PlantFUNCO: integrative functional genomics database reveals clues into duplicates divergence evolution

Víctor Roces[1], Sara Guerrero[1], Ana Álvarez[1], Jesús Pascual[1*#], Mónica Meijón[1*#]

[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, pascualjesus@uniovi.es

# Contributed equally

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 PlantFUN(ctional)CO(nservation) database, which is constituted by several tools and two main resources: inter-species chromatin states and functional genomics conservation scores, presented and analysed in this work for 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, providing 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. Taking all the above into account, PlantFUNCO aim to leverage data diversity and extrapolate molecular mechanisms findings from different model organisms to determine the extent of functional conservation, thus, deepening our understanding of how plants epigenome and functional non-coding genome 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. Chromatin states 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 elements 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 a relative 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 interrogate sequence-conserved loci of interest across a wide range of species, and their functions are determined by perturbing their homologous 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 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 chracterisation. 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, inter-species chromatin states 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 of the database and found that chromatin state 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 and provided 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, transposable element (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**; **supplementry 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 the three species. However, many CS definitions exhibit species-specific nuances at different levels, which could actually reflect how epigenomic complexity has evolved in plants. The various degrees of CS divergence were determined based on CS chromatin modifications composition (**fig. 1**, **top panel**) and genomic distribution (**fig. 1**, **bottom panel**). Ranging from less to more divergent: 1) states which shared 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 inter-species 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*. 2) Landscapes whose chromatin modifications 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 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 marks tracks for all species under study to test our states 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 inter-species approach, we further evaluated wether 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) 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 for the three species in TF motifs and genetic variability annotation modules (**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 inter-species 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 wether 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 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 out 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 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 prediciting 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 between the species. By querying the LECIF scores, we sought to identify genomic regions with a high degree of functional tracks 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, chromatin states 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* GWAS significant 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; bar plots**). This could be explained by a balanced significant SNP distribution in the *A. thaliana* genome due to its architecture and higher number of GWAS, 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 inter-species 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 out of 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 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 (PEGxHeat) 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 mitochondrial 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 monitorized 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 genes in the redox state, DAB staining quantification was performed to measure hydrogen peroxide levels. Although both stressors agreed in the WT, *aox1d* relevant increase of hydrogen peroxide levels, and *aox1c* was not significant; *aox1a* trends were not congruent. In *aox1a*, 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 epigenome/functional non-coding genome. This database is derived from two resources presented and analysed in this work for three well-established plant models. We generated inter-species CS using hiHMM (**fig. 1**). While this flexible framework provides a consistent definition of CS across multiple genomes, making the extrapolation of intra-species 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 inter-species perspective (**figs. 3 and 4**; **supplementary fig. S3 and S4**). All the 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 ECIF score elucidated functional genomic cross-species agreement without being correlated with other comparative genomics sources (**fig. 6**). This probably reflects a complementary side of evolution. Despite the greater divergence between plants models compared to metazoans (Ho et al., 2014; Kwon and Ernst, 2021), both resources 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 inter-species 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 wether 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 relation described 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 were 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* could retain the ancestral AOX function because it is marked as functionally conserved with the distanly 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* would indicate that *AOX1C* as 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-/neofunctionalisation (Panchy, Lehti-shiu, and Shiu, 2016) in AOX genes, and we characterised the unknown *AOX1C* as almost stress-insensitive during the seedling stages. 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 inter-species CS resources is to conduct intra-species 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 elements prediction. 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 orthologes 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 plants epigenome and functional non-coding genome have fascinatingly 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 intra-species 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.

Inter-species chromatin states definition and annotation

We applied hiHMM (Sohn et al., 2015) to jointly infer multiple species chromatin states (CS) using common chromatin modifications 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 a total of 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 colocalisation 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) assesment 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 inter-species 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).

However, in the model developed by Cusack et al. (2021) multiple transformations and interpretations of the same feature were included; thus, all the 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 the 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 (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 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 4º C in darkness. The in vitro culture of seeds was carried out in 12x12 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 21º C, 8 h dark 18º 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 (ThermoFisherScientific) was added to the initial plates and seedlings were subjected to 37º 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 decribed 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).

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) 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). We are grateful to Prof. James Whelan (Zhejian University) for kindly sharing the *aox* mutant lines used in this study.

Conflict of interest

The authors declare that there are no conflicts 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, analysed and 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.

Ashe A, Colot V, Oldroyd BP. 2021. How does epigenetics influence the course of evolution ? *Philosophical Transactions B*, *376*(20200111). https://doi.org/10.1098/rstb.2020.0111

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

Clercq I De, Vermeirssen V, Aken O Van, Vandepoele K, Murcha M W, Law S R, Inzé A, Ng S, Ivanova A, Rombaut D, et al. 2013. The Membrane-Bound NAC Transcription Factor ANAC013 Functions in Mitochondrial Retrograde Regulation of the Oxidative Stress Response in Arabidopsis. *The Plant Cell*, *25*(September), 3472–3490. https://doi.org/10.1105/tpc.113.117168

Cusack S A, Wang P, Lotreck S G, Moore B M, Meng F, Conner J K, Krysan PJ, Lehti-Shiu MD, Shiu-Han 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, Guo E, Dider S, Zhang J, Bridge C, et al. 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, Batut P, Chaisson M, 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

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

689 Expósito-Alonso M, Drost H, Burbano H A, Weigel D. 2020. The Earth BioGenome
690 project : opportunities and challenges for plant genomics and conservation. *Plant*
691 *Journal*, *102*, 222–229. https://doi.org/10.1111/tpj.14631

692 Ezoe A, Shirai K, Hanada K. 2021. Degree of Functional Divergence in Duplicates Is
693 Associated with Distinct Roles in Plant Evolution. *Molecular Biology and Evolution*, *38*(4),
694 1447–1459. https://doi.org/10.1093/molbev/msaa302

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

698 Fuchs P, Bohle F, Lichtenauer S, Ugalde M, Araujo E F, Mansuroglu B, Ruberti C,
699 Wagner S, Müller-Schüssele J, Meyer AJ, et al. 2022. Reductive stress triggers
700 ANAC017-mediated retrograde signaling to safeguard the endoplasmic reticulum by
701 boosting mitochondrial respiratory capacity. *The Plant Cell*, *34*, 1375–1395.
702 https://doi.org/10.1093/plcell/koac017

703 Giraud E, Ho L H M, Clifton R, Carroll A, Estavillo G, Tan Y, Howell KA, Ivanova A,
704 Pogson BJ, Millar AH, et al. 2008. The Absence of ALTERNATIVE OXIDASE1a in
705 Arabidopsis Results in Acute Sensitivity to Combined. *Plant Physiology*, *147*(June), 595–
706 610. https://doi.org/10.1104/pp.107.115121

707 Giraud E, Aken O Van, Ho L H M, Whelan J. 2009. The Transcription Factor ABI4 Is a
708 Regulator of Mitochondrial Retrograde Expression of. *Plant Physiology*, *150*(July), 1286–
709 1296. https://doi.org/10.1104/pp.109.139782

710 Goodstein D M, Shu S, Howson R, Neupane R, Hayes R D, Fazo J, Mitros T, Dirks W,
711 Hellsten U, Putnam N, et al. 2012. Phytozome : a comparative platform for green plant
712 genomics. *Nucleic Acids Research*, *40*(November 2011), 1178–1186.
713 https://doi.org/10.1093/nar/gkr944

714 Hazarika R R, Serra M, Zhang Z, Zhang Y, Schmitz R J, Johannes F. 2022. Molecular
715 properties of epimutation hotspots. *Nature Plants*, *8*(February), 146–156.
716 https://doi.org/10.1038/s41477-021-01086-7

Ho J W K, Jung Y L, Liu T, Alver B H, Lee S, Ikegami K, Sohn K, Minoda A, Tolstorukov MY, Appert A, et al. 2014. Comparative analysis of metazoan chromatin organization. *Nature*, *512*(7515), 449–452. https://doi.org/10.1038/nature13415

