1. Introduction

Mammalian transcriptional regulation is achieved through the action of approximately 1500 transcription factors (TFs) [1] that bind to cis-regulatory elements (e.g. promoters, enhancers) found proximal and distal to genes. TFs bind to specific motifs within regulatory elements, but have limited specificity in their DNA-binding preferences resulting in many more motif matches within a genome than actual binding sites for a given factor [2,3]. With this limited motif specificity, it is puzzling how TFs achieve the target specificity necessary to control gene expression [4–7].

A widely held view is that TFs achieve greater target specificity by forming partnerships with other TFs and cofactors. It has been estimated, for instance, that 75% of TFs possibly form dimeric complexes [8]. Once partnered, TF complexes often show preferences for binding longer motifs (composite motifs, or cis-regulatory modules) composed of specific configurations of both binding sites separated by a fixed spacing. As these are less likely to occur randomly in genomic sequence, they offer a means by which binding specificity may arise between TFs and regulatory elements. Composite motifs have been discovered for numerous TF complexes, including FOX:ETS [9], SOX:OCT [10,11] and HOX:PBX [12], a classic example of TF cooperativity. The characteristic spacings within these elements reflect the unique structural mechanisms by which each complex cooperatively binds DNA. For some complexes, these mechanisms have been elucidated via experimentally determined
motifs thus represent evolutionarily conserved and structurally overlapping binding sites that are physically consistent with composite motifs, mostly novel and mostly composed of

inferred their three-dimensional structures and cooperativity functional importance. Structural information was then used to infer evolutionary information to detect signals that are likely to be of non-coding elements (CNEs), and therefore made use of evolutionary abundances of co-occurring binding sites indicative of the binding preferences of TF complexes.

Traditional discovery of TF–TF cooperative interactions has typically involved detailed experimental dissection of one or a small number of promoter sequences [15–17]. Selected cases are further studied by crystallography, which has resulted in a steady but slow growth of TF complexes in the Protein Data Bank (PDB; approx. five unique structures per year, approx. 100 unique structures in total; electronic supplementary material, figure S1). This is probably a tiny fraction of the combinatorial interactions that take place in vivo, a taste of which can be obtained, for example, through mammalian two-hybrid screening [18].

Computational methods have potential for accelerating the discovery of TF complexes and their motif preferences. These methods typically perform large-scale, sequence-based, statistical analyses of regulatory DNA and use position weight matrix (PWM) models of TF binding preferences to scan for co-occurring TF binding sites indicative of cooperative interactions [19,20]. The basis for such methods is that while motif pairs occur randomly in genomic sequence, motif pairs for interacting TFs are expected to occur at a greater rate throughout regulatory regions than that expected by chance, and so can be detected by statistical approaches.

While sequence-based computational methods are capable of generating many hypotheses regarding associations between TF motifs, they do not explore the mechanistic and structural detail of TF–TF cooperativity investigated by structural studies. Nor can they incorporate the physical likelihood of adjacent and overlapping motif configurations for which the necessary data resides outside the base pairs themselves. Combining these two perspectives could not only help avoid physically infeasible sequence-based predictions, but also uncover general structural principles underlying TF complexes that may be difficult to see from isolated examples.

In this work, we have combined a structural perspective on protein–DNA interactions with genomic motif co-occurrence analysis to detect and analyse TF–TF cooperativity on a large scale. This represents the first large-scale, protein structure-aided analysis of non-coding DNA. For 300 TFs, we combined sequence binding preferences in the form of PWMs with three-dimensional data obtained from structures of DNA-bound TFs. We then screened the human genome for co-occurring motifs at specific spacings that are enriched in conserved non-coding elements (CNEs), and therefore made use of evolutionary information to detect signals that are likely to be of functional importance. Structural information was then used to model the physical validity of predicted complexes, and infer their three-dimensional structures and cooperativity mechanisms. Using this approach, we discovered 422 composite motifs, mostly novel and mostly composed of overlapping binding sites that are physically consistent with simultaneous binding of a TF pair to the same bases. Such motifs thus represent evolutionarily conserved and structurally supported TF complex binding sites, and may be key sites for enhancer activity.

Using available ChIP-seq data, we support our predictions and identify specific TF–TF interaction partners. A resource for these predictions at http://bejerano.stanford.edu/complex has potential to accelerate future studies of TF cooperativity and cis-regulatory function.

2. Results

(a) Genomic screen for transcription factor complexes using 3DPWM co-occurrence

We developed a computational approach to identify evolutionarily conserved, physically realistic and genomically abundant spacings of co-occurring binding sites indicative of the binding preferences of TF complexes.

We first paired 223 structural interactions of individual TFs bound to DNA from the PDB with 300 models of sequence binding preferences (PWMs) from Uniprobe, JASPAR, TransFac and literature (the list of PWMs and their sources is available in the electronic resource) [21–23]. This resulted in 300 3DPWMs for 267 vertebrate TFs (figure 1a and see §4) across 22 major DNA-binding domain (DBD) families (figure 1b).

Each 3DPWM includes the per base sequence preferences as well as the per base binding areas in each strand and groove of DNA, resulting in a novel, highly detailed representation of TF binding preferences.

We then considered all TF1 × TF2 × 2 orientations × (−25 bp upstream through overlap region to +25 bp downstream), yielding a total of 6 548 947 rigid arrangements for 45 150 different TF motif pairs (figure 1c and see §4). 3DPWM structural information was used to estimate the physical plausibility of each arrangement. Structurally unrealistic arrangements, which are especially common for highly overlapping motif arrangements, were identified as those requiring excessive (greater than 20 Å²) simultaneous binding of both factors to the same DNA surface (same base in the same groove and strand; figure 1d).

Such arrangements were omitted from further analysis. Arrangements involving simultaneous binding to the same base pair in opposite grooves or strands, however, were allowed. Of all considered arrangements, 5 685 158 (87%) were estimated to be physically feasible, and were subjected to further analysis (figure 1c).

We then searched for evolutionarily conserved matches to the structurally realistic motif arrangements using the PRISM phylogenetic footprinting method (see §4) [24]. We compared the number of matches in putative cis-regulatory DNA (conserved elements) with the number in genomic background DNA (unconserved elements that are at least 500 kb from known gene bodies and not in putative regulatory domains of genes annotated for TF and/or early developmental activity as defined by GREAT [25]), and identified arrangements enriched in conserved regions (binomial test). Arrangements that were statistically abundant in the background were also removed (greatly improving our false discovery rate (FDR); see Discussion). 29 745 arrangements passed the above filters and were found to be significantly and uniquely enriched in conserved non-coding DNA (Bonferroni-corrected p-value ≤ 0.05; see §4 and figure 1c).

