LETTER

doi:10.1038/nature15518

DNA-dependent formation of transcription factor pairs alters their binding specificity

Arttu Jolma[1], Yimeng Yin[1], Kazuhiro R. Nitta[1], Kashyap Dave[1], Alexander Popov[2], Minna Taipale[1], Martin Enge[1], Teemu Kivioja[3], Ekaterina Morgunova[1] & Jussi Taipale[1,3]

Gene expression is regulated by transcription factors (TFs), proteins that recognize short DNA sequence motifs[1–3]. Such sequences are very common in the human genome, and an important determinant of the specificity of gene expression is the cooperative binding of multiple TFs to closely located motifs[4–6]. However, interactions between DNA-bound TFs have not been systematically characterized. To identify TF pairs that bind cooperatively to DNA, and to characterize their spacing and orientation preferences, we have performed consecutive affinity-purification systematic evolution of ligands by exponential enrichment (CAP-SELEX) analysis of 9,400 TF–TF–DNA interactions. This analysis revealed 315 TF–TF interactions recognizing 618 heterodimeric motifs, most of which have not been previously described. The observed cooperativity occurred promiscuously between TFs from diverse structural families. Structural analysis of the TF pairs, including a novel crystal structure of MEIS1 and DLX3 bound to their identified recognition site, revealed that the interactions between the TFs were predominantly mediated by DNA. Most TF pair sites identified involved a large overlap between individual TF recognition motifs, and resulted in recognition of composite sites that were markedly different from the individual TF's motifs. Together, our results indicate that the DNA molecule commonly plays an active role in cooperative interactions that define the gene regulatory lexicon.

The set of rules by which a DNA sequence can be converted into knowledge of spatial and temporal expression patterns of a protein has been difficult to decipher[1–3]. This is in part because, in mammals, more than 1,000 TFs recognizing over 200 different short DNA motifs participate in interpreting gene regulatory information[7–10]. In addition, TFs also interact with each other, and many TFs bind DNA as homo- or heterodimers. A pair of TFs can bind to multiple different DNA motifs, as the recognition sites of individual TFs can occur in different orientations and/or spacings relative to each other[11–13]. Most known heterodimeric interactions occur between two TFs of the same structural family, but several cases where TFs of different structural classes bind cooperatively have also been identified[4,6].

To chart the prevalence of co-operative interactions between TFs in the presence of DNA, we developed a novel method, CAP-SELEX, in which specific DNA sequences that interact with two different TFs at the same time are selected from a library of random sequences. CAP-SELEX only detects a specific type of interaction involving three macromolecules, where the two tested proteins both bind to the same DNA in a sequence-specific manner (Fig. 1a). Compared to existing methods such as SELEX-seq[14] and universal protein binding microarrays[7,15], CAP-SELEX respectively allows higher throughput and interrogation of larger sequence space. A total of 100 streptavidin-binding peptide (SBP) tagged TF1 proteins were used in the assay together with 94 (3× Flag-tagged) TF2 proteins to characterize 9,400 potential interactions (Fig. 1a and Supplementary Table 1). TFs were selected to cover a wide variety of structural classes and individual binding specificities. To allow bacterial expression, unstructured regions and amino- and carboxy-terminal sequences that do not correspond to known protein domains were removed from the constructs.

Co-operative signals were detected from CAP-SELEX enriched sequences (Extended Data Fig. 1) either as novel composite sites that combine partial specificities of the two TFs (using Autoseed[16]), or as orientation and spacing preferences between the individual TF's motifs (Supplementary Table 2). This analysis revealed that 55% (55/100) of TF1 and 70% (66/94) of TF2 proteins were 'active' in the CAP-SELEX assay, as indicated by identification of the expected pair of monomeric sites in at least five experiments, or a heterodimeric site in at least one experiment (Extended Data Fig. 2 and Supplementary Table 1). To test reproducibility of the assay, 10 TF1 proteins were run again against all TF2 proteins. In most cases, the recovered motifs were very similar. In addition, we validated 10 TF–TF pairs using purified full-length proteins (Extended Data Fig. 3).

Of the 3,630 tested interactions between active TF1–TF2 pairs, 315 (8.7%) displayed cooperative binding. This result is likely an underestimate, as enrichment of both expected motifs was not observed in many cases (83% of all tested pairs, not shown). The interactions were not limited to those between related TFs, and also occurred commonly between TFs from different structural families. Only 5% of all active TF1 and TF2 pairs appeared to bind to DNA independently of each other, as indicated by the presence of both expected motifs without strong orientation and spacing preferences (Extended Data Fig. 2).

