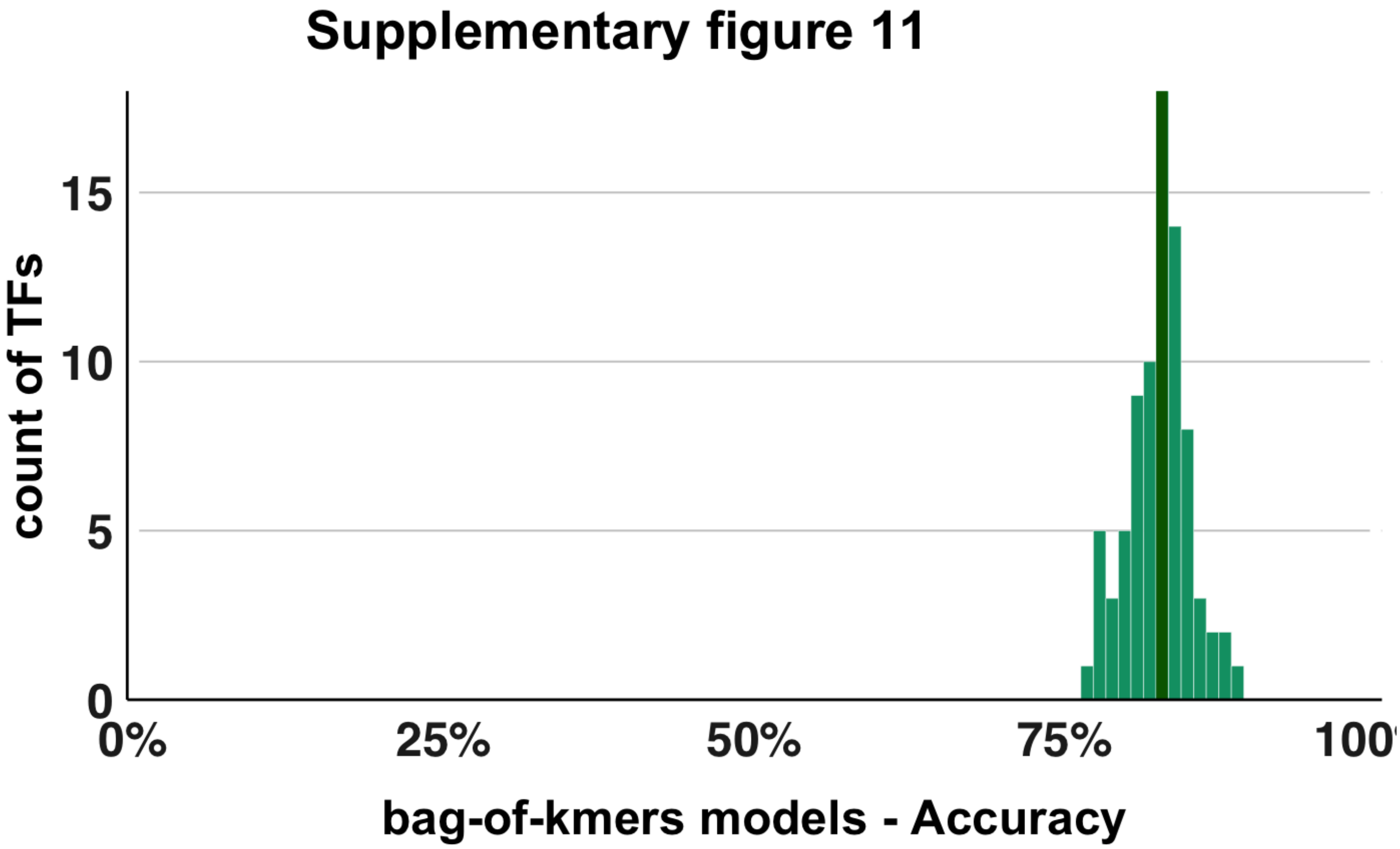
Code ▾

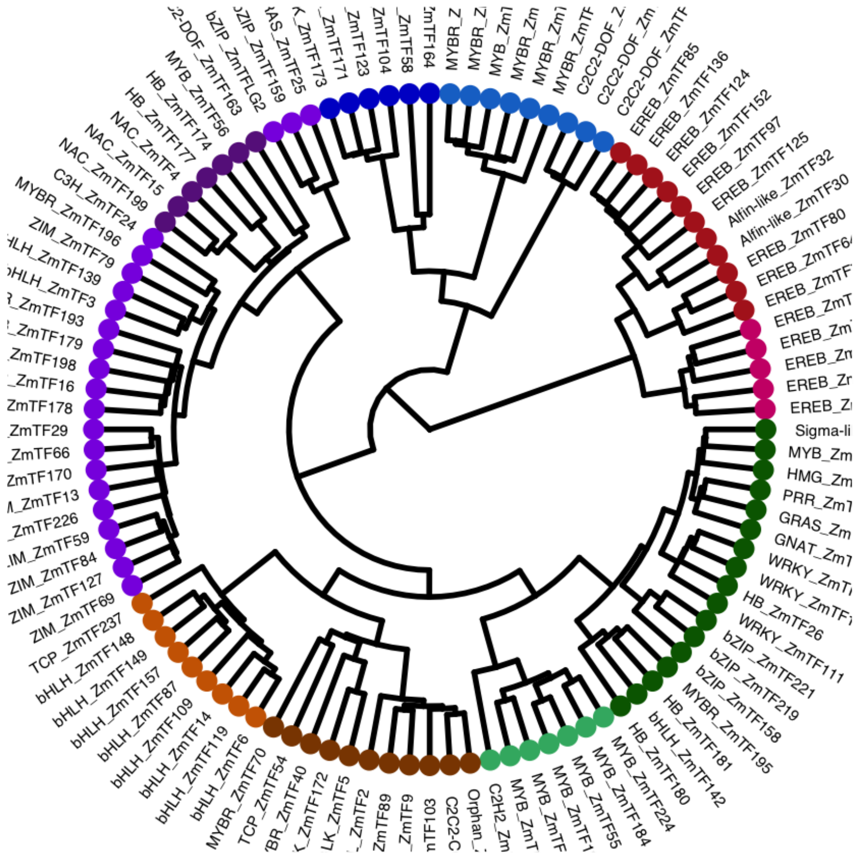
Rules of TF binding

Next, we assayed the potential for predicting individual TF binding using either sequence information, or co-localization information (Fig 1A). To model TF binding from sequence we applied a machine learning approach (i.e., “bag-of-k-mers”, [1]) to discriminate TF binding regions from other regions in the genome, which resulted in reliable models for all the TFs (5 fold cross-validation, average accuracy for each TF > 70%, Supplementary figure 11).