Jamge B, Lorkovi Z J, Axelsson E, Osakabe A, Shukla V, Yelagandula R, Akimcheva S, Kuehn AL, Berger F. 2023. Histone variants shape chromatin states in Arabidopsis. *ELife*, *12*(RP87714), 1–26. https://doi.org/10.7554/eLife.87714.3

Kliebenstein D J. 2019. Questionomics : Using Big Data to Ask and Answer. *The Plant Cell*, *31*(July), 1404–1405. https://doi.org/10.1105/tpc.19.00344

Kumar S, Suleski M, Craig J M, Kasprowicz A E, Sanderford M, Li M, Li M, Stecher G, 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, Pan Y, Wang Y, Zou D, Zhang Z, et al. 2023. GWAS Atlas : an updated knowledgebase integrating more curated associations in plants and animals. *Nucleic Acids Research*, *51*(October 2022), 969–976. https://doi.org/10.1093/nar/gkac924

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

Lu Z, Marand A P, Ricci W A, Ethridge C L, Zhang X, Schmitz R J. 2019. The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nature Plants*, *5*(December), 1250–1259. https://doi.org/10.1038/s41477-019-0548-z

Maher K A, Bajic M, Kajala K, Reynoso M, Pauluzzi G, West D A, Zumstein K, Woodhouse M, Bubb K, Dorrity MW, et al. 2018. Profiling of Accessible Chromatin Regions across Multiple Plant Species and Cell Types Reveals Common Gene Regulatory Principles and New Control Modules. *The Plant Cell*, *30*(January), 15–36. https://doi.org/10.1105/tpc.17.00581

Marand A P, Eveland A L, Kaufmann K, Springer N M. 2023. cis -Regulatory Elements in Plant Development , Adaptation , and Evolution. *Annual Review of Plant Biology*, *74*, 111–137. https://doi.org/10.1146/annurev-arplant-070122-030236

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

751 Monroe J G, Srikant T, Carbonell-bejerano P, Becker C, Lensink M, Exposito-alonso M,
752 Klein M, Hildebrandt J, Neumann M, Kliebenstein D, et al. 2022. Mutation bias reflects
753 natural selection in Arabidopsis thaliana. *Nature*, *602*(3), 101–105.
754 https://doi.org/10.1038/s41586-021-04269-6

755 Oh Khim G G, Leary B M O, Signorelli S, Millar A H. 2022. Alternative oxidase (AOX)
756 1a and 1d limit proline- induced oxidative stress and aid salinity recovery in Arabidopsis.
757 *Plant Physiology*, *188*, 1521–1536. https://doi.org/10.1093/plphys/kiab578

758 Oh Khim G G, Kumari V, Millar A H, Leary B M O. 2023. Alternative oxidase 1a and 1d
759 enable metabolic flexibility during Ala catabolism in Arabidopsis Research Article. *Plant*
760 *Physiology*, *192*(4), 2958–2970. https://doi.org/10.1093/plphys/kiad233

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

764 Panchy N, Lehti-shiu M, Shiu S. 2016. Evolution of Gene Duplication in Plants. *Plant*
765 *Physiology*, *171*(August), 2294–2316. https://doi.org/10.1104/pp.16.00523

766 Papatheodorou I, Moreno P, Manning J, George N, Fexova S, Fonseca N A, Füllgrabe
767 A, Green M, Huang N, Huerta L, et al. 2020. Expression Atlas update : from tissues to
768 single cells Anja F ullgrabe. *Nucleic Acids Research*, *48*(October 2019), 77–83.
769 https://doi.org/10.1093/nar/gkz947

770 Ram F, Ryan D P, Bhardwaj V, Kilpert F, Richter A S, Heyne S, Dündar F, Manke T.
771 2016. deepTools2 : a next generation web server for deep-sequencing data analysis.
772 *Nucleic Acids Research*, *44*(April), 160–165. https://doi.org/10.1093/nar/gkw257

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

776 Schmitz R J, Grotewold E, Stam M. 2022. Cis-regulatory sequences in plants : Their
777 importance , discovery , and future challenges. *The Plant Cell*, *34*, 718–741.
778 https://doi.org/10.1093/plcell/koab281

779 Schneider C A, Rasband W S, Eliceiri K W. 2012. NIH Image to ImageJ : 25 years of
780 Image Analysis. *Nature Methods*, *9*(7), 671–675. https://doi.org/10.1038/nmeth.2089

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

784 Sohn K A, Ho J W K, Djordjevic D, Jeong H H, Park P J, Kim J H. 2015. HiHMM: Bayesian
785 non-parametric joint inference of chromatin state maps. *Bioinformatics*, *31*(13), 2066–
786 2074. https://doi.org/10.1093/bioinformatics/btv117

787 Strodtkotter I, Padmasreea K, Dinakara C, Spetha B, Niazi P S, Wojtera J, Voss I, Do
788 PT, Nunes-Nesi A, Fernie AR, et al. 2009. Induction of the AOX1D Isoform of Alternative
789 Oxidase in A . thaliana T-DNA Insertion Lines Lacking Isoform AOX1A Is Insufficient to
790 Optimize Photosynthesis when Treated with Antimycin A. *Molecular Plant*, *2*(2).
791 https://doi.org/10.1093/mp/ssn089

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

795 Tan G, Polychronopoulos D, Lenhard B. 2019. CNEr : A toolkit for exploring extreme
796 noncoding conservation. *PLoS Computational Biology*, *15*((8)), 1–16.
797 ttps://doi.org/10.1371/journal.pcbi.1006940

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

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

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

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

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. https://doi.org/10.1186/s13059-021-02572-z

Yocca A E, Edger P P. 2022. Current status and future perspectives on the evolution of cis -regulatory elements in plants. *Current Opinion in Plant Biology*, *65*(102139). https://doi.org/10.1016/j.pbi.2021.102139

Yoshida K, Noguchi K. 2009. Differential Gene Expression Profiles of the Mitochondrial Respiratory Components in Illuminated Arabidopsis Leaves. *Plant and Cell Physiology*, *50*(8), 1449–1462. https://doi.org/10.1093/pcp/pcp090

Zhang Y, Liu T, Meyer C A, Eeckhoute J, Johnson D S, Bernstein B E, Nusbaum C, Myers RM, Brown M, Li W, et al. 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, Yang M, Bishop J, Teng Y, Cao Y, Beall B D, Li S, Liu T, Fang Q, Fang Q, et al. 2022. Identification and functional validation of super-enhancers in Arabidopsis thaliana. *PNAS*, *119*(48), 1–11. https://doi.org/10.1073/pnas.

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 L, Xie L, Zhang Q, Ouyang W, Deng L, Guan P, Ma M, Li Y, Zhang Y, Xiao Q, et al. 2020. Integrative analysis of reference epigenomes in 20 rice varieties. *Nature Communications*, *11*(2658), 1–16. https://doi.org/10.1038/s41467-020-16457-5

Figure legends

Fig. 1. Inter-species chromatin states definition. Top panel: From left to right chromatin state definitions, abbreviation, species relation, track composition (emission probability) and genome coverage based on 10 common chromatin modifications. Chromatin states with “>” indicate definitions transitioning between species. Darkblue colors in relation heatmap highlight 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 with non-common chromatin modifications. Heatmaps depicting significant ($p < 0.05$) genomic overlap-enrichment (odds ratio) of inter-species states with different annotation modules. From top to bottom: non-common chromatin-binding proteins and histone modifications/variants. 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). Each module is constituted by three 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. Minor gene biotypes are represented by different symbols.

Fig. 4. Inter-species chromatin states annotation with conservation, genetic variability and transcription factor motifs modules. Heatmaps depicting significant ($p < 0.05$) genomic overlap-enrichment (odds ratio) of inter-species states with different annotation modules. From top to bottom: conservation covered by PhastCons elements and pairwise conserved non-coding elements (CNEs), transcription factor (TF) motifs illustrated by TF binding sites (BS) according to PlantRegMap categories and genetic variability represented by significant SNPs in GWAS. 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.