Of the interacting TF pairs, 162 had only one preferred site, whereas 153 pairs displayed more than one spacing and/or orientation (Fig. 1b, c). Analysis of pairs of motifs that occurred in the same orientation revealed that the stringency of their spacing was dependent on the motif-to-motif distance. Most TF pairs whose motifs overlapped preferred just one (negative) spacing between the motifs. In contrast, if the most enriched motif pair had a gap, two or more spacings were more commonly observed. Most longer-range interactions where the gap between the motifs was 3 base pairs (bp) or more displayed a relatively wide, ±2 bp, spacing preference, similar to that reported previously[17] (Fig. 1d and Extended Data Fig. 4). Many TF pairs displayed both kinds of interactions, with one orientation preferring stringent short-range interactions, and the other orientation(s) preferring the more relaxed long-range interactions (not shown).

In cases where two or more motif spacings were allowed in the same orientation, the differences between the observed motif-to-motif distances were in general very small (Fig. 1e, >73% were only 1 bp), indicating that the mechanism of the TF–TF cooperativity is very sensitive to the relative distance and/or angle between the bound TFs. The promiscuous nature of the cooperativity was highlighted by the fact that most cases where more than one preferred mode of binding was observed involved multiple motif orientations (Fig. 1f). Two orientations was most common, whereas fewer cases of three or four orientations were observed, in part because pairs with one or two TFs with

[1]Department of Biosciences and Nutrition, Karolinska Institutet, SE 141 83, Sweden. [2]European Synchrotron Radiation Facility, 38043 Grenoble, France. [3]Genome-Scale Biology Program, University of Helsinki, P.O. Box 63, FI-00014, Finland.

