrative functional genomics database constituted by several tools and two main resources, interspecies CSs and functional genomics conservation scores, for the well-known plant models *Arabidopsis thaliana*, *Oryza sativa* and *Zea mays*. To illustrate how the results derived from the generated resources could be functionally relevant, we developed an application for the database and found that CS information improved the paralogous degree of functional divergence predictions. Lastly, we validated the redundancy predictions based on the phenotypic effects of alternative oxidase (AOX) gene knockout mutants under several stressors to provide insights into the evolution of these genes.

# Results

## Characterisation of shared and species-specific chromatin states

We generated a universal CS map annotation from 10 common chromatin modifications (greatest number of tracks found simultaneously available) (**Supplementary Fig. S1**) using hiHMM software for three widely studied model plant species: *A. thaliana*, *O. sativa* and *Z. mays*. We focused our analysis on a model with 16 CSs (see **Methods**). In turn, 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**).

The co-occurrence of chromatin modification pairs exists between these species, but there are clearly specific patterns in both CSs and correlation analyses (**Fig. 1**; **Supplementary Fig. S2**). Despite the diversity of data, we found some conserved chromatin definitions, such as Bivalent TSS/Promoter CS1, which is strongly linked to all active marks with very low enrichment in H3K27me3 and without the clear presence of heavy repressive marks, such as 5mC and H3K9me2; and Active CS6, which is established in gene bodies and mainly constituted by H3K36me3, H3K4me2, H3K4me3 and H3K9ac in all species. However, many CS definitions exhibit species-specific nuances at different levels, which reflect how epigenomic complexity has evolved in plants. The various degrees of CS divergence were determined based on the CS chromatin modification composition (**Fig. 1**, **top panel**) and genomic distribution (**Fig. 1**, **bottom panel**). Ranging from less to more divergent: 1) states shared a genomic distribution and were constituted by chromatin modifications with the same roles but covered with different chromatin modifications, such as Heretochromatin 1 strong CS11 and Heterochromatin 2 weak CS12 (**Fig. 1**). Repressive modifications, which were also pinpointed in the correlation analysis with the highest interspecies variance (**Supplementary Fig. S2**), suggested two distinct types of heterochromatin across species, requiring H3K27me3 for strong and H3K9me2 for weak definitions in *A. thaliana*. However,they were not necessary in *O. sativa* or *Z. mays*. Additionally, there were 2) landscapes whose chromatin modifications and genomic distribution gradually transitioned between species. A good case representing this is 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 the distribution changed towards the TES. 3) Ultimately, the divergent region CS10 had totally different chromatin modifications and genomic distribution profiles. 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-binding proteins and histone mark tracks for all species under study to test our state definitions (**Fig. 2**). There was evidence supporting our interpretation of the states for each species under study. For example, RNA polymerase II (Pol2) was significantly located in all active and several bivalent states, and there was enrichment of the well-known H3K9-demethylase (IBM1) and transposon-methylase (CMT3) over heterochromatic states in *A. thaliana*. Most of the transcription factors (TFs) observed in heterochromatin states were related to flowering, an organ missed in our collection, and cell cycle/division functions, which have been previously described as present in chromatin barriers and strictly under control, with low expression levels (Feng and Michaels, 2015; Velay, Méteignier, and Laloi, 2022). 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 located 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 only alluded to *Z. mays,* as *O. sativa* CS7 was absent of any repressive mark; therefore, this could potentially increase the CS7 relationship between both Poaceae family members. We decided to be conservative and maintain our initial interpretation because H3K9me3 data were not available for all species.

