nature genetics

ARTICLES

https://doi.org/10.1038/s41588-019-0411-1

Similarity regression predicts evolution of transcription factor sequence specificity

Samuel A. Lambert[1], Ally W. H. Yang[2], Alexander Sasse[1], Gwendolyn Cowley[3], Mihai Albu[2], Mark X. Caddick[3], Quaid D. Morris[1,2,4,5,6], Matthew T. Weirauch[7,8] and Timothy R. Hughes[1,2,9*]

Transcription factor (TF) binding specificities (motifs) are essential for the analysis of gene regulation. Accurate prediction of TF motifs is critical, because it is infeasible to assay all TFs in all sequenced eukaryotic genomes. There is ongoing controversy regarding the degree of motif diversification among related species that is, in part, because of uncertainty in motif prediction methods. Here we describe similarity regression, a significantly improved method for predicting motifs, which we use to update and expand the Cis-BP database. Similarity regression inherently quantifies TF motif evolution, and shows that previous claims of near-complete conservation of motifs between human and *Drosophila* are inflated, with nearly half of the motifs in each species absent from the other, largely due to extensive divergence in C2H2 zinc finger proteins. 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.

To understand the function of noncoding DNA—for example, in gene regulation—it is essential to know the potential TFs that can bind to any sequence. Libraries of experimentally derived TF motifs, most typically position weight matrices (PWMs)[1], are widely used and encompass at most a few thousand motifs and are oriented mainly towards well-studied TFs in human and model systems (for example, JASPAR)[2]. However, hundreds of eukaryotic genomes have now been sequenced, and the analysis of gene expression and corresponding sequences in regulatory regions can be performed in almost all of them. To enable such analyses, we previously described Cis-BP, a database of measured and 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 cut-offs for each DBD type established on the basis of replicate experiments and pairwise comparisons of motifs from different proteins with homologous DBD types. The cut-off prediction method yielded an 89% precision (with undetermined recall) on these data.

It may be possible to improve both precision and recall by adapting predictions to specific DBDs. One approach is to use known 'specificity residues' or to prioritize DNA-contacting residues when measuring amino acid sequence similarity. Other approaches including affinity regression[4], which predicts affinity to DNA/RNA *k*-mers on the basis of amino acid *k*-mer composition of proteins. Affinity regression was applied to only two families, however: homeodomain TFs and RNA recognition motif (RRM)-containing RNA-binding proteins. DBD-specific 'recognition codes' have also been described, which predict binding motifs on the basis of DNA-contacting residues, for Cys_2His_2 (C2H2) zinc finger (ZF) and homeodomain proteins[5–7]. It is unclear whether and how these methods will extend to the approximately 100 other types of DBDs.

Improving motif predictions would help to answer questions regarding the degree of diversity and evolution of eukaryotic TF motifs. TF-binding specificities are thought to be highly conserved between *Drosophila* and mammals[8]; however, the specificity residues for C2H2 ZF proteins are often very different even among *Drosophila* species (and very different from those in mammals), suggesting a much faster pace of change for some TF families[9,10]. TF motif diversification has also been observed in other lineages, including plants and fungi, indicating that TF evolution occurs broadly, in parallel to the more established *cis*-regulatory turnover[11]. Measuring motif diversification cannot be addressed by simply examining the orthology patterns of whole genes and proteins, because there are examples in which one-to-one orthologs (which would be expected to have conserved motifs) in fact have different motifs[10,12,13] and, conversely, there are entire families (in which members might be expected to diverge in their binding sites) that have identical motifs (for example, the core sequence for E2F and regulatory factor X (RFX) TFs, as well as the rigid specificity of plant WRKY TFs)[3,14]. Indeed, how motif diversification is dictated by protein structure and mechanisms of DNA binding is largely unknown, except in a few cases[13,15,16].

We reasoned that developing a system for determining both the similarity and dissimilarity of TF motifs would provide uniform and unbiased estimates of the conservation of TFs and their DNA-binding functions. Here, we describe such a 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

Similarity regression predicts TF motif similarity. We developed a homology-based system that, when calculating sequence

[1]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [2]Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada. [3]Institute of Integrative Biology, University of Liverpool, Liverpool, UK. [4]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [5]Canadian Institutes For Advanced Research (CIFAR) Artificial Intelligence Chair, Vector Institute, Toronto, Ontario, Canada. [6]Ontario Institute of Cancer Research, Toronto, Ontario, Canada. [7]Divisions of Biomedical Informatics and Developmental Biology, Center for Autoimmune Genomics and Etiology (CAGE), Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [8]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA. [9]CIFAR, Toronto, Ontario, Canada. *e-mail: t.hughes@utoronto.ca