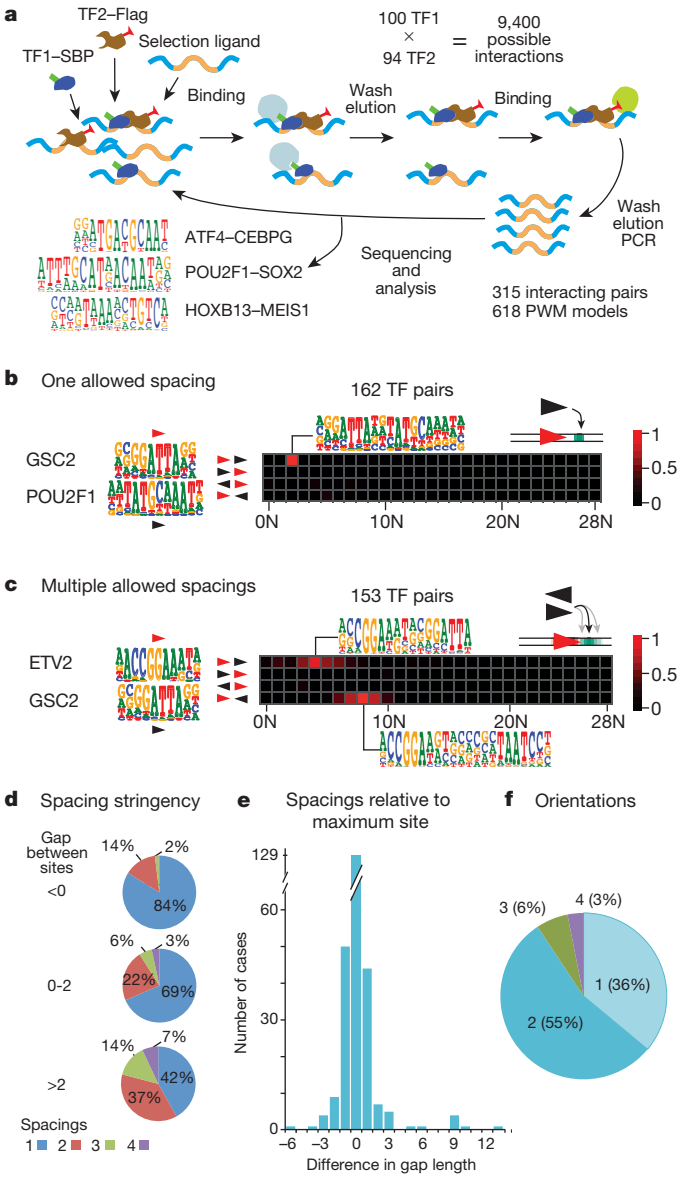


Figure 1 | CAP-SELEX reveals DNA-mediated TF–TF interactions. **a**, Schematic description of CAP-SELEX. A TF1–TF2–DNA complex is formed (top left) and subjected to two consecutive affinity purifications, followed by amplification of DNA and sequencing. The entire process is repeated three times, and the cooperative complexes are then detected from the sequences. CAP-SELEX derived PWM models for the indicated previously known[26–28] TF complexes are also shown. **b**, An example of a TF–TF pair preferring a single spacing and orientation. Heatmap shows counts (divided by max) of representative 6-mers for the TFs. **c**, A TF pair preferring two different orientations, with relatively flexible spacing in both orientations. Logos for the strongest cases are also shown. **d**, High stringency of closely packed TF–TF sites. The pie charts show the number of allowed spacings in a single orientation, binned according to the motif-to-motif distance (gap) of the strongest-bound site (maximum). **e**, TFs that can bind to sites with multiple spacings prefer very closely spaced sites. Histogram shows difference in gap length between the most strongly enriched motif spacing (normalized to 0) and other identified motif spacings. **f**, Most TF–TF pairs with multiple cooperatively bound sites allow more than one orientation. Pie chart shows frequency of TF–TF pairs binned according to the number of allowed orientations. Only pairs with multiple preferred motifs are included.

palindromic recognition sequences can only have two or one orientations, respectively. These results indicate that TF–TF cooperativity is widespread and not just mediated by the highly specific protein–protein interactions observed in previously described canonical heterodimers.

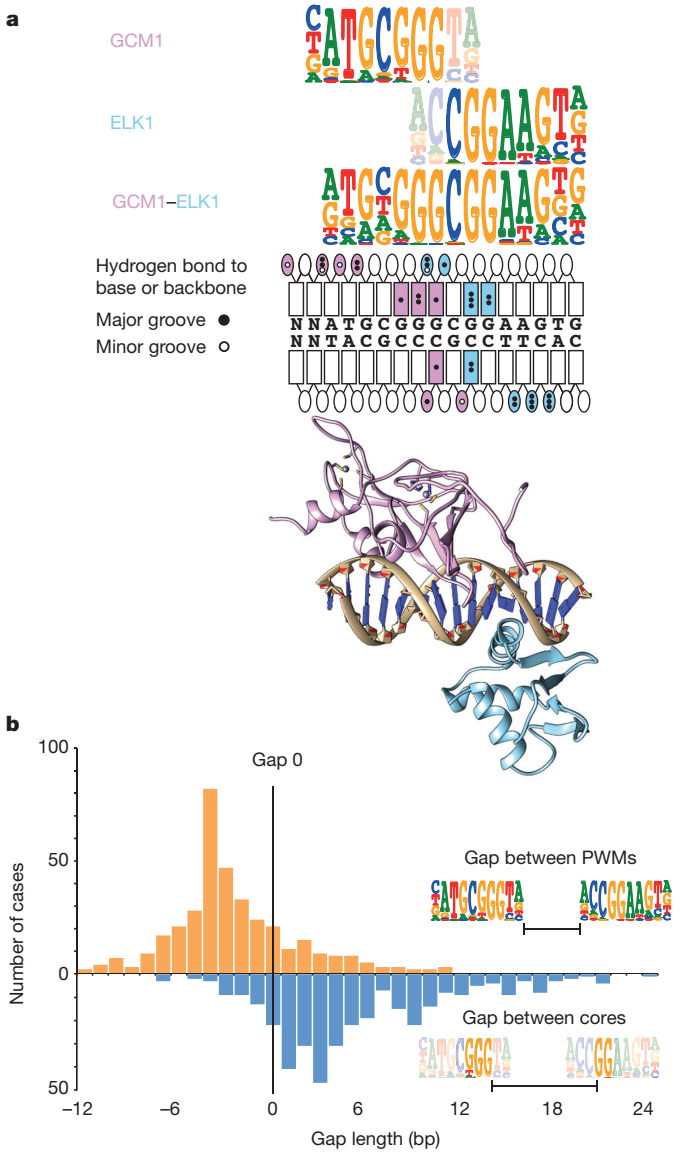


Figure 2 | Overlapping composite TF motifs with novel specificity. **a**, An example of a TF pair binding to an overlapping composite site. Top, a composite GCM1–ELK1 logo aligned to the individual logos. Middle, DNA–protein contacts for GCM1 (purple) and ELK1 (light blue) in the composite site, predicted based on GCM1–DNA and ELK1–DNA structures[29,30]. Dots indicate the number of hydrogen bonds between the TF and DNA backbone (ovals) and bases (rectangles) occurring via major (filled circle) and minor (open circle) grooves. Bottom, a schematic model of GCM1–ELK1 heterodimer. DNA is shown as idealized B-DNA. **b**, TFs prefer to bind to sites where their core motifs are closely spaced. Histogram of gaps observed between all full width motifs (core plus flank) and core motifs. Gap widths were counted for all TF pairs identified in this study for which structural data was available. Examples of calculating distance using full PWMs (above *x* axis) and core motifs (below *x* axis) are also shown.