Taking advantage of the interspecies approach, we further evaluated whether 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**; **Supplementary Figs. S3 and S4**; **Supplementary Table S1**). A decreasing trend in gene functional convergence (KO and GO) and the proportion of orthologous relationships was identified, following the order active > bivalent > heterochromatin, illustrated by CS6 > CS1 > CS11, respectively (the first state of each functional group was selected for representation). CS10 represented a divergent state corresponding to heterochromatic, bivalent and active states in *A. thaliana*, *O. sativa* and *Z. mays*, respectively. Additionally, most of the PhastCons elements’ genomic overlaps were located in the active and bivalent states (**Fig. 4**). Conserved non-coding elements (CNEs) co-localisation in the same states for *A. thaliana* and the greater number of CNE-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 most of the states enriched in conserved TF binding sites (BS) were active and bivalent in *A. thaliana* and *O. sativa*, we did not observe a constrained pattern in TF motifs and genetic variability annotation modules for any of the species (**Fig. 4**). In opposition to conservation, these results could indicate that CS information is still useful because significant overlaps were detected, but it probably reflects species-specific features in genetic variability and TF motif contexts.

Taken together, these discoveries introduce a single plant interspecies CS annotation as a resource to provide conservation and diversity evolutionary epigenomic information for future research.

## Chromatin state features improve predictions of paralogs’ functional divergence

To exemplify an application of the generated resource, we reproduced two previously published models predicting *A. thaliana* genetic redundancy (Cusack et al., 2021; Ezoe, Shirai, and Hanada, 2021), including CS information, to determine which of the feature categories (such as evolutionary properties, gene expression patterns, protein sequence properties, epigenetic modification, and CS) could be relevant regulators of paralogs’ functional divergence. To the best of our knowledge, *A. thaliana* is 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 similarity and distance metrics between both CS profiles and fed these data to the abovementioned models (**Fig. 5A**; see **Methods**).

For the models developed by Ezoe, Shirai, and Hanada (2021) (**Fig. 5B–E**), we first checked whether 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. 5B**). High and low CCSM values were significantly associated with high and low diversified pairs, respectively (P-value = 3.4e-15, two-tailed Wilcoxon rank sum test). Despite the epigenomic features tested in the reference that did not pass this threshold, our CS metric even joined the two best explanatory variables Ka/Ks (protein divergence rate) and Re/Ks (gene expression similarity rate) in terms of relative importance (**Fig. 5C**; see **Methods**). These results indicate the need to use integrative metrics when predicting genome elements. Logistic regression models (see **Methods**) using different sets 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. 5D**). 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+FD+PPI+GO) 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 the potential validation of paralogous pairs of DFD in upcoming studies and to minimise the erroneous assignment of high and low diversified duplicates, we calculated 5% FDR as a threshold. DFD stringent thresholds were 0.93 and 0.46 for high and low diversified pairs, respectively (**Fig. 5E**). **Supplementary Table S3** contains labelled genome-wide predictions with additional filters to assist paralog redundancy experimental verification (see **Methods**).

In contrast, the models developed by Cusack et al. (2021) (**Fig. 5F–I**) categorised redundancy into different definitions, covering a plethora of features with distinct transformations. Consequently, we opted to incorporate all CS metrics to model redundancy for each definition, resulting in four different sets: RD4 (extreme redundancy, where single mutants have no abnormal phenotype, and the double mutant is lethal; without CS information), RD4C (with CS information), RD9 (inclusive redundancy, general definition that also included RD4 gene pairs; without CS information) and RD9C (with CS information). Analysis of models without CS information (RD4 and RD9) revealed that the number of variables and the relative importance of the six feature categories largely corroborated the discoveries in the reference (**Fig. 5F**). In summary, the ranking from best to worst, based on median importance ranks in those categories for RD4/RD9-based models (without CS information), 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), with gene expression being the category with the highest number of variables in both cases. These findings validated the reproducibility of the models and ensured rigorous interpretation of subsequent results. Considering RD4C/RD9C-based models (with CS information), the CS feature category was sixth/second in importance rankings and emerged as the first in terms of the number of variables for both cases. This suggests that CS information is more valuable when predicting general (RD9 definition gene pairs) than extreme redundancy (RD4 definition gene pairs). This notion was further verified when comparing SVM models (see **Methods**) with different sets using AUC-ROC and AU-PRC values (**Fig. 5G,H**). While CS data notably improved predictions for general redundancy (RD9C vs RD9, AUC-ROC = 0.665 vs 0.634, AU-PRC = 0.651 vs 0.603), it also reduced the values for the extreme definition (RD4C vs RD4, AUC-ROC = 0.807 vs 0.842, AU-PRC = 0.795 vs 0.825). Finally, we observed that the intersection with the highest number of features was common to all sets, suggesting that the core predicting power remained constant for all models, thereby ensuring accurate comparisons between all mentioned models (**Fig. 5I**).