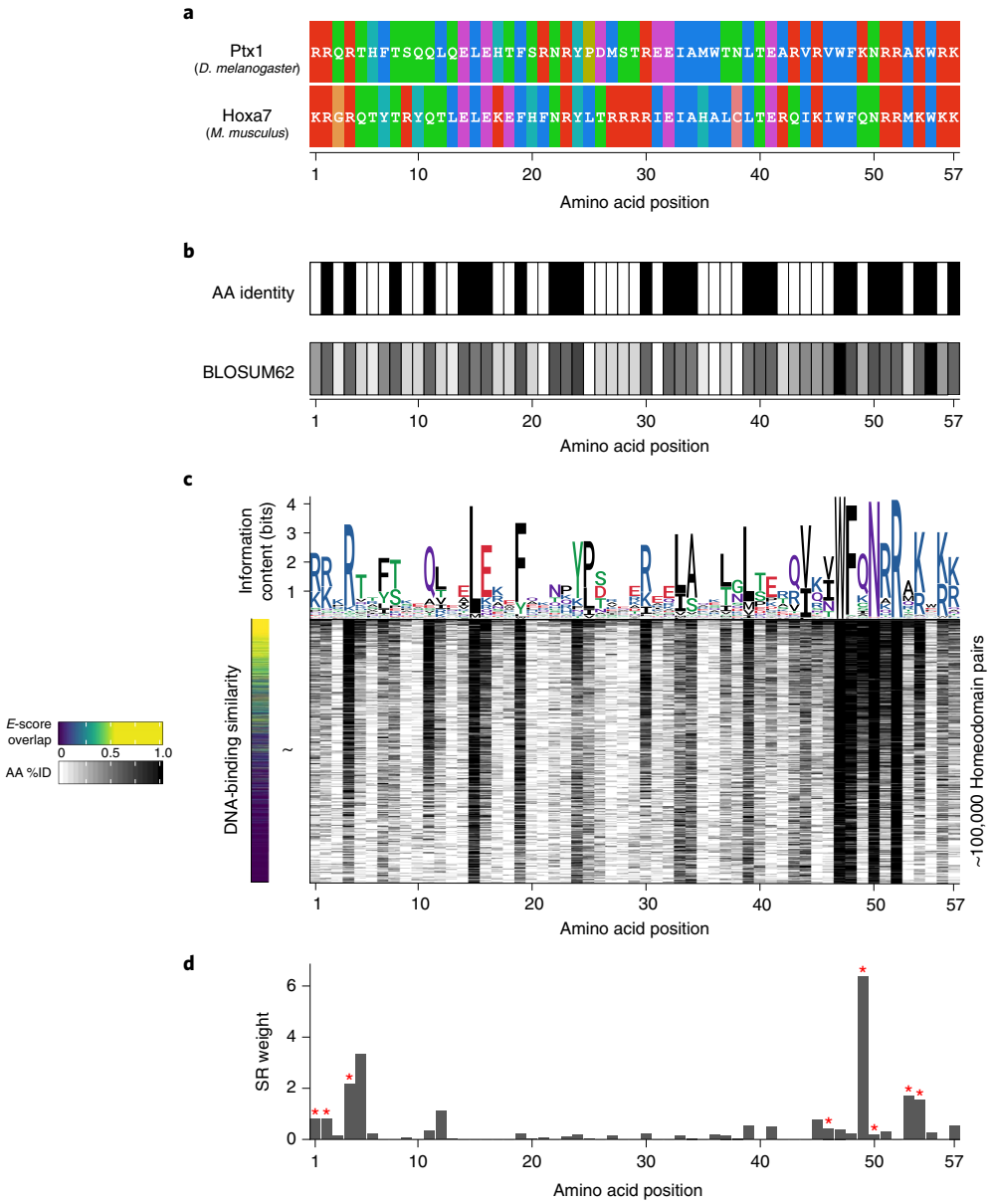

Fig. 1 | Overview of the similarity regression method. Similarity regression uses TF protein similarity to predict the similarity in TF sequence specificities. The procedure and results are outlined for homeodomain TFs as an example. **a**, The DBD sequence of each TF is aligned to the Pfam HMM as a common reference to generate a global alignment. Amino acids (AA) shown are colored according to standard clustal colors for two homeodomains. **b**, For each pair of TFs, the amino acid similarity is measured at each position of the alignment, recording whether the two residues are identical or similar (BLOSUM62 substitution score). This procedure is repeated for every pair of homeodomain TFs with PBM data. **c**, Regression is performed for 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 positional protein similarity as independent *X* values (right, identity shown). Sequence diversity among the TFs is represented here for reference, plotted as a sequence logo (in bits) above the protein similarity matrix. %ID, percent identity. **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[37] are indicated by an asterisk.

similarity in DBDs, assigns greater importance to positions of the DBDs (for example, base-contacting or specificity residues) that have more impact on sequence specificity. To do this, we assigned a weight to each residue when calculating similarity between two DBDs of the same class (for example, C2H2 ZF, ETS or Forkhead). Figure 1 displays the overall procedure. 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 hidden Markov model (HMM) (Fig. 1a–c), whereas the dependent variable is the similarity in DNA sequence preference (Fig. 1c). The weights are the coefficients learned over all pairs for each DBD class (Fig. 1d). We tested four variations of this scheme, including two different regression approaches (linear and logistic) and two different representations of sequence similarity (identity and BLOSUM62 substitution scores). Here, we trained the regression models to learn highly overlapping 8-base oligomer *E*-score preferences obtained from universal protein-binding microarrays (PBMs)[17], as cataloged in Cis-BP[3]. These scores are comparable among different studies, thus circumventing the potentially confounding impact of motif derivation[18]; however, we note that the model could be trained on any metric of motif identity or similarity. Each variation of the model generates a different set

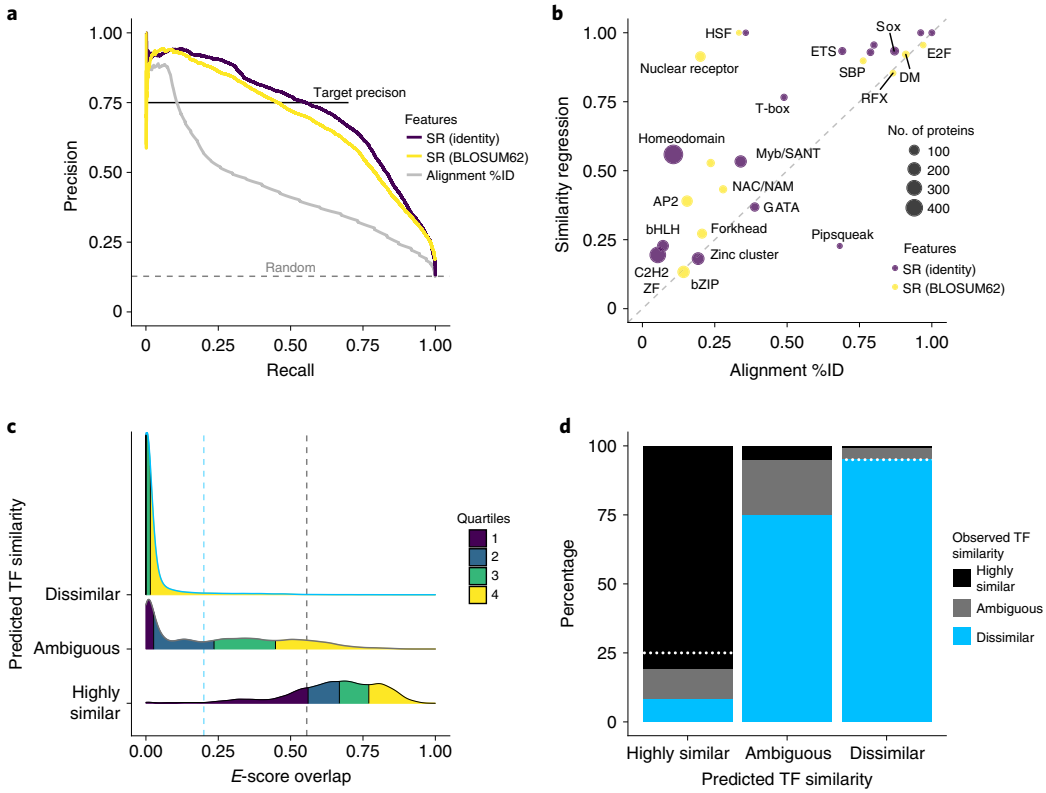

Fig. 2 | Similarity regression classification of TFs for highly similar or dissimilar sequence specificities. a, Precision/recall curves for homeodomains are shown for three prediction methods: alignment percent identity, similarity regression using amino acid identity (SR (identity)), and similarity regression using BLOSUM similarity (SR (BLOSUM62)), on held-out data across all cross-validation folds. 'Positives' are pairs of TFs with highly similar specificities (*E*-score overlap >twenty-fifth percentile of replicate experiments). 'Negatives' are all other pairs. **b**, Scatter plot comparing recall values (predicting highly similar specificities at 75% precision threshold) for similarity regression versus percent identity, for each TF family. The best of the four similarity regression models is shown. Points are sized according to the number of PBM experiments used for training and colored according to the amino acid features used in each model. Domain abbreviations are taken from Pfam[47], where full names can be found. **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 for the highly similar and dissimilar classes (that is, thresholds were chosen to achieve these levels on training data).

of weights, which are selected by leave-one-out cross-validation (Supplementary Fig. 1). Using the same cross-validation, we select the best model for each DBD class among these four models (linear or logistic regression, using amino acid identity or similarity) or, as a fifth possibility, the simple alignment percent identity that is currently used by Cis-BP[3]. We refer to this procedure as similarity regression. Application of similarity regression to TF families for which DBDs are present in arrays (for example, C2H2 ZFs) is explained in Supplementary Fig. 2.

Similarity regression has several advantages over previous approaches. It 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 1d illustrates the weights for the well-studied homeodomain class, which has established specificity residues in DNA-contacting positions[5]. Weights for all eukaryotic DBD families with similarity regression models are given in Supplementary Data 1 (models for homeodomains and C2H2 ZFs are shown in Supplementary Fig. 3a,b, respectively). These weights correspond to known mechanisms of DNA recognition: there is a strong relationship between similarity regression model weight and DNA contact frequency (Supplementary Fig. 3c). For example, similarity regression 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 that contact the minor groove, consistent with structural data[19], whereas GAL4/zinc cluster proteins, the dimerization of which is organized along the DNA backbone[20,21], receive high weights in backbone-contacting residues (Supplementary Fig. 3c).

Similarity regression shows notable improvement in recall at identical precision values over percent identity alone. This improvement is notable for homeodomains (Fig. 2a) and other families with a large amount of PBM data (summarized in Fig. 2b). In these precision/recall curves, positives are those pairs of proteins with *E*-score overlap that exceeds the twenty-fifth percentile of experimental replicates (the same threshold was used previously[3]) and negatives are all other pairs. Thus, our precision/recall curves are likely underestimates, because our stringent negative definition includes highly similar experiments that are just below the threshold.

