Similarity Regression predicts evolution of transcription factor sequence specificity

Samuel A. Lambert[1], Ally Yang[2], Alexander Sasse[1], Gwendolyn Cowley[3], Mihai Albu[2], Mark X. Caddick[3], Quaid D. Morris[1,2,4], Matthew T. Weirauch[5], and Timothy R. Hughes[1,2,6,§]

[1]Department of Molecular Genetics, University of Toronto, Toronto ON M5S 1A8, Canada

[2]Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto ON M5S 3E1, Canada

[3]Institute of Integrative Biology, University of Liverpool, The Biosciences Building, Crown Street, Liverpool, L69 7ZB, UK

[4]Department of Computer Science, University of Toronto, Toronto, ON M5T 3A1, Canada

[5]Center for Autoimmune Genomics and Etiology (CAGE) and Divisions of Biomedical Informatics and Developmental Biology, Cincinnati Children's Hospital Medical Center, Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA

[6]Canadian Institutes For Advanced Research, Toronto, Canada

[§]To whom correspondence should be addressed:
t.hughes@utoronto.ca
416-946-8260

ABSTRACT

Transcription factor (TF) binding specificities (motifs) are essential to the analysis of noncoding DNA and gene regulation. Accurate prediction of the sequence specificities of TFs is critical, because the hundreds of sequenced eukaryotic genomes encompass hundreds of thousands of TFs, and assaying each is currently infeasible. There is ongoing controversy regarding the efficacy of motif prediction methods, as well as the degree of motif diversification among related species. Here, we describe Similarity Regression (SR), a significantly improved method for predicting motifs. We have updated and expanded the Cis-BP database using SR, and validate its predictive capacity with new data from diverse eukaryotic TFs. SR inherently quantifies TF motif evolution, and we show that previous claims of near-complete conservation of motifs between human and Drosophila are grossly inflated, with nearly half the motifs in each species absent from the other. We conclude that diversification in DNA binding motifs is pervasive, and present a new tool and updated resource to study TF diversity and gene regulation across eukaryotes.

INTRODUCTION

To understand the function of noncoding DNA, e.g. in gene regulation, it is essential to know the potential transcription factors (TFs) that can bind to any sequence. Libraries of experimentally-derived TF motifs, most typically position weight matrices [1], are widely used, and encompass at most a few thousand motifs, oriented mainly towards well-studied TFs in human and model systems (e.g. JASPAR) [2]. Hundreds of eukaryotic genomes have now been sequenced, however, and analysis of gene expression and corresponding sequences in regulatory regions can be performed in virtually all of them. To enable such analyses, we previously described Cis-BP, a database of predicted TF motifs for 59,998 TFs from 340 sequenced eukaryotes [3]. The predictions in Cis-BP were made by simple amino acid sequence identity between DNA-binding domains (DBDs), with cutoffs for each DBD type established on the basis of replicate experiments and pairwise comparisons of motifs from different proteins with homologous DBD types.

The initial Cis-BP system was clearly a first approximation, as it did not utilize known “specificity residues”, or prioritize DNA-contacting residues, and yielded an estimated 89% precision (with undetermined recall). Other approaches have been developed to predict motifs, including Affinity Regression (AR) [4], which predicts affinity to DNA/RNA k-mers on the basis of amino acid k-mer composition of proteins. AR was applied to only two families, however -

homeodomain TFs and RRM-containing RNA-binding proteins. Global "recognition codes" have also been described, which predict binding motifs on the basis of DNA-contacting residues, for C2H2 zinc finger and homeodomain proteins [5-7]. It is unclear whether and how these methods will extend to the ~100 other types of DBDs.

More generally, there is uncertainty in the degree of diversity and evolution of eukaryotic TF motifs. It has been claimed that TF binding specificities are highly conserved between *Drosophila* and mammals [8], but at the same time, the specificity residues for C2H2 zinc finger proteins are very different just among mammals [9]. There are numerous examples of TF diversification in other lineages (e.g. plants and fungi), indicating that TF evolution occurs in parallel to better-established *cis*-regulatory turnover [10]. The degree of divergence of TF motifs is an important question, as it impacts the degree to which gene regulation mechanisms are conserved. How motif diversification relates to protein structure and mechanisms of DNA binding is also largely unknown, except in a few cases [11-13].

We reasoned that developing a system for determining both similarity and dissimilarity of TF motifs would provide uniform and unbiased estimates of the extent to which TF motifs are conserved. Here, we describe an improved system, its incorporation into the Cis-BP database, validation experiments in several eukaryotes, and use of the system to broadly describe TF motif evolution across eukaryotes.

RESULTS

Improved classification of TFs as having similar or different sequence specificity on the basis of protein sequence identity

We improved our previous homology-based system [3] to account for the fact that some positions of the DBDs (e.g. base-contacting or “specificity” residues) have a stronger impact on sequence specificity than others. To do this, we assigned a weight to each residue when calculating similarity between two DBDs of the same class (C2H2, ETS, Forkhead, etc). **Figure 1** displays the overall scheme. We use regression to assign the weights: for each pair of proteins, the independent variables are the binary vector of amino acid similarity at each position of an alignment to the Pfam HMM (**Figure 1**), while the dependent variable is the similarity in DNA sequence preference. The weights are the coefficients learned over all pairs for each DBD class (**Figure 1D**). We tested four variations of this scheme, including two different regression approaches (Linear and Logistic) and two different representations of sequence similarity (identity vs. BLOSUM62). Here, we trained the regression models to learn highly-overlapping 8-mer E-score preferences obtained from universal Protein Binding Microarrays (PBMs [14]); cataloged in Cis-BP [3]. These scores are comparable among different studies, thus circumventing the potentially confounding impact of motif derivation [15]; we note, however, that the scheme could be trained on any metric of motif identity or similarity. Each variation of the scheme generates a different set of weights, which are selected by leave-one-out cross-validation. The best model for each DBD class is among these four (and simple amino acid identity) is also chosen in the same cross-validation. We refer to this procedure as “Similarity Regression” (SR). Application of SR to TF families whose DBDs are present in arrays (e.g. C2H2 ZFs) is explained in **Figure S1**.

SR offers several advantages over previous approaches. For one, it inherently identifies residues that are informative regarding DNA sequence specificity. The weights obtained are highly biased towards DNA contacting regions and “specificity residues”, if known. **Figure 1** illustrates weights for the well-studied homeodomain class, which has established specificity residues in DNA contacting positions [5], also **Figure S3A**). Weights for all eukaryotic DBD families are given in **Supplementary Data 1** (shown for Homeodomains and C2H2 ZFs in **Figures S3 A** and **B** respectively). These weights correspond to known mechanisms of DNA recognition: there is a strong relationship between SR model weight and DNA contact frequency (**Figure S3C**). In addition, SR pinpoints known binding modes: for most TFs, weights are higher in the residues that contact the major groove, which is predominant among TFs. For Sox proteins, however, the weights are much higher in residues contacting the minor groove, consistent with structural data [16], while GAL4/Zinc Cluster proteins, whose dimerization is organized along the DNA backbone [17, 18], receive high weights in backbone contacting residues (**Figure S3C**).

A second advantage of SR is that, relative to overall similarity cutoffs, it confers a dramatic improvement in recall (i.e. total number of positive predictions) at identical precision values (displayed for Homeodomains in **Figure 2A)**, particularly for families with a large amount of PBM data (summarized in **Figure 2B**). In these precision/recall (PR) curves, *positives* are pairs of proteins with E-score overlap that exceeds the 25th percentile of experimental replicates (the same threshold employed in [3]), and *negatives* are all other pairs (note that the use of all other pairs underestimates the predictive potential because it includes experiments that are highly similar but below the stringent threshold utilized).

SR also outperforms the AR method [4] in many cases. SR predicts similarity in DNA sequence specificity of two proteins, while AR directly predicts preferences of TFs/RBPs to individual DNA or RNA sequences on the basis of their protein sequences. Nonetheless, the two can be compared by using SR to predict 8-mer preferences from proteins that should have highly similar sequence preferences (see **Methods** for details). Using an identical training set (i.e. the same experiments on the same proteins), SR slightly outperformed AR when predicting Z-score profiles for 315 held-out constructs across 19 TF families using either the single most similar protein ($p < 0.01$, **Figure S4A**), or by predicting the Z-scores as a composite of up to five most similar proteins ("Top 5", $p < 0.01$, **Figure**

S4A). SR has the added benefit that it abstains from making poor predictions for dissimilar proteins, whereas AR makes a prediction for every protein, without an associated quality metric (**Figure S4**). Compared to AR, SR “Highly Similar” has higher correlation to the measured 8-mer Z-score profiles than AR using the NN ($p < 0.01$), or Top 5 predictions $p < 0.0001$). These outcomes hold for most (albeit not all) individual TF families analyzed in isolation. For example, while SR performs equivalently to AR for Zinc cluster TFs, it scores higher for the Homeodomains and C2H2 ZF families (**Figure S4B-D**).

Importantly, Similarity Regression can also be used to predict whether two proteins are highly unlikely to share DNA sequence specificity: employing the same learned weights described above, a threshold can be identified below which proteins will almost always bind very different sequences. In this analysis, we defined different sequence preferences to be an overlap of 20% or less among the highly preferred 8-mers. We allowed some overlap because many families bind a characteristic sequence “core” (e.g. many homeodomains bind TAAT-like sequences, even though their most highly-preferred 8-mers differ among family members). For each DBD type, we set an SR score threshold using a negative predictive value (NPV), at which 95% of pairs of proteins at that similarity score indeed have different sequence preferences. As shown in **Figure S5A**, SR outperformed unweighted alignments at discriminating these pairs with dissimilar sequence preference (increased specificity at the same NPV).

The two score thresholds obtained (one that predicts identity in sequence preferences, and the other that predicts difference in sequence preferences) are typically very different, such that there is a middle ground we refer to as “ambiguous”. **Figure 2C** shows that the ambiguous score range is, in fact, predictive of intermediate 8-mer overlap for Homeodomain TFs; the same phenomenon is observed in other TF families (data not shown). In all subsequent analyses, we therefore use the weighted models to classify all pairs of proteins sharing the same DBD type as either “Highly Similar”, “Ambiguous”, or “Dissimilar.” Multi-class accuracy of SR models and their improvement over %ID approaches are summarized by the Matthews Correlation Coefficient (MCC, **Figure S5B**), showing that SR outperforms unweighted alignments in all but four TF families.

Validation of sequence specificity classifications using new PBM data

To confirm that the models correctly classify previously unseen proteins, we generated new PBM data for 340 TFs representing multiple eukaryotic kingdoms, with a particular focus on *Cannabis sativa* (a medicinal plant), *Caenorhabditis briggsae* (a nematode), *Aspergillus nidulans* and *Neurospora crassa* (model fungi), and also 15 human TFs (**Table S1**). These TFs were selected on the basis of at least one of two different criteria: first, to increase the number of experimentally determined motifs for TFs in these species of interest, and second, to obtain novel motifs by analyzing proteins that are dissimilar to TFs with known motifs. We used these data as a validation set to test how well SR models measure the similarity of TF sequence specificity on unseen data (**Figure 2D**). The TF similarity classifications for the newly analyzed proteins are correctly predicted for 81.2% of the predicted Highly Similar and 95.2% of the predicted Dissimilar pairs, regardless of their level of similarity to other proteins in Cis-BP, confirming that the models are accurate with independent data. Indeed, there is an overall correlation between SR score and 8-mer overlap between the held-out data and the most similar training construct (by SR score) for each TF family's SR model (**Figure S6**, median $R^2 = 0.63$). **Figure 3** provides examples of conservation and divergence of motifs in the new data.

New TF similarity predictions, motifs, and genomes improve Cis-BP

To capitalize on the increased recall of SR relative to unweighted alignments, we implemented the method in Cis-BP, which compiles known TF motifs and tracks homology relationships among similar TFs. Since Cis-BP was described in 2014, both the number of sequenced eukaryotes and the number of known motifs has roughly doubled. We therefore updated Cis-BP, which now includes 741 genomes (previously 340) and 11,493 experimentally determined motifs, corresponding to 4,560 distinct proteins (previously 6,559 motifs for 3,202 distinct proteins), and implemented SR across all 392,333 known and putative eukaryotic TFs. We also updated many other properties of the database (e.g. genome builds and DBD models) (see **Online Methods**).

The incorporation of SR to Cis-BP increases the number of TFs with predicted motifs by more than 25,000 compared to our previous method, at the same expected precision - a 16% overall increase, on identical genomes, DBDs, and motifs). Coverage of numerous TF families is increased dramatically (**Figure S7A**). For instance, 10 TF families more than doubled their motif coverage, including Zinc cluster TFs (123% increase) and Sox (162% increase), the second and seventh most abundant families in Cis-BP respectively. The average species now has 7% more TFs with motifs (experimental and predicted), yielding an average motif coverage of 41% (with 75% for human) (**Figure S7B**) and a total coverage of 158,606 out of 392,333 eukaryotic TFs (40.4%). This updated Cis-BP database can be found at http://cisbp.ccbr.utoronto.ca/, where TF annotations, motifs, and PBM data compiled from our lab and other public databases can be accessed, and downloaded. In addition to increased coverage, the new build, which contains many more genomes, also reveals many new families of TFs with still-unknown sequence specificity.

Evolution of sequence specificity across Eukarya

Finally, we used the motif predictions and the Cis-BP update to gain an overview of TF motif conservation and divergence over eukaryotic evolution. We focused on 84 species with well-annotated genomes (present in Ensembl and/or Uniprot, species listed in **Figure S7B**). For each protein, we identified the protein with the highest SR model score as described above in each other species, and recorded the classification (i.e. highly similar, ambiguous, dissimilar). If there is no protein with the same DBD type in the other species, then the TF is labeled as “DBD not shared” with the other species. Thus, there are four possible labels for each TF/species comparison, and they are mutually exclusive.

Figure 4 shows that eukaryotic kingdoms display qualitatively similar trends in the proportion of TFs within each of the categories above, with respect to divergence time. At ~100 Mya (e.g. origin of placental mammals, and eudicot plants), ~75% of motifs are conserved (highly similar) and an additional ~5% are potentially conserved (ambiguous category). But at 900 Mya (origin of metazoans), only ~60% are conserved or potentially conserved; a similar proportion is obtained for the origin of fungi (~1055 mya). Within the plant kingdom (~1160 Mya), only slightly more motifs are conserved or potentially conserved (~65%). Across kingdoms (e.g. between fungi and metazoan), most DBDs are not shared [19], and are thus not comparable. Even among those that are comparable (i.e. DBD families that are present in both), the majority have dissimilar or ambiguous motifs.