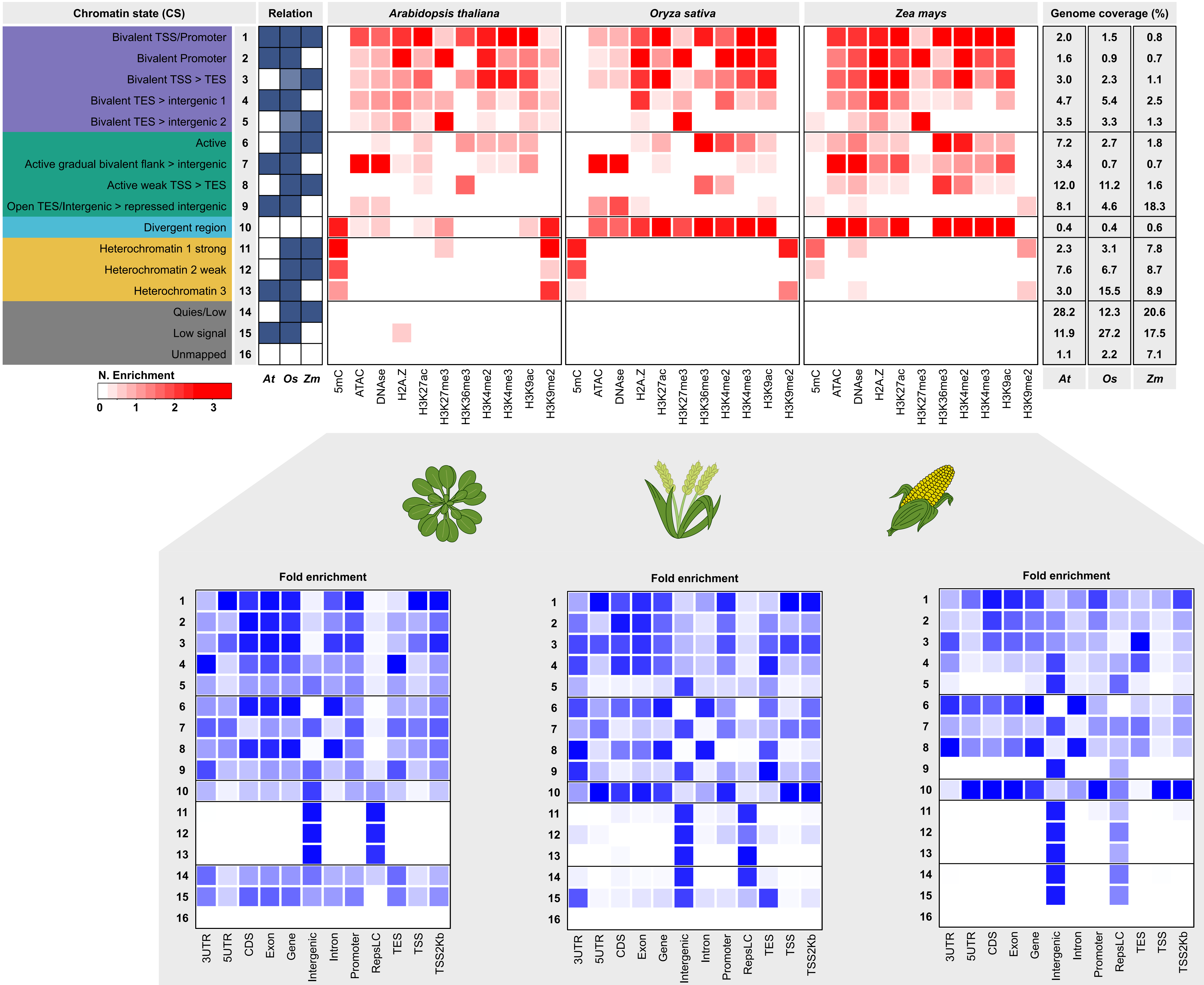
Fig. 5. Predictive models of paralogs degree of functional divergence including chromatin states metrics. A) Chromatin states metrics were obtained dividing promoter and genes in a fixed number of windows, calculating frequency and presence vectors and computing several distance and similarity coefficients between genes from the same paralog pair comparing equivalent vector types (see **Methods**). **B-E)** Results reproducing Ezoe, Shirai, and Hanada, 2021 models including CS metrics. **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)** Relative importance in explanatory variables. The relative importance was inferred based on the logistic regression algorithm. **D)** 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 area 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. **E)** 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). Ka/Ks = protein divergence sequence rate, Re/Ks = gene expression similarity rate, FD = number of shared functional domains, GO = number of shared gene ontologies, PPI = protein-protein interactions. **F-I)** Results reproducing Cusack et al., 2021 models including CS metrics. **F)** 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. **G)** 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. **H)** AUC-ROC and AU-PRC for the heldout tests for models built with each redundancy definition/feature set. **I)** 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. 6. Functional genomics conservation (LECIF) score overview and downstream analyses. This figure is constituted by 4 panels (**overview (A), *Arabidopsis thaliana* (B, E, H), *Oryza sativa* (C, F, I),** and ***Zea mays* (D, G, J)**). **A)** 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. ***Arabidopsis thaliana* (B, E, H), *Oryza sativa* (C, F, I),** and ***Zea mays* (D, G, J) panels** illustrate LECIF-score downstream analyses for *A. thaliana (At), O. sativa (Os)* and *Z. mays (Zm)*, respectively. Each of this panels are divided into two sides according to the two remaining target species and three description modules: **B,C and D)** Genetic variability as genomic overlap-enrichment of GWAS significant SNPs over regions

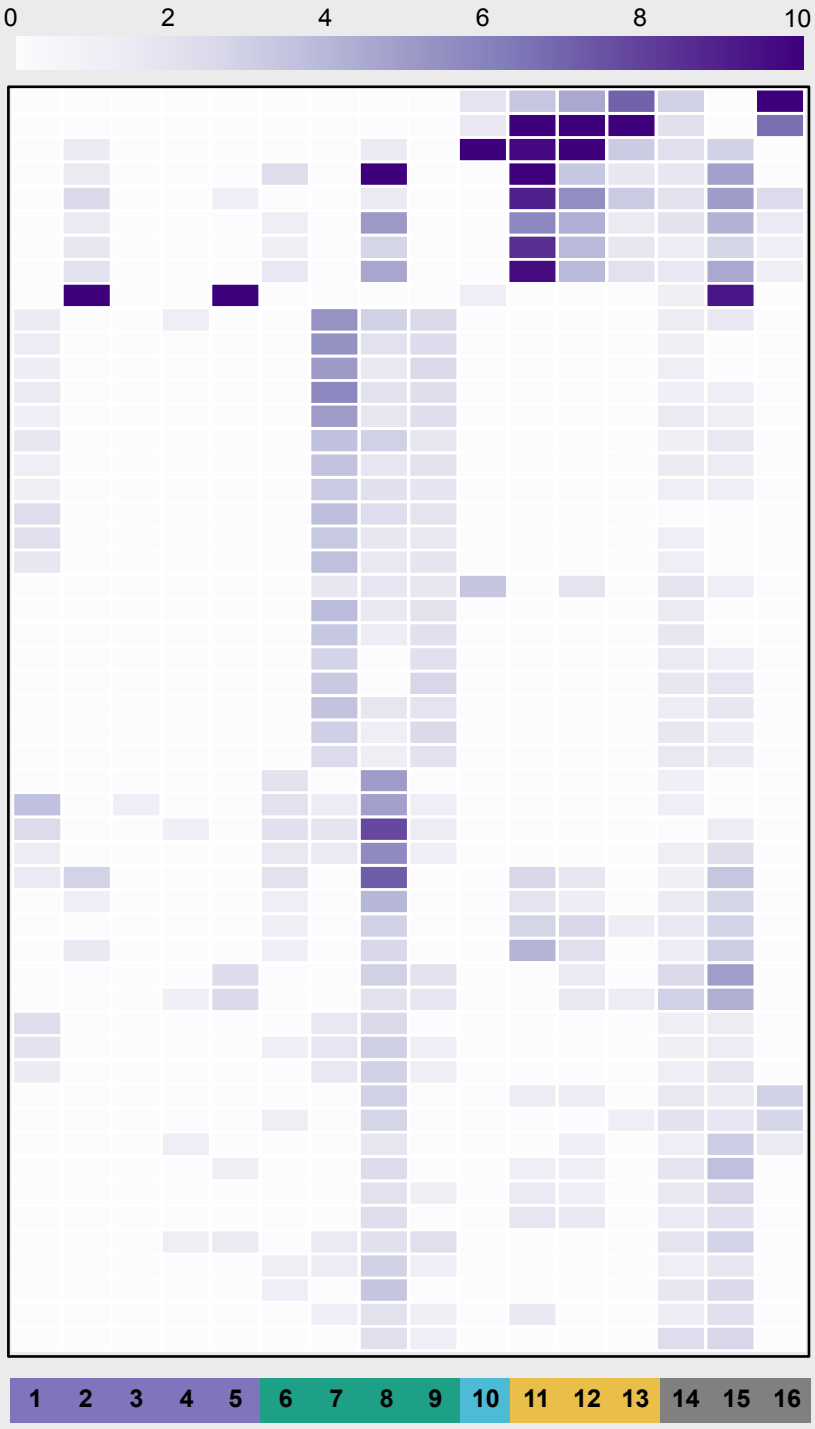
divided into five bins based on LECIF scores. Black bars indicate significance ($p < 0.05$). **E, F and G)** Chromatin states module with genome-wide (histogram) and state-specific (violinplot) LECIF scores distribution. Additionally, this module is covered by chromatin state similarity between high/low (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). Chromatin state similarity was computed using the Dice coefficient. **H, I and J)** 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. Coverage (%) referes to the aligning regions overlap. PlantFUNCO DB is available at https://rocesv.github.io/PlantFUNCO.