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 scores and the database

Evolutionary functional (epi)genomics is an emerging field of study with a growing body of literature reporting the massive generation of functional genomics data; however, the determinants underlying these processes are still not well understood due to 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 the LECIF algorithm (Kwon and Ernst, 2021). LECIF was applied to integrate epigenomic, CSs, whole genome alignments and transcriptomic information for all pairwise comparisons. By querying the LECIF scores, we sought to identify genomic regions with a high degree of functional track convergence and, therefore, similar phenotypic properties (**Fig. 6A**).

To research elements highlighted by LECIF, we characterised the genome distribution of the scores over genetic variability, CSs and conservation modules. In all comparisons, the LECIF score density decreased in centromeres due to the lower number of alignments in these regions (**Supplementary Fig. S5**). As mentioned previously, we did not find a constrained pattern in the genetic variability module. Although both *Z. mays* contrasts and *O. sativa* vs *Z. mays* genome-wide association study (GWAS)-significant single nucleotide polymorphisms (SNPs) were enriched in regions with high functional conservation, neither *A. thaliana* contrast reflected any enrichment and *O. sativa* vs *A. thaliana* was enriched in regions with low LECIF scores (**Fig. 6B–D; barplots**). This could be explained by a balanced significant SNP distribution in the *A. thaliana* genome due to its architecture and higher number of GWASs, more similarity 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. 6E–G; histograms**). To validate that the LECIF score displays the 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 a higher mean LECIF score than the other states (**Fig. 6E–G; violin plots**). We investigated cross-species CS similarity for different ranges of the LECIF score (**Fig. 6E–G; line plots**). As the LECIF score increased, cross-species CS agreement was gradually higher in the active, bivalent and heterochromatin functional groups. This pattern was not fulfilled for divergent and quies/no signal states because similarity was not expected by definition and the absence of epigenetic regulation, respectively. To provide further proof, we analysed CS annotations in regions where functional genomics (LECIF) and comparative genomics (PhyloP) scores disagreed (**Fig. 6E–G; horizontal grouped bar plots**). Specifically, for pairs of regions where the LECIF score was high (percentile rank > 60) and the PhyloP score was low (percentile rank < 40), we computed CS similarity. We appreciate that such pairs were more likely to exhibit convergent states for all groups and vice versa.

Next, we evaluated the relationships between functional/comparative genomics scores and annotations more deeply (**Fig. 6H–J; box plots**). As we studied distantly related species, the scores of annotations with a high coverage percentage in aligning regions, such as PhastCons/PhyloP (Tian, Yang, Meng, Jin, and Gao, 2020) sequence-based conservation, would be influenced by the high negative:positive weights ratio. We found that regions overlapping the PhastCons elements did not have a greater average LECIF score compared to the genome-wide distribution, and the LECIF score was not correlated with the 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 likely to be functionally conserved.

In summary, these reports suggest that plant LECIF scores can capture functional conservation without being correlated with other comparative genomics and sequence constraint scores. We expect the LECIF score and interspecies CS to be useful tools for unifying and extrapolating molecular mechanism discoveries using different model systems; thus, we developed an integrated hub called PlantFUN(ctional)CO(nservation) to provide interactive user-friendly functionalities for further requests (**Fig. 6A**; see **Methods**). The PlantFUNCO database is available at <https://rocesv.github.io/PlantFUNCO/>.