Much of the divergence in motifs occurs in a small number of TF families (**Figure 5 and Figure S8**), but these families have a large number of members, and in general are already known for their lineage-specific expansions: C2H2 zinc fingers in metazoa, Nuclear Hormone Receptors in nematodes, and Myb proteins in plants. The SR analysis thus underscores DNA sequence specificity as a mode of diversification following duplication of these proteins. Many other families appear rigid in their DNA binding motifs, however, and presumably diversify in function by other mechanisms (e.g. bZIP and bHLH proteins are able to diversify through changes in heterodimerization partners) [20, 21].

One striking example of C2H2 diversification is counter to a previous claim in the literature, but is supported by extensive experimental data. A previous study [8] claimed that there is near-perfect conservation of binding motifs for TFs between human and *Drosophila*. This discrepancy appears to be due to the fact that the Nitta study was highly biased towards families that do not diversify, while C2H2 zinc fingers - the largest class of TFs in both species – were represented by only a few examples. SR predicts that the vast majority of C2H2s ZF proteins do not have conserved motifs (**Figure 6A**), and existing experimental data confirm this prediction (**Figure 6B** and **6C**). Even those that have 1-to-1 orthology relationships often differ substantially in their DNA binding specificity, illustrating that simple orthology alone can be a poor predictor of shared motifs (**Figure 6B**). As a control, C2H2 proteins predicted by SR to have highly similar motifs between human and Drosophila do display highly similar motifs in the experimental data (**Figure S9**), even though they were obtained using different techniques (primarily HT-SELEX [8, 22] vs. Bacterial 1-hybrid [23, 24]).

DISCUSSION

We anticipate that SR will contribute to our understanding of TF function in several ways. First, it presents several advantages in the task of predicting motifs. Like simple homology (i.e. percent identity), the score it produces serves as a confidence measure that can be used to avoid incorrect predictions. At the same time, the dramatically increased recall (i.e. coverage) of SR, relative to percent identity, provides a large increase in the number of predicted motifs, which are now included in our update of the Cis-BP database.

Second, the weights (i.e. coefficients) produced by SR are often highest for known specificity residues and DNA contacting positions. Thus, unstudied positions with high weights represent candidates for new determinants of TF sequence specificity. Together with structural data, these weights may also shed new light on biophysical aspects of DNA binding.

SR can also predict when proteins are unlikely to share sequence preferences. To our knowledge, prior to this study, there has been no systematic examination of the overall degree of *trans*-regulatory change among eukaryotes. Our analyses lend strong support to the notion that *cis*-regulatory turnover is accompanied by alterations to *trans*-regulators, even over relatively short timescales (<100 My), and that these changes are concentrated in large families with established patterns of diversification. This study is the first major analysis of

TF sequence specificity for both *Cannabis* and *Aspergillus*, and both the outputs of SR and the new data generated highlight the diversity of DNA binding motifs in both the plant and fungal lineages. Despite less diversity in the specificity residues of individual C2H2-ZF domains of fungi, relative to metazoa [13], proteins containing these domains contribute substantially to diversification of motifs in fungi, presumably due to the fact that multiple C2H2 domains can be combined in different ways. Myb domains also contribute substantially in multiple lineages (both plants and fungi). The GAL4/ZnClus domain proteins, which have also expanded in fungi, have largely conserved monomeric binding specificity in their DBDs, and thus more likely contribute to TF diversification by alterations in spacing and orientation of dimeric sites as homo or heterodimers [25].

SR also confirms the extreme diversity of motifs in the C2H2-ZF family. C2H2-ZFs are the fastest evolving TF family in the recent human lineage [26], and SR indicates (and experimental data confirm) that their sequence specificities are largely distinct from those in Drosophila, even among their clear orthologs. Intriguingly, in *Drosophila* species, even 1-1 orthologs of C2H2 TFs frequently differ in specificity residues, and these differences are predicted to impact DNA sequence preferences [27]. In human, there is strong evidence that retroelement silencing by KRAB-containing C2H2-ZFs plays a role in their evolution; it is unclear what the driving force is, outside of tetrapods, to which the KRAB domain is restricted.

Knowing the sequence specificities of TFs is an important first step in their characterization. Overall, we anticipate that SR and the results it produces will represent a major advance in our understanding of the function and evolution of both TFs and gene regulatory mechanisms.

METHODS

Similarity Regression (SR). SR is formulated as a regression task where the dependent variable (Y) is a metric of similarity in DNA sequence specificity between pairs of proteins (see below), and the independent variables (the feature vector X) are identity or similarity in amino acid residues at each individual position of the aligned DBDs, for the same pairs of proteins. To make the alignment, each instance of a DBD is aligned to its corresponding Pfam HMM using the semi-global method implemented in *aphid* [28], recording match positions (i.e. positions present in the HMM). An example alignment of two homeodomain sequences is presented in **Figure 1A**. At each position of the aligned sequences, either identity (as binary values) or similarity (BLOSUM62 substitution score [29]) is recorded (**Figure 1B**), yielding the feature vector for each the TF pair. For TF families that have DBDs present in arrays (mainly C2H2 ZFs and Myb/SANT) the best un-gapped and overlapping pairwise alignment of DBD arrays (**Figure S1A**) is found by selecting the alignment offset with the maximum amino acid identity. For a multi-DBD alignment, the feature vector is generated by the average score (identity or similarity) in each position of the DBD alignment from all DBD arrays, normalizing by the DBD length of the longest protein (**Figure S1B**).

In the analyses described, the metric of similarity in DNA sequence specificity between pairs of proteins (Y) is calculated from the 8-mer PBM data as the fraction of high-scoring 8-mers (E-score > 0.45) that are shared between two TFs (i.e. intersection/union for two experiments, referred to as “E-score overlap”). For each TF family, E-score overlaps that exceed the 25th percentile of experimental replicates (the same threshold employed in [3]) are taken as having “Highly Similar” specificities. The Highly Similar labels are used as positives for training logistic SR models (see next paragraph), and also for evaluating the performance of SR (e.g. by Precision-Recall (PR) analysis). E-score overlaps less than 0.2 are taken as having “Dissimilar” specificities, allowing some overlap because many families bind a characteristic sequence while their highest-affinity 8-mers differ. The Dissimilar labels are used as negatives to define the score threshold below which TFs are unlikely to share specificities in a Negative Predictive Value (NPV) analysis.

For each TF family, we trained four SR models that varied in the representation of protein similarity (identity or BLOSSUM substitution score) and in the representation of the data (either linear or logistic regression models). Each regression model is trained in R [30] using *glmnet* [31] constrained to fit positive regression coefficients, selecting the optimal Ridge (L_2) regularization strength using cross-validation (CV). Since the data consist of pairs, normal *k*-fold cross-validation is invalid, as random training and test splits would not be independent. To solve this problem, we train the models using leave-one-TF-out CV (testing on data points made from a single TF’s comparisons) which we implemented using the *caret* package [32]. This performance measure can be interpreted as how well an SR model generalizes to unseen TFs, and is used to select the optimal regularization parameters and score thresholds for each regression model.

An outline of the SR model generation and selection for Homeodomain TFs is presented in **Figure S2**. First the optimal regularization strength is selected using the CV procedure implemented in *caret*, yielding a selected model for each feature/output combination. For each regression model, and the unweighted alignment identity method, two thresholds are derived to predict TFs with Highly Similar, or Dissimilar specificities. To select these thresholds, the predictions on held-out data from each CV fold are combined and compared with their known TF similarity labels. To identify TFs with Highly Similar sequence specificities (E-score overlap > TF family replicate threshold) a Precision-Recall (PR) curve is generated on the held-out data, and a score threshold is selected from the curve such that it yields 75% precision (a heuristic identical to that in our previous study [3]). A threshold for Dissimilar specificities is derived by finding a Negative Predictive Value (NPV) cutoff that classifies 95% of TFs below that score threshold as having truly dissimilar specificities (E-score overlap < 0.2). For each threshold, the recall of positive and negative predictions was recorded to evaluate the improvement of SR models over unweighted alignments. The Highly Similar and Dissimilar thresholds are then applied to the predictions to classify each TF pair in the held-out data as having Highly Similar, Ambiguous or Dissimilar specificities for each SR model (and for the unweighted alignment method). The best SR model is then selected by comparing the 3-class predictions to ground truth labels and selecting the model with the best Matthews Correlation Coefficient (MCC), a metric of multi-class classification accuracy that is sensitive to class imbalance. This process yields a single final SR model for each TF family, composed of a weight vector (i.e. coefficients for X values, which are the selected measure of protein similarity), as well as two thresholds for the dependent variable (Y) that are used to predict whether two TFs have Highly Similar, Ambiguous, or Dissimilar sequence specificities.

Comparing SR weights with known DNA-contacting residues. We used the DNAproDB database [33] to compare the SR weights with known protein-DNA contacts. DNAproDB catalogues DNA–protein complexes present in the Protein Data Bank [34], annotating the amino acid residues that contact the DNA backbone and bases in the major and minor grooves. We transferred these annotations to our models by first extracting all the protein sequences in DNAproDB and identified DBDs using *hmmscan* and the same Pfam HMM models and thresholds as Cis-BP. We then parsed the nucleotide-residue interactions for each structure into backbone, major, and minor groove interactions (scored using DNAproDB recommended Buried solvent Accessible Surface Area (BASA), hydrogen bond, and van der Waals interaction thresholds), and associated them with the position of the residue in the DBD alignment. We represented the interactions as a Contact Frequency for each type of DNA contact, by normalizing the number of nucleotide-residue interactions that occurred in each position of the DBD by the number of protein-DNA structures containing that

DBD. Correspondence between SR weights and the three classes of DNA contacts were evaluated using partial correlations, which assess the correlation between each contact type after removing the effects of the other two contacts on the SR weights.

Comparison of SR with Affinity Regression. Affinity Regression (AR) predicts Z-scores of DNA 8-mers from short peptides in the protein sequence. Here, we implemented a softcoded python version of AR, ensuring similar performance on the original data reported in [4], and using identical constructions of the protein and DNA features. A single AR model for each TF family was trained using the same data as the corresponding SR model, and the number of informative components selected after dimensionality reduction was set to capture 90% of the singular values' weights. To predict the Z-scores of uncharacterized/tested transcription factors, AR determines their protein K-mer vectors to predict the similarities of the held-out TF to all characterized protein profiles in the training set. AR uses these similarities to reconstruct the Z-score profiles weighted by the predicted similarities: (1), using either the nearest (NN) or Top 5 nearest neighbours; and (2), a geometrical reconstruction from the span of the training vectors, proposed and applied in the Affinity Regression paper. AR was applied to the new TFs present in the new PBM data from this study.

We used three means to predict the Z-score profile for each held-out TF using SR: (1) copying the Z-score profile from the protein with the highest SR score

(NN, or Nearest Neighbor); (2) combining the Z-score profiles of the five proteins with the highest SR scores ("Top 5 NNs"), weighting the Z-scores for each of the five by the corresponding SR score; and (3) combining the Z-scores from all TFs in the training set that are predicted by SR to have Highly Similar specificities (SR Highly Similar), weighting the Z-scores for each of them by the corresponding SR score. We evaluated the accuracy of SR and AR predictions using the Pearson correlation coefficient (PCC) between the predicted Z-score profile and the experimental Z-scores. We used paired Wilcoxon signed-rank tests to identify significant differences in mean PCC ranks between Z-score reconstruction methods.

Updates to the Cis-BP database. We performed extensive updates to the Cis-BP database, encompassing changes to both the data and the methodologies. Build 2.0 of Cis-BP now contains data for 741 species (up from 340) (http://cisbp-dev.ccbr.utoronto.ca/ - development version: 1.98d, user name: reviewer, password: checkCisbp). In addition to adding new species, updated genome builds were incorporated for all existing species, where available. Each of these updates includes the latest available protein sequences, protein and gene IDs, gene names, and gene aliases. Further, the set of human TFs contained in Cis-BP now matches the set of 1,639 curated TFs provided in our recent review [26]. DBD scans were performed using updated Pfam HMM models [35], including models for EBF1 (COE1_DBD), FLYWCH, and ICP4 (Herpes_ICP4_N). We also removed models for DP and SART-1, which are now known to not bind DNA with specificity. A total of 1,358 new motifs were obtained from 38 different sources, including 541 HT-SELEX motifs obtained for human TFs from methylated and unmethylated DNA [36], 534 DAP-seq motifs for *Arabidopsis thaliana* [37], 248 HT-SELEX Drosophila melanogaster motifs [8], and 221 ChIP-exo and ChIP-seq-derived C2H2 zinc finger motifs [38]. Existing motif sources such as UNIPROBE [39], Transfac [40], JASPAR [41], and HOCOMOCO [42] were also updated to include data from the latest database builds. In addition to these improvements in the database contents, this update of Cis-BP also incorporates several methodological advances. First, when two predicted DBDs overlap in a given protein, only the DBD with the most significant HMMER p-value is retained. Second, matches to the Pfam Myb/SANT domain are now further subclassified into Myb (which binds DNA specifically; also contains Myb-like sequences which are also likely to bind DNA), or SANT (which does not bind DNA specifically). In brief we scored each Myb/SANT domain with the Myb (PS51294), Myb-like (PS50090), and SANT (PS51293) specific PROSITE [43] models and annotated domains by the profile with the highest score. This procedure is now applied to both remove SANT-only containing proteins (which are not TFs), and remove SANT domains from proteins that contain both Myb and SANT domains. Third, we removed 1-1 orthologs (reciprocal best BLAST hits) of metazoan proteins with false-positive human TFs derived from a recent curation effort [26]. Finally, motif inferences in Cis-BP are now performed using the Similarity Regression approach described in this manuscript, as opposed to the original method, which was based on amino acid identity.