Fig. 7. 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 $p < 0.05$. “ns”: $p > 0.05$; “*”: $p < 0.05$; “**”: $p < 0.01$; “***”: $p < 0.001$; “****”: $p < 0.0001$. KW = Kruskall-Wallis.

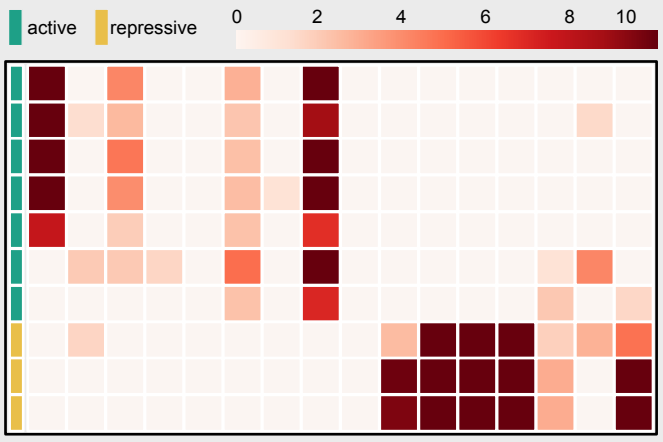


Chromatin state (CS)
Relation
Arabidopsis thaliana
Oryza sativa
Zea mays
Genome coverage (%)
Bivalent TSS/Promoter 1 2.0 1.5 0.8
Bivalent Promoter 2 1.6 0.9 0.7
Bivalent TSS > TES 3 3.0 2.3 1.1
Bivalent TES > intergenic 1 4 4.7 5.4 2.5
Bivalent TES > intergenic 2 5 3.5 3.3 1.3
Active 6 7.2 2.7 1.8
Active gradual bivalent flank > intergenic 7 3.4 0.7 0.7
Active weak TSS > TES 8 12.0 11.2 1.6
Open TES/Intergenic > repressed intergenic 9 8.1 4.6 18.3
Divergent region 10 0.4 0.4 0.6
Heterochromatin 1 strong 11 2.3 3.1 7.8
Heterochromatin 2 weak 12 7.6 6.7 8.7
Heterochromatin 3 13 3.0 15.5 8.9
Quies/Low 14 28.2 12.3 20.6
Low signal 15 11.9 27.2 17.5
Unmapped 16 1.1 2.2 7.1
N. Enrichment
0 1 2 3
At Os Zm
5mC ATAC DNAse H2A.Z H3K27ac H3K27me3 H3K36me3 H3K4me2 H3K4me3 H3K9ac H3K9me2
Fold enrichment
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
3UTR 5UTR CDS Exon Gene Intergenic Intron Promoter RepsLC TES TSS TSS2Kb