Some TF pairs displayed strong orientation and spacing preferences, without major changes in either motif (Fig. 1b, c). However, in a large number of cases (207), the specificity of the pair of TFs differed markedly from that expected from the individual motifs (Fig. 2a). These differences were observed when the two TFs were close to each other. To understand the mechanism of the altered specificity, we analysed available structures for the studied TFs and their paralogues (Supplementary Data Set 2). Based on the analysis, 95% of the complexes are consistent with either a completely DNA-mediated mechanism, or a DNA-facilitated mechanism, where the interaction between the proteins is scaffolded by DNA and limited to few amino-acid contacts; only 5%

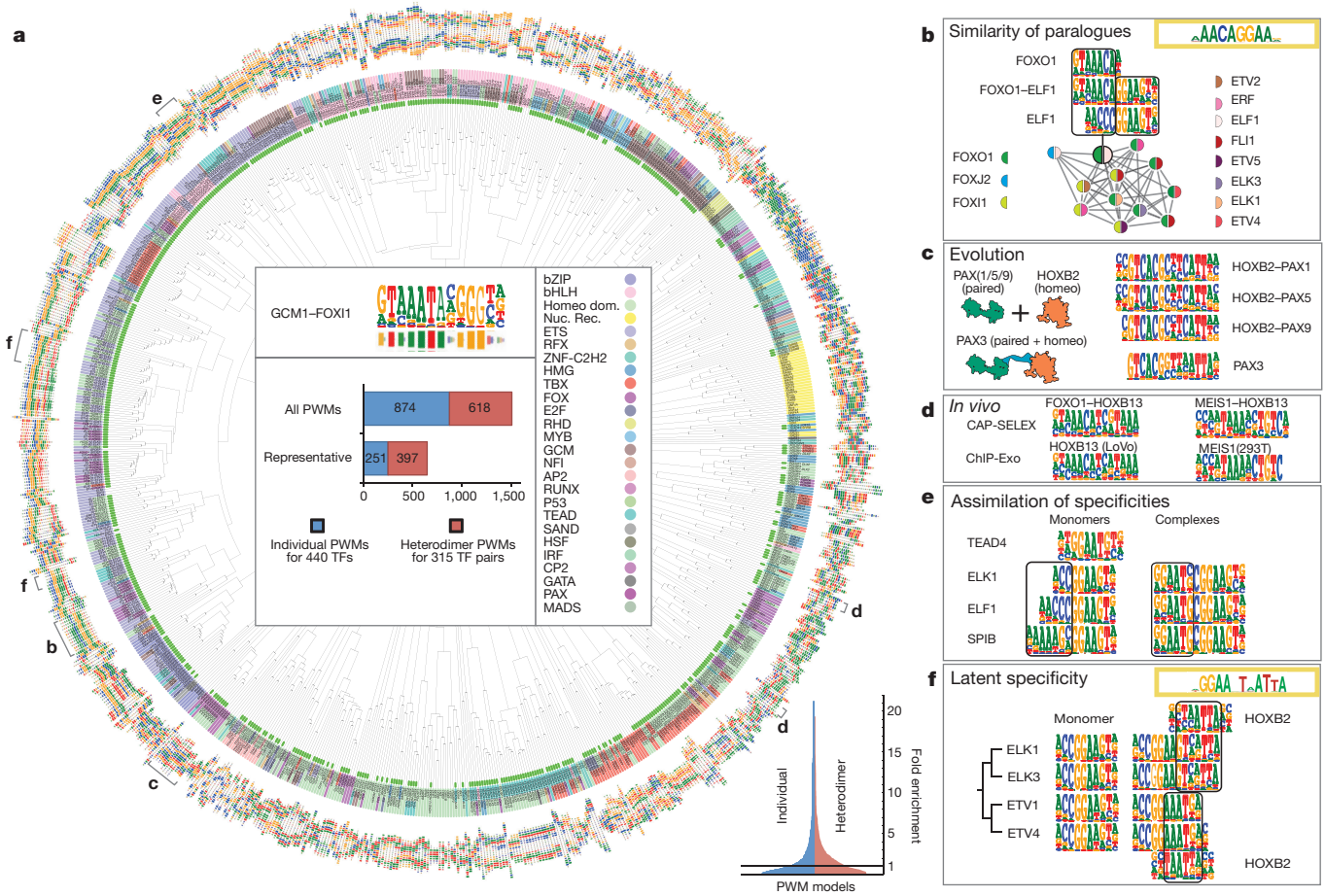


Figure 3 | All identified TF–TF interactions. a, PWM motif similarities between the heterodimer motifs (green bars) and monomeric and homodimeric representative motifs from ref. 8. Barcode logos for each factor are shown, and background colour of name indicates TF structural family. Center of dendrogram shows comparison of sequence and barcode logos, the colour key, and the number of all PWMs and representative PWMs. Inset (bottom right), fold enrichment of matches of the motifs in known TF clusters from human colon cancer cells[19]. **b**, A network representation of the very similar heterodimeric sites formed between multiple FOX and ETS proteins. Note that a similar site is recognized when the FOX protein is either used as TF1 (FOXO1, FOXJ2) or TF2 (FOXI1). Similar conserved motif is also shown[20]. **c**, HOXB2–PAX1, HOXB2–PAX5 and HOXB2–PAX9 heterodimer sites are similar to a site for PAX3, which contains both Pax- and homeodomains. **d**, FOXO1–HOXB13 and MEIS1–HOXB13 heterodimers validated by ChIP-exo. **e**, Binding of TEAD4 together with the indicated ETS TFs makes their divergent flanking recognition sequences (left box) more similar to each other (right box). **f**, HOXB2 reveals latent specificity of the TFs indicated. Inset shows conserved genomic motif[20] that is similar to the ELK1–HOXB2 motif.

of the complexes appeared to form extensive protein–protein interaction surfaces (Extended Data Fig. 5). We next generated position weight matrices (PWMs) that included information about hydrogen bond contacts between the TF amino acids, and DNA bases and backbone (Supplementary Table 3). Alignment of pairs of such contact-annotated PWMs to their respective composite models revealed that the changes in binding specificity mostly affected base positions that are recognized by the TFs via contacts to the DNA backbone. In contrast, 'core' bases directly read via hydrogen bonds were rarely affected (Fig. 2b and Extended Data Fig. 6).