Predicting TF motif conservation across species. To evaluate motif conservation between species we used the TF annotations and DBD sequences from Cis-BP (v2.0). For each pair of species analyzed, we used SR to predict TF similarity for all pairs of TFs from the same TF family. To calculate the conservation of each TF in each species relative to a second species, we report the maximum SR score among all TFs in the second species, and the resulting similarity classification. If the TF was from a family that is not shared between species (e.g. DBD families that are clade-specific) we assume that the motif is not conserved, and report the TF as uncomparable with the label "DBD not shared". We obtained the time to the last common ancestor (Divergence Time) from the TimeTree database [44].

To identify the most similar proteins between human and Drosophila we employed BLASTP [45] with default settings, using full-length TF sequences present in Cis-BP. The closest TF in each species (BLAST NN) was identified using the minimum E-value, and reciprocal best BLAST NNs (putative 1-1 orthologs) were recorded.

DNA Binding Doman (DBD) Cloning. 350 novel *A. nidulans* transcription factor binding domains were been selected for analysis and 180 were successfully cloned into the expression vector (pTH6838) and validated by sequencing. These were cloned using RNA extracted from the wild-type *A. nidulan*s strain (FGSC

A4). cDNA was generated by RT-PCR using random hexameric primers. Proof-reading KOD Hot Start DNA polymerase was used to amplify the DBD-coding region, and extracted from a 1% agarose gel using a Silica Bead DNA Gel Extraction Kit (Thermo Fisher Scientific Inc., 2013). Double digests were performed using the restriction endonucleases AscI (10U/µL) (Thermo Fisher Scientific Inc., 2013) and SbfI-HF (20U/µL) (New England Biolabs, 2014). The fragments were ligated into the expression vector using T4 DNA ligase (New England Biolabs, 2014). Constructs were verified by Sanger sequencing (GATC Biotech. 2014). Other DBDs were cloned by previously reported procedures [3].

Protein Binding Microarrays (PBMs). PBM laboratory methods were performed as described previously [15, 46]. Each DBD-encoding plasmid was analyzed in duplicate on two different arrays with differing probe sequences. 8-mer Z- and E-scores were calculated as previously described [14]. We deemed experiments successful if at least one 8-mer had an E-score > 0.45 on both arrays, the complimentary arrays produced highly correlated E- and Z-scores, and the complimentary arrays yielded similar PWMs based on the PWM_align algorithm [15]. Motifs shown (and deposited in Cis-BP) for each TF are chosen by cross-replicate evaluation of three motif derivation methods (PWM_align, PWM_align_Z, and BEEML-PBM) [3, 47].

Data and software availability. New PBM data and motifs are deposited in GEO (accession number: GSE121420, reviewer password: mpcvcwwkjhgjvip), and Cis-BP (http://cisbp.ccbr.utoronto.ca/). The SR code, and examples, are made available on GitHub (https://github.com/smlmbrt/SimilarityRegression).

ACKNOWLEDGMENTS

We are grateful to Xiaoting Chen and Mario Pujato for computational support. SAL was funded by a Natural Sciences and Engineering Research Council of Canada Doctoral Fellowship. TRH holds the Billes Chair of Medical Research at the University of Toronto. This work was supported by Canadian Institutes of Health Research grant (FDN-148403) to TRH, and National Institutes of Health (NIH) R01 NS099068 and R01 GM055479, Lupus Research Alliance "Novel Approaches", CCRF Endowed Scholar, and CCHMC CpG Award 53553 to MTW. . <Ask Quaid, and Mark for funding and people>.

AUTHOR CONTRIBUTIONS

SAL, MTW, and TRH conceived of the study and oversaw it to completion. SAL analyzed the data, made the figures, and performed all computational analyses except, that AS re-implemented the Affinity Regression pipeline and applied it to new data. QDM guided the computational and statistical analyses. MA, SAL, and MTW maintained and updated the Cis-BP database. GC and MXC produced the clones for Aspergillus PBM experiments. AWY produced the remainder of the clones and performed all PBM experiments. SAL and TRH wrote the manuscript with feedback and approval from all authors.

FIGURE LEGENDS

Figure 1. Overview of the Similarity Regression (SR) method. SR uses TF protein similarity to predict the similarity in TF sequence specificities. The procedure and results are outlined in this figure using Homeodomain TFs as an example. (**A**) First, each TF's DBD sequence is aligned to the Pfam HMM as a common reference to generate a global alignment. Amino acids shown are coloured according to standard clustal colours for two homeodomains. (**B)** For each pair of TFs, amino acid similarity is measured at each position of the alignment, recording whether the two residues are identical or similar (BLOSSUM62 substitution score). This procedure is repeated for every pair of Homeodomain TFs with PBM data. (**C**) Regression is performed over a matrix in which each row is a pair of Homeodomain TFs, with the similarity of their DNA sequence specificities (E-score overlap) as the dependent Y variables (left), and protein similarity scores as independent X values (right). Sequence diversity among the TFs is represented here for reference, plotted as a logo above the protein similarity matrix. (**D**) The regression outputs a weight vector that indicates how much amino acid similarity in each position of the DBD contributes to DNA-binding similarity. Known specificity residues [48] are represented by an asterisk.

Figure 2. SR classification of TFs as having Highly Similar or Dissimilar sequence specificities. (A) Precision-Recall curves for Homeodomains are shown for three prediction methods: simple DBD %ID, SR using AA identity, and SR using BLOSUM similarity, on heldout data across all CV folds. *Positives* are pairs of TFs with Highly Similar specificities (E-score overlap > 25th percentile of replicate experiments), and *Negatives* are all other pairs. (**B)** Scatter plot comparing recall values (predicting Highly Similar specificities at 75% Precision threshold) for SR vs. simple %ID, for each TF family. The best of the four SR models is shown. Points are sized according to the number of PBM experiments used for training. **(C)** Smoothed density estimates for Homeodomain E-score overlaps in each predicted TF similarity class. Densities are filled according to the quartiles of the data. Vertical dashed lines indicate the E-score overlap thresholds used to define Dissimilar (blue line), and Highly Similar (black line) TF specificities in the initial data. **(D)** Percentage of actual TF similarities within each predicted TF similarity class, for new PBM data. White dotted lines show expected percentages the for Highly Similar and Dissimilar classes (i.e. thresholds were chosen to achieve these levels on training data).

Figure 3. New PBM data from the medicinal plant *Cannabis sativa*, and model fungi *Aspergillus nidulans* and *Neurospora crassa* for TFs with conserved and dissimilar motifs. Nearest neighbours for each new TF with PBM data were identified by finding the most similar TF (by SR score) with a motif from either *Arabidopsis thaliana* (for *C. sativa*), or *Saccharomyces cerevisiae* (for the fungi *A. nidulans*, and *N. crassa*). Motifs for (**A**) Myb/SANT

TFs from *C. sativa,* (**B**) C2H2 ZF TFs from *N. crassa*, and (**C**) TFs from five other TF families in *A. nidulans* are shown, with a neighbour-joining tree scaled by DBD amino acid identity in (A) and (B). The coloured bar represents predicted motif similarity. See **Figure S5** for a comparison between SR predicted similarity and NN TF similarity for all new PBM data.

Figure 4. Conservation of TF motifs within major eukaryotic kingdoms. The average percentage of TFs whose closest TF in the other species is Conserved (SR classifies as Highly Similar), Likely Conserved (SR Ambiguous), or Diverged (SR Dissimilar, and unshared DBDs) was calculated for each pair of species from the same kingdom (species and kingdoms are listed in **Figure S6B**). Each point represents the average percentage of TFs within each category, for each pair of species (i.e. average of species A vs. species B, and B vs. A), plotted against divergence time in millions of years. Divergence time is plotted on a square root scale to visualize differences between closely related species. Lines show a LOESS regression fit.

Figure 5. Motif divergence of TF families in metazoans and plants. (**A**) Nested pie charts showing the percentage of human TFs whose closest TF in other metazoans is Highly Similar, Ambiguous, Dissimilar, or Not Shared, for the 11 most abundant metazoan DBDs. The outer ring of each pie chart shows the proportion of human TFs in each SR-predicted similarity class relative to the other species; the inner ring shows the proportion of TFs for the other species, relative to human. (**B**) Motif similarity between *Arabidopsis thaliana* and other plants, for the 13 most abundant plant DBDs.

Figure 6. TF motif conservation between human and *Drosophila melanogaster*. A) Percentage of all TFs in human or Drosophila (as indicated) that fall into each SR motif similarity class. Stacked bar plots indicate TF family. (**B,C**) Experimentally determined motifs for individual *Drosophila* and human C2H2 zinc finger TFs, shown in pairs that correspond to the BLASTP best hit (Drosophila query to human database). Reciprocal best BLASTP matches (putative 1-to-1 orthologs) are indicated with bidirectional arrows. **(B)** shows pairs predicted to be Dissimilar by SR; (**C**) shows pairs predicted to be Highly Similar by SR.

Figure S1. Application of SR to TFs with an array of DBDs. (**A**) DBDs are first aligned to find the best ungapped and internal (maximizing amino acid identity) alignment. Examples of permissible alignment configurations are shown. (**B**) Alignments are then scored by calculating positional protein similarity features in each finger of a DBD array (e.g. C2H2 ZFs), and combined into a single representation by averaging the features by the length of the longest DBD array.

Figure S2. Additional SR model building and selection details. Four SR models are made for each TF family, and compared to unweighted alignment identity to identify the best SR model. The best model is selected after cross-validation, and threshold selection by Matthews Correlation Coefficient (MCC).

Figure S3. Comparison of SR weights to known DNA contacting residues. (A) Homeodomain, or **(B)** C2H2 ZF SR weights are compared to DNAproDB contact frequencies for DNA backbone, major and minor groove contacts, using partial correlations. TF amino acid sequence diversity (for the SR model training sequences) is displayed, for reference (above). **(C)** Partial correlations for all TF families with structural information in DNAproDB [33] are displayed and coloured according to the statistical significance, as $-\log_{10}$(p-value).

Figure S4. Comparison of predicted Z-score profiles for SR, AR, and DBD %ID. (**A**) Individual points show the Pearson Correlation Coefficient of predicted *vs*. actual Z-score profiles for 315 TFs (those among the 340 that have SR models), for the reconstruction methods tested. Reconstruction methods are grouped by whether they are a mixture of one (Nearest Neighbour), or multiple (Z-score reconstructions) TF profiles, as indicated by grey bars above. Points are coloured by TF family (see legend). (**B-D**) Individual results for the three most abundant TF families in the test set are plotted separately: (**B**) Zinc cluster, (**C**) Homeodomain, and (**D**) C2H2 ZFs.

Figure S5. Comparison of SR to DBD %ID at predicting TF pairs with Dissimilar specificities. (**A**) Scatter plot comparing the fraction of all dissimilar TF pairs captured by the 95% NPV threshold (Specificity). (**B**) Scatter plot showing Matthews correlation coefficient, which summarizes multi-class classification accuracy (for Highly Similar, Ambiguous, and Dissimilar TF sequence specificity) classification accuracy. In both panels, points are sized according to the number of PBM experiments used for training.

Figure S6. Comparison of SR scores with experimentally determined similarity in DNA sequence specificity, for new PBM data. Predicted TF similarity (SR score) and actual DNA-binding similarity (PBM E-score overlap) are plotted for each new PBM experiment, vs the most similar (by SR score) TF in the training set. Results are displayed for each TF family with more than three TFs. Linear fit is shown, with correspodingR^2 value. Points are coloured by their actual TF similarity based on family-specific E-score overlap thresholds.

Figure S7. Increase in percentage of TFs with a predicted motif in Cis-BP (SR vs. %ID). (**A**) The percentage of TFs with a "direct" (i.e. experimentally determined) (black bars), or predicted (grey bars) motif are plotted for the 50 largest TF families in Cis-BP. Increase in percentage due to SR models is shown by red bars. Total number of TFs encompassed is shown at right. (**B**) Motif coverage in well-studied eukaryotes, plotted as in panel **A**. Relationships between the species are represented by divergence time (million years ago) obtained from the TimeTree database [44]. The major clades of fungi, metazoans, and plants are coloured in red, blue, and green respectively**.**

Figure S8. Motif divergence of TF families in fungi. Classifications of motif similarity are shown as in **Figure 5**. The outer ring of each pie chart represents *Saccharomyces cerevisiae* TFs similarities with respect to the species it's being compared to (displayed along the phylogeny). The inner ring represents the compared species similarities with respect to *S. cerevisiae*. Branch length is the divergence time between species (millions of years).