## Experimental validation of potential divergent duplicates

To illustrate that the functional uses of the database could be translated into solutions for complex biological problems, we focused on the experimental validation of mitochondrial alternative oxidase (AOX) redundancy in *A. thaliana*. Although these pairs did not pass the stringent threshold (>0.93/<0.46; **Fig. 5E**), they presented high enough DFD values to be considered high divergent paralogs (*AOX1A*–*AOX1C*: 0.77, *AOX1A*–*AOX1D*: 0.72, *AOX1C*–*AOX1D*: 0.89; **Fig. 7**). We assessed AOX redundancy by monitoring root phenotypes under two stressors, considering the previously described roles of these genes in response and retrograde signalling (Fuchs et al., 2022). Two of the five paralogs are not root expressed (Papatheodorou et al., 2020), simplifying the system and evaluating the seedling stages. The DFD of duplicates can be inferred based on the phenotypes of knockout plants. When single knockouts exhibit abnormal phenotypes related to the wild type (WT, Col-0) under a specific condition, the duplicates are not compensated for by the other gene copies; thus, they are assumed to be functionally divergent (Ezoe, Shirai, and Hanada, 2021).

Seedling phenotypes followed the same pattern for the 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. 7**). Under drought/heat (PEG×Heat) stress, significant differences were also observed, with two exceptions: *aox1c* root length and *aox1a* hypocotyl length. We established an additional stress assay using Antimycin A (AA), a complex III inhibitor that can be tolerated in plants due to electron bypass via AOX, but not when the activity of these genes is supressed/diminished (Strodtkotter et al., 2009). Only root length was monitored because of the small size of *aox1a* seedlings. Again, significant changes were found for all AOX genotypes measured in root length and root:hypocotyl ratio. The greater p-values for hypocotyl length in drought/heat and no significance in AA suggest a general stress hypocotyl elongation mechanism in these mutants. In view of the roles of the AOX gene in the redox state, DAB staining quantification was performed to measure hydrogen peroxide levels. Although both stressors agreed in the WT, *aox1d* showed a relevantincrease, *aox1c* was not significant and *aox1a* trends were not congruent. In *aox1a*, the hydrogen peroxide content change was not meaningful for drought/heat, 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 it was covered by active CS and marked with high LECIF scores compared to *O. sativa*.

In brief, these findings validated our high divergence predictions and set a scenario in which *AOX1A* appeared to retain the ancestral function, allowing the understanding of the remaining AOX gene redundancy in relation to this reference.

## Discussion

We introduced PlantFUNCO, a database that allows for further inspection of the crosstalk between evolution and epigenomics/functional genomics. This database is derived from two resources presented and analysed in this work for three well-established plant models. We generated interspecies CS using hiHMM (**Fig. 1**). While this flexible framework provides a consistent definition of CS across multiple genomes, making the extrapolation of intraspecies analyses between them easier, the stack approach allows for an understanding of the potential epigenomic regulation over several tissues/conditions, such as differentiating constitutively active/repressive regions (Vu and Ernst, 2022). CS links with different types of evolutionary information set a foundation for the epigenomics interspecies perspective (**Figs. 3 and 4**; **Supplementary Figs. S3 and S4**). All approaches have trade-offs; thus, this resource should be considered complementary to and not a replacement for other single species/condition annotations. We obtained functional genomic conservation scores using LECIF. In accordance with the abovementioned framework, LECIF can handle very diverse datasets and take advantage of them to quantify functional conservation. Plant LECIF scores elucidated functional genomics cross-species agreement without being correlated with other comparative genomics sources (**Fig. 6**). This reflects a complementary side of evolution. Despite the greater divergence between plant models compared to metazoans (Ho et al., 2014; Kwon and Ernst, 2021), both resource results are congruent with a higher plant epigenomic/functional complexity probed by more states with species-specific features and lower LECIF scores.