Chromatin-binding proteins

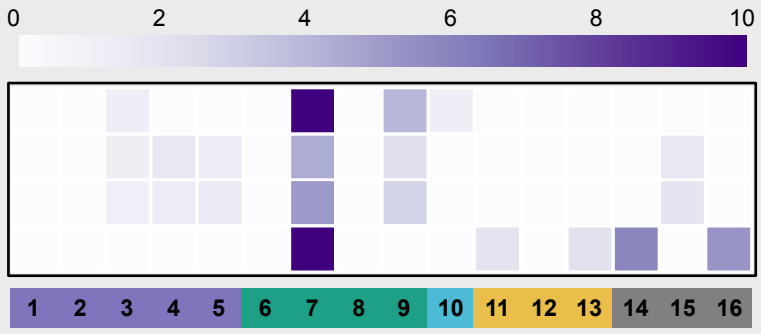


Histone marks

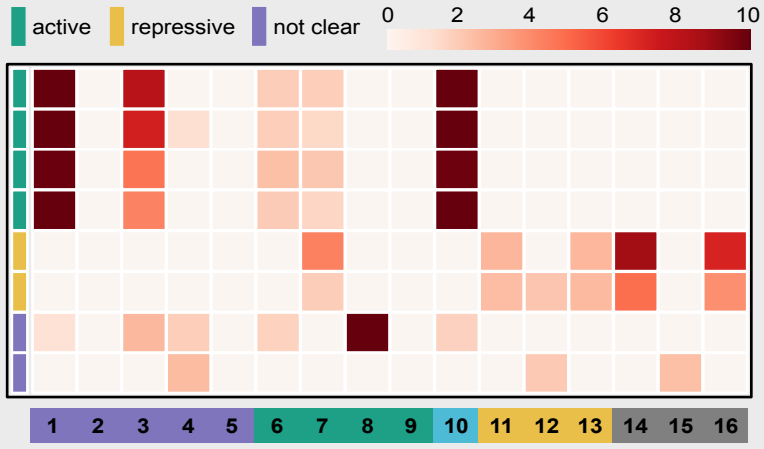


Arabidopsis thaliana

Chromatin-binding proteins



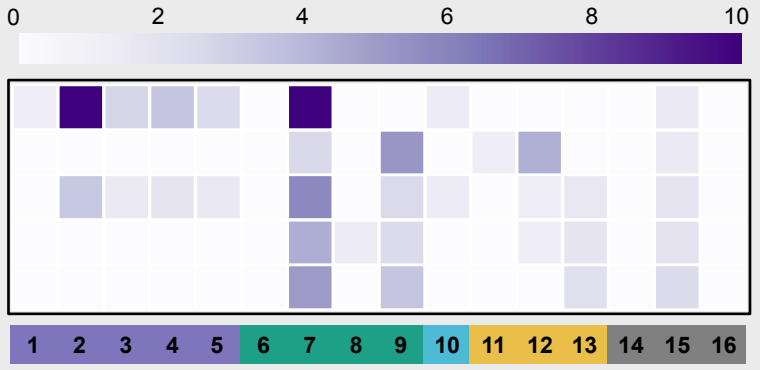
Histone marks



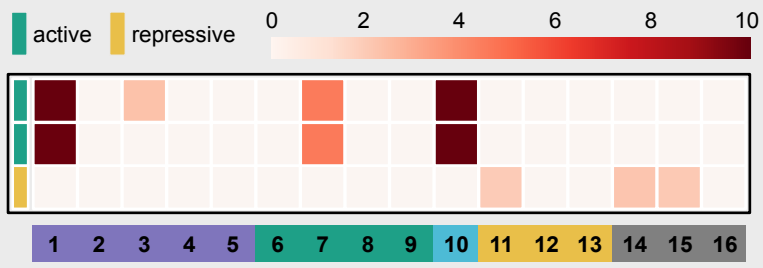
Oryza sativa

Zea mays

Chromatin-binding proteins



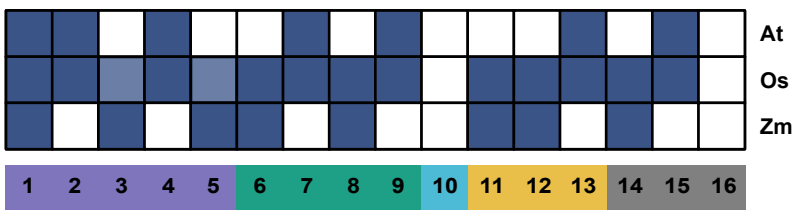
Histone marks



Chromatin States

1	Bivalent TSS/Promoter	4	Bivalent TSS > intergenic 2	8	Open TES/Intergenic > repressed intergenic	12	Heterochromatin 3
2	Bivalent Promoter	5	Active	9	Divergent region	13	Quies/Low
3	Bivalent TSS > TES	6	Active gradual bivalent flank > intergenic	10	Heterochromatin 1 strong	14	Low signal
4	Bivalent TES > intergenic 1	7	Active weak TSS > TES	11	Heterochromatin 2 weak	15	Unmapped

Relation



CS1
At
Os
Zm
protein coding
transposable element
orthologous
non orthologous
CS6
CS10
CS11
Both
Arabidopsis thaliana (At)
Oryza sativa (Os)
Zea mays (Zm)
antisense long non coding RNA
novel transcribed region
pseudogene
antisense RNA
pre tRNA / tRNA
small nuclear RNA
long non coding RNA / lincRNA
transposable element
small nucleolar RNA
miRNA primary transcript / miRNA
other RNA
ribosomal RNA

Bivalent TSS/Promoter
Bivalent Promoter
Bivalent TSS > TES
Bivalent TES > intergenic 1
Bivalent TES > intergenic 2
Active
Active gradual bivalent flank > intergenic
Active weak TSS > TES
Open TES/Intergenic > repressed intergenic
Divergent region
Heterochromatin 1 strong
Heterochromatin 2 weak
Heterochromatin 3
Quies/Low
Low signal
Unmapped

CS

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

Relation

At
Os
Zm

0 2 4 6 8 10

0 2 4 6 8 10

0 1 2 3 4 5

Conservation

PhastCons
CNEsvsAt
CNEsvsOs
CNEsvsZm

0 10

0 10

0 10

TF motifs

Conserved TFBS
Wide TFBS
Functional TFBS

0 1 2 3 4 5

0 0.5 1 1.5 2

0 1 2 3 4 5 6 7

Genetic variability

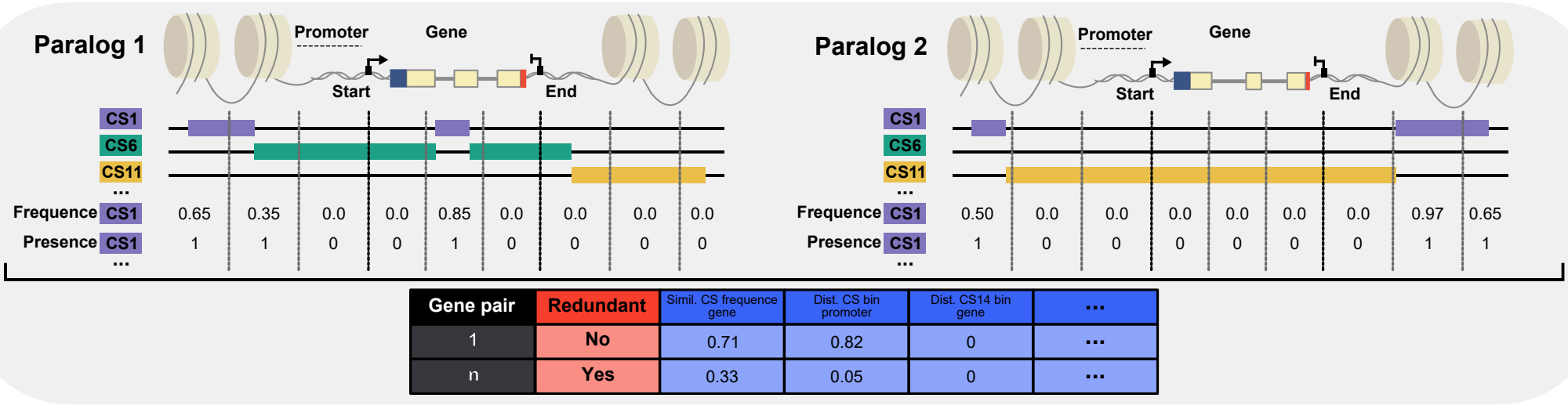
GWAS-SNPs

Arabidopsis thaliana

Oryza sativa

Zea mays

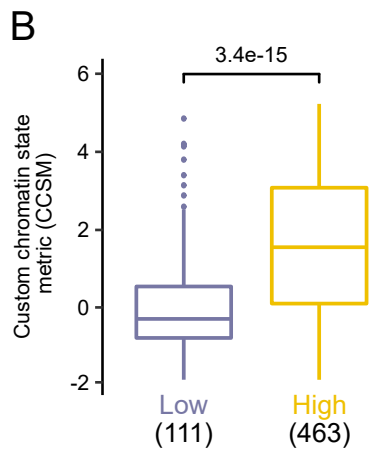
A



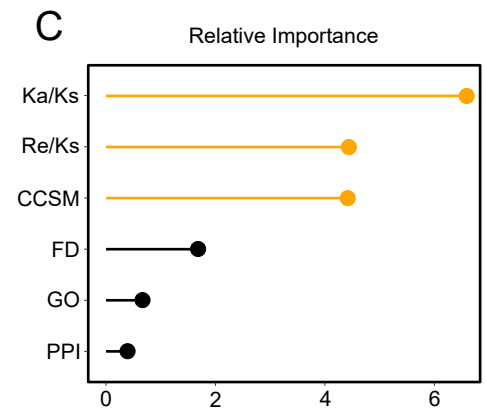
Gene pair	Redundant	Simil. CS frequence gene	Dist. CS bin promoter	Dist. CS14 bin gene	...
1	No	0.71	0.82	0	...
n	Yes	0.33	0.05	0	...