Importantly, similarity regression can also be used to predict whether two proteins are highly unlikely to share DNA sequence specificity: using 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 oligomers. We allowed some overlap because many families bind a characteristic sequence 'core'. For example, many homeodomains bind TAAT-like sequences, even

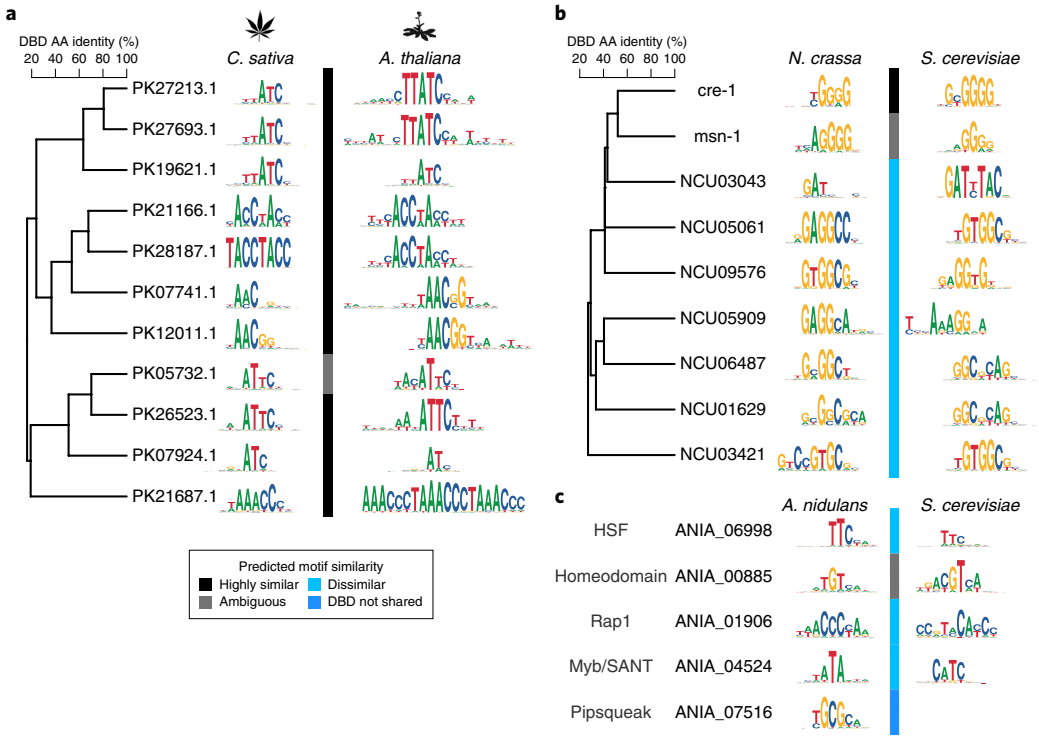

Fig. 3 | PBM data from the plant *C. sativa* and model fungi *A. nidulans* and *N. crassa* for TFs with conserved and dissimilar motifs. NNs for each new TF with PBM data were identified by finding the most similar TF (by similarity regression score) with a motif from either *Arabidopsis thaliana* (for *C. sativa*) or *Saccharomyces cerevisiae* (for *A. nidulans* and *N. crassa*). **a**–**c**, Motifs for Myb/SANT TFs from *C. sativa* (**a**), C2H2 ZF TFs from *N. crassa* (**b**) and TFs from five other TF families in *A. nidulans* (**c**) are shown, with a neighbor-joining tree scaled by DBD alignment percent identity in **a** and **b**. The colored bar represents predicted motif similarity. See Supplementary Fig. 5 for a comparison between similarity regression-predicted similarity and NN TF similarity for all new PBM data. Silhouettes of each species are displayed, adapted from Phylopic (http://phylopic.org) under a Creative Commons license (https://creativecommons.org/licenses/by-sa/3.0/).

though their most highly preferred 8-base oligomers differ among family members. For each DBD type, we set a similarity regression 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 Supplementary Fig. 4a, similarity regression outperformed percent identity at discriminating these dissimilar pairs by achieving a higher recall at the same NPV.

In all subsequent analyses, we use similarity regression to classify all pairs of proteins that share the same DBD type as 'highly similar' if their similarity regression score is over the positive pair threshold, 'dissimilar' if it is below the negative pair threshold or 'ambiguous' if it is between the two thresholds. The ambiguous score range can, in some cases, be quite large; Fig. 2c shows that it is predictive of intermediate 8-base oligomer overlap for homeodomain TFs. Similar phenomena are observed in other TF families (data not shown). Multi-class accuracy of the similarity regression models, and their improvement over percent identity, is summarized by the Matthews correlation coefficient in Supplementary Fig. 4b. Similarity regression outperforms percent identity in all but four TF families.

New PBM data validate motif similarity classifications. To confirm that the models correctly classify previously unseen proteins, we generated new PBM data for 340 TFs, including 15 human TFs and other TFs that represent multiple eukaryotic kingdoms, with a particular focus on *Cannabis sativa* (a medicinal plant), *Caenorhabditis briggsae* (a nematode), *Aspergillus nidulans* and *Neurospora crassa* (model fungi) (Supplementary Table 1). These TFs were selected to increase the number of experimentally determined motifs for TFs in these species of interest, to obtain novel motifs by analyzing proteins that are dissimilar to TFs with known motifs, or both. We used these data as a validation set to test how well similarity regression models measure the similarity of TF sequence specificity on unseen data (Fig. 2d). The TF similarity classifications for the newly analyzed proteins are correct for 81.2% of the highly similar and 95.2% of the dissimilar pairs. These results held over a range of percent identity to other proteins in Cis-BP, confirming that the models are accurate with independent data. Indeed, there is an overall correlation between similarity regression score and *E*-score overlap between the held-out data and the most similar training construct (by similarity regression score) for the similarity regression model of each TF family (Supplementary Fig. 5, median $R^2 = 0.63$). Figure 3 provides examples of conservation and divergence of motifs in the new data.

Comparison to alternative motif prediction methods. We also investigated whether similarity regression could accurately predict motifs by comparing motif predictions with generalist methods (for example, affinity regression[4] applied to all TF families) and domain-specific recognition codes. Similarity regression predicts similarity in DNA sequence specificity of two proteins, whereas affinity regression directly predicts preferences of TFs or RNA-binding proteins to individual DNA or RNA sequences on the basis of their protein sequences. Nonetheless, the two can be compared using similarity regression to predict 8-base oligomer preferences from proteins that should have highly similar sequence preferences, and similarity regression outperforms affinity regression in many cases. Using an identical training set (that is, the same experiments on the same proteins), similarity regression slightly outperformed affinity regression when predicting 8-base oligomer *Z*-score profiles

for our 315 held-out constructs across 19 TF families (see Methods for details) using either the single most similar protein (nearest neighbor (NN), $P<0.05$; Supplementary Fig. 6a) or by predicting the Z-scores as a composite of up to five most similar proteins (top 5, $P<0.01$; Supplementary Fig. 6a). Similarity regression has the added benefit that it does not make predictions for dissimilar proteins when the predictions are poor, whereas affinity regression makes a prediction for every protein, without an associated quality metric. In comparison to affinity regression, similarity regression predictions that use only highly similar TF pairs ($n=104$) have a higher correlation to the measured Z-score profiles than affinity regression using the NN ($P<0.01$) or top 5 predictions ($P<0.0001$) (Supplementary Fig. 6). These outcomes hold for most (although not all) individual TF families analyzed in isolation. For example, whereas similarity regression performs equivalently to affinity regression for zinc cluster TFs, it scores higher for the homeodomains and C2H2 ZF families (Supplementary Fig. 6b–d).

We also compared similarity regression predictions to those produced by state-of-the-art recognition code algorithms for C2H2 ZF and homeodomain TFs, which directly predict PWMs from DBD sequences. One of the C2H2 ZF predictors uses support vector machines[22] whereas the other uses random forests (ZFModels)[23]; the homeodomain predictor (PreMoTF)[5] also uses random forests. To compare to these specialist prediction tools, we converted the similarity regression-predicted Z-score profiles to PWMs using a simple alignment of the top-scoring 8-base oligomers (PWMalign)[18]. We compared the motif similarity between predicted motifs to those derived from the held-out PBM experiments described above, which included 34 C2H2 ZF and 17 homeodomain proteins. Supplementary Fig. 7a shows that, in most cases, similarity regression motif predictions for C2H2 ZFs are more similar to the experimental motifs than the motifs predicted by either of two recognition codes, regardless of whether similarity regression predictions were filtered to be highly similar. For homeodomains, there was no significant difference between predictions made by the recognition code and those made by similarity regression based on multiple (that is top 5 or highly similar) neighbors ($P>0.05$); however, PreMoTF outperforms similarity regression NN-based motif predictions ($P<0.05$) (Supplementary Fig. 7b).