Putative TF complex motifs were then constructed for each significant motif arrangement by aligning predicted
Figure 1. Computational approach for prediction of transcription factor (TF) complexes using genomic and structural data. (a) Procedure for derivation of 3DPWMs as outlined using the crystal structure of a forkhead (FOX) DNA-binding domain (DBD) bound to DNA (PDB structure 3C07). A 3DPWM is constructed by mapping structural binding information (DNA-bound surface area per base, strand and groove) calculated from the TF–DNA interface within the crystal structure, and mapping it onto a PWM representing the sequence-binding preferences. See §4 for additional details. (b) The 3DPWM library covers 22 of 25 major DBD families in the human genome. Red bars indicate the size of each DBD family and blue bars indicate the number of TFs for which 3DPWMs were created. (c) Flow diagram outlining the various statistical, structural and genomic filters used to produce a conservative, high-quality set of predictions with low a false discovery rate. See §4 and electronic supplementary material, methods for details on each filter. (d) Structural overlap between aligned 3DPWMs estimates physical clash between TFs. For each possible binding site arrangement, the total binding area shared by both factors is computed in four categories: major groove of strand one (M1), minor groove of strand two (M2), minor groove of strand one (m1) and minor groove of strand two (m2). Some overlapping binding site arrangements (e.g. right arrangement) are physically possible as they require zero (or minimal, see §4) overlap of binding area. Others (e.g. left arrangement) require extensive binding of both factors to the same bases in the same groove and strand, which indicates steric clash. This approach is extended to all pairwise comparisons of 3DPWMs for all arrangements. (e) For all non-clashing pairwise arrangements (cyan triangles mark clashing arrangements) between 3DPWMs, TF binding site co-occurrences are predicted in conserved non-coding elements (CNEs) versus unconserved elements. Genomic statistical outliers (i.e. arrangements with excess abundance in unconserved elements) are removed (magenta crosses). The statistical significance (p-value) of the co-occurrence pattern is then computed using a binomial test. All instances of the predicted binding sites in CNEs that share the identified, significant co-occurrence pattern are aligned, and the alignment is used to generate a TF complex motif representing the binding preferences of the inferred TF complex. (f–h) Some statistically significant predictions produce unlikely complex motifs, which we conservatively prune: (f) an example of complex motifs that are too similar to known TF motifs; (g) an example of a complex motif with unrealistic binding affinity; and (h) motifs that have a very high compositional bias (AT-rich motifs). All of these unlikely motifs are removed. Please refer to electronic supplementary material, methods for additional details and figure S2 for justification of chosen thresholds with respect to known complex motifs.
This procedure defined, predicted motifs using a clustering procedure (see electronic supplementary material, table S3), we grouped together similar preferences yield similar predictions (see electronic supplementary material, figure S3) shows that the results are generally robust to slight parameter changes.

A parameter sensitivity analysis (see electronic supplementary material, figure S2) were largely absent in the curated positive control set (see electronic supplementary material, figure S2).

A high similarity to characterized individual TF motifs. [26] references a crystal structure that is not available in the PDB.

Table 1. The top 25 predicted TF complex motifs in the form of position weight matrices, labelled by their DNA-binding domain family constituents. Seven previously known cases are indicated. The four letter codes following the references are PDB structure identifiers of TF complexes with binding sites that match the associated motifs. [26] references a crystal structure that is not available in the PDB.

<table>
<thead>
<tr>
<th>TF complex motif</th>
<th>complex name</th>
<th>p-value</th>
<th>fold</th>
<th>known dimers</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT...CCAT...G...AAC</td>
<td>RXF→(-5)RXF+</td>
<td>0</td>
<td>37.86</td>
<td>[27] 1DP7</td>
</tr>
<tr>
<td>AACAGGGAA</td>
<td>FOX+(0)ETS+</td>
<td>1.24×10−205</td>
<td>10.88</td>
<td>[9]</td>
</tr>
<tr>
<td>CTTACAAAAATCA</td>
<td>PBX+(−1)HOX9+</td>
<td>2.66×10−122</td>
<td>7.21</td>
<td>[28] 1PUF</td>
</tr>
<tr>
<td>GATATATC</td>
<td>GATA+(0)GFI+</td>
<td>1.23×10−117</td>
<td>5.77</td>
<td></td>
</tr>
<tr>
<td>GTGC...AA...AAC</td>
<td>RXF→(-5)RXF+</td>
<td>2.45×10−102</td>
<td>30.28</td>
<td>[29]</td>
</tr>
<tr>
<td>CA...TG...TTA</td>
<td>BHLH+(4)BARHL−</td>
<td>8.68×10−92</td>
<td>4.31</td>
<td></td>
</tr>
<tr>
<td>T...T...CAT...AAATG</td>
<td>NANO−(3)OCT+</td>
<td>1.19×10−70</td>
<td>4.15</td>
<td></td>
</tr>
<tr>
<td>GAAA...TCATATG</td>
<td>NFAT+(−1)CUX+</td>
<td>7.95×10−62</td>
<td>6.60</td>
<td></td>
</tr>
<tr>
<td>CA...TG...CA...TG</td>
<td>BHLH+(5)BHLH−</td>
<td>5.13×10−53</td>
<td>3.84</td>
<td></td>
</tr>
<tr>
<td>G...ACAGGA</td>
<td>GR+(0)ETS+</td>
<td>3.76×10−71</td>
<td>6.00</td>
<td></td>
</tr>
<tr>
<td>CA...TG...TTTGC</td>
<td>BHLH+(0)NFAT−</td>
<td>1.18×10−60</td>
<td>4.15</td>
<td></td>
</tr>
<tr>
<td>TGGTT...CAG...TG</td>
<td>BHLH+(4)FOX+</td>
<td>2.14×10−62</td>
<td>6.24</td>
<td></td>
</tr>
<tr>
<td>CA...TG...TTT...TTT</td>
<td>BHLH+(0)IRF−</td>
<td>1.28×10−60</td>
<td>5.15</td>
<td></td>
</tr>
<tr>
<td>GGAATG...G...A</td>
<td>ETS+(1)BZIP−</td>
<td>5.18×10−59</td>
<td>4.81</td>
<td></td>
</tr>
<tr>
<td>ACA...TTTTG</td>
<td>SOX+(6)SOX−</td>
<td>2.30×10−58</td>
<td>6.09</td>
<td>[30]</td>
</tr>
<tr>
<td>G...ACA...TTG...C</td>
<td>GR+(3)GR−</td>
<td>2.28×10−57</td>
<td>12.88</td>
<td>[31] 1R4R</td>
</tr>
<tr>
<td>G...AAAATCA</td>
<td>NFAT+(0)PBX−</td>
<td>6.02×10−55</td>
<td>3.70</td>
<td></td>
</tr>
<tr>
<td>AACAAAAATCA</td>
<td>SOX+(−1)PBX−</td>
<td>1.18×10−51</td>
<td>9.11</td>
<td></td>
</tr>
<tr>
<td>CA...TG...TTT...TTT</td>
<td>BHLH+(−1)IRF−</td>
<td>4.41×10−50</td>
<td>6.83</td>
<td></td>
</tr>
<tr>
<td>AATG...TTTTG</td>
<td>NANO+(5)SOX−</td>
<td>1.47×10−49</td>
<td>3.81</td>
<td></td>
</tr>
<tr>
<td>GGAATGAAAA</td>
<td>ETS+(−3)IRF−</td>
<td>1.19×10−48</td>
<td>6.68</td>
<td>[26] structure</td>
</tr>
<tr>
<td>CATTATCA</td>
<td>BZIP−(−4)SOX+</td>
<td>2.31×10−44</td>
<td>5.17</td>
<td></td>
</tr>
<tr>
<td>GGAATTCTA</td>
<td>ETS+(0)SOX+</td>
<td>5.06×10−42</td>
<td>4.35</td>
<td></td>
</tr>
<tr>
<td>AGACAGGGAA</td>
<td>SMAD+(1)ETS+</td>
<td>5.32×10−42</td>
<td>5.29</td>
<td></td>
</tr>
<tr>
<td>TAAT...ATTTCC</td>
<td>HOX+(1)ETS+</td>
<td>1.16×10−41</td>
<td>3.88</td>
<td></td>
</tr>
</tbody>
</table>