Degree of Functional Divergence in Duplicates Is Associated with Distinct Roles in Plant Evolution (Ezoe *et al.*, 2021)

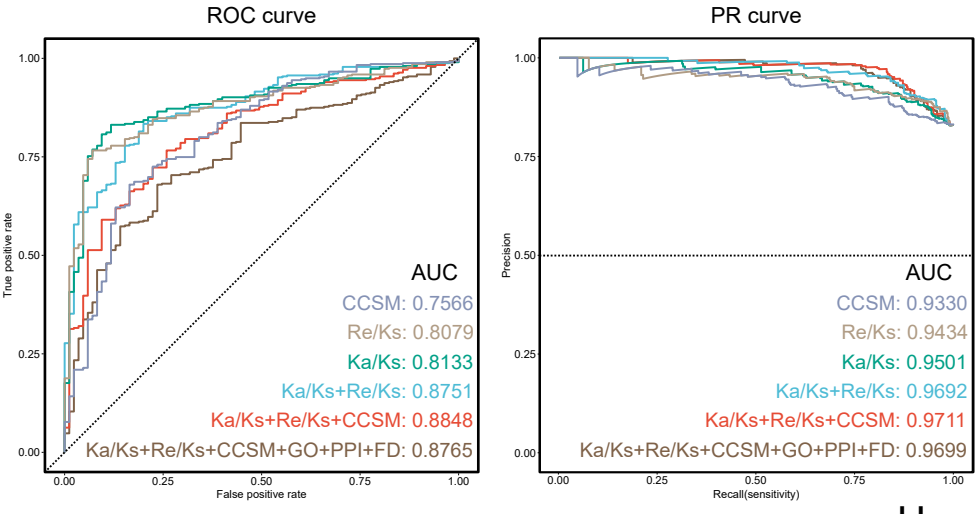
B



C

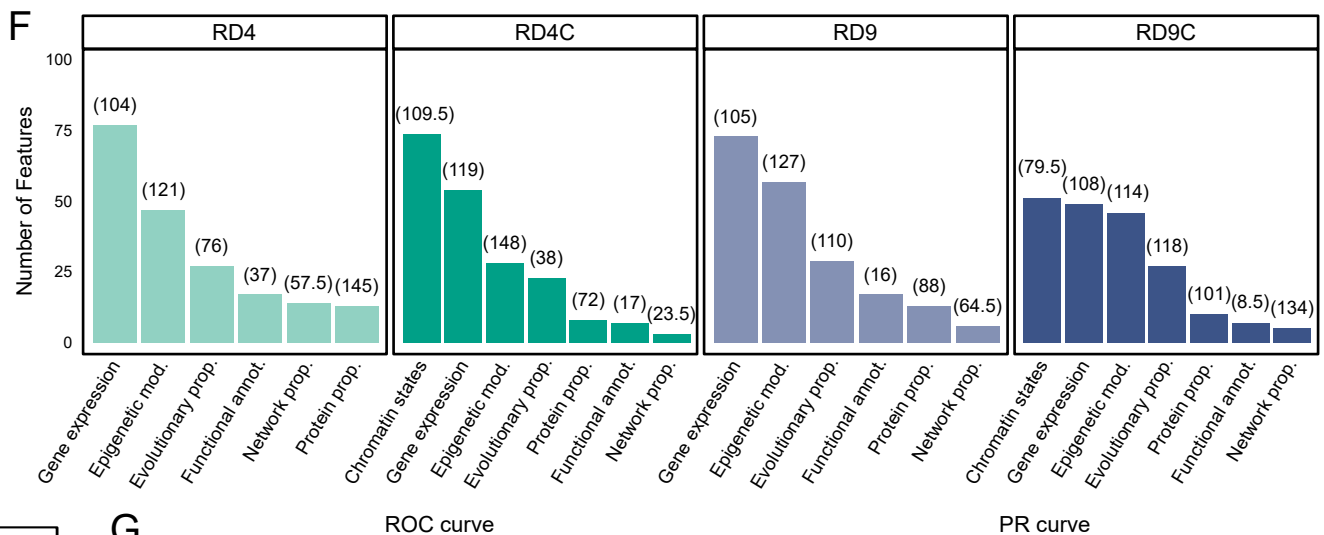


D

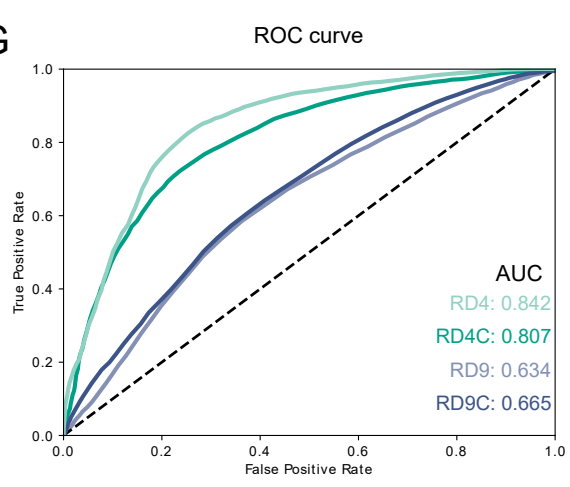


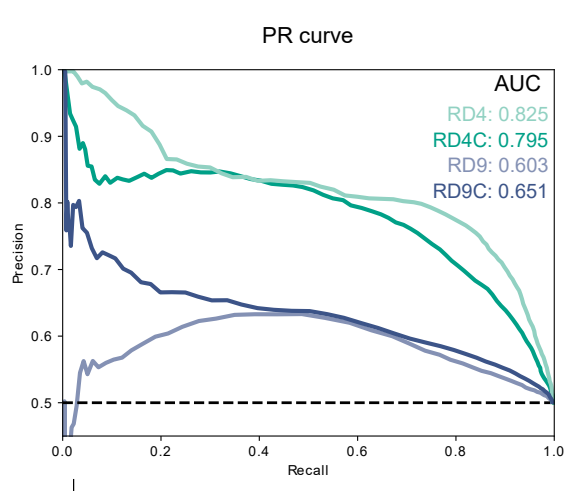
Predictive Models of Genetic Redundancy in *Arabidopsis thaliana* (Cusack *et al.*, 2021)