A major focus of this study was to illustrate the application of the generated resources. Due to the holistic approach adopted and exploiting that our interspecies CS could differ between constitutively active/repressive regions, we replicated two previously published models predicting paralogous functional divergence in *Arabidopsis* (Cusack et al., 2021; Ezoe et al., 2021), including our CS information. We determined whether CS similarity could be a determinant of duplicates’ degree of functional divergence under the initial hypothesis that two paralogs covered by different state profiles are more likely to present distinct functions. Although the models are far from perfect, useful information about gene features can be extrapolated. These models independently reported CS information as relevant, and including this type of data improved general redundancy predictions (**Fig. 5**). This shows an example of how PlantFUNCO’s integrative resources can be effectively employed to predict genomic elements.

An important goal of a database is to functionally translate applications into solutions to explain complex biological mechanisms; thus, we decided to check the redundancy predictions of AOX genes. DFD values were high enough to be considered, and earlier AOX research made their context of high biological interest. Briefly, past reports mainly focused on the dominant isoform *AOX1A* (Giraud et al., 2008), which has a partial redundancy relationship with *AOX1D* (Strodtkotter et al., 2009), but current literature is not congruent with the use of single *aox1a* or double *aox1a-aox1d* mutants to discover causal drivers of retrograde signalling/metabolism/stress response (Giraud et al., 2009; Clercq et al., 2013; Oh Khim et al., 2022; Oh Khim et al., 2023). Additionally, more AOX isoforms exist, but their relationships are still not addressed. The abnormal seedling growth observed in control and mock conditions for all tested single mutants (*aox1a*, *aox1c*, *aox1d*) (**Fig. 7**) validated the high functional divergence predicted by PlantFUNCO since in case of redundancy, other duplicates could rescue these phenotypes (Ezoe, Shirai, and Hanada, 2021). Our findings suggest that the dominant isoform *AOX1A* can retain the ancestral AOX function because it is marked as functionally conserved with the distantly related *O. sativa* and is the only one covered by an active CS; thus, all redundancy relationships can potentially be compared to this gene*.* Considering that oxidative stress was more severe than drought/heat conditions, we found putative evidence of a probable stress-dependent partial non-mutual redundancy of *AOX1D* to *AOX1A*. Although *AOX1D* could partially alleviate *aox1a* raw hydrogen peroxide content under drought/heat (no significance), during more severe oxidative conditions, *AOX1D* would not be enough to supply the *AOX1A* function (significant) (Strodtkotter et al., 2009). It is defined as a potential non-mutual relationship because, in all cases, *aox1d* phenotypes remained significant. Finally, non-meaningful differences in raw hydrogen peroxide content for both stressors and WT-like root lengths under drought/heat in *aox1c* indicate that *AOX1C* is a non-stress-responsive gene. This could agree with the previously described *AOX1C* AA expression insensitivity (Yoshida and Noguchi, 2009), but we still found significant differences in root length in our severe oxidative assay. Compared to other genotypes, the p-value was close to not significant; thus, *AOX1C* may only be related to stress under severe conditions and could probably be defined as almost non-stress-responsive. In summary, stress seems to be a crucial evolutionary force driving sub-/neo-functionalisation (Panchy, Lehti-shiu, and Shiu, 2016) in AOX genes, and we characterised the unknown *AOX1C* asalmost stress-insensitive during the seedling stage. Furthermore, extra attention should be taken when using double AOX mutants to identify the causal determinants of biological processes because all AOX genes evaluated appeared to be functionally divergent during early development.