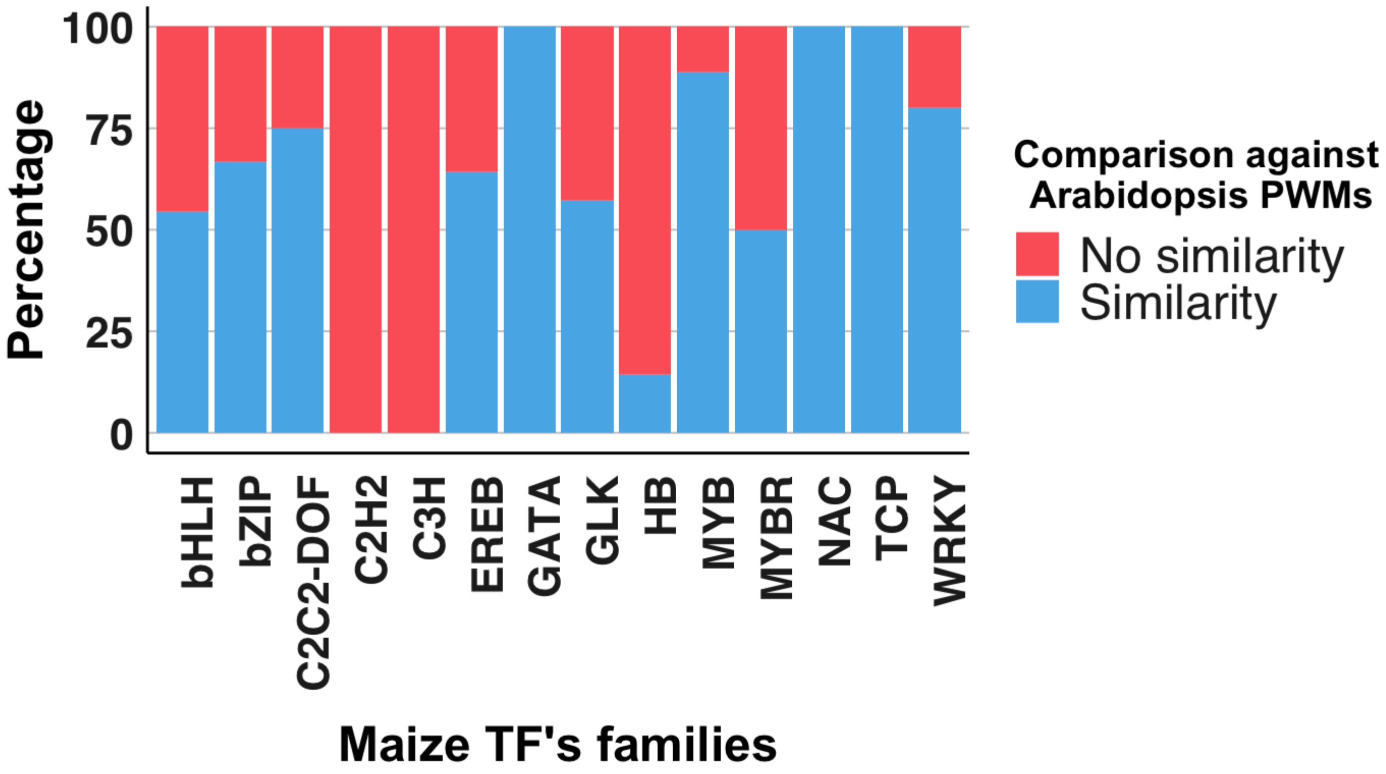
Using average k-mer weights obtained from the “bag-of-k-mers”, we derived a distance matrix among TFs, and a dendogram to summarize sequence similarity relationships (Fig 4A). After removal of singleton families, we observed that for 85% of them, the majority of their members (>= 50%) belong to the same group in the dendogram (Fig 4A).

4A



This observation prompted us to evaluate conservation of TF sequence preferences across species, as TFs families are well conserved across the plant lineage. Using top predictive k-mers for each TF, we examined their similarity to Arabidopsis PWMs[2]. After removal of families that did not have counterpart (or were poorly represented), in the Arabidopsis collection, 50 out of 81 (61%) of the evaluated TFs preferentially matched PWMs to their corresponding family in Arabidopsis (Fig 4B) (TOMTOM (*missing citation*) p-value < 0.001). At family level, we identified 11 out of 14 families that show overall sequence conservation, with >= 50% of their members binding to sequences alike to their Arabidopsis counterpart (Fig 4B).