F

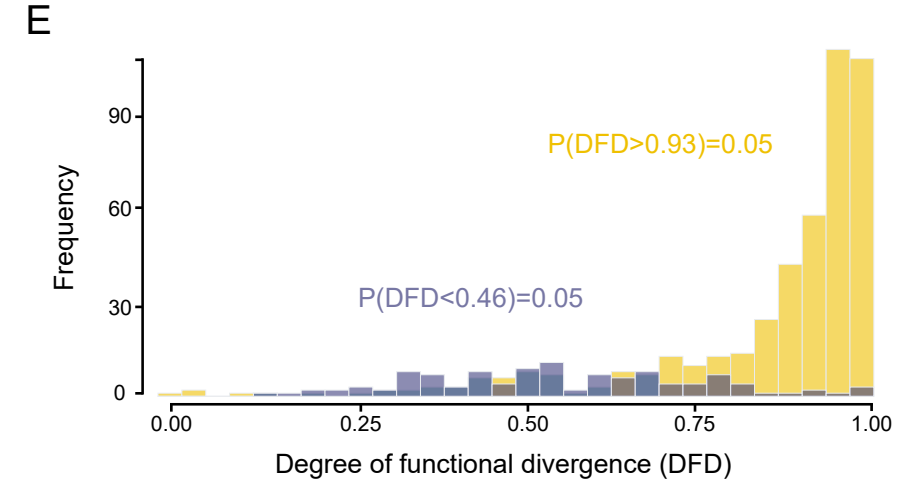


G

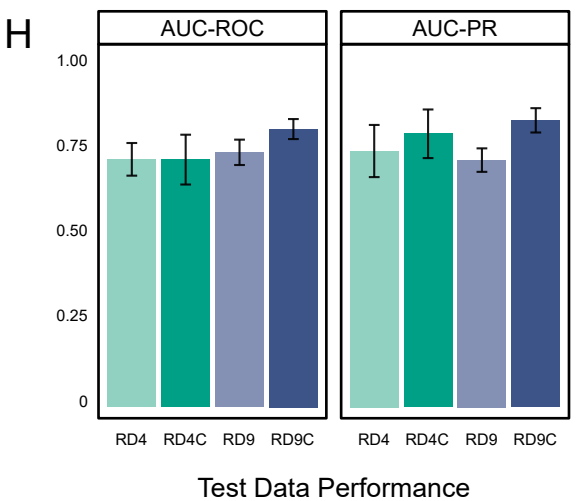




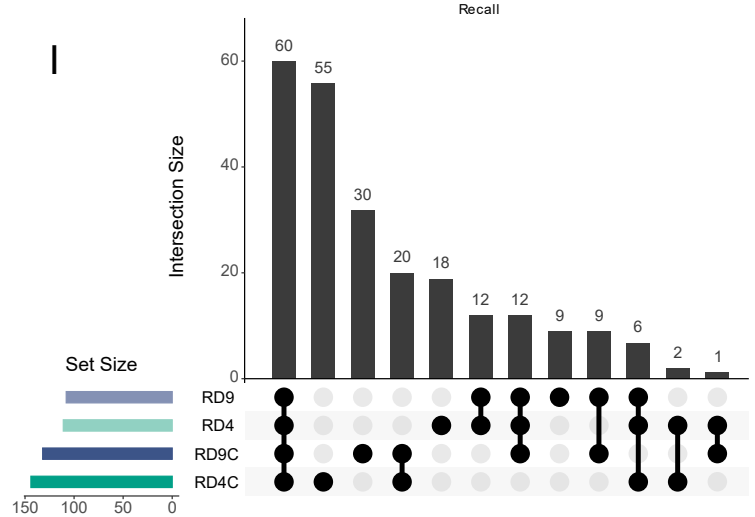
E



H



I



A

Whole Genome Alignments
.axtNet

Baseline RNA-seq
.bigwig

10 COMMON MARKS
PCSD + SRA

H. ACETYLATION	H. METHYLATION
H3K9ac H3K27ac	H3K4me2 H3K27me2 H3K4me3 H3K36me3 H3K9me2
H. VARIANT	**DNA METHYLATION**
H2A.Z	5mC

CHROMATIN ACCESS

At Os Zm
Functional genomics conservation score

Inter species-CS
Paralog modelling

LECIF-score

HYPOTHESIS
Highlight regions with high phenotypic simmilarity

Duplication event
Zm1 Os1 At1 At2

one2many gene pairs ancestral function?

RESOURCE
large scale analyses
biological applications

Tracks and applications available in a database

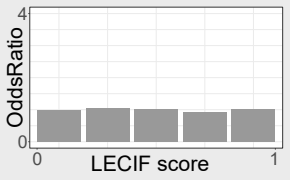
f(x)