binding sites and generating a PWM. In addition to structural clash, three additional methods were used to remove potential artefacts (figure 1f–h). We removed TF complex motifs with a high similarity to characterized individual TF motifs (see electronic supplementary material, figure S2b), extremely low nucleotide variation per base (see electronic supplementary material, figure S2c) and high compositional bias (see electronic supplementary material, figure S2d) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2e) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2f) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2g) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2h) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2i) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2j) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2k) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2l) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2m) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2n) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2o) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2p) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2q) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2r) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2s) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2t) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2u) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2v) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2w) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2x) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2y) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S2z) because these properties were largely absent in the curated positive control set (see electronic supplementary material, figure S3) shows that the results are generally robust to slight parameter changes.

Finally, as related TFs with similar DNA-binding preferences yield similar predictions (see electronic supplementary material, table S3), we grouped together similar predicted motifs using a clustering procedure (see electronic supplementary material, methods). This procedure defined, for each predicted complex, a pattern represented by (TF1[+/−]/ spacer:TF2[+/−]). Here, TF1 and TF2 correspond to specific DBD families or subfamilies, the [+/−] correspond to the relative orientation of the two motifs and (spacer) corresponds to the number of bases between the two motifs.

Even following removal of structural clashes and potential sequence artefacts, and motif clustering, a final set of 422 highly significant predictions remained. These are 422 predicted TF complex binding motifs that are significantly enriched in CNEs and encode binding sites for physically realistic TF–TF complexes. The predictions represent a relatively small fraction of the total search space (1.3%), conservatively estimated at 33 469 clustered arrangements (see §4).

Shown in table 1 are the top-scoring predictions ranked by p-value and named with respect to the consensus binding motifs of the TF components (see electronic supplementary material, table S4).
lexes were predicted (tables S1 and S2) jointly, 18/38 (47%) of known TF complexes (see electronic supplementary material, detectable through analysis of overlapping structural binding mechanisms is not only permitted by the method, but is also sides of DNA [29]. As explored in a later section, this structural mechanism is accounted for, because energy minimization and rotamer adjustment significantly reduced or completely eliminated clashes in all initially clashing models (data not shown).

## Table 2

<table>
<thead>
<tr>
<th>set</th>
<th>no. found</th>
<th>p-value</th>
<th>fold</th>
</tr>
</thead>
<tbody>
<tr>
<td>known TF complex (total: 14)</td>
<td>9</td>
<td>1.40 × 10⁻¹⁴</td>
<td>50.99</td>
</tr>
<tr>
<td>known TF complex structures (total: 24)</td>
<td>9</td>
<td>8.18 × 10⁻¹²</td>
<td>29.74</td>
</tr>
<tr>
<td>combined set (total: 38)</td>
<td>18</td>
<td>1.20 × 10⁻²⁴</td>
<td>37.57</td>
</tr>
</tbody>
</table>

(b) Overlap of predicted transcription factor complexes with known complexes

To evaluate the quality of the CNE-enriched predictions, we compared them with positive control sets of known TF complex motifs and TF–TF binding site arrangements in crystal structures whose binding sites had high-scoring matches to TF motifs in our PWM library. We also assessed their overlap with TF–TF interaction data, and computed FDRs.

First, the predictions contain a considerable enrichment of known TF complex motifs. Out of a curated set of 14 TF complex motifs identified from the literature (see §4 and electronic supplementary material, table S1), nine are present within the set of 422 predictions with identical spacings, orientations, and TF partners (table 2; p = 1.4 × 10⁻¹⁴; 51-fold enrichment; see §4). Correctly identified composite motifs for SOX : OCT, HOX : PBX and FOX : ETS complexes, for instance, can be found within the top 25 predictions (table 1).

