# Results

## Characterization of shared and species-specific chromatin states

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

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

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

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

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

## Chromatin states features improve predictions of paralogs functional divergence

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

Párrafo Ezoe

Parrafo Cusack

Mini conclusion

## Defining functional genomics conservation score

## Experimental validation of potential divergent duplicates

## Discussion

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

Not replace and wide range of genomic prediction properties for the community. Diversity compared to mammals. thus highlighting plant kingdom epigenomic complexity. Deep and narrow vs shallow and broad - community.

integrative features needed for genomic elements and patterns discovery

Despite these pairs do not pass the stringent threshold, they presented high enough DFD values to be considered high divergent paralogs. We decided to assess AOX redundancy by monitoring root phenotypes under two different stresses, considering previously described roles of these genes in response and retrograde-signalling (cita); 2/5 paralogs are not root expressed, simplifying the system and evaluation in seedling stages.

DB and future of evolutionary epigenomics.