New TF similarity predictions improve Cis-BP. To capitalize on the increased recall of similarity regression relative to percent identity, 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 (updated from 340) and 11,491 experimentally determined motifs that correspond to 4,559 distinct TFs (updated from 6,559 motifs for 3,202 distinct TFs) and implemented similarity regression across all 392,333 known and putative eukaryotic TFs. We also updated many other properties of the database (for example, genome builds and DBD models) (Methods).

The incorporation of similarity regression in Cis-BP increases the number of TFs with predicted motifs by more than 6,000 compared to our previous method, at the same expected precision (a 4.2% overall increase, on identical genomes, DBDs and motifs). Coverage of numerous TF families is increased markedly (Supplementary Fig. 8a). For instance, ten TF families more than doubled their motif coverage, including zinc cluster (123% increase) and Sox (162% increase) TFs, the second and seventh most abundant families in Cis-BP, respectively. The average species now has 4% more TFs with motifs (experimental and predicted), yielding an average motif coverage of 37% (with 73% for human) (Supplementary Fig. 8b) and a total coverage of 165,030 out of 392,333 eukaryotic TFs (42%). This updated

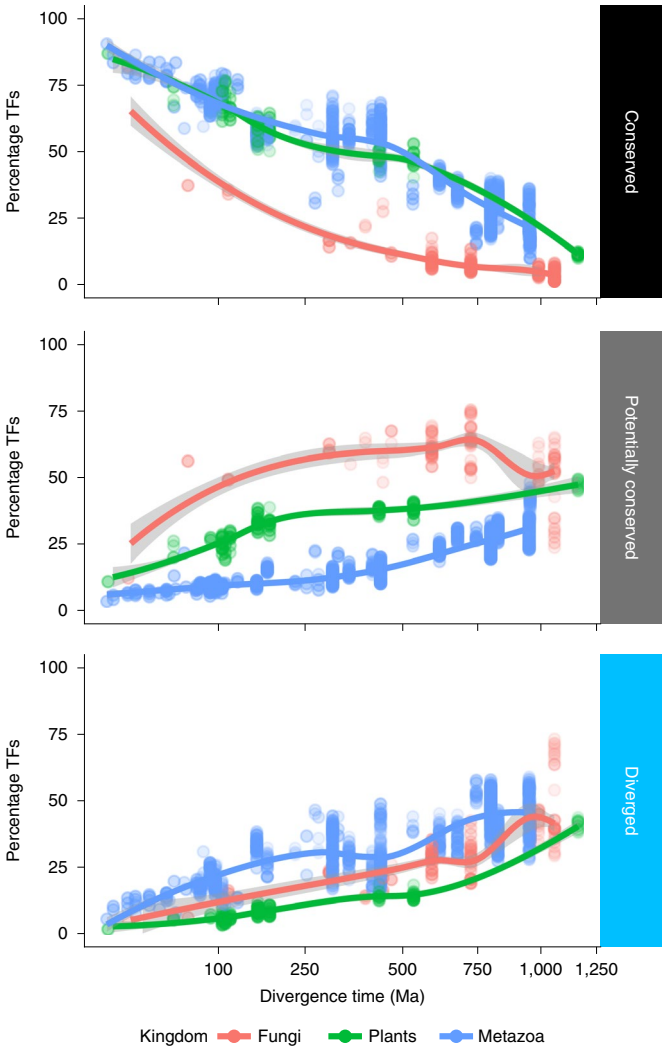

Fig. 4 | Conservation of TF motifs within major eukaryotic kingdoms. The average percentage of TFs for which the closest TF in the other species is conserved (similarity regression classifies as highly similar), potentially conserved (ambiguous) or diverged (dissimilar and unshared DBDs) was calculated for each pair of species from the same kingdom (48 metazoan, 15 plant and 15 fungal species; Supplementary Fig. 7b). Each point represents the average percentage of TFs within each category, for each pair of species (that is, average of species *X* versus species *Y* and *Y* versus *X*), 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 and shading show the LOESS regression fit and 95% confidence interval, respectively. Ma, million years ago.

Cis-BP database can be found at http://cisbp.ccbr.utoronto.ca/, where TF annotations, motifs and PBM data compiled from our laboratory and other public databases can be accessed and downloaded. In addition to increased coverage, the new build—which contains many more genomes—also identifies many new families of TFs with still-unknown sequence specificity.

Evolution of TF 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 are listed in Supplementary Fig. 8b). We included all TF DBDs in this analysis, regardless of conservation level or patterns of orthology, to gain