In addition to the overlap with known TF complex motifs, we found a significant association between the identified CNE-enriched binding site spacings and those observed in available crystal structures of TF–TF–DNA complexes. Out of a dataset of 24 binding site configurations extracted from structures of TF–TF–DNA interactions derived from the PDB (see electronic supplementary material, table S2), nine were detected among the 422 predictions (table 2; p = 8.18 × 10⁻¹², 30-fold enrichment, see §4). For example, the predicted RFX : RFX spacing and motif (the top-ranked prediction) is identical to that observed in a crystal structure of two RFX domains binding to an overlapping element on opposite sides of DNA [29]. As explored in a later section, this structural mechanism is not only permitted by the method, but is also detectable through analysis of overlapping structural binding footprints contained within 3DPWMs.

Finally, when considering the two independently curated sets of known complexes (see electronic supplementary material, tables S1 and S2) jointly, 18/38 (47%) of known TF complexes were predicted (p = 1.2 × 10⁻²⁵; 38-fold enrichment; see §4). Thus, overall, the method has a significant ability to detect known cooperative TF–DNA interactions.

For each predicted TF complex motif, we then chose the top-scoring individual motif pair as a representative of the group, and asked whether the specific pair was observed to physically interact using mammalian two-hybrid interaction data (Fantom and Corum) [18,32]. Among the 422 predicted complexes, 36 top-scoring TF pairs were found to interact physically (see electronic supplementary material). This is highly significant (p < 0.001) as 1000 datasets of 422 random TF pairs did not yield a single two-hybrid match.

(c) False discovery rate of transcription factor complex prediction

To estimate the random probability of obtaining significant predictions through the screen, we shuffled the sequence binding preferences of each 3DPWM and repeated the entire pipeline to obtain a null distribution of p-values (see §4). The null distribution of p-values was used to estimate a harsh FDR for each predicted complex motif, yielding an average FDR of 18% for our entire set of 422 predictions.

(d) Structural modelling and analysis of predicted TF–TF–DNA complexes

Given the significant detected relationship between genomic motif co-occurrence and TF–TF structural data, we used our computational framework to analyse general structural features among predicted complexes and infer three-dimensional structures.

Structural information in the form of 3DPWMs was first used to assess the proximity of TFs to their partners on DNA for all 422 complexes. This involved estimating the number of base pairs each TF would have to move before clashing with its partner (see §4). This calculation revealed a strong tendency for TFs to co-bind as close as physically plausible, with most predictions occurring within a few base pairs of the predicted clash borders (figure 2a). Consistent with this finding, a majority of the predicted TF complex motifs contain some sequence overlap between binding sites (326, 77%). These results suggest tightly packed interactions between DBDs on DNA, as observed previously in the interferon enhanceosome and other cooperative TF complexes [13,29]. Random-shuffled motifs predicted fewer close interactions suggesting that this finding is not due to a bias towards short inter-motif spacings (figure 2a). In addition, there is no significant difference in median FDR between complexes with/without overlapping motifs (see electronic supplementary material, figure S4a).

To study the predicted complexes in greater structural detail, we generated three-dimensional models of predicted complexes by superimposing pairs of DNA-bound TFs according to the genomically detected TF binding site arrangements (figure 2b–d and electronic supplementary material, figure S5). Structural models (326) were generated using a novel automated approach (see §4), and selected cases were refined manually. While TFs may clash at their footprints on DNA (detected by 3DPWM overlap), they may also clash at regions outside of the DNA-binding interface. Thus, we further analysed the three-dimensional models to measure the degree of physical clash between TF structures, identifying only nine (2.8%) with an excessive amount of clash (see §4). However, even these may be structurally viable if flexibility (e.g. DNA bending, TF conformational changes) is accounted for, because energy minimization and rotamer adjustment significantly reduced or completely eliminated clashes in all initially clashing models (data not shown).
(e) Structural mechanisms of transcription factor interaction in transcription factor complexes

From these three-dimensional models, it was then possible to infer structural mechanisms potentially underlying TF–TF cooperativity. Three recurring mechanisms (figure 2b–d and electronic supplementary material, figure S5) of combinatorial interactions were identified: (i) direct protein–protein interactions between DBDs; (ii) the so-called through-DNA effects [27] in which two DBDs may synergize by binding overlapping sites in DNA (same base pair in an opposite groove or strand); and (iii) indirect interactions possibly mediated by cofactors. All three mechanisms were computationally evaluated within all 326 models.

Forty per cent of the models were found to exhibit direct protein–protein interactions between DBDs, identified as atomic contacts less than 5 Å. We then used the 3DPWM information to identify which predictions are consistent with through-DNA synergy, and conservatively defined through-DNA synergy as overlap of protein–DNA contacts (i.e. both TFs bind the same DNA base pair but in opposite grooves or strands) in four or more base pairs. Quite strikingly, 38% of the predicted complexes were found to contain four or more bases of through-DNA synergy. Concordantly, through-DNA interactions were also detected in 28% of positive controls. Direct and through-DNA interactions were not mutually exclusive, however, as 19% of all models exhibited both types of synergy mechanisms. The remaining set of 41% displayed no evidence of direct protein–protein or through-DNA interactions, but rather may participate in indirect interactions mediated by cofactors. The median FDR of motifs associated with through-DNA synergy (15%) did not differ greatly from non-synergy motifs (19%; electronic supplementary material, figure S4b), which suggests that the ratios are not due to an inflated number of false predictions in any particularly category.

(f) Promiscuous partnering across transcription factor DNA-binding domain families

For combinatorial regulation to take place, TFs must partner with a range of additional TFs and cofactors, a phenomenon seen in recent TF–TF interactome studies [18]. While the mechanisms of promiscuous partnering are largely unclear, sequence and structural insights into this phenomenon can be obtained from our predictions.

Of 253 possible partnerships between 22 DBD families (at any spacing or orientation, including same-DBD-pairings), we detected 103 (41%), most of which are novel. This is
visualized in a DBD family interaction matrix shown in figure 3a, which reflects a considerable degree of promiscuous partnering between TF–TF families. One such family known to undergo promiscuous partnering is the MH1 family composed of SMAD factors. SMAD factors exhibit weak DNA-binding on their own and typically bind DNA in conjunction with a variety of TF partners [33], but the exact binding mechanisms are unclear.