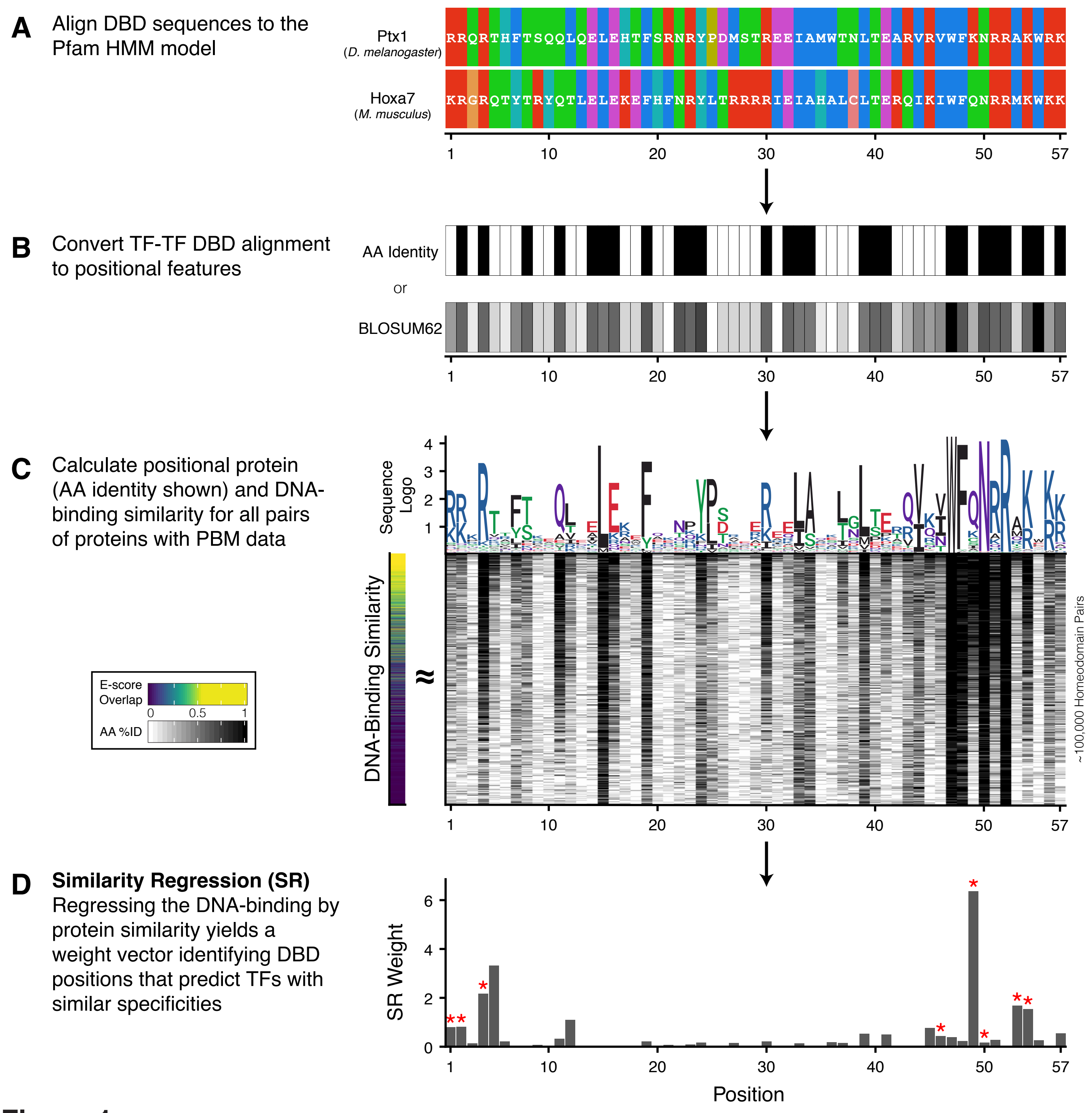
Figure S9. Motif similarity between corresponding *Drosophila* and human TFs (highest scoring BLASTP hits with Drosophila as query). Motif similarity was calculated between PWMs with experimentally determined motifs, using MoSBAT [49]. (**A**) The maximum motif similarity for all pairs of human and fly TFs (i.e. considering that there are often multiple motifs per TF) is displayed as a boxplot, according to the SR predicted TF similarity for each NN pair. (**B**) Similar plot as panel (A), but only HT-SELEX data is used in the analysis.

REFERENCES

1. Stormo, G.D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16-23 (2000).
2. Mathelier, A. et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **44**, D110-115 (2016).
3. Weirauch, M.T. et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443 (2014).
4. Pelossof, R. et al. Affinity regression predicts the recognition code of nucleic acid-binding proteins. *Nat Biotechnol* **33**, 1242-1249 (2015).
5. Christensen, R.G. et al. Recognition models to predict DNA-binding specificities of homeodomain proteins. *Bioinformatics* **28**, i84-89 (2012).
6. Persikov, A.V. et al. A systematic survey of the Cys2His2 zinc finger DNA-binding landscape. *Nucleic Acids Res* **43**, 1965-1984 (2015).
7. Najafabadi, H.S. et al. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat Biotechnol* (2015).
8. Nitta, K.R. et al. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife* **4** (2015).
9. Liu, H., Chang, L.H., Sun, Y., Lu, X. & Stubbs, L. Deep vertebrate roots for mammalian zinc finger transcription factor subfamilies. *Genome Biol Evol* **6**, 510-525 (2014).
10. Lynch, V.J. & Wagner, G.P. Resurrecting the role of transcription factor change in developmental evolution. *Evolution; international journal of organic evolution* **62**, 2131-2154 (2008).
11. Sayou, C. et al. A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity. *Science* **343**, 645-648 (2014).
12. McKeown, A.N. et al. Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module. *Cell* **159**, 58-68 (2014).
13. Najafabadi, H.S. et al. Non-base-contacting residues enable kaleidoscopic evolution of metazoan C2H2 zinc finger DNA binding. *Genome Biol* **18**, 167 (2017).
14. Berger, M.F. et al. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* **24**, 1429-1435 (2006).
15. Weirauch, M.T. et al. Evaluation of methods for modeling transcription factor sequence specificity. *Nat Biotechnol* **31**, 126-134 (2013).
16. Love, J.J. et al. Structural basis for DNA bending by the architectural transcription factor LEF-1. *Nature* **376**, 791-795 (1995).
17. Marmorstein, R., Carey, M., Ptashne, M. & Harrison, S.C. DNA recognition by GAL4: structure of a protein-DNA complex. *Nature* **356**, 408-414 (1992).
18. King, D.A., Zhang, L., Guarente, L. & Marmorstein, R. Structure of a HAP1-DNA complex reveals dramatically asymmetric DNA binding by a homodimeric protein. *Nat Struct Biol* **6**, 64-71 (1999).

19. de Mendoza, A. et al. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proc Natl Acad Sci U S A* **110**, E4858-4866 (2013).
20. Grove, C.A. et al. A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors. *Cell* **138**, 314-327 (2009).
21. Reinke, A.W., Baek, J., Ashenberg, O. & Keating, A.E. Networks of bZIP protein-protein interactions diversified over a billion years of evolution. *Science* **340**, 730-734 (2013).
22. Jolma, A. et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res* **20**, 861-873 (2010).
23. Noyes, M.B. et al. A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. *Nucleic Acids Res* **36**, 2547-2560 (2008).
24. Zhu, L.J. et al. FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res* **39**, D111-117 (2011).
25. MacPherson, S., Larochelle, M. & Turcotte, B. A fungal family of transcriptional regulators: the zinc cluster proteins. *Microbiol Mol Biol Rev* **70**, 583-604 (2006).
26. Lambert, S.A. et al. The Human Transcription Factors. *Cell* **175**, 598-599 (2018).
27. Nadimpalli, S., Persikov, A.V. & Singh, M. Pervasive variation of transcription factor orthologs contributes to regulatory network evolution. *PLoS Genet* **11**, e1005011 (2015).
28. Wilkinson, S. The 'aphid' package for analysis with profile hidden Markov models. R package version 1.1.0. (2018).
29. Henikoff, S. & Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915-10919 (1992).
30. Team, R.C. R: A language and environment for statistical computing. (2016).
31. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software* **33**, 1-22 (2010).
32. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of statistical software* **28** (2008).
33. Sagendorf, J.M., Berman, H.M. & Rohs, R. DNAproDB: an interactive tool for structural analysis of DNA-protein complexes. *Nucleic Acids Res* **45**, W89-W97 (2017).
34. Berman, H.M. et al. The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242 (2000).
35. Finn, R.D. et al. The Pfam protein families database. *Nucleic Acids Res* **38**, D211-222 (2010).
36. Yin, Y. et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356** (2017).
37. O'Malley, R.C. et al. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* **165**, 1280-1292 (2016).

38. Barazandeh, M., Lambert, S.A., Albu, M. & Hughes, T.R. Comparison of ChIP-Seq Data and a Reference Motif Set for Human KRAB C2H2 Zinc Finger Proteins. *G3* **8**, 219-229 (2018).
39. Hume, M.A., Barrera, L.A., Gisselbrecht, S.S. & Bulyk, M.L. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res* **43**, D117-122 (2015).
40. Matys, V. et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34**, D108-110 (2006).
41. Khan, A. et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res* **46**, D1284 (2018).
42. Kulakovskiy, I.V. et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res* **46**, D252-D259 (2018).
43. Sigrist, C.J. et al. PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings in bioinformatics* **3**, 265-274 (2002).
44. Kumar, S., Stecher, G., Suleski, M. & Hedges, S.B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol* **34**, 1812-1819 (2017).
45. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).
46. Lam, K.N., van Bakel, H., Cote, A.G., van der Ven, A. & Hughes, T.R. Sequence specificity is obtained from the majority of modular C2H2 zinc-finger arrays. *Nucleic Acids Res* **39**, 4680-4690 (2011).
47. Zhao, Y. & Stormo, G.D. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat Biotechnol* **29**, 480-483 (2011).
48. Noyes, M.B. et al. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* **133**, 1277-1289 (2008).
49. Lambert, S.A., Albu, M., Hughes, T.R. & Najafabadi, H.S. Motif comparison based on similarity of binding affinity profiles. *Bioinformatics* **32**, 3504-3506 (2016).



A Align DBD sequences to the Pfam HMM model

B Convert TF-TF DBD alignment to positional features

C Calculate positional protein (AA identity shown) and DNA-binding similarity for all pairs of proteins with PBM data

D **Similarity Regression (SR)** Regressing the DNA-binding by protein similarity yields a weight vector identifying DBD positions that predict TFs with similar specificities

Figure 1.

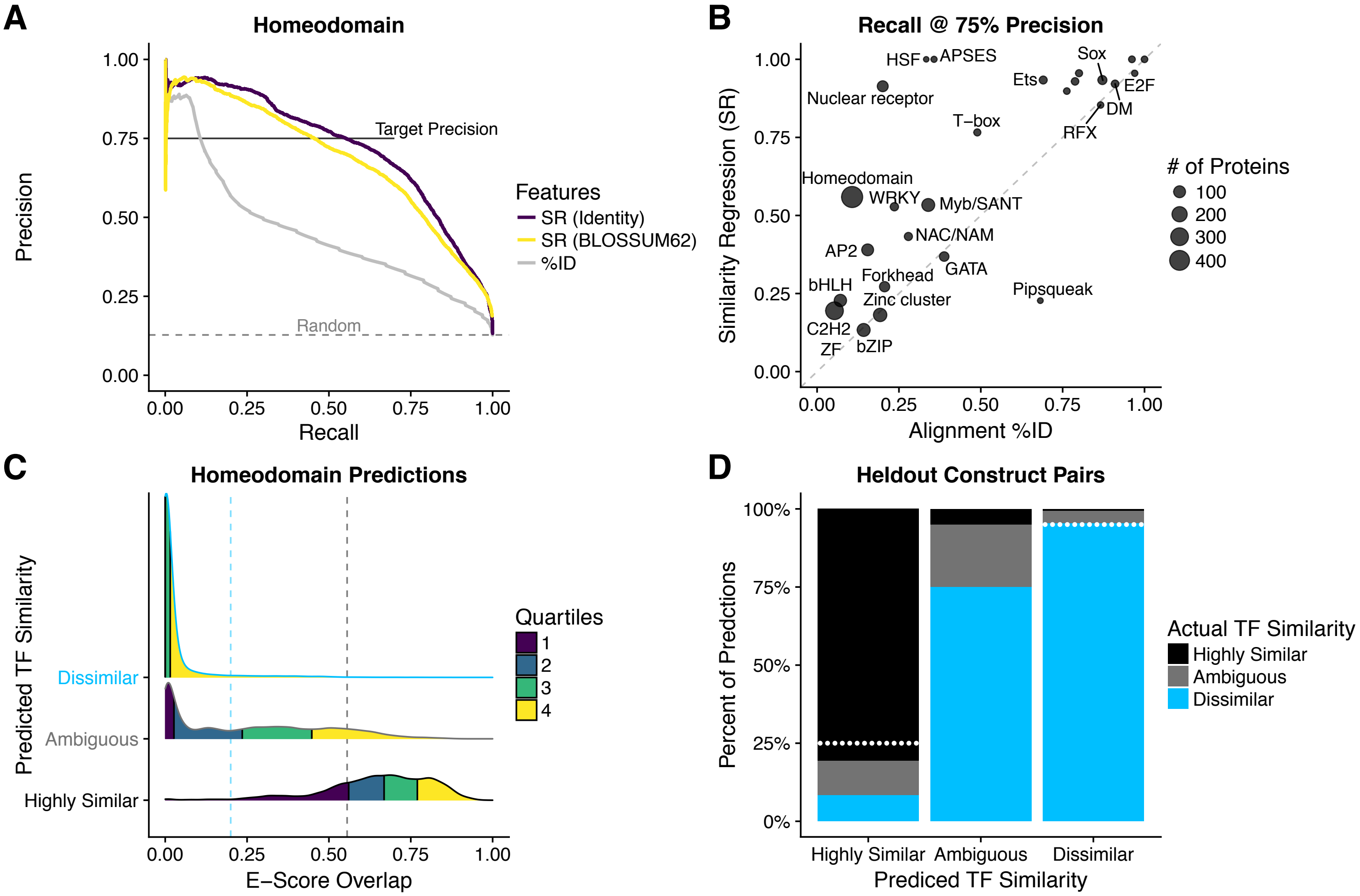

A
Homeodomain
Precision
Recall
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Target Precision
Random
Features
SR (Identity)
SR (BLOSSUM62)
%ID
B
Recall @ 75% Precision
Similarity Regression (SR)
Alignment %ID
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
HSF
APSES
Sox
Ets
E2F
DM
RFX
Nuclear receptor
T−box
Homeodomain
WRKY
Myb/SANT
NAC/NAM
AP2
GATA
Forkhead
bHLH
Zinc cluster
Pipsqueak
C2H2
ZF
bZIP
# of Proteins
100
200
300
400
C
Homeodomain Predictions
Predicted TF Similarity
Dissimilar
Ambiguous
Highly Similar
E−Score Overlap
0.00
0.25
0.50
0.75
1.00
Quartiles
1
2
3
4
D
Heldout Construct Pairs
Percent of Predctions
100%
75%
50%
25%
0%
Highly Similar
Ambiguous
Dissimilar
Prediced TF Similarity
Actual TF Similarity
Highly Similar
Ambiguous
Dissimilar

Figure 2.