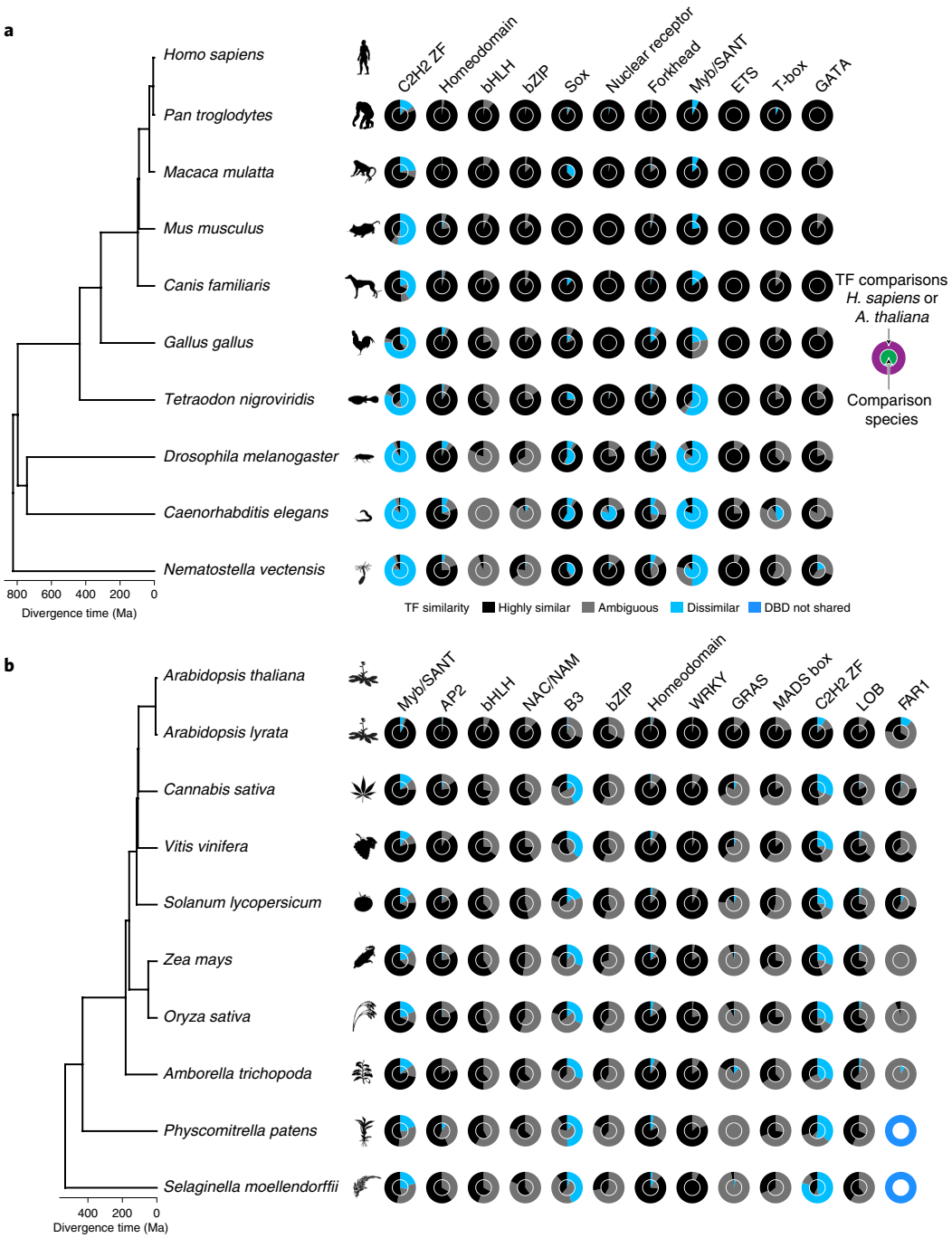

Fig. 5 | Motif divergence of TF families in metazoans and plants. a, Nested pie charts showing the percentage of human TFs for which the closest TF in other metazoans is highly similar, ambiguous, dissimilar or not shared, for the 11 most abundant metazoan DBDs. The outer ring of each pie chart shows the proportion of human TFs in each similarity regression-predicted similarity class relative to the other species; the inner ring shows the proportion of TFs for the other species, relative to humans. **b**, Motif similarity between *A. thaliana* and other plants, for the 13 most abundant plant DBDs. Silhouettes of each species are displayed, adapted from Phylopic (http://phylopic.org) and/or are available under a Creative Commons license (https://creativecommons.org/publicdomain/zero/1.0/).

a complete picture of motif conservation. For each protein, we identified the protein with the highest similarity regression model score as described above in each other species and recorded the classification (that is, highly similar, ambiguous or dissimilar). If there is no protein with the same DBD type in the other species, the TF is labeled 'DBD not shared' with the other species. Thus, there are four possible labels for each TF–species comparison that are mutually exclusive.

Figure 4 shows that eukaryotic kingdoms display qualitatively similar trends in the proportion of TFs within each of the categories, with respect to divergence time. Around 100 million years ago (for example, origin of placental mammals and eudicot plants), approximately 75% of motifs are conserved (highly similar) and an additional 5–25% are potentially conserved (ambiguous category) for metazoans and plants, respectively. However, around 900 million years ago (the origin of metazoans), only around 60% are conserved

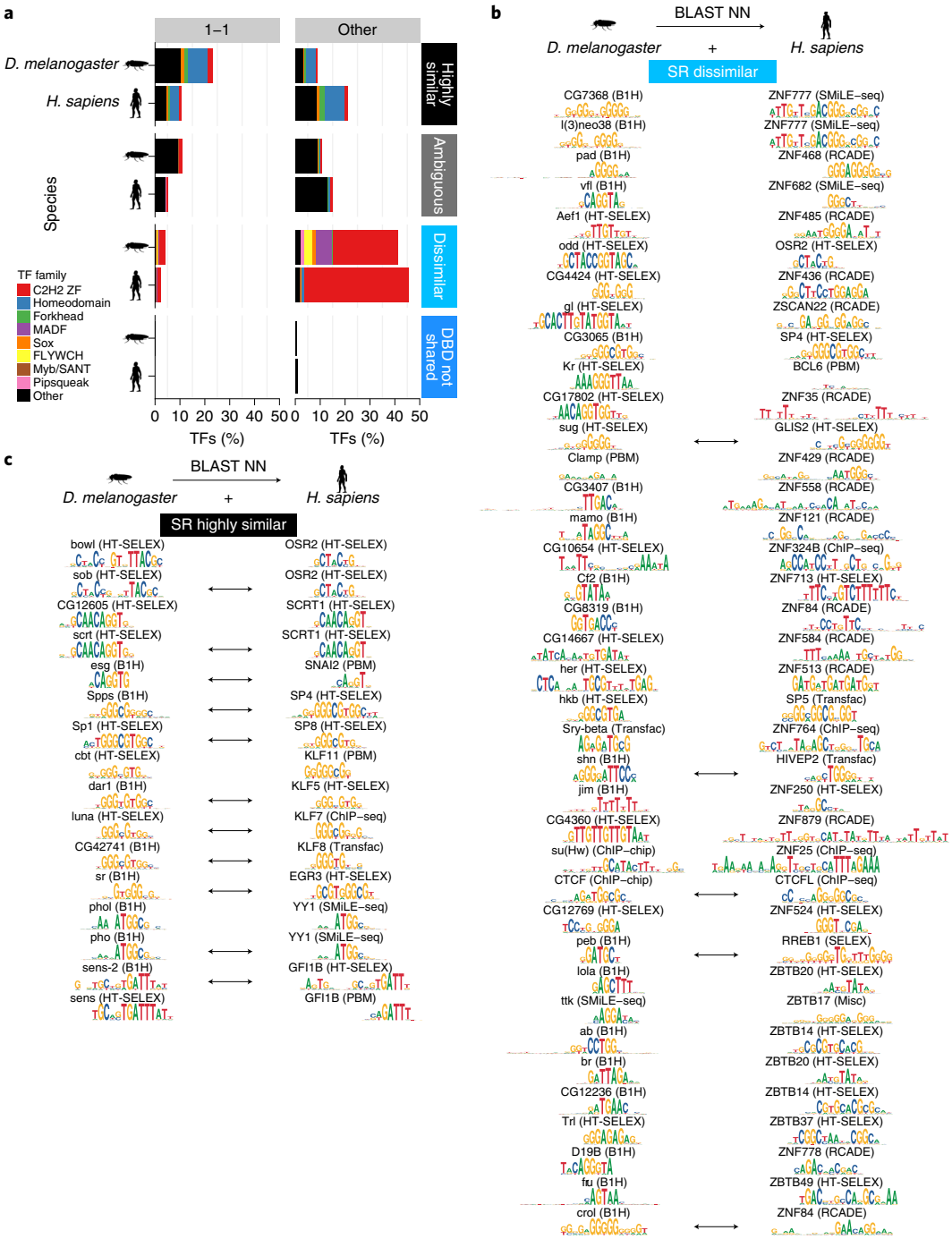

Fig. 6 | TF motif conservation between human and *Drosophila melanogaster*. **a**, Percentage of TFs in human or *Drosophila* (as indicated) that fall into each similarity regression motif similarity class. TF conservation is partitioned by whether the TF has a one-to-one ortholog (reciprocal best BLAST hit) or all other orthology relationships. Colors of the stacked bar plots indicate TF family. **b**,**c**, Experimentally determined motifs for individual *Drosophila* and human C2H2 ZF TFs, shown in pairs that correspond to the BLASTP best hit (*Drosophila* query to human database). Reciprocal best BLASTP matches (putative one-to-one orthologs) are indicated with bidirectional arrows. **b**, Pairs predicted to be dissimilar by similarity regression are shown. **c**, Pairs predicted to be highly similar by similarity regression are shown.ChIP-seq, chromatin immunoprecipitation with high-throughput sequencing; ChIP-chip, ChIP with DNA microarray; SMiLE-seq, selective microfluidics-based ligand enrichment followed by sequencing; RCADE, recognition code-assisted discovery of regulatory elements. Silhouettes of each species are displayed, adapted from Phylopic (http://phylopic.org) under a Creative Commons license (https://creativecommons.org/licenses/by-sa/3.0/).

or potentially conserved; a similar proportion is obtained for the origin of fungi (around 1,055 million years ago). Within the plant kingdom (approximately 1,160 million years ago), only slightly more motifs are conserved or potentially conserved (around 65%). Across kingdoms (for example, between fungi and metazoan), most DBDs are not shared[24] and are thus not comparable. Even among those that are comparable (that is, 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 (Fig. 5, Supplementary Fig. 9); however, these families have a large number of members and are in general already known for their lineage-specific expansions: C2H2 ZF in metazoans[9], nuclear

hormone receptors in nematodes[25,26] and Myb proteins in plants[27]. The similarity regression analysis thus underscores DNA sequence specificity as a mode of diversification following duplication of these proteins. However, many other families appear rigid in their DNA-binding motifs and have presumably diversified in function by other mechanisms (for example, basic-leucine zipper (bZIP) and basic helix–loop–helix (bHLH) proteins are able to diversify through changes in heterodimerization partners)[28,29].