Figure 3b shows the top seven predicted SMAD:cofactor complexes. The inferred three-dimensional structures of genomically predicted TF complexes show a single TF family forming multiple partnerships through diverse interaction interfaces and mechanisms. Direct interactions involving diverse protein–protein interfaces and bound positions on DNA are evident for 6/7 of the SMAD complexes. For instance, despite an apparent wide separation between binding sites for the ETS + (7)SMAD + complex, the two TFs interact in a side-by-side fashion along the same face of DNA. A through-DNA interaction is evident for the predicted SMAD + {1}ETS + complex, which exhibits opposite strand binding within the same major groove (figure 3b).

Figure 3. Detected interactions between DNA-binding domain families and selected structural models of SMAD complexes. (a) All detected interactions between DBD families visualized as an interaction matrix. Light blue squares indicate novel interactions in the set of 422 predictions, dark blue squares indicate predicted interactions that are verified by known literature, grey squares indicate interactions verified by known literature that are not in our set of 422 predictions. (b) The SMAD TFs containing the MH1 domain were found to display promiscuity of TF–TF interactions. According to structural modelling, SMADs interact with other domains either via a range of protein–protein interaction sites between DNA binding domains or via through-DNA effects. See electronic supplementary material, figure S5 for examples of different classes of TF–TF interactions. Note the two very different SMAD:ETS configurations.

Figure 4. Validation of predicted TF complexes and identification of specific TF partners using ChIP-seq data. The number of validated complexes correctly predicted is highly significant compared with a null model obtained from analysis of 10 000 random complexes.

(g) Validation of transcription factor complexes and identification of specific transcription factor partners using ChIP-seq data

While the TF complexes detected within conserved non-coding elements are significantly enriched in known cases, we sought to identify further evidence to support novel predictions. Additional data may also be required in some cases to discern the specific constituents of a predicted TF complex as motifs can be highly similar within a given TF family. We therefore applied our co-occurrence analysis to publicly available ChIP-seq data (see electronic supplementary material, table S5) to identify TF complex motifs uniquely enriched in genomic regions mutually bound by two factors (see §4). In total, we evaluated pairwise intersections of ChIP datasets for 52 human and 31 mouse TFs in the same context and found 19/51 (37%) ‘validatable’ TF complex motifs (7/12 known, 12/39 novel) were significantly enriched (p < 1 × 10⁻⁵). Simulated sets of random TF spacings only yielded a maximum of 8/51 (16%) validations over 10 000 iterations (see §4 and figure 4)
experimental studies of TF combinatorial regulation. Such motifs are invaluable starting points for future erythroleukaemia cells (see electronic supplementary material, ETSþf).embryonic stem cell datasets for OCT4 and NANOG, and a
motif for a SPI1:FOS complex in K562 erythroleukaemia cells (see electronic supplementary material, figure S6). Inferred three-dimensional structure of two TCF3 dimers bound to the predicted motif in (b). The structure is based on PDB ID 2QL2 (E47/NEUROD1).

Figure 5. Detection of a BHLH+{5}BHLH+ complex in genomic regions bound by Tcf3 (BHLH) in mouse T-cell ChIP-seq data. (a) Motif co-occurrence plots for genomic predictions (above) and Tcf3-bound genomic regions from mouse T-cell ChIP-seq. (b) Composite motif discovered through the initial screen corresponding to the top enriched arrangements observed in (a). (c) Inferred three-dimensional structure of two TCF3 dimers bound to the predicted motif in (b). The structure is based on PDB ID 2QL2 (E47/NEUROD1).

To illustrate how ChIP-seq data may be used to help disambiguate constituents of predicted TF complexes, we outline the following example. The initial genomic screen predicted a BHLH+{5}BHLH+ motif. Further examination of ChIP-seq data for TCF3 (a bHLH factor) in mouse T-cells suggests that this motif is a potential binding site for two adjacent TCF3 dimers (figure 5). The genomic and ChIP-seq motif co-occurrence patterns match (arrangement is most enriched in both cases; figure 5a,b). Furthermore, the inferred three-dimensional model (figure 5c) reveals a potentially cooperative complex that may involve side-by-side protein interactions with further DNA bending. This motif arrangement was also significantly enriched in overlapping mouse B-lymphocyte ChIP-seq data for Tcf3 and Max factors (see electronic supplementary material, figure S6). Thus, it may be a more general mechanism for DNA-mediated cooperativity between pairs of bHLH dimers.

Numerous additional TF complexes discovered within specific ChIP-seq datasets are functionally intriguing. These include a novel NANOG–{3}OCT+ motif enriched in embryonic stem cell datasets for OCT4 and NANOG, and a ETS+{1}BZIP+ motif for a SPI1:FOS complex in K562 erythroleukaemia cells (see electronic supplementary material, figure S6). Such motifs are invaluable starting points for future experimental studies of TF combinatorial regulation.

(h) Distal and proximal associated occurrence of transcription factor complexes
Enhancers and promoters may preferentially bind different regulatory factors owing to their distinct requirements as distal and proximal regulatory elements. To investigate the potential role of TF complex binding sites in these different classes of regulatory sequences, we analysed our dataset for putative enhancer-associated and promoter-associated TF complexes.

Proximal gene promoters were defined as regions 500 bp upstream of all (alternative) gene transcription start sites. By design, only a tiny fraction (0.84%) of CNEs used in our genomic screen overlap these promoters. Thus, the 422 complex predictions we obtain are representative of more distal cis-regulatory elements, such as enhancers. We can however use genome-wide prediction to assess the abundance of these predicted complexes in proximal regions. Note that to assure their quality our predictions continue to rely on cross species conservation, resulting in over twice as many distal bases overlapping binding site predictions than proximal promoter bases.

With the above caveats in mind, the top five most abundant distal complexes that were not observed in proximal promoters were BHLH +{4}NANOG–, BHLH +{6}HOX9–, BHLH–{4}PBX+, HOX +{3}ETS+ and SOX +{1}PBX–. Interestingly, this short list is dominated by bHLH TFs, which have been shown to play a role in DNA bending and may facilitate DNA-looping by enhancers [34]. Indeed, the bHLH factors are known to pair with various known developmental homeodomain proteins [35–37] through which they may allow enhancers precise spatio-temporal control during development [38]. Conversely, the top five most abundant complexes in promoters were ETS +{2}ETS+, RXF –{4}RXF+, BZIP +{1}ETS–, FOX +{0}ETS+ and ETS +{1}NR–. In this case, the complexes abundant in promoters appear to be dominated by ETS factors. This finding is in agreement with previous studies reporting the presence of consensus ETS binding motifs in proximal promoters of housekeeping genes with ‘redundant’ binding of multiple ETS factors [39,40].