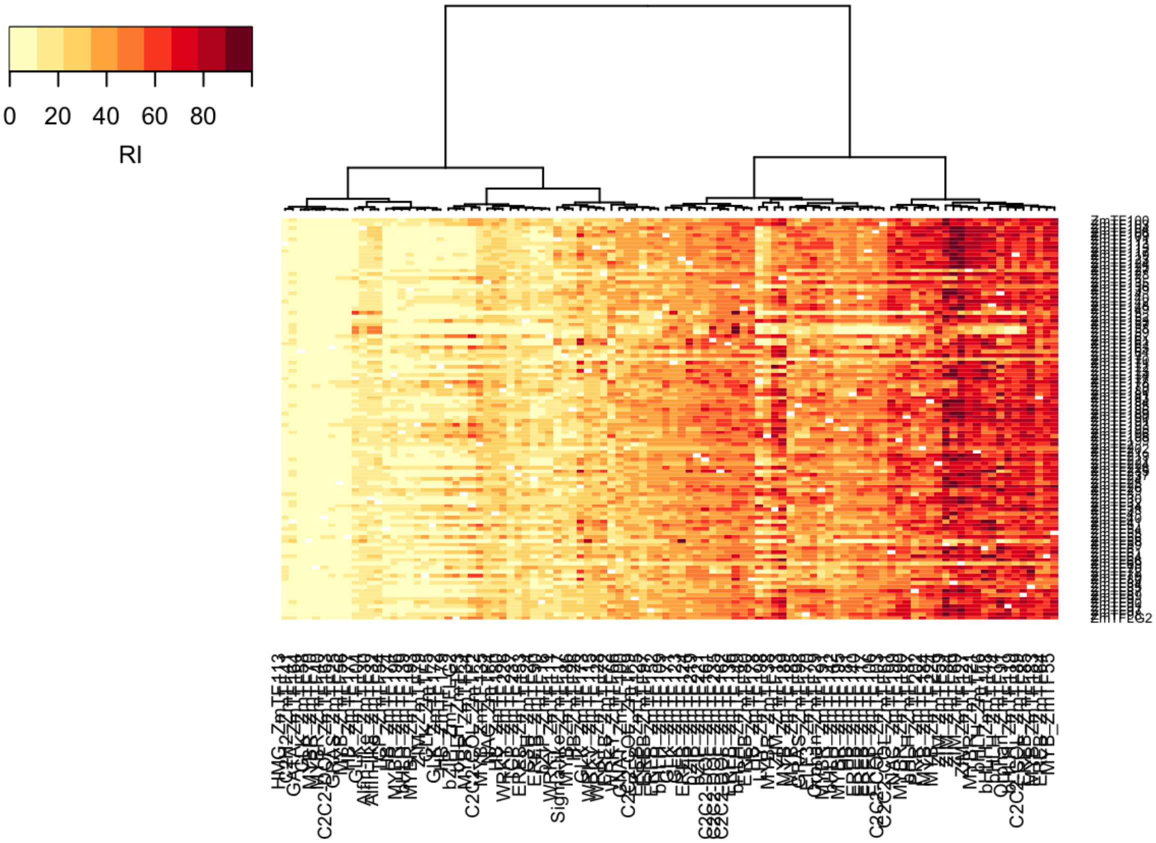
4B



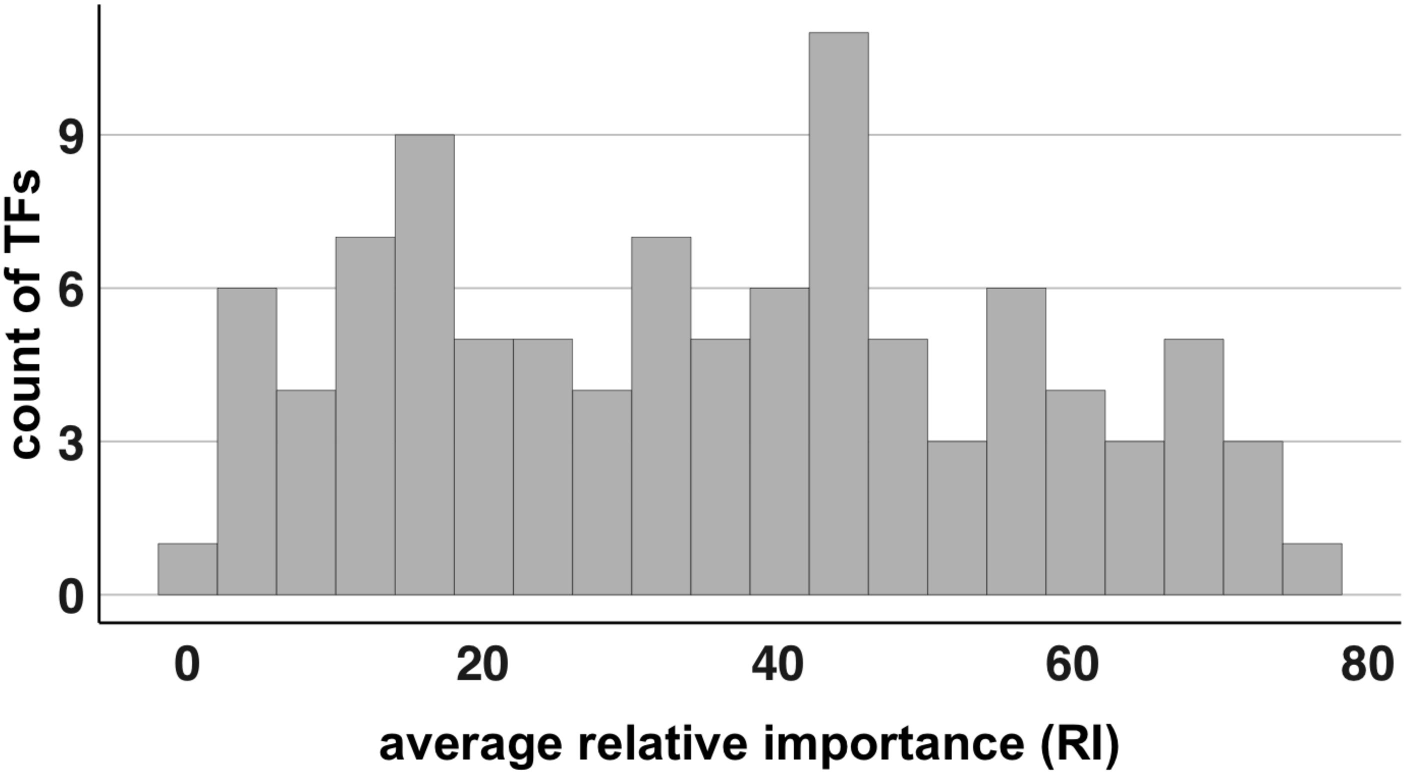
The clustering of Maize TFs with members of the same family based on sequence preferences, and the “motif” similarity between Maize and Arabidopsis TFs, appear driven by similarities among DNA-binding domains, which could favor functional redundancy, as a backup mechanism to maintain system robustness or diversify the regulatory network.