One notable 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 only a few TF-binding motifs have diverged in sequence specificity between human and *Drosophila*. The discrepancy appears to be due to the fact that the HT-SELEX data in this previous study were highly biased towards TF families that have not diversified. In particular, it included only a small minority of C2H2 ZFs, which represent the largest class of TFs in both species. Similarity regression predicts that the majority of C2H2 ZF proteins do not have conserved motifs (Fig. 6a), even between TFs that have putative one-to-one orthology as defined by BLAST. Experimental data confirm our similarity regression predictions (motifs for C2H2 TFs are shown in Fig. 6b,c; motif similarities for all TF comparisons are shown in Supplementary Fig. 10). Examples of orthologous C2H2 TFs, even one-to-one orthologs, differing substantially in their DNA-binding specificity are shown in Fig. 6b, illustrating that simple orthology alone can be a poor predictor of shared motifs. As a control, BLAST NNs predicted by similarity regression to have highly similar motifs between human and *Drosophila* do display more similar motifs in the experimental data than TFs with predicted ambiguous or dissimilar specificities (Supplementary Fig. 10), even when they were obtained using different techniques (primarily high-throughput SELEX (HT-SELEX)[8,30] compared to bacterial one-hybrid (B1H) assays[31,32]).

Discussion

We anticipate that similarity regression will contribute to understanding of TF function in several ways. First, it presents several advantages in the task of predicting motifs. Like simple homology (that is, alignment percent identity), the score it produces serves as a confidence measure that can be used to avoid incorrect predictions. At the same time, the increased recall (that is, coverage) of similarity regression, relative to percent identity, provides a substantial increase in the number of predicted motifs, which are now included in our update of the Cis-BP database. Similarity regression can also be adapted to predict new motifs, by combining binding data from related proteins. These motifs score favorably relative to both a related general-purpose prediction method (affinity regression), as well as state-of-the-art recognition codes for specific DBD families (C2H2 ZFs and homeodomains). We do note that the motifs for the homeodomain-specific prediction tool PreMoTF[5] scored more highly than similarity regression among the held-out homeodomains for which there was no highly similar protein (that is, a protein with a high similarity regression score) in the training set. Thus, although the use of the similarity regression score as a confidence metric can be seen as an advantage, this outcome also highlights a disadvantage of making predictions solely on the basis of similarity among proteins: a well-formulated recognition code has the potential to make accurate predictions for completely novel proteins. Unfortunately, such recognition codes do not exist for the vast majority of DBD types.

Second, the weights (that is, coefficients) produced by similarity regression 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—which could, in turn, contribute to the long-term objective of developing accurate recognition codes that do not rely on as much experimental data as similarity regression does.

Similarity regression can also predict when proteins are unlikely to share sequence preferences, thus enabling 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 between species, even over relatively short timescales (<100 million years). TFs with divergent motifs are concentrated in TF families with established patterns of lineage-specific expansions, although changes in one-to-one orthologs do occur as have previously been observed[10,12,13]. This study provides an extensive analysis of TF sequence specificity for both *Cannabis* and *Aspergillus*, and both the outputs of similarity regression and the newly generated data highlight the diversity of DNA-binding motifs in both the plant and fungal lineages. Despite lower diversity in the specificity residues of individual C2H2 ZF domains in fungi, relative to metazoa[16], proteins that contain 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 to motif divergence in multiple lineages (both plants and fungi). The GAL4/zinc cluster domain proteins, which have expanded in fungi, have largely conserved monomeric binding specificity in their DBDs, and are thus more likely contribute to TF diversification by alterations in spacing and orientation of dimeric sites as homo- or heterodimers[33].

Similarity regression also confirms the extreme diversity of motifs in the C2H2 ZF family in metazoa. C2H2 ZFs are the fastest evolving TF family in the recent human lineage[34] and similarity regression indicates—and experimental data confirm—that their sequence specificities are largely distinct from those in *Drosophila*, even among clear orthologs. Our findings differ from a previous conclusion that TF-binding specificities are highly conserved between *Drosophila* and mammals[8], mainly because the vast majority of the C2H2 ZF proteins were absent from the HT-SELEX data in the previous study. This absence could be due to low success rates in HT-SELEX (and other in vitro assays), due to long binding sites or other factors. Importantly, most C2H2 ZFs are bona fide TFs, which bind specific DNA sequences in vivo and/or in vitro[7,34]. Notably, among *Drosophila* species, even one-to-one orthologs of C2H2 TFs frequently differ in specificity residues and these differences are predicted to influence DNA sequence preferences[10]. The bulk of motif differences occur in TFs with more complex orthology patterns, however. In mammals, there is strong evidence that the need to recognize new retroelements for silencing by KRAB-containing C2H2 ZFs has played a part in their evolution[35], but the KRAB domain is restricted to tetrapods and it is unclear what forces are driving C2H2 ZF motif diversification in other lineages. Even in humans, most C2H2 ZF proteins do not appear to bind to retroelements[36] and presumably have other functions.

Knowing the sequence specificities of TFs is an important first step in their characterization. Overall, we anticipate that similarity regression 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.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41588-019-0411-1.

Received: 13 November 2018; Accepted: 4 April 2019;
Published online: 27 May 2019

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–D115 (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–i89 (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.* **33**, 555–562 (2015).
8. Nitta, K. R. et al. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife* **4**, e04837 (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. Nadimpalli, S., Persikov, A. V. & Singh, M. Pervasive variation of transcription factor orthologs contributes to regulatory network evolution. *PLoS Genet.* **11**, e1005011 (2015).
11. Lynch, V. J. & Wagner, G. P. Resurrecting the role of transcription factor change in developmental evolution. *Evolution* **62**, 2131–2154 (2008).
12. Baker, C. R., Tuch, B. B. & Johnson, A. D. Extensive DNA-binding specificity divergence of a conserved transcription regulator. *Proc. Natl Acad. Sci. USA* **108**, 7493–7498 (2011).
13. Sayou, C. et al. A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity. *Science* **343**, 645–648 (2014).
14. Morgunova, E. et al. Structural insights into the DNA-binding specificity of E2F family transcription factors. *Nat. Commun.* **6**, 10050 (2015).
15. 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).
16. Najafabadi, H. S. et al. Non-base-contacting residues enable kaleidoscopic evolution of metazoan C2H2 zinc finger DNA binding. *Genome Biol.* **18**, 167 (2017).
17. Berger, M. F. et al. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* **24**, 1429–1435 (2006).
18. Weirauch, M. T. et al. Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.* **31**, 126–134 (2013).
19. Love, J. J. et al. Structural basis for DNA bending by the architectural transcription factor LEF-1. *Nature* **376**, 791–795 (1995).
20. Marmorstein, R., Carey, M., Ptashne, M. & Harrison, S. C. DNA recognition by GAL4: structure of a protein–DNA complex. *Nature* **356**, 408–414 (1992).
21. 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).
22. Persikov, A. V. & Singh, M. De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. *Nucleic Acids Res.* **42**, 97–108 (2014).
23. Gupta, A. et al. An improved predictive recognition model for Cys2-His2 zinc finger proteins. *Nucleic Acids Res.* **42**, 4800–4812 (2014).
24. de Mendoza, A. et al. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proc. Natl Acad. Sci. USA* **110**, E4858–E4866 (2013).
25. Narasimhan, K. et al. Mapping and analysis of *Caenorhabditis elegans* transcription factor sequence specificities. *eLife* **4**, e06967 (2015).
26. Robinson-Rechavi, M., Maina, C. V., Gissendanner, C. R., Laudet, V. & Sluder, A. Explosive lineage-specific expansion of the orphan nuclear receptor HNF4 in nematodes. *J. Mol. Evol.* **60**, 577–586 (2005).
27. Stracke, R., Werber, M. & Weisshaar, B. The R2R3-MYB gene family in *Arabidopsis thaliana*. *Curr. Opin. Plant Biol.* **4**, 447–456 (2001).
28. Grove, C. A. et al. A multiparameter network reveals extensive divergence between *C. elegans* bHLH transcription factors. *Cell* **138**, 314–327 (2009).
29. 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).
30. Jolma, A. et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* **20**, 861–873 (2010).
31. 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).
32. 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–D117 (2011).
33. MacPherson, S., Larochelle, M. & Turcotte, B. A fungal family of transcriptional regulators: the zinc cluster proteins. *Microbiol. Mol. Biol. Rev.* **70**, 583–604 (2006).
34. Lambert, S. A. et al. The human transcription factors. *Cell* **175**, 598–599 (2018).
35. Ecco, G., Imbeault, M. & Trono, D. KRAB zinc finger proteins. *Development* **144**, 2719–2729 (2017).
36. Schmitges, F. W. et al. Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Res.* **26**, 1742–1752 (2016).
37. Noyes, M. B. et al. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* **133**, 1277–1289 (2008).
47. Finn, R. D. et al. The Pfam protein families database. *Nucleic Acids Res.* **38**, D211–D222 (2010).