(i) Complexes web resource
Our predictions have been made freely available via a web resource at http://bejerano.stanford.edu/complex. The resource provides access to the predicted TF complex motifs, their FDRs and suggestions of potential TF partners and binding sites, which can be easily assessed for gene targets and functional enrichment computed via the PRISM framework (see electronic supplementary material, figure S7) [24,25]. We hope that the predictions available in this resource will accelerate future structural and genomic studies of TF cooperativity.

3. Discussion
We developed an approach to detect TF complexes by analysis of conserved non-coding elements in the human genome. The method combines sequence, structural, and evolutionary information to predict TF complexes, their binding preferences, genomic binding sites and structural mechanisms. The predictions are significantly enriched in known TF complex motifs, spacings observed in crystal structures of TF–DNA complexes, and protein–protein interaction
The developed approach stands to identify combinatorial interactions involved in biological processes that are under selective pressure and are genomically abundant in a high-throughput manner. This is complementary to traditional, gene-centric studies of single promoter or regulatory sequences from which even rare TF combinatorial interactions are discovered, though more slowly and serendipitously. On the one hand, our method requires abundant genome-wide signals and will certainly miss complexes used more sparingly. This is illustrated by the 13/20 known complexes we were not able to predict that were not enriched in CNEs (see electronic supplementary material, tables S1 and S2). On the other hand, the rapidity and extent of the confident complex predictions we make cannot be matched using comparable amounts of time and money. Again, a combined approach is probably most valuable, using our screen as a guide to discover many (but by no means all) novel complexes.

In addition to detecting TF complexes and binding preferences, the genomic approach used in this study was useful in surveying structural principles underlying TF cooperativity by observing the DNA binding footprints of complexes preserved by evolution. Through three-dimensional modelling, we detected evidence of direct protein–protein and through-DNA interactions between DBDs, many of which are novel interaction mechanisms. Further interactions probably mediated through additional domains or cofactors that do not necessarily bind DNA themselves or facilitated through DNA bending [46] could also be seen. In reality, a complex may use a combination of direct, through-DNA and indirect interactions to achieve synergistic binding. Of particular interest is that the approach re-discovered atomic-resolution structural details identified in previous small-scale studies, providing broader genome-wide support. For instance, the distribution of spatial proximity between factors and the detection of TF cooperativity mediated through DNA and not solely through protein–protein interactions mirrors that seen in the highly studied interferon enhanceosome [13,14] and other ternary protein–DNA complexes [26,27]. The prevalence of detected through-DNA synergy and direct protein–protein interactions suggests that enhanceosome-like structures are far more widespread throughout regulatory sequences than currently known based on the limited set of structures available in the PDB.

While a more flexible mode of combinatorial regulation may well exist at larger inter-motif distances [47], our analysis provides considerable evidence that enhanceosome-like binding is widespread in regulatory regions, and represents an enriched feature of conserved non-coding elements. Additionally, although not a focus of this study, higher-order structures can be predicted in principle using the same approach, by iteratively performing co-occurrence searches based on the discovered complex motifs. This is supported by our discovery of tetrameric complexes (e.g. P53 tetramer) based on dimeric motifs that were included in our initial motif dataset. Extension of the method to discovery of higher-order structures will be a focus of future work.

Ultimately, our predictions provide approximately 400 novel complexes, and thousands of binding sites that are starting points for future studies of both functional and structural aspects of combinatorial regulation. For example, our predictions may be fed to PRISM to infer TF complex functions [24]. Predictions of our 422 TF complexes, their motifs, potential TF partners and genomic binding sites are publicly accessible at http://bejerano.stanford.edu/complex.
4. Material and methods

(a) Structural analysis of co-occurring binding sites using structure-annotated position weight matrices (3DPWMs)

(i) Constructing 3DPWMs and defining structural clash

We obtained 300 PWMS describing the binding preferences of 267 TFs from UniProbe, TransFac, JASPAR and the literature (the list of PWMS and their sources is available in the electronic resource [21–23]). TF-DNA complexes (223) matching these TFs were derived from 172 total three-dimensional structures obtained from the PDB. Using surface area calculations [48], we measured the TF-bound, per base surface area in each strand and groove, resulting in a four-component vector, which was aligned to the PWM to generate a 3DPWM (see below for details). This resulted in a dataset of 300 structure-annotated PWMS (3DPWMs; figure 1a).

Because numerous known TF complex motifs involve overlapping binding sites, it is important to determine which arrangements of overlapping binding sites are and are not physically realistic. The clash score C for a given PWM arrangement j is therefore computed as the sum of overlapping surface area (same base per groove) across the aligned 3DPWMs (figure 1d). Because a single 3DPWM was allowed to have DNA footprint data from several PDB structures, the final clash score was defined conservatively as the maximum clash for all PDB structure pairings.

(ii) Combining PWMS (one-dimensional) with structural data (three-dimensional) to create 3DPWMS

PDB structures (172) were split into 223 chains containing separate TF–DNA interactions. Chains for obligate dimers (e.g. bHLHs, leucine-zippers) were merged together to be consistent with the binding preferences observed in their respective PWMS. For each complex, a structural bioinformatics approach implementing a Voronoi algorithm [48] was used to measure the solvent-accessible surface area of each atom in the double stranded DNA alone and, separately, the TF–DNA complex. The per-atom difference in solvent-accessible surface area was then used to calculate, for each base pair position:

- total surface area ($\tilde{A}^2$) bound in strand 1, major groove (M1 in figure 1, and elsewhere);
- total surface area ($\tilde{A}^2$) bound in strand 1, minor groove (m1);
- total surface area ($\tilde{A}^2$) bound in strand 2, major groove (M2); and
- total surface area ($\tilde{A}^2$) bound in strand 2, minor groove (m2).

The major and minor groove atom definitions were based on those used in the PDA software [49].