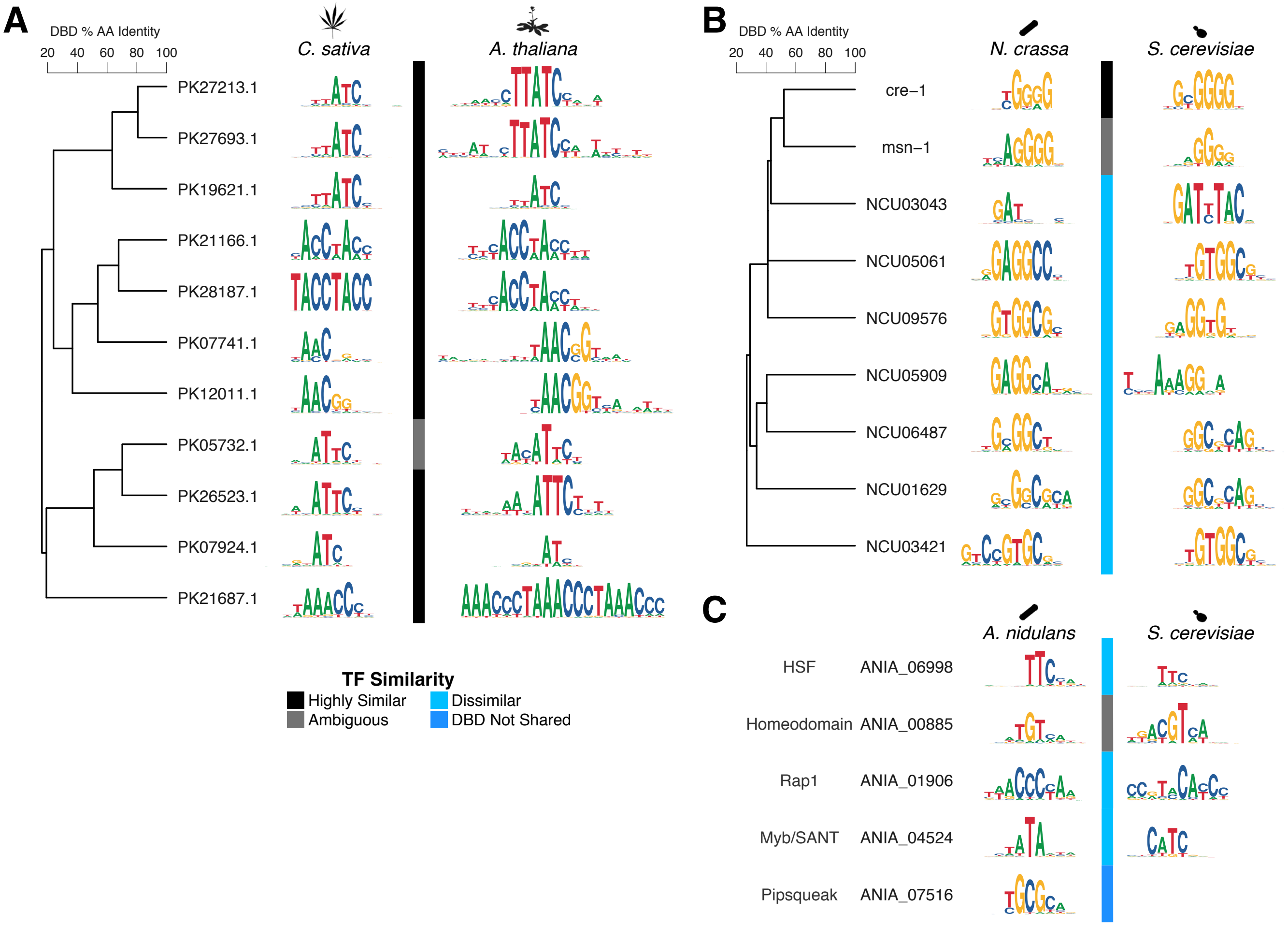

A
DBD % AA Identity
20 40 60 80 100
C. sativa
A. thaliana
PK27213.1
PK27693.1
PK19621.1
PK21166.1
PK28187.1
PK07741.1
PK12011.1
PK05732.1
PK26523.1
PK07924.1
PK21687.1
TF Similarity
Highly Similar
Ambiguous
Dissimilar
DBD Not Shared
B
DBD % AA Identity
20 40 60 80 100
N. crassa
S. cerevisiae
cre−1
msn−1
NCU03043
NCU05061
NCU09576
NCU05909
NCU06487
NCU01629
NCU03421
C
A. nidulans
S. cerevisiae
HSF ANIA_06998
Homeodomain ANIA_00885
Rap1 ANIA_01906
Myb/SANT ANIA_04524
Pipsqueak ANIA_07516

Figure 3.

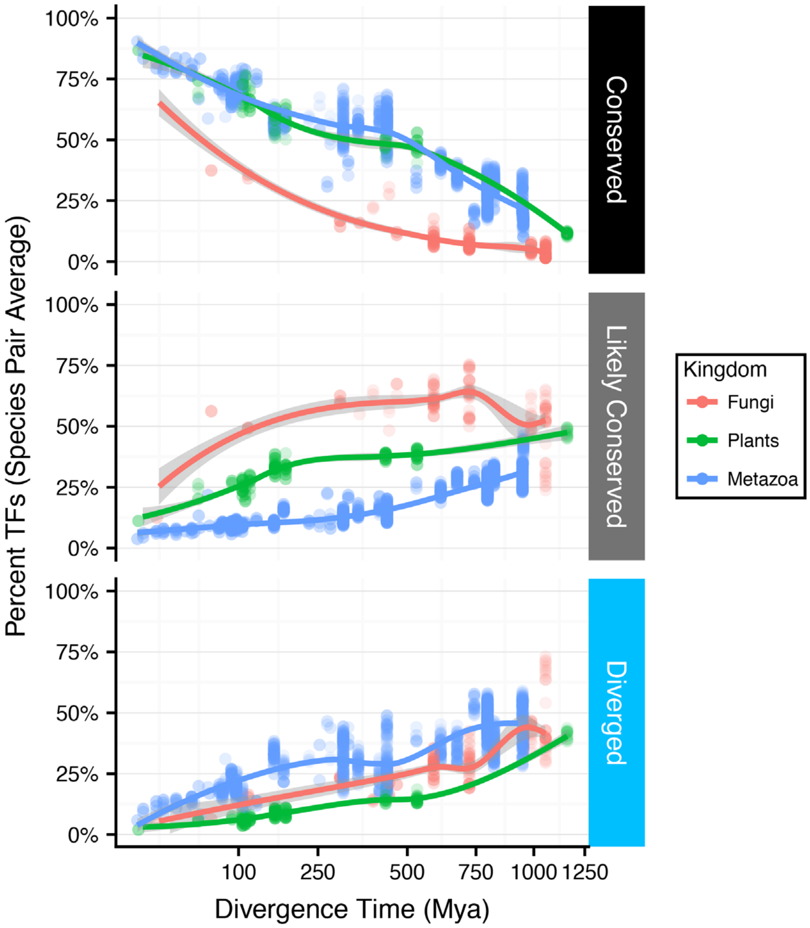

Percent TFs (Species Pair Average)
100%
75%
50%
25%
0%
Conserved
Likely Conserved
Diverged
Kingdom
Fungi
Plants
Metazoa
100
250
500
750
1000
1250
Divergence Time (Mya)

Figure 4.

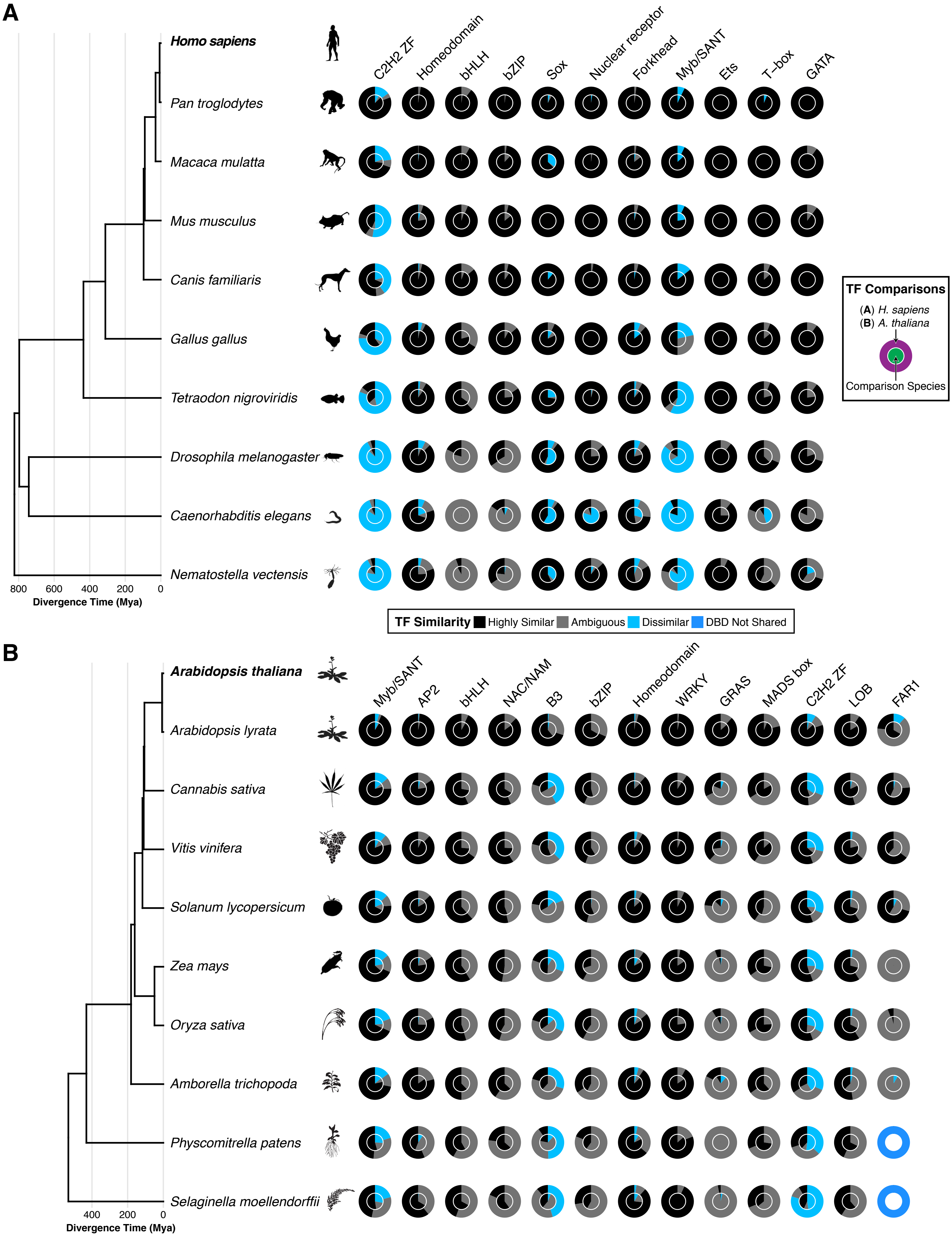

A
Homo sapiens
Pan troglodytes
Macaca mulatta
Mus musculus
Canis familiaris
Gallus gallus
Tetraodon nigroviridis
Drosophila melanogaster
Caenorhabditis elegans
Nematostella vectensis
800
600
400
200
0
Divergence Time (Mya)
C2H2 ZF
Homeodomain
bHLH
bZIP
Sox
Nuclear receptor
Forkhead
Myb/SANT
Ets
T-box
GATA
TF Comparisons
(A) H. sapiens
(B) A. thaliana
Comparison Species
TF Similarity
■ Highly Similar
■ Ambiguous
■ Dissimilar
■ DBD Not Shared
B
Arabidopsis thaliana
Arabidopsis lyrata
Cannabis sativa
Vitis vinifera
Solanum lycopersicum
Zea mays
Oryza sativa
Amborella trichopoda
Physcomitrella patens
Selaginella moellendorffii
400
200
0
Divergence Time (Mya)
Myb/SANT
AP2
bHLH
NAC/NAM
B3
bZIP
Homeodomain
WRKY
GRAS
MADS box
C2H2 ZF
LOB
FAR1

Figure 5.