Arabidopsis thaliana

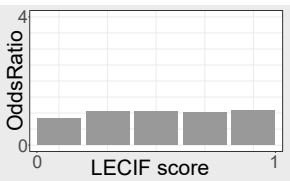
Genetic Variability

B Enrichment of At GWAS variants in LECIF binned regions

Oryza: top side



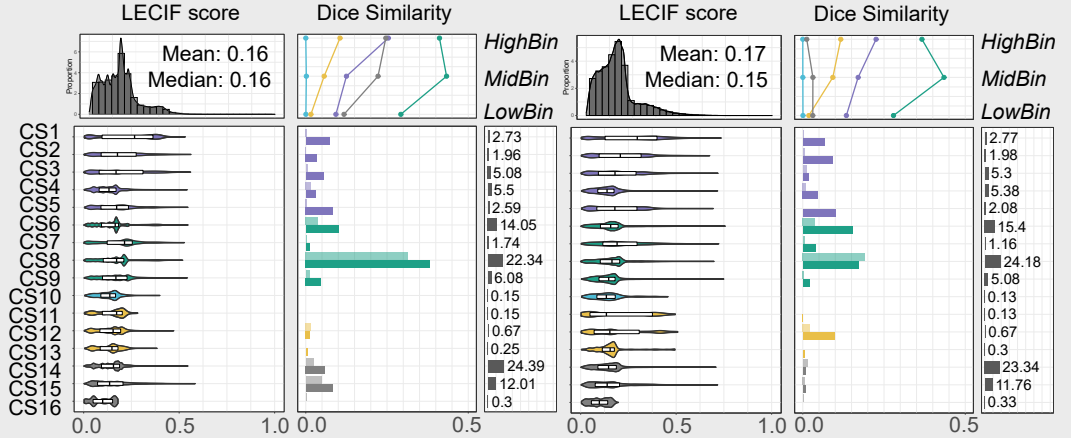
Zea: bot side



Chromatin States

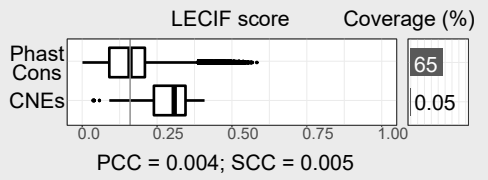
E

Oryza: left side | Zea: right side



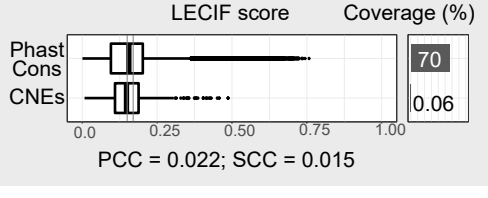
H **Comparative Genomics Conservation**

Oryza: top side



PCC = 0.004; SCC = 0.005

Zea: bot side



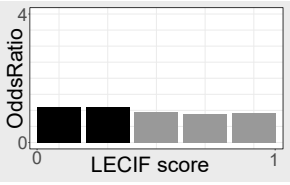
PCC = 0.022; SCC = 0.015

Oryza sativa

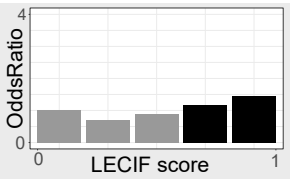
Genetic Variability

C Enrichment of Os GWAS variants in LECIF binned regions

Arabidopsis: top side



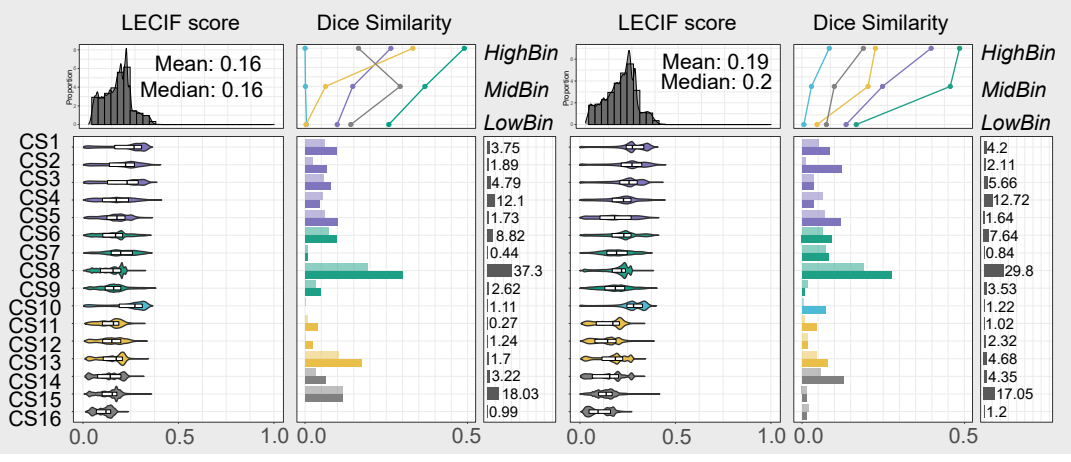
Zea: bot side



Chromatin States

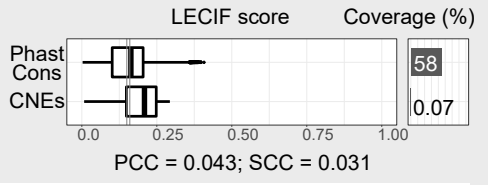
F

Arabidopsis: left side | Zea: right side



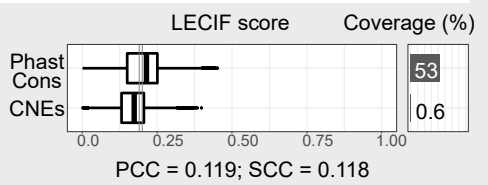
I **Comparative Genomics Conservation**

Arabidopsis: top side



PCC = 0.043; SCC = 0.031

Zea: bot side



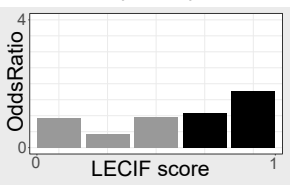
PCC = 0.119; SCC = 0.118

Zea mays

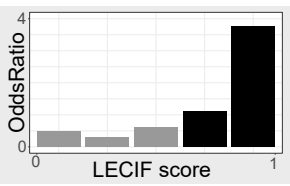
Genetic Variability

D Enrichment of Zm GWAS variants in LECIF binned regions

Arabidopsis: top side



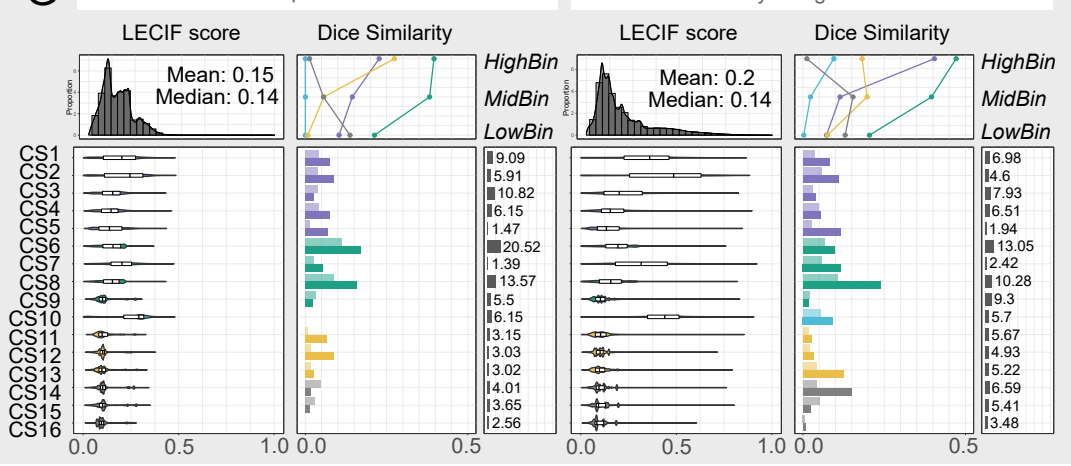
Oryza: bot side



Chromatin States

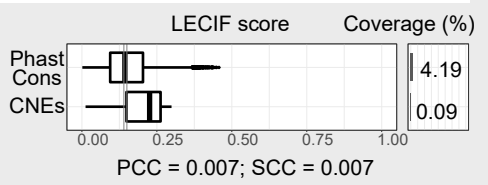
G

Arabidopsis: left side | Oryza: right side



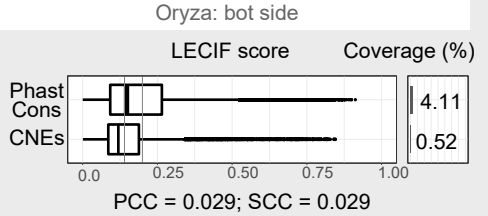
J **Comparative Genomics Conservation**

Arabidopsis: top side

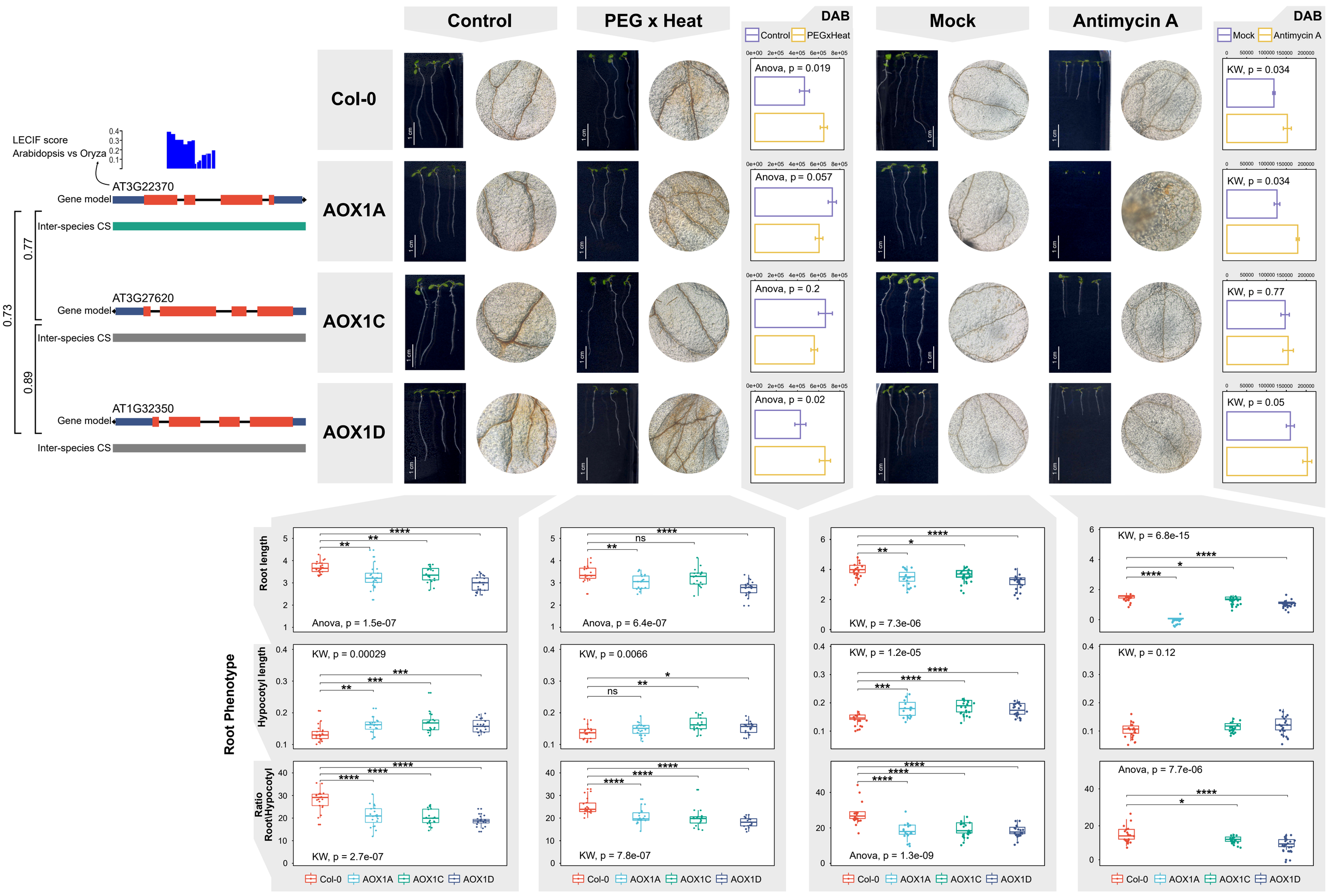


PCC = 0.007; SCC = 0.007

Oryza: bot side



PCC = 0.029; SCC = 0.029


Control
PEG x Heat
DAB
Control
PEGxHeat
Mock
Antimycin A
DAB
Mock
Antimycin A
LECIF score
Arabidopsis vs Oryza
0.4
0.3
0.2
0.1
AT3G22370
Gene model
Inter-species CS
0.77
AT3G27620
Gene model
Inter-species CS
0.73
0.89
AT1G32350
Gene model
Inter-species CS
Col-0
AOX1A
AOX1C
AOX1D
1 cm
0e+00 2e+05 4e+05 6e+05 8e+05
Anova, p = 0.019
Anova, p = 0.057
Anova, p = 0.2
Anova, p = 0.02
0 50000 100000 150000 200000
KW, p = 0.034
KW, p = 0.034
KW, p = 0.77
KW, p = 0.05
Root Phenotype
Root length
Anova, p = 1.5e-07
Anova, p = 6.4e-07
KW, p = 7.3e-06
KW, p = 6.8e-15
Hypocotyl length
KW, p = 0.00029
KW, p = 0.0066
KW, p = 1.2e-05
KW, p = 0.12
Ratio RootHypocotyl
KW, p = 2.7e-07
KW, p = 7.8e-07
Anova, p = 1.3e-09
Anova, p = 7.7e-06
Col-0
AOX1A
AOX1C
AOX1D