Acknowledgements

We thank Xiaoting Chen and Mario Pujato for computational support. S.A.L. was funded by a Natural Sciences and Engineering Research Council of Canada Doctoral Fellowship. T.R.H. holds the Billes Chair of Medical Research at the University of Toronto. This work was supported by a Canadian Institutes of Health Research grant (FDN-148403) and a Natural Sciences and Engineering Research Council of Canada grant (RPGIN-2016-05643) to T.R.H., National Institutes of Health (NIH) grants R01 AR073228, R01 NS099068 and R01 GM055479, Lupus Research Alliance 'Novel Approaches', CCRF Endowed Scholar and CCHMC CpG Award 53553 to M.T.W. and a Canadian Institutes of Health Research Operating grant (MOP-125894) to Q.D.M. and T.R.H.

Author contributions

S.A.L., M.T.W. and T.R.H. conceived the study and oversaw it to completion. S.A.L. analyzed the data, made the figures and performed all computational analyses except for experiments for which A.S. reimplemented the affinity regression pipeline and applied it to new data. Q.D.M. guided the computational and statistical analyses. M.A., S.A.L. and M.T.W. maintained and updated the Cis-BP database. G.C. and M.X.C. produced the clones for *Aspergillus* PBM experiments. A.W.H.Y. produced the remainder of the clones and performed all PBM experiments. S.A.L. and T.R.H. wrote the manuscript with feedback and approval from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0411-1.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to T.R.H.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019

Methods

Similarity regression. Similarity regression is formulated as a regression task in which 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 $\mathbf{X}$) are the identity and similarity of 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[38], recording match positions (that is, positions that are present in the HMM). An example alignment of two homeodomain sequences is presented in Fig. 1a. At each position of the aligned sequences, either identity (as binary values) or similarity (BLOSUM62 substitution score[39]) is recorded (Fig. 1b), yielding the feature vector for each TF pair. For TF families that have DBDs present in arrays (mainly C2H2 ZFs and Myb) the best ungapped and overlapping pairwise alignment of DBD arrays (Supplementary Fig. 2a) 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 (Supplementary Fig. 2b).

In the analyses described, the metric of similarity in DNA sequence specificity between pairs of proteins (Y) is calculated from the 8-base oligomer PBM data as the fraction of high-scoring 8-base oligomers ($E > 0.45$) that are shared between two TFs (that is intersection/union for two experiments, referred to as 'E-score overlap'). For each TF family, E-score overlaps that exceed the twenty-fifth percentile of experimental replicates (the same threshold was used previously[3]) are taken as having 'highly similar' specificities. The highly similar labels are not only used as positives for training logistic similarity regression models (see next paragraph), but also for evaluating the performance of similarity regression (for example, precision/recall analysis). E-score overlaps that are 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-base oligomers differ. The dissimilar labels are used as negatives to define the score threshold below which TFs are unlikely to share specificities in a NPV analysis.

For each TF family with sufficient PBM data, we trained four similarity regression models that varied in the representation of protein similarity (identity or BLOSUM substitution score) and in the representation of the data (either linear or logistic regression models). Each regression model is trained in R[40] using glmnet[41] constrained to fit positive regression coefficients, selecting the optimal ridge (L_2) regularization strength using cross-validation. Because 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 cross-validation (testing on data points made from the comparisons of a single TF), which we implemented using the caret package[42]. This performance measure can be interpreted as how well a similarity regression 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 similarity regression model generation and selection for homeodomain TFs is presented in Supplementary Fig. 1. First, the optimal regularization strength is selected using the cross-validation procedure implemented in caret, yielding a selected model for each feature–output combination. For each regression model, and the percent 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 cross-validation fold are combined and compared with their known TF similarity labels. To identify TFs with highly similar sequence specificities (E-score overlap > the TF family replicate threshold (the twenty-fifth percentile of experimental replicates as previously described[3])), a precision/recall 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 the one used in the previous study[3]). A threshold for dissimilar specificities is derived by finding a NPV cut-off that classifies 95% of TFs below that score threshold as having truly dissimilar specificities (E-score overlap of <0.2). For each threshold, the recall of positive and negative predictions was recorded to evaluate the improvement of similarity regression models over percent identity. If thresholds could not be derived for a TF family a global threshold of 70% identity for predicting highly similar TFs (identical to our previous study[3]) and a threshold of 25% identity for predicting dissimilar TFs (selected to yield 95% NPV) were derived. 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 similarity regression model (and for the percent identity method). The best similarity regression model is then selected by comparing the three-class predictions to ground-truth labels and selecting the model with the best Matthews correlation coefficient, a metric of multi-class classification accuracy that is sensitive to class imbalance[43]. This process yields a single final similarity regression model for each TF family, composed of a weight vector (that is, 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 similarity regression weights with known DNA-contacting residues. We used the DNAproDB database[44] to compare the similarity regression weights with known protein–DNA contacts. DNAproDB catalogues DNA–protein complexes present in the Protein Data Bank[45], 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 of the protein sequences in DNAproDB and identified DBDs using hmmscan[46] and the same Pfam HMM models[47] and thresholds as the Cis-BP database. We then parsed the nucleotide–residue interactions for each structure into backbone, major and minor groove interactions (using DNAproDB-recommended buried solvent accessible surface area, 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 that contained that DBD. Correspondence between similarity regression weights and the three classes of DNA contacts were evaluated using partial Pearson correlations, which assessed the correlation between each contact type after removing the effects of the other two contacts on the similarity regression weights.

Comparison of similarity regression with affinity regression. Affinity regression predicts Z-scores of DNA 8-base oligomers from short peptides in the protein sequence. Here, we implemented a soft-coded Python version of affinity regression, ensuring similar performance on the previously published, original data[4] and using identical constructions of the protein and DNA features. A single affinity regression model for each TF family was trained using the same data as the corresponding similarity regression model and the number of informative components selected after dimensionality reduction was set to capture 90% of the weight of each singular value. To predict the Z-scores of uncharacterized/tested transcription factors, affinity regression determines their protein k-mer vectors to predict the similarities of the held-out TF to all characterized protein profiles in the training set. Affinity regression uses these similarities to reconstruct the Z-score profiles by a similarity-weighted sum of Z-score profiles using either the NN or top 5 NNs and a geometrical reconstruction from the span of the training vectors, proposed and applied in the previously published study that describes affinity regression[4]. Affinity regression was applied to the new TFs that were present in the PBM data from this study.