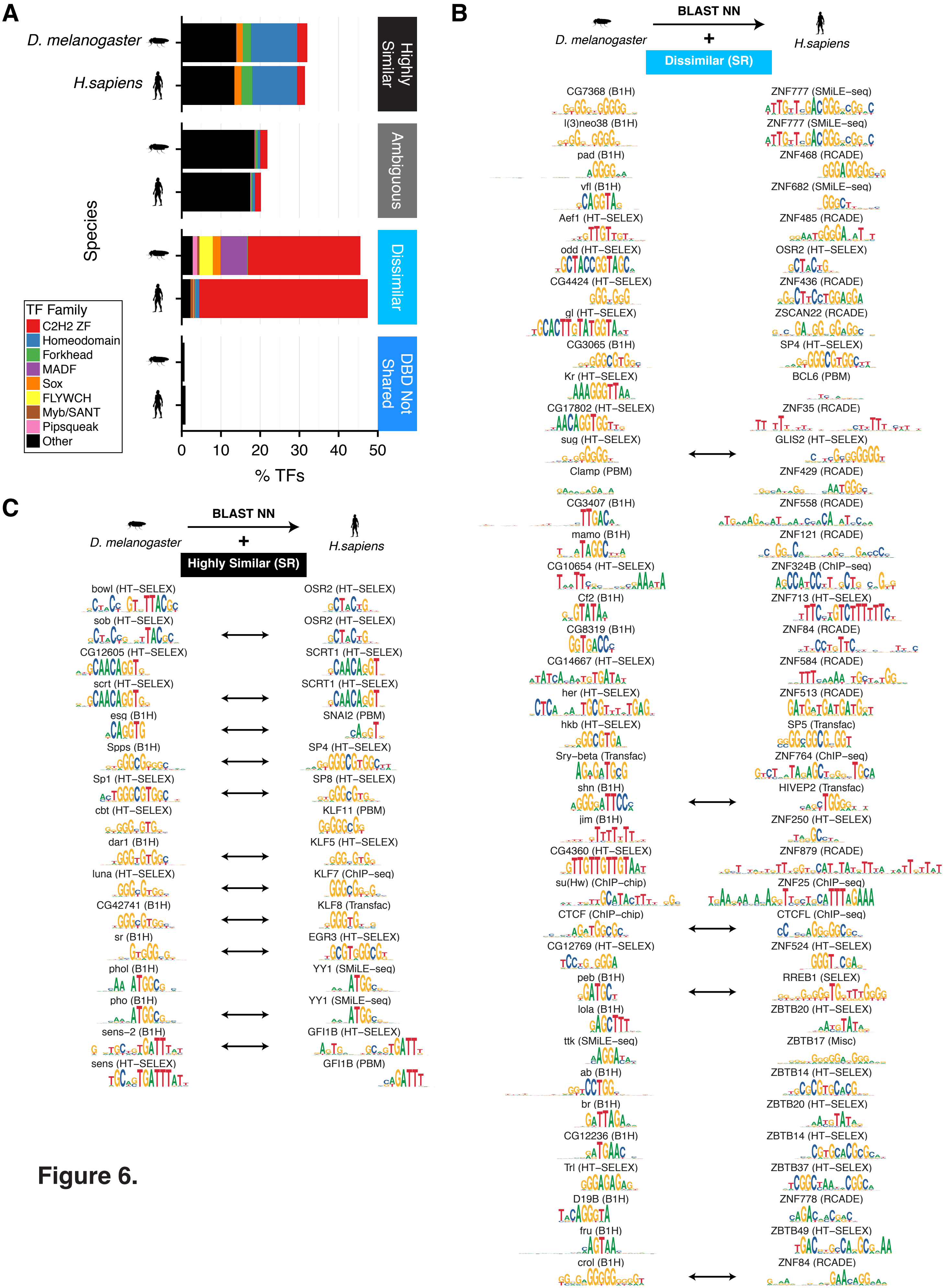

A
D. melanogaster
H.sapiens
Species
Highly Similar
Ambiguous
Dissimilar
DBD Not Shared
TF Family
🟥 C2H2 ZF
🟦 Homeodomain
🟩 Forkhead
🟪 MADF
🟧 Sox
🟨 FLYWCH
🟫 Myb/SANT
Pipsqueak
⬛ Other
0
10
20
30
40
50
% TFs
B
D. melanogaster
BLAST NN
+
Dissimilar (SR)
H.sapiens
CG7368 (B1H)
ZNF777 (SMiLE−seq)
l(3)neo38 (B1H)
ZNF777 (SMiLE−seq)
pad (B1H)
ZNF468 (RCADE)
vfl (B1H)
ZNF682 (SMiLE−seq)
Aef1 (HT−SELEX)
ZNF485 (RCADE)
odd (HT−SELEX)
OSR2 (HT−SELEX)
CG4424 (HT−SELEX)
ZNF436 (RCADE)
gl (HT−SELEX)
ZSCAN22 (RCADE)
CG3065 (B1H)
SP4 (HT−SELEX)
Kr (HT−SELEX)
BCL6 (PBM)
CG17802 (HT−SELEX)
ZNF35 (RCADE)
sug (HT−SELEX)
GLIS2 (HT−SELEX)
Clamp (PBM)
ZNF429 (RCADE)
CG3407 (B1H)
ZNF558 (RCADE)
mamo (B1H)
ZNF121 (RCADE)
CG10654 (HT−SELEX)
ZNF324B (ChIP−seq)
Cf2 (B1H)
ZNF713 (HT−SELEX)
CG8319 (B1H)
ZNF84 (RCADE)
CG14667 (HT−SELEX)
ZNF584 (RCADE)
her (HT−SELEX)
ZNF513 (RCADE)
hkb (HT−SELEX)
SP5 (Transfac)
Sry−beta (Transfac)
ZNF764 (ChIP−seq)
shn (B1H)
HIVEP2 (Transfac)
jim (B1H)
ZNF250 (HT−SELEX)
CG4360 (HT−SELEX)
ZNF879 (RCADE)
su(Hw) (ChIP−chip)
ZNF25 (ChIP−seq)
CTCF (ChIP−chip)
CTCFL (ChIP−seq)
CG12769 (HT−SELEX)
ZNF524 (HT−SELEX)
peb (B1H)
RREB1 (SELEX)
lola (B1H)
ZBTB20 (HT−SELEX)
ttk (SMiLE−seq)
ZBTB17 (Misc)
ab (B1H)
ZBTB14 (HT−SELEX)
br (B1H)
ZBTB20 (HT−SELEX)
CG12236 (B1H)
ZBTB14 (HT−SELEX)
Trl (HT−SELEX)
ZBTB37 (HT−SELEX)
D19B (B1H)
ZNF778 (RCADE)
fru (B1H)
ZBTB49 (HT−SELEX)
crol (B1H)
ZNF84 (RCADE)
C
D. melanogaster
BLAST NN
+
Highly Similar (SR)
H.sapiens
bowl (HT−SELEX)
OSR2 (HT−SELEX)
sob (HT−SELEX)
OSR2 (HT−SELEX)
CG12605 (HT−SELEX)
SCRT1 (HT−SELEX)
scrt (HT−SELEX)
SCRT1 (HT−SELEX)
esg (B1H)
SNAI2 (PBM)
Spps (B1H)
SP4 (HT−SELEX)
Sp1 (HT−SELEX)
SP8 (HT−SELEX)
cbt (HT−SELEX)
KLF11 (PBM)
dar1 (B1H)
KLF5 (HT−SELEX)
luna (HT−SELEX)
KLF7 (ChIP−seq)
CG42741 (B1H)
KLF8 (Transfac)
sr (B1H)
EGR3 (HT−SELEX)
phol (B1H)
YY1 (SMiLE−seq)
pho (B1H)
YY1 (SMiLE−seq)
sens−2 (B1H)
GFI1B (HT−SELEX)
sens (HT−SELEX)
GFI1B (PBM)

Figure 6.

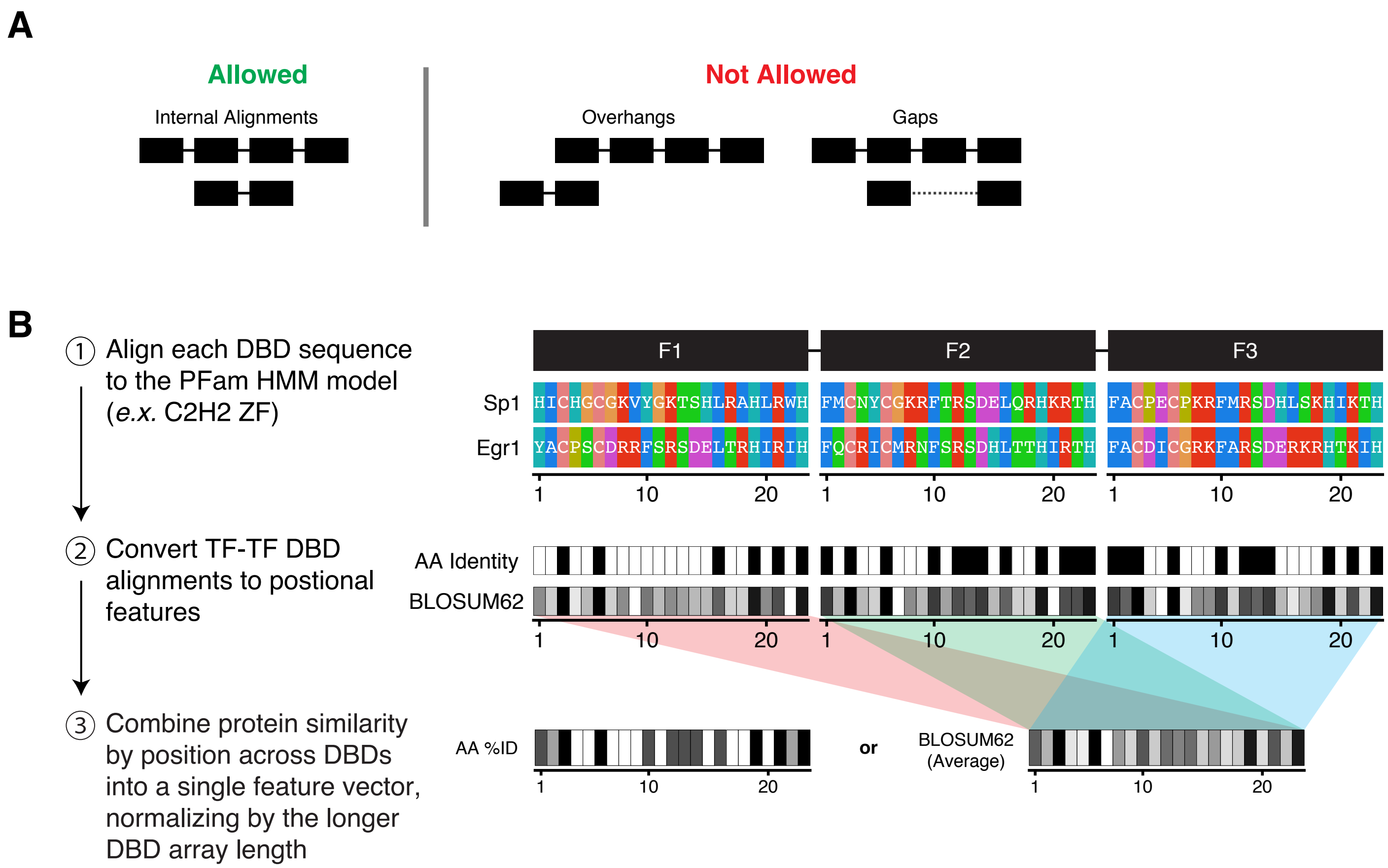
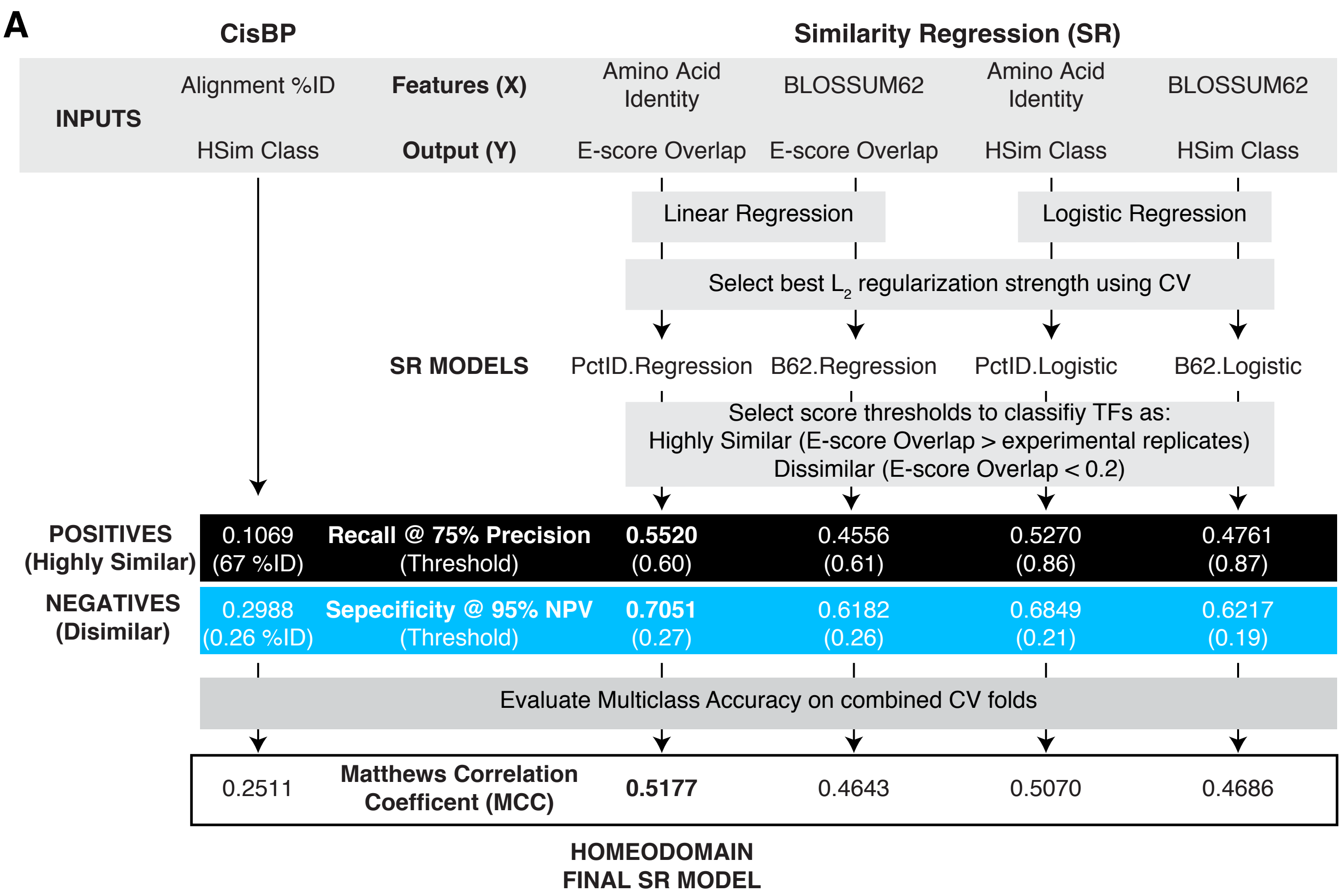

A
Allowed
Internal Alignments
Not Allowed
Overhangs
Gaps
B
① Align each DBD sequence to the PFam HMM model (*e.x.* C2H2 ZF)
② Convert TF-TF DBD alignments to postional features
③ Combine protein similarity by position across DBDs into a single feature vector, normalizing by the longer DBD array length
F1
F2
F3
Sp1
HICHGCGKVYGKTSHLRAHLRWH
FMCNYCGKRFTRSDELQRHKRTH
FACPECPKRFMRSDHLSKHIKTH
Egr1
YACPSCDRRFSRSDELTRHIRIH
FQCRICMRNFSRSDHLTTHIRTH
FACDICGRKFARSDERKRHTKIH
1
10
20
AA Identity
BLOSUM62
AA %ID
or
BLOSUM62
(Average)

Figure S1.



	CisBP		Similarity Regression (SR)			
INPUTS	Alignment %ID	**Features (X)**	Amino Acid Identity	BLOSSUM62	Amino Acid Identity	BLOSSUM62
	HSim Class	**Output (Y)**	E-score Overlap	E-score Overlap	HSim Class	HSim Class

Linear Regression

Logistic Regression

Select best L_2 regularization strength using CV

		SR MODELS	PctID.Regression	B62.Regression	PctID.Logistic	B62.Logistic

Select score thresholds to classifiy TFs as:
Highly Similar (E-score Overlap > experimental replicates)
Dissimilar (E-score Overlap < 0.2)

POSITIVES (Highly Similar)	0.1069 (67 %ID)	**Recall @ 75% Precision** (Threshold)	**0.5520** (0.60)	0.4556 (0.61)	0.5270 (0.86)	0.4761 (0.87)
NEGATIVES (Disimilar)	0.2988 (0.26 %ID)	**Sepecificity @ 95% NPV** (Threshold)	**0.7051** (0.27)	0.6182 (0.26)	0.6849 (0.21)	0.6217 (0.19)

Evaluate Multiclass Accuracy on combined CV folds

0.2511	**Matthews Correlation Coefficent (MCC)**	**0.5177**	0.4643	0.5070	0.4686

HOMEODOMAIN FINAL SR MODEL

Figure S2.