To model TF binding from co-localization, we adopted a machine learning approach to learn non-linear dependencies among TFs used in the ENCODE project [3]. In brief a co-localization model requires as input: a co-localization matrix for each TF (i.e., “focus TF”) with values of peak intensity for all the overlapping peaks that correspond to remaining TFs (i.e., partner TFs); and a randomized version of the matrix [4]. The output, is a set of combinatorial rules that can predict TF binding (10 models with independent randomized matrices, average accuracy for each TF > 95%), and can be used to derive statistics to inform about the “combinatorial potential” of each TF. We scored the relative importance (RI) of each partner TF for the joint distribution of the set of peaks for a focus TFs; RI can be seen as a summary of the influence of a partner TF with respect to a given focus TF (Supplementary figure 12), and the average RI as a global summary (Fig 4C).

Supplementary figure 12

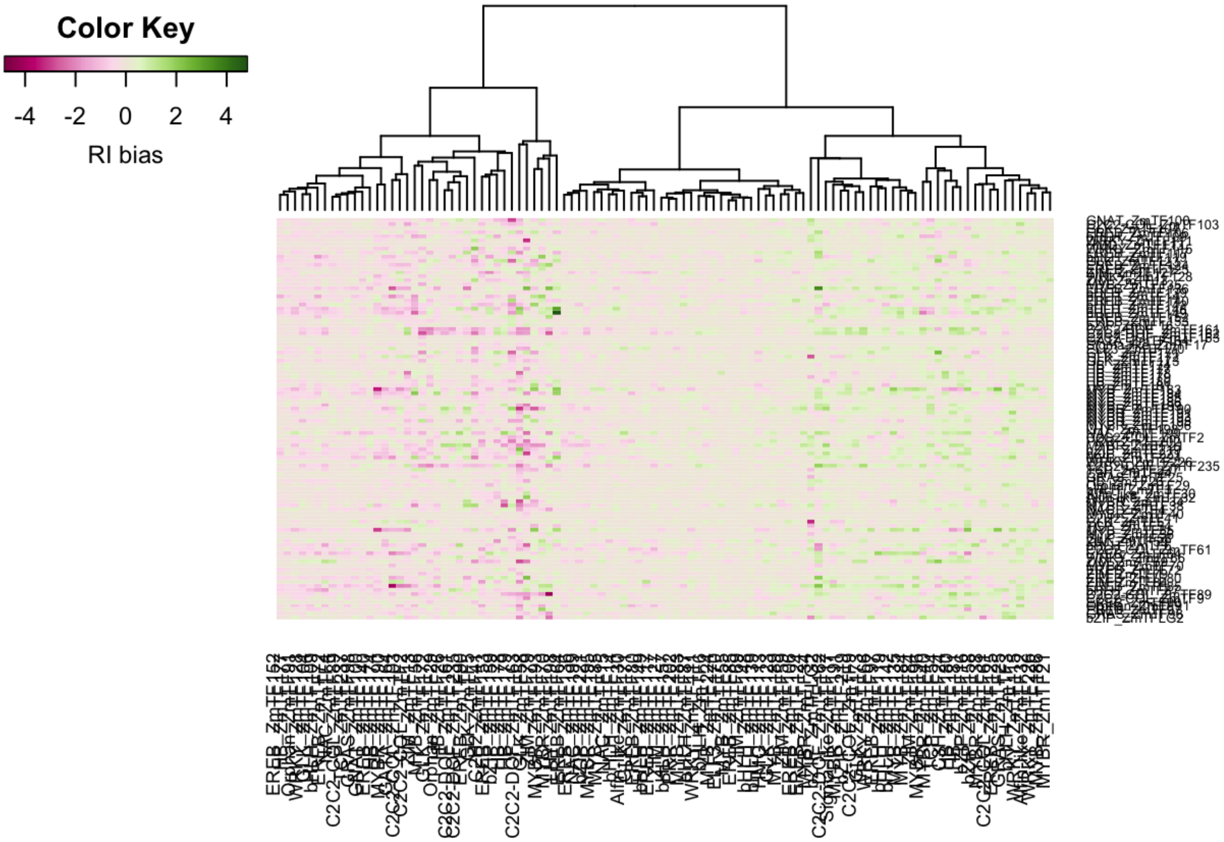


4C

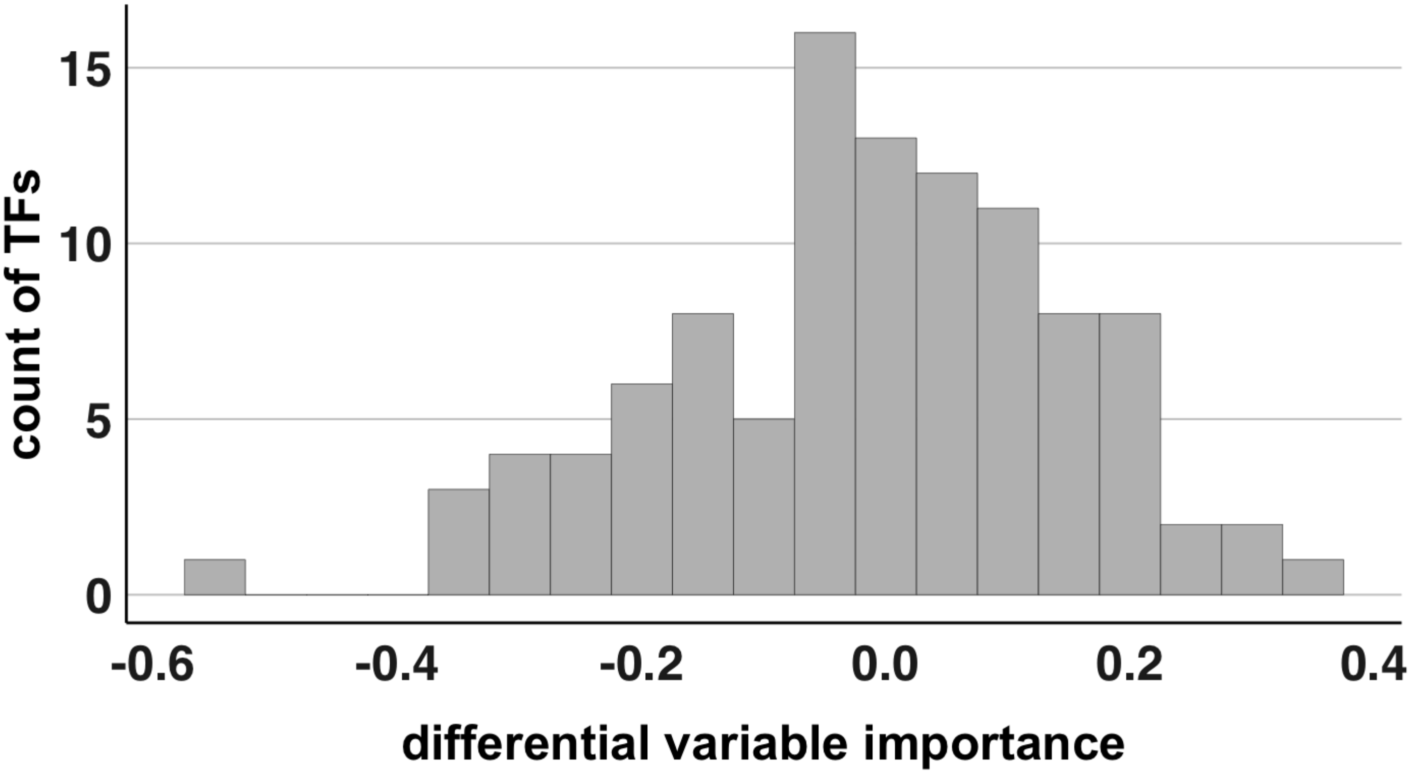


In addition, we derived a differential importance (DI) score for each TF, that describe how the RI changes between subsets of peaks (i.e., genic proximal and distal) (Supplementary figure 13), and calculated the average DI to score genome-wide bias (Fig 4D).

Supplementary figure 13



4D



From examination of the RI we identified TFs that have, in general, a low combinatorial potential, as well as examples of TFs that are important partners to predict a large number of focus TFs (Supplementary figure 12). But we found that most of the TFs were important for specific focus