We used three means to predict the 8-base oligomer Z-score profile for each held-out TF using similarity regression by first copying the Z-score profile from the protein with the highest similarity regression score (that is, the single NN), after which the Z-score profiles of the five proteins with the highest similarity regression scores (top 5) were combined, weighting the Z-scores for each of the five by the corresponding similarity regression or percent identity score and by then combining the Z-scores from all TFs in the training set that are predicted by Similarity regression to have highly similar specificities (similarity regression highly similar), weighting the Z-scores for each of these values by the corresponding similarity regression score. We evaluated the accuracy of similarity regression, percent identity and affinity regression predictions using the Pearson correlation coefficient between the predicted Z-score profile and the experimental Z-scores. We used two-sided paired Wilcoxon signed-rank tests to identify significant differences in mean Pearson correlation coefficient ranks between Z-score reconstruction methods.

Comparison of similarity regression with recognition codes. We used 'PWM_align'[18] to convert the similarity regression and percent identity-predicted Z-score profiles (described above) into PWMs. We used published webservers with default settings to obtain predicted motifs (homeodomains: PreMoTF[5], http://stormo.wustl.edu/PreMoTF/); C2H2 predictions, linear-expanded support vector machine method[22] (http://zf.princeton.edu/) and ZFModels[5] (http://stormo.wustl.edu/ZFModels/). We measured motif similarity between the predicted motifs and experimentally determined motifs using MoSBAT energy scores[48] (parameters: $N = 100{,}000$, $L = 25$ nt). We used two-sided paired Wilcoxon signed-rank tests to identify significant differences in mean MoSBAT energy score ranks between motif prediction 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 (increased from 340) (http://cisbp.ccbr.utoronto.ca/). 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 identifiers, gene names and gene aliases. Furthermore, the set of human TFs contained in Cis-BP now matches the set of 1,639 curated TFs provided in a recently published review[34]. DBD scans were performed using updated Pfam HMM models[47], 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 to 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[49], 534 DNA affinity purification sequencing (DAP-seq)

motifs for *Arabidopsis thaliana*[50], 248 HT-SELEX *D. melanogaster* motifs[8] and 221 ChIP-exo (ChIP–seq with increased resolution from exonuclease treatment) and ChIP–seq-derived C2H2 ZF motifs[51]. Existing motif sources such as UNIPROBE[52], Transfac[53], JASPAR[54] and HOCOMOCO[55] 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 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 to DNA specifically and also contains Myb-like sequences that are also likely to bind DNA) or SANT (which does not bind to DNA specifically). In brief, we scored each Myb/SANT domain with the Myb (PS51294), Myb-like (PS50090) and SANT (PS51293) specific PROSITE[56] models and annotated domains by the profile with the highest score. This procedure is now applied to 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 one-to-one orthologs (reciprocal best BLAST hits) of metazoan proteins with false-positive human TFs derived from a recent curation effort[34]. 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 percent identity.

Predicting TF motif conservation across species. To evaluate motif conservation between species, we used the TF annotations and DBD sequences from Cis-BP (version 2.0). For each pair of species analyzed, we used similarity regression 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 similarity regression 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 (for example, 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[57].

To identify the most similar proteins between human and *Drosophila*, we used BLASTP[58] 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 one-to-one orthologs) were recorded. We measured motif similarity between BLAST NNs with experimentally determined motifs using MoSBAT energy scores[48] (parameters: $N = 200{,}000$, $L = 50$ nt).

DBD cloning. In total, 350 novel *A. nidulans* TF DBDs were selected for analysis and 180 were successfully cloned into the expression vector (pTH6838) and validated by sequencing. These were cloned using RNA that was extracted from the wild-type *A. nidulans* strain (FGSC A4). cDNA was generated by RT–PCR using random hexamer primers. Proof-reading KOD Hot Start DNA polymerase was used to amplify the DBD-coding region and flanking regions up to 50 amino acids long and products were extracted from a 1% agarose gel using a Silica Bead DNA Gel Extraction Kit (Thermo Fisher Scientific). Double digests were performed using the restriction endonucleases AscI (10 U μl^{-1}) (Thermo Fisher Scientific) and SbfI-HF (20 U μl^{-1}) (New England Biolabs). The fragments were ligated into the expression vector using T4 DNA ligase (New England Biolabs). Constructs were verified by Sanger sequencing (GATC Biotech). Other DBDs were cloned using previously reported procedures[3].

PBMs. PBM laboratory methods were performed as described previously[18,59]. Each DBD-encoding plasmid was analyzed in duplicate on two different arrays with differing probe sequences. The 8-base oligomer *Z*- and *E*-scores were calculated as previously described[17]. We deemed experiments successful if at least one 8-base oligomer had $E > 0.45$ on both arrays, the complementary arrays produced highly correlated *E*- and *Z*-scores and yielded similar PWMs based on the PWM_align algorithm[18]. 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)[3,60].

Statistics and experimental design. Two-sided paired Wilcoxon signed-rank tests were used to identify significant differences between motif prediction methods. Distributions were summarized with box plots where appropriate and described in the relevant figure legends. Additional details of the experimental design and data are included in the Nature Research Reporting Summary.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

New PBM data and motifs are deposited in GEO (accession number GSE121420) and the Cis-BP database (v.2.0; http://cisbp.ccbr.utoronto.ca/).

Code availability

The Similarity Regression code, and examples, are available on GitHub (https://github.com/smlmbrt/SimilarityRegression).

References

38. Wilkinson, S. P. aphid: an R package for analysis with profile hidden Markov models. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btz159 (2019).
39. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA* **89**, 10915–10919 (1992).
40. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2013);http://www.R-project.org/
41. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
42. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **28**, 1–26 (2008).
43. Gorodkin, J. Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.* **28**, 367–374 (2004).
44. 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).
45. Berman, H. M. et al. The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
46. HMMER: biosequence analysis using profile hidden Markov models (Howard Hughes Medical Institute, 2015); http://hmmer.org/
48. 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).
49. Yin, Y. et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017).
50. O'Malley, R. C. et al. Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell* **165**, 1280–1292 (2016).
51. 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 (Bethesda)* **8**, 219–229 (2018).
52. 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–D122 (2015).
53. Matys, V. et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).
54. 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).
55. 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).
56. Sigrist, C. J. et al. PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.* **3**, 265–274 (2002).
57. 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).
58. 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).
59. 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).
60. Zhao, Y. & Stormo, G. D. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat. Biotechnol.* **29**, 480–483 (2011).

nature research

Corresponding author(s): Hughes, TR

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Universal Protein Binding Microarray (PBM) Analysis Suite (http://the_brain.bwh.harvard.edu/PBMAnalysisSuite/indexSep2017.html) python Dependancies: pandas, biopython R Dependancies (packages at: https://cran.r-project.org/): caret (v6), glmnet (v 2.0-13), PRROC (v1.3), aphid (v 1.0.1), seqinr (v3.4.5)
Data analysis	The SR code, and examples, are made available on GitHub (https://github.com/smlmbrt/SimilarityRegression). hmmscan (v3.2.1, part of the HMMER toolkit, http://hmmer.org/) Affinity Regression (https://bitbucket.org/leslielab/affreg) PreMoTF (http://stormo.wustl.edu/PreMoTF/) ZFModels (http://stormo.wustl.edu/ZFModels/) C2H2 SVM Method (http://zf.princeton.edu/)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

New PBM data and motifs are deposited in GEO (accession number: GSE121420), and Cis-BP (http://cisbp.ccbr.utoronto.ca/).

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We successfully 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. 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. For A. nidulans TF 350 novel DBDs were selected for analysis and 180 were successfully cloned into the expression vector (pTH6838) and validated by sequencing. Sample size was not predetermined, but is large enough to test the SR and other prediction algorithms on unseen data from relevant/highly-abundant TF families.
Data exclusions	No data passing our PBM success criteria was excluded from the analysis.
Replication	Successful replication of SR model accuracy was evaluated using the new PBM data, and motif similarity calculated from other motifs included in the human vs. fly analysis.
Randomization	Randomization was not necessary in this study as samples were allocated to different groups (TF similarity) based on uniform rules learned during cross-validation.
Blinding	Blinding was not necessary as SR models are developed automatically using labeled training data based on established thresholds. Performance metrics were calculated without human intervention.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☐	☒ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Unique biological materials

Policy information about availability of materials

Obtaining unique materials	TF clones used for PBMs are available upon direct request to the corresponding author.