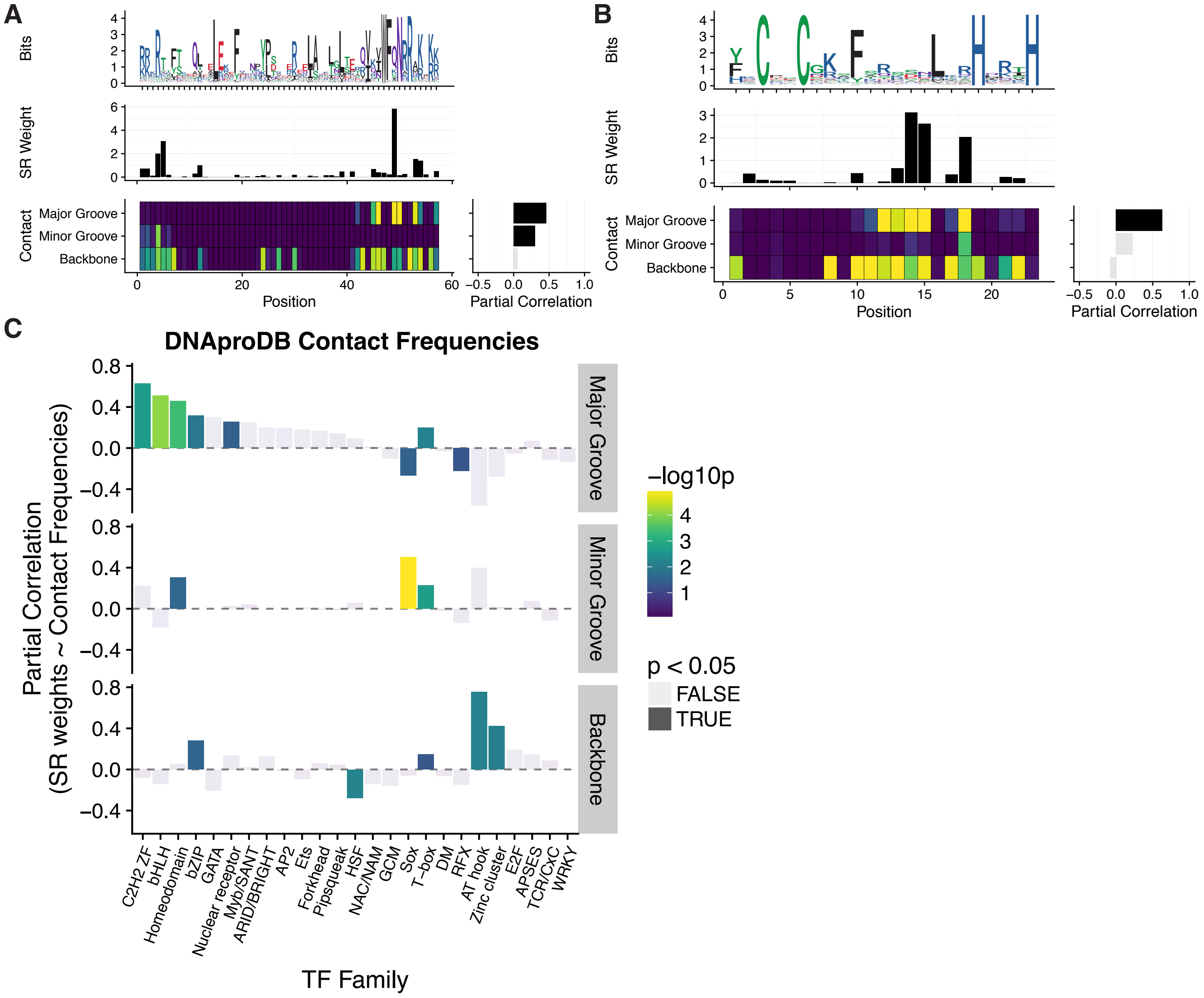

A
Bits
SR Weight
Contact
Major Groove
Minor Groove
Backbone
Position
Partial Correlation
B
Bits
SR Weight
Contact
Major Groove
Minor Groove
Backbone
Position
Partial Correlation
C
DNAproDB Contact Frequencies
Partial Correlation
(SR weights ~ Contact Frequencies)
Major Groove
Minor Groove
Backbone
−log10p
p < 0.05
FALSE
TRUE
C2H2 ZF
bHLH
Homeodomain
bZIP
GATA
Nuclear receptor
Myb/SANT
ARID/BRIGHT
AP2
Ets
Forkhead
Pipsqueak
HSF
NAC/NAM
GCM
Sox
T−box
DM
RFX
AT hook
Zinc cluster
E2F
APSES
TCR/CxC
WRKY
TF Family

Figure S3.

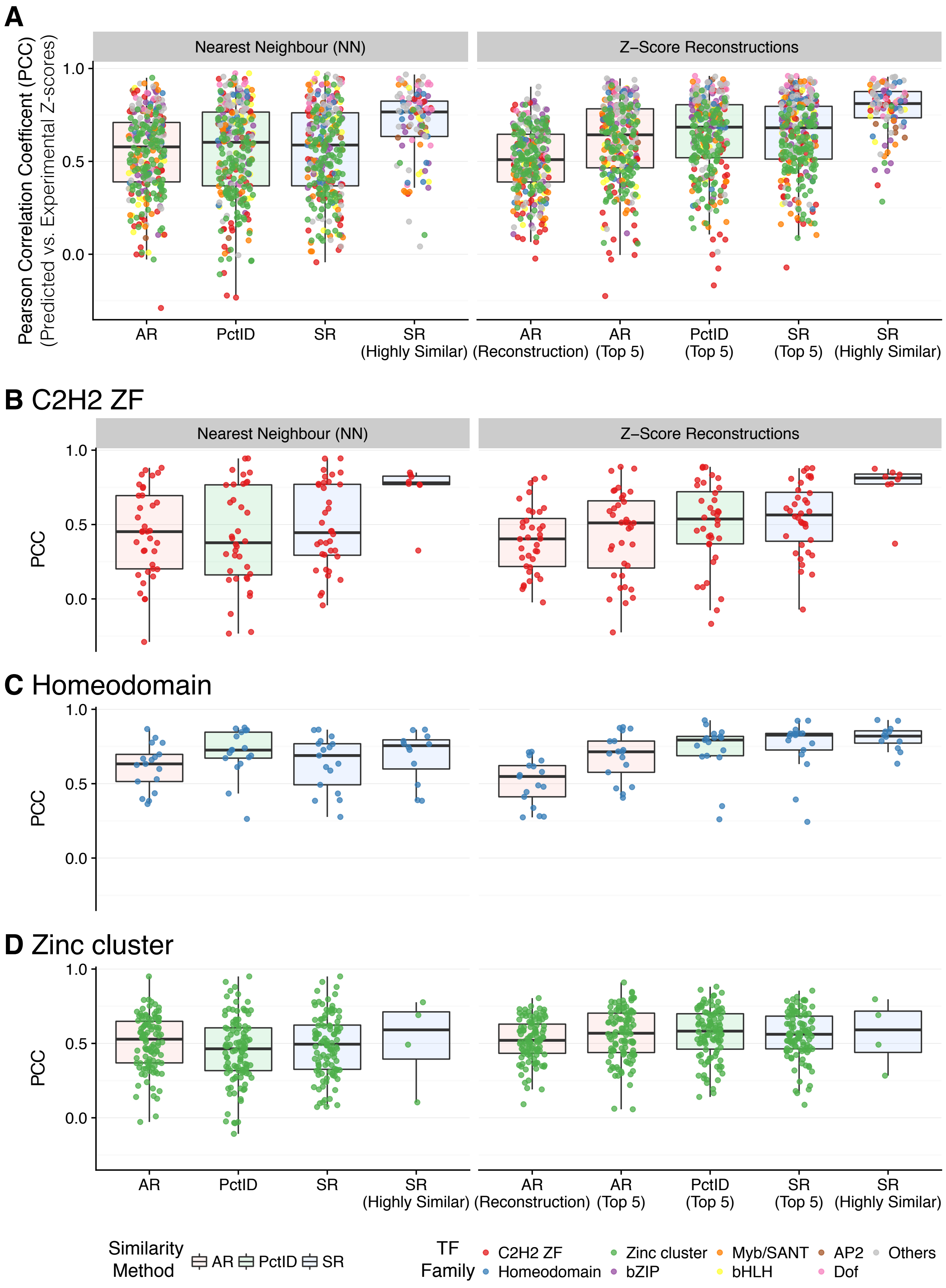

A
Nearest Neighbour (NN)
Z–Score Reconstructions
Pearson Correlation Coefficent (PCC)
(Predicted vs. Experimental Z-scores)
1.0
0.5
0.0
AR
PctID
SR
SR (Highly Similar)
AR (Reconstruction)
AR (Top 5)
PctID (Top 5)
SR (Top 5)
SR (Highly Similar)
B C2H2 ZF
PCC
C Homeodomain
D Zinc cluster
Similarity Method
AR
PctID
SR
TF Family
C2H2 ZF
Homeodomain
Zinc cluster
bZIP
Myb/SANT
bHLH
AP2
Dof
Others

Figure S4.

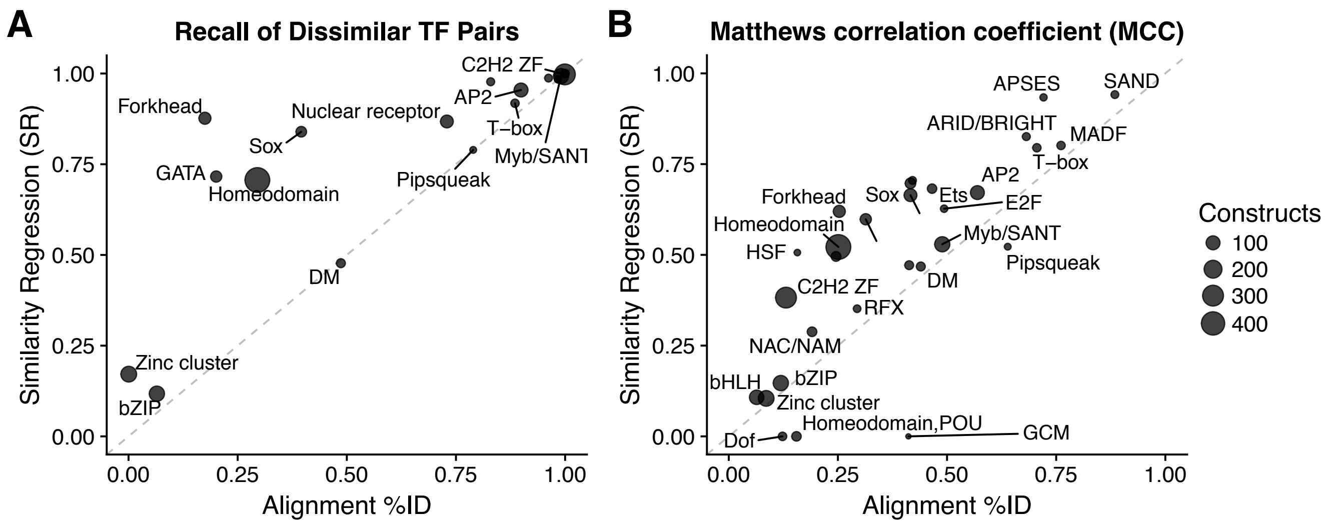


Figure S5.

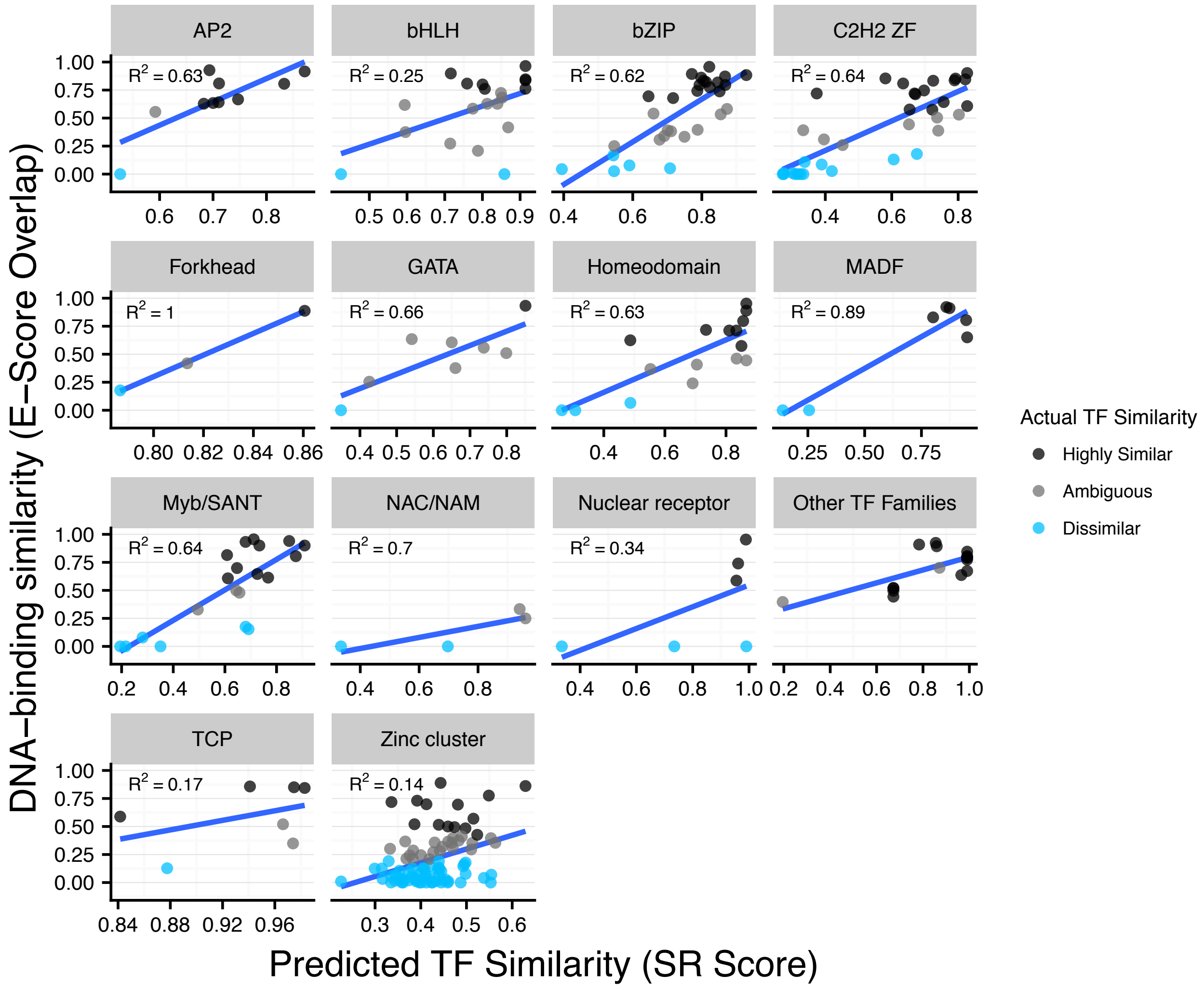


Figure S6.