The four-component vectors (collectively referred to as the three-dimensional footprint) defined above were mapped to the DNA sequence extracted from the PDB structure. An ‘anchor’ PWM was aligned to the TF-bound sequence segment, and the maximum-scoring alignment of the forward and reverse-complement alignments was kept. A PWM was allowed to be an ‘anchor’ for several PDB structures. Only high-scoring anchor alignments (greater than or equal to 0.5) were used, and each was verified manually. PWMS of homologous TFs were used if direct PWM mappings were not available. Related PWMS from the same DBD family were also aligned to the anchor PWM and kept if the alignment score was greater than or equal to 0.6, and multiple anchor PWMS were allowed to map to related PWMS if they satisfied the alignment score. This resulted in a dataset of 300 structure-annotated PWMS (3DPWMS) consisting of three-dimensional TF–DNA complexes matched to one-dimensional binding preferences (PWM). The 300 3DPWMS cover 22 DBD families and 267 mammalian TFs. There are more 3DPWMS than TFs owing to some TFs having multiple one-dimensional sequence binding preferences.

(iii) Definition of clashing and non-clashing 3DPWM arrangements

For any possible arrangement of a pair of 3DPWMs, the structural data can be used to infer the amount of binding area overlap in the major and minor grooves for each strand. Because TFs can bind to the same base pair on different strands or in different grooves [27,50], but will probably occupy the same physical space (clash) if they bind the same base in the same strand/groove, we penalize the latter case. The clash score C for a given PWM arrangement j is therefore computed as the sum of overlapping surface area (same base per groove) across the aligned 3DPWMS in the following way:

$$C_{ij}^{A,B} = \sum_{i=1}^{n} \min (M1_i^A, M1_i^B) + \min (M2_i^A, M2_i^B) + \min (m1_i^A, m1_i^B) + \min (m2_i^A, m2_i^B),$$

where $C_{ij}^{A,B}$ is clash score for arrangement j using PDB structure A and B, i is alignment position, A is the DNA footprint for first TF from PDB structure A, B is the DNA footprint for second TF from PDB structure B, M1 is the binding area ($\tilde{A}^2$) in major groove, strand 1, M2 is the binding area ($\tilde{A}^2$) in major groove, strand 2, m1 is the binding area ($\tilde{A}^2$) in minor groove, strand 1, m2 is the binding area ($\tilde{A}^2$) in minor groove, strand 2. Because a single 3DPWM was allowed to have DNA footprint data from several PDB structures, the final clash score was defined conservatively as:

$$\zeta_j = \max_{A,B} C_{ij}^{A,B}.$$

In the above equation, A and B represent all combinations of DNA footprints being used to represent the two 3DPWMS. The maximum clash value was used to produce the most conservative predictions.

It is believed that statistically significant, but structurally clashing predictions of overlapping 3DPWMS occur owing to similarity between the one-dimensional sequence binding preferences of TFs and do not represent true complexes. Our method allows a particular arrangement of two 3DPWMS if there is minimal clash score $\zeta_j < 20 \text{Å}^2$ total clash since a value of 20 $\text{Å}^2$ distinguishes known complexes from the full distribution of scores and, in principle, this value allowed some degree of structural rearrangement of the TF partners upon binding as a complex, as well as structural variation in binding between homologous DBDs. Specifically, 863 789 (approx. 13.2%) potential complexes (figure 1c) had scores exceeding this value, but only 1 (2.6%) of the positive controls (see electronic supplementary material, figure S2a) had scores exceeding this value. The filter was even stronger for the final set of predictions (post-clustering), where 36% of predictions (none of which were positive controls) would have had scores exceeding this value were this filter not used.

(iv) Construction and analysis of three-dimensional structural models

A set of representative structures was selected for each DBD family based on 3DPWM mappings. To construct a TF–TF–DNA complex model, each individual TF–DNA structure was aligned via the DNA molecule to a linear template DNA structure fragment along the binding site in the specified orientation using PyMol. For manually selected cases shown in figures, we manually aligned each case using the DNA available in the original structure to better preserve its characteristics (e.g. angle of bending). Protein–protein interactions were identified between DBDs by detecting atom–atom contacts within 5 $\text{Å}$. Physically clashing
models were identified as defined by Tress et al. [51] as those containing more than four severe clashes (alpha C–C distances less than 1.9 Å) or more than 50 bumps (alpha C–C distances between 1.9 and 3.6 Å).

(b) Defining genomic datasets
The set of CNEs was built in a similar scheme to that described by Lowe et al. [52]. We combined five sources of CNEs for the human reference assembly hg18, including insertion and deletion resistant elements [53], phastCons placental mammal and vertebrate most conserved tracks [54], regions conserved in five near species (chimpanzee, macaque, mouse, rat, dog) by Multiz multiple alignment and regions conserved in at least two distant species (chicken, zebra finch, lizard, frog, tetraodon, fugu, medaka, stickleback, zebrasiped and lamprey) by UCSC reciprocal best chains. To ensure quality of the set as true CNEs, we removed all regions that were in simple repeats, segmental duplications or 100 bp up or downstream of potential coding exons and required the regions to be at least 50 bases long and syntenic to human and mouse (see electronic supplementary material for detailed UCSC track names).

For the set of unconserved likely non-functional elements, we selected genomic regions not including seven sources of potentially functional interesting regions: genes, potential regulatory regions of developmental genes, conserved elements, subtelomeric, pericentromeric regions, assembly gaps and chromosomes X, Y and M (see electronic supplementary material for detailed UCSC track names).

(c) Identification of conserved binding sites
The binding sites used to infer TF complexes were derived using the phylogenetic footprinting method developed by the PRISM approach (figure 1 in [24]). In brief, PRISM identifies conserved binding sites and evaluates the statistical significance of the observed conservation score by its comparison with an empirical distribution of conservation scores for shuffled motifs in similarly conserved genomic neighbourhoods.

(d) Performing the genomic screen
For each pair of 3DPWMs, we identified conserved binding sites using PRISM [24] in all sequences in the set of CNEs. For each pair of 3DPWMs, we analysed the overlapping region (corresponding to the length of the two 3DPWMs) and 0–25 bases up and downstream of the overlapping region for both relative orientations. When analysing co-occurring motifs for the same factor in the same relative orientation, only the overlapping region and 0–25 bases downstream of this were analysed to avoid double counting.

For all non-clashing arrangements of 3DPWM pairs (figure 1d), we assessed whether the identified spacing arrangement of binding sites is also conserved by evolution in multiple species using the UCSC 44 way alignment (hg18.multiz44way). We required presence of the motif pair at the same arrangement within 20 bases of the orthologous region in at least five species in the eutherian mammal subset of the UCSC 44 way alignment (hg18.multiz44way), with a total branch length of at least two substitutions per site. We performed the equivalent predictions in unconserved elements without the conservation constraints.