While we expect PlantFUNCO to be useful, we acknowledge certain limitations. Owing to our data collection design, the main goal of interspecies CS resources is to conduct intraspecies analyses while leveraging the advantage of having additional layers of interpretation, including direct correspondence between CS and conservation/divergence relationships established across species. Direct cross-species comparisons of equivalent loci or CSs should be undertaken only in conjunction with plants’ LECIF scores, as this algorithm is explicitly designed to handle highly diverse datasets. There may be states/regions that are functionally conserved but have low scores/agreement in the database since the evidence was not present in our collection. While the interpretation of the resources generated is less ambiguous due to the broad-shallow perspective adopted, we also perceived that PlantFUNCO is limited by the input functional genomics resolution and does not provide direct information about which particular tracks/conditions supported the evidence. The results promoted the potential application of PlantFUNCO to further test new hypotheses in the context of duplicate evolution and other genomic element predictions. For example, as CSs are determinants of paralogs’ functional divergence and LECIF scores highlight regions with high phenotypic similarity, it could be possible to identify genes that are more likely to retain ancestral functions if high scores are found between orthologues in distantly related species (**Fig. 6A**). Here, we focused on *A. thaliana*, *O. sativa* and *Z. mays*, which are widely used models in plant science research with substantial high-quality publicly available data. Given the increasing availability of epigenomics and functional genomics datasets, the utility of PlantFUNCO will continue to grow and serve as an additional resource to simplify functional conservation annotations for a more diverse set of species, such as *Chlamydomonas reinhardtii*, *Marchantia polymorpha* and *Solanum lycopersicum*. Overall, PlantFUNCO aims to leverage data diversity and extrapolate findings from different models to determine the extent of molecular conservation, thus deepening our understanding of how plant epigenomics and functional genomics have evolved.

# 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 chromatin modifications published in recent 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 include only 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 an 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 the bowtie2 algorithm (Langmead and Salzberg, 2012). Mapped reads with a MAPQ > 30 were used to secure the 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 subsequent analyses, we performed peak calling (narrow and broad), signal track building, correlation, and formatting with MACS2 and deepTools (Zhang et al., 2008; Ram et al., 2016). 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 peak calling, and the arguments *--nomodel --shift 75 --extsize 150* were added for ATAC- and DNase-seq file processing. Additional information detailing intraspecies correlations and variance can be found in **Supplementary Table S1**. To guarantee the reproducibility of the analysis, a docker was created, and it is available at <https://hub.docker.com/r/rocesv/plantina-chiplike>.

## Interspecies chromatin state definition and annotation

We applied hiHMM (Sohn et al., 2015) to jointly infer multiple species of CSs using common chromatin modification signal tracks from several tissues as input. Signal tracks consisted of scaled log2 (fold enrichment + 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; thus, the number of states is learned from the training data instead of a pre-specified value by the user. The model inferred that 15 CSs with unmappable regions added *a posteriori* as the 16th state to avoid any bias in the segmentation. We defined CSs based on the co-localisation of chromatin modifications and overlap enrichments of different genomic features using ChromHMM (Ernst and Kellis, 2017).

To further improve the interpretability of the states, additional annotations and descriptions were performed. The annotation was based on significant overlap enrichments using the LOLA package (Sheffield and Bock, 2016) and was divided as follows: 1) assessment of the presence of other epigenomic features employing non-common liftover information in PCSD; 2) conservation covered by PhastCons elements in PlantRegMap and pairwise CNEs; 3) transcription factor binding motifs collected in PlantRegMap (Tian et al., 2020); 4) genetic variability represented by significant SNPs compiled in GWAS-ATLAS and AraGWAS (Togninalli et al., 2020; Liu et al., 2023). The description involved KEGG-Orthology(KO)/Gene-Ontology(GO) enrichments using clusterProfiler/REVIGO, respectively, and gene biotype-orthology correspondence using inParanoid information stored in Phytozome (Goodstein et al., 2012).

## Modelling paralogs’ degrees of functional divergence

We reproduced two published models that predict genetic redundancy in *A. thaliana* paralogs (Cusack et al., 2021; Ezoe et al., 2021), including our interspecies CS 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 included a third type of vector, with each element having 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 the philentropy package (Drost, 2018).

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

However, in the model developed by Cusack et al. (2021), multiple transformations and interpretations of the same feature were included; thus, all distance state metrics were considered. Only the available extreme (RD4) and inclusive (RD9) redundancy gene pair sets were analysed, 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 using random forest top 200 best transformed variables (determined by feature importance) for sets without (RD4–RD9) and with (RD4C–RD9C) chromatin information. The C value for the SVM algorithm was set as a hyperparameter during tuning. To measure SVM performance using different feature sets, we calculated AUC-ROC and AU-PRC values. All analyses were conducted using the pipeline implemented and developed by the authors (<https://github.com/ShiuLab/ML-Pipeline>).