TFs, for instance HB66 showed a low RI for all the focus TFs, except for NAC109 (Supplementary figure 12). The stack of DIs for all focus TFs (Supplementary figure 13) as well as the distribution of the average DI (Fig 4D) indicated that for several TFs the relative importance is different between proximal and distal regions. The differences are modest but consistent, and a similar number of TFs were bias towards the proximal and distal regions. Taken together, our co-localization model suggested a large combinatorial space for TF binding sites that likely favors the occurrence of specific combinations –and perhaps specific functions– for different genomic contexts.

We have generated a compendium of regulatory regions consisting of 104 maize TF in-vivo binding profiles, that massively overlaps with open chromatin regions determined with ATAC-seq. The depth and breadth of the data made possible the first system-view of how TFs are organized in monocots leaves, with important implications for understanding the detailed mechanisms and general architecture of the regulatory network that determines molecular and complex plant traits. We have identified over two million TF binding sites that make-up for ~2% of the genome. These sites have low sequence diversity, suggesting that regulatory interactions are under selection, and are genuine functional. TF binding sites were enriched in cis-expression QTLs, and GWAS hits for several traits which illustrate how understanding regulatory mechanisms is crucial to interpreting functional variants. Using this dataset, we have constructed a graph that provides regulatory hypothesis for 50% of the genes in