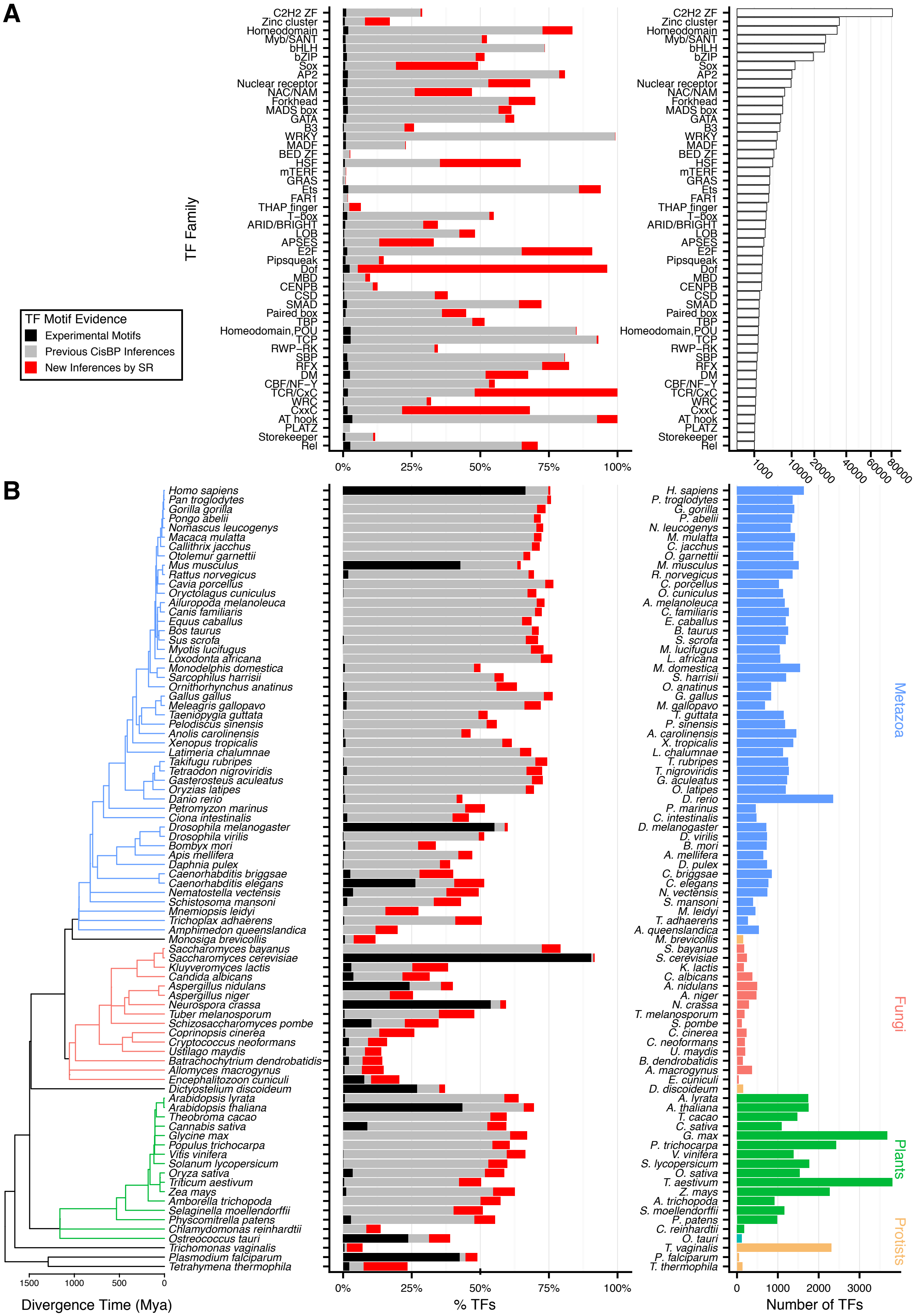

A
TF Family
C2H2 ZF
Zinc cluster
Homeodomain
Myb/SANT
bHLH
bZIP
Sox
AP2
Nuclear receptor
NAC/NAM
Forkhead
MADS box
GATA
B3
WRKY
MADF
BED ZF
HSF
mTERF
GRAS
Ets
FAR1
THAP finger
T-box
ARID/BRIGHT
LOB
APSES
E2F
Pipsqueak
Dof
MBD
CENPB
CSD
SMAD
Paired box
TBP
Homeodomain,POU
TCP
RWP-RK
SBP
RFX
DM
CBF/NF-Y
TCR/CxC
WRC
CxxC
AT hook
PLATZ
Storekeeper
Rel
0%
25%
50%
75%
100%
1000
10000
20000
40000
60000
80000
TF Motif Evidence
Experimental Motifs
Previous CisBP Inferences
New Inferences by SR
B
Homo sapiens
Pan troglodytes
Gorilla gorilla
Pongo abelii
Nomascus leucogenys
Macaca mulatta
Callithrix jacchus
Otolemur garnettii
Mus musculus
Rattus norvegicus
Cavia porcellus
Oryctolagus cuniculus
Ailuropoda melanoleuca
Canis familiaris
Equus caballus
Bos taurus
Sus scrofa
Myotis lucifugus
Loxodonta africana
Monodelphis domestica
Sarcophilus harrisii
Ornithorhynchus anatinus
Gallus gallus
Meleagris gallopavo
Taeniopygia guttata
Pelodiscus sinensis
Anolis carolinensis
Xenopus tropicalis
Latimeria chalumnae
Takifugu rubripes
Tetraodon nigroviridis
Gasterosteus aculeatus
Oryzias latipes
Danio rerio
Petromyzon marinus
Ciona intestinalis
Drosophila melanogaster
Drosophila virilis
Bombyx mori
Apis mellifera
Daphnia pulex
Caenorhabditis briggsae
Caenorhabditis elegans
Nematostella vectensis
Schistosoma mansoni
Mnemiopsis leidyi
Trichoplax adhaerens
Amphimedon queenslandica
Monosiga brevicollis
Saccharomyces bayanus
Saccharomyces cerevisiae
Kluyveromyces lactis
Candida albicans
Aspergillus nidulans
Aspergillus niger
Neurospora crassa
Tuber melanosporum
Schizosaccharomyces pombe
Coprinopsis cinerea
Cryptococcus neoformans
Ustilago maydis
Batrachochytrium dendrobatidis
Allomyces macrogynus
Encephalitozoon cuniculi
Dictyostelium discoideum
Arabidopsis lyrata
Arabidopsis thaliana
Theobroma cacao
Cannabis sativa
Glycine max
Populus trichocarpa
Vitis vinifera
Solanum lycopersicum
Oryza sativa
Triticum aestivum
Zea mays
Amborella trichopoda
Selaginella moellendorffii
Physcomitrella patens
Chlamydomonas reinhardtii
Ostreococcus tauri
Trichomonas vaginalis
Plasmodium falciparum
Tetrahymena thermophila
1500
1000
500
0
Divergence Time (Mya)
0%
25%
50%
75%
100%
% TFs
H. sapiens
P. troglodytes
G. gorilla
P. abelii
N. leucogenys
M. mulatta
C. jacchus
O. garnettii
M. musculus
R. norvegicus
C. porcellus
O. cuniculus
A. melanoleuca
C. familiaris
E. caballus
B. taurus
S. scrofa
M. lucifugus
L. africana
M. domestica
S. harrisii
O. anatinus
G. gallus
M. gallopavo
T. guttata
P. sinensis
A. carolinensis
X. tropicalis
L. chalumnae
T. rubripes
T. nigroviridis
G. aculeatus
O. latipes
D. rerio
P. marinus
C. intestinalis
D. melanogaster
D. virilis
B. mori
A. mellifera
D. pulex
C. briggsae
C. elegans
N. vectensis
S. mansoni
M. leidyi
T. adhaerens
A. queenslandica
M. brevicollis
S. bayanus
S. cerevisiae
K. lactis
C. albicans
A. nidulans
A. niger
N. crassa
T. melanosporum
S. pombe
C. cinerea
C. neoformans
U. maydis
B. dendrobatidis
A. macrogynus
E. cuniculi
D. discoideum
A. lyrata
A. thaliana
T. cacao
C. sativa
G. max
P. trichocarpa
V. vinifera
S. lycopersicum
O. sativa
T. aestivum
Z. mays
A. trichopoda
S. moellendorffii
P. patens
C. reinhardtii
O. tauri
T. vaginalis
P. falciparum
T. thermophila
Metazoa
Fungi
Plants
Protists
0
1000
2000
3000
Number of TFs

Figure S7.

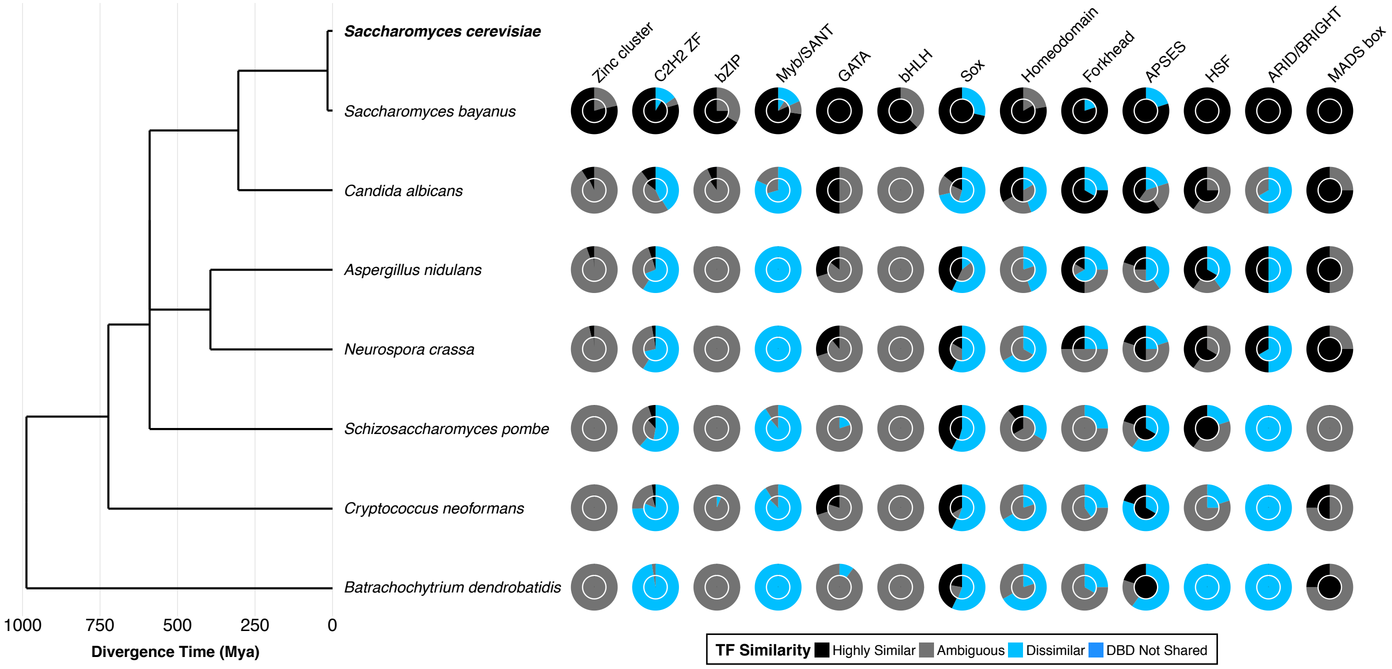


Figure S8.

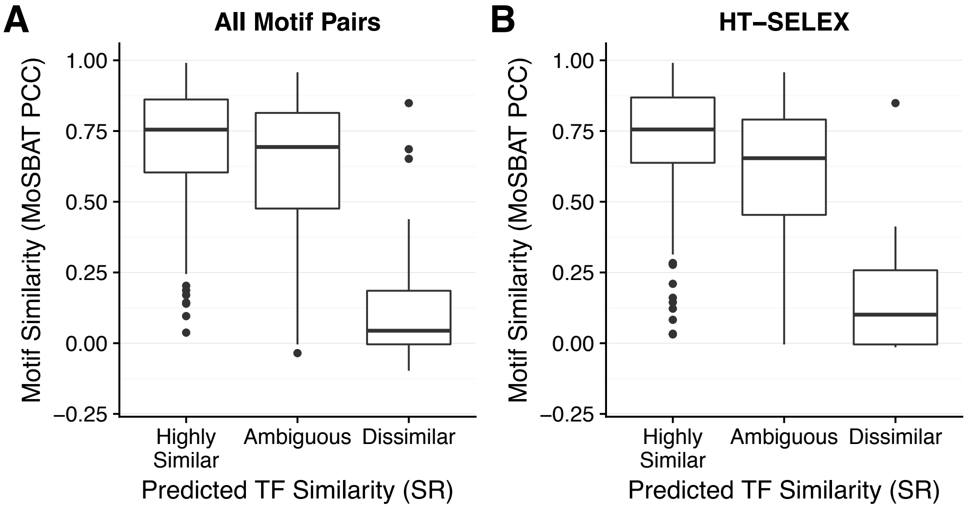


Figure S9.