## Genome-wide redundancy 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 a 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. *Arabidopsis thaliana* genes (longest sequence) were used as queries to search for self-match homologues with DIAMOND v2 (E-value=1e-04) (Buchfink, Reuter, and 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 and Hanada (2021). An additional table is provided with filters, such as the same second closest paralog and expression under stress and in the 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 a knockout and validated by genotyping before use (Fuchs et al., 2022). We characterised the *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 (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 (all primers are available in **Supplementary Table S3**).

For stress evaluation, *aox1a*, *aox1c* and *aox1d* seeds were surface sterilised in 2.8% hypochlorite solution and washed several times with sterile water; they were stratified for 3 days at 4°C in darkness. The in vitro culture of seeds was carried out in 12 × 12 plates (Greiner) containing 50 mL of MS medium, pH 5.8, 1% (w/v) sucrose and 0.8% (w/v) agar, and they were vertically placed under a long-day photoperiod (16 h light 21°C, 8 h dark 18°C) for control conditions. To avoid a position effect, the four genotypes (Col-0 as WT, *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 (ThermoFisher Scientific) was added to the initial plates and seedlings were subjected to 37°C stress for 1 h every day at the same hour, gradually increasing and decreasing the temperature. For the antimycin A (AA) treatment, 50 μM AA (Sigma-Aldrich) was added to the initial plates; control conditions were set 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, and Eliceiri, 2012) in at least 12 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 and O’Brien (2012); DAB quantification was carried out using ImageJ.

## RNA-seq data processing

The sequence quality of RNA-seq libraries was evaluated by FastQC and multiQC (Andrews, 2013; Ewels, Lundin, and 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 the reference genome and minimum/maximum intron size according to species. Bigwig files were obtained using the *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 and <100 MYA) *distance* arguments were performed using the CNEr package interface (Tan, Polychronopoulos, and Lenhard, 2019). This was followed by format conversion, chain building, and processing using lavToPsl, maf-convert, axtChain and chainMergeSort. RepeatFiller (Osipova, Hecker, and Hiller, 2019) was applied to the chains to improve the identification of conserved non-coding elements (CNEs). After RepeatFiller, we executed ChainCleaner (Suarez, Langer, Ladde, and Hiller, 2017) to improve alignment specificity and chains were then converted into alignment nets using Hillerlab chainNet and netToAxt. Finally, Axt files were used as input for the pairwise identification of CNEs using the CNEr package with 45 identity/50 length windows while considering the difference in whole genome duplication history between these species, as described by Ren et al. (2018).

To take advantage of previously processed epigenetic tracks in PCSD that are not included in our initial collection (not common for all species), we executed another WGA pipeline to lift over these files to the new reference assemblies. In summary, we used near as a *distance* argument and skipped the RepeatFiller-ChainCleaner step because we aligned the same species, and 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; it is available at <https://hub.docker.com/r/rocesv/compcnes>.

## Functional genomics conservation score

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

## Database resource

We developed a 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) a search section with interactive tables to retrieve gene- or superenhancer-level (Zhao et al., 2022) functional and comparative genomics information; 2) a Shiny application to compute LOLA genomic overlap enrichments of user query bed files over CSs and LECIF/PhyloP binned scores; and 3) a JBrowse2 genome browser (Diesh et al., 2023). PlantFUNCO is available at <https://rocesv.github.io/PlantFUNCO>.

# Data availability

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

# Acknowledgements

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

# Conflict of interest

The authors declare that there are no conflicts of interest.

# Author’s contributions

VR and MM conceived of 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, analysed and interpreted the data, and wrote the manuscript. JP and MM supervised the study. All authors revised, read and approved the final manuscript.