the maize genome. The architecture of the maize leaf regulatory network has a similar topology (scale-free) to others real-world network, with TFs acting as target hubs, and topological modules for which we inferred biological functions. At this stage, an important limitation in our model is the lack of long-range chromatin interaction information to annotate the distal TF binding sites, which account for 30% of our data. However, as other approaches, such as ChIA-Pet and HiChIP, are currently being generated in the community, incorporation of these data will unravel a more complete view with an additional layer of regulatory hypothesis that will enrich the topology of the graph. Finally, we generated quantitative and interpretable models of the data, that indicates substantial redundancy among TF families, and a large number of possible combinations of TF binding site that are key to specificity.

1. Mejia-Guerra, M. K., & Buckler, E. S. (2019). A k-mer grammar analysis to uncover maize regulatory architecture. BMC Plant Biology, 19(1), 103. (http://doi.org/10.1186/s12870-019-1693-2)
2. O'Malley, R. C., Huang, S.-S. C., Song, L., Lewsey, M. G., Bartlett, A., Nery, J. R., et al. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell, 165(5), 1280–1292. (http://doi.org/10.1016/j.cell.2016.04.038)
3. Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., et al. (2012). Architecture of the human regulatory network derived from ENCODE data. Nature, 489(7414), 91–100. (http://doi.org/10.1038/nature11245)
4. Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. The Annals of Applied Statistics, 2(3), 916–954. (http://doi.org/10.1214/07-AOAS148)