Because the unconserved elements are used as a null model for assessing significance of the observed counts in the set of CNEs, we also aimed to avoid all arrangements that were potentially enriched (statistical outliers) in the set of unconserved elements (figure 1e). Outliers were detected by computing the median of counts at all arrangements, and masking arrangements for which the observed counts were greater than or equal to three times the inter-quartile range from the upper quartile (greater than or equal to 3 × IQR + Q3) in the unconserved regions [55].

The statistical significance of a potential TF complex pairing with spacer ‘j’ is calculated (figure 1e,f) using the binomial test with the following parameters:

(i) \( n \) —total number of non-overlapping instances of a TF complex pairing with spacer ‘j’ in CNEs;
(ii) \( N \) —the total number of non-overlapping instances of a TF complex pairing for all spacers in CNEs;
(iii) \( p_j \) —probability of observing a TF complex pairing with spacer ‘j’ in the unconserved regions (using \( n/N \) in unconserved elements).

The test calculates the p-value (\( P_j \)) of a TF complex spacer ‘j’ as the probability of observing ‘n’ or more TF complex pairings for spacer ‘j’ give there are ‘N’ total co-occurrences of the TF complex pairing for all spacers.

\[ P_j = \sum_{i=n}^{N} p_j^i (1 - p_j)^{N-i}. \]

A p-value of 1 × 10⁻⁶ was used as a significance cut-off, as the Bonferroni-corrected p-value for \( \alpha = 0.05 \), with over five million total evaluated patterns. We constructed PWMs for each discovered TF complex motif by aligning all instances of the complex identified in CNEs and calculating base frequencies for each column of the alignment using seqLogo [56] (figure 1c). Additionally, we filtered TF complex motifs that are too similar to known TF motifs, had low nucleotide variation (K-mer entropy) or were compositionally biased (low Watson–Crick entropy) to avoid prediction artefacts (see electronic supplementary material, methods and figure S2b–d).

(e) Assessing enrichment of Protein Data Bank structures and literature curated motifs in predictions
A set of 14 TF complex motifs was retrieved from the literature. A second set of 24 TF–TF–DNA structures and their patterns was constructed from the PDB and one published structure absent from the PDB [26]. For each motif, we required that the observed TF complex motif arrangement (orientation and spacer) be supported by either a TF complex structure or multiple, validated binding sequences.

For all predicted TF complex motif clusters, the top-scoring motif was chosen as the representative TF complex motif arrangement. The number of representative TF complex motif arrangements matching those defined in the two control sets above was counted. A highly conservative null model was used, in which the pattern preferences are assumed for a given DBD family. This is not always the case and is an underestimate of the true number of possible patterns. Counting overlapping spacings and those ± 25 bp flanking the two shortest motifs per domain family (because they have the smallest overlap region) resulted in a lower bound estimate of 33 469 possible TF complex arrangements. Using this null model, we calculated the p-value for the observed number of TF complex motifs matching those in the control sets using the hypergeometric statistic.

(f) Assessing the false discovery rate of the identified transcription factor complex motifs
The FDR of identified TF complex motifs was obtained by creating a library of 3DPWMs with their one-dimensional sequence binding affinity shuffled (see electronic supplementary material for details on shuffling approach). The shuffled library was
used to predict false arrangements that achieved statistical significance using the same pipeline as for the real motifs. The FDR for a given p-value cut-off was calculated as:

\[ \text{no. of significant complexes identified using the shuffled library} / \text{no. of complexes identified using the real library} \]

(g) Assessing significance of validating factors using intersections of ChIP datasets in the same context

For a pair of ChIP sets, we extracted all regions in the genome that are overlapped by both sets, and passed the regions to 100 bases from the middle of each overlapping region. TF binding sites for the two assayed factors were then predicted within these regions requiring a minimum PWM sequence score of 0.8 (matching at least 80% of the information content). We then screened for statistically significant TF complex motif arrangements using 3DPWMs and all other parameters as described in §4d.

As a null comparison, for each significant TF complex prediction obtained from the screen, we sampled a random motif spacing with the same TF partners and orientation. We sampled only spaces that were not statistically significant in the standard screen or screen without clash and outlier filters to avoid true positives in our null distribution. To avoid over-counting of redundant patterns, the resampled dataset was made to consist of only unique TF1\[+/-\]/TF2\[+/-\] patterns. The number of total ChIP-validated TF spacings was then counted. The maximum number of ChIP-validated cases found in 10 000 iterations was 8. Because we achieve a validation rate of 19, our validated results are significant with p-value \( < 1 \times 10^{-5} \).

Acknowledgements. We thank Romain Madelaine, Philippe Mourrain, Omri Amirav-Dory and Axel Brunger for their technical assistance with different aspects of the project and Terence Capellini for critical evaluation of our manuscript. Finally, we thank members of the Bejerano Laboratory (particularly Shoa Clarke, Geetu Tuteja and Jim Notwell) for insightful discussions and help. The authors declare that they have no conflict of interest. H.G., A.C.D. and G.B. developed the genome-wide TF complex detection method. H.G. focused on genomic sequence analyses and A.C.D. on structural analyses. H.G. and A.C.D. co-wrote the manuscript. A.M.W. implemented the conserved binding site prediction algorithm and provided in-depth feedback on the project and the manuscript. G.B. supervised and guided the study and manuscript preparation. All authors read and approved the final manuscript.

Funding statement. H.G. is supported by National Science Foundation Fellowship DGE-1147470. A.C.D. is supported by a Natural Sciences and Engineering Research Council of Canada Postdoctoral Fellowship (PDF). A.M.W. is supported by a Bio-X Stanford Interdisciplinary Graduate Fellowship. G.B. is supported by NIH grants R01HG005058 and R01HD059862 and KAUST. G.B. is a Packard Fellow and Microsoft Research Fellow. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References


15. Wadman IA, Osada H, Gritz GG, Aguillnick AD, Westphal H, Forster A, Rabbits TH. 1997 The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NIK proteins. EMBO J. 16, 3145 – 3157. (doi:10.1093/emboj/16.11.3145)


53. Natella M. 2010 NIST/SEMATECH e-handbook of statistical methods. NIST/SEMATECH. 25(01)00336-7)