In total, we recovered 618 PWM models describing the specificities of the TF–TF pairs. To globally analyse the collection, we identified distinctly different 'representative' motifs from a combination of our data set and previous high-throughput SELEX (HT-SELEX) data[8]. Out of all representative motifs, 61% were TF pair motifs identified in this study (Fig. 3 and Extended Data Fig. 7), suggesting that a very large fraction of TF specificity space is defined by TF heterodimers. A dendrogram displaying the similarity of the heterodimeric motifs revealed that paralogous proteins often shared the same partners, and bound to similar heterodimeric motifs (Fig. 3a). A total of 63 of such motif groups were identified, representing 239 TF pairs. For example, many ETS factors formed complexes with forkhead proteins and with the posterior homeodomain TFs HOXD12 and HOXB13, binding to highly similar composite sites (Fig. 3b and Supplementary Table 2). Furthermore, PAX proteins containing only a paired domain interacted with homeodomain-containing partners, binding to sites that were similar to those recognized by PAX proteins that include both paired domains and homeodomains. This suggests that this site predates the joining of the paired domain and homeodomain to the same gene (Fig. 3c).

HOXB13 also formed complexes with forkhead and MEIS proteins. The motifs recovered using CAP-SELEX were also enriched by HOXB13 ChIP-exo[18]. The preferred dimer partner of HOX13 was cell-line specific, suggesting that TFs dimerize with different proteins in different cell types (Fig. 3d). The inclusion of HOXB13 and MEIS1 dimer motifs also improved prediction of the corresponding ChIP-seq peak positions (Extended Data Fig. 5 and Supplementary Table 4).

The novel motifs were enriched in ChIP-seq-identified TF cluster sequences[19] (Fig. 3a); a larger fraction (52%) of the novel motifs were enriched compared to the monomer motifs (34%, Supplementary Tables 2 and 4). Furthermore, comparison to motifs discovered *de novo*[20] and our independent analysis revealed that many (24%) of the motifs were enriched in mammalian conserved sequences (Supplementary Table 2 and Extended Data Figs 5 and 8). A total of 390 of 618 motifs were found enriched in human TF clusters and/or mammalian conserved sequences. Both of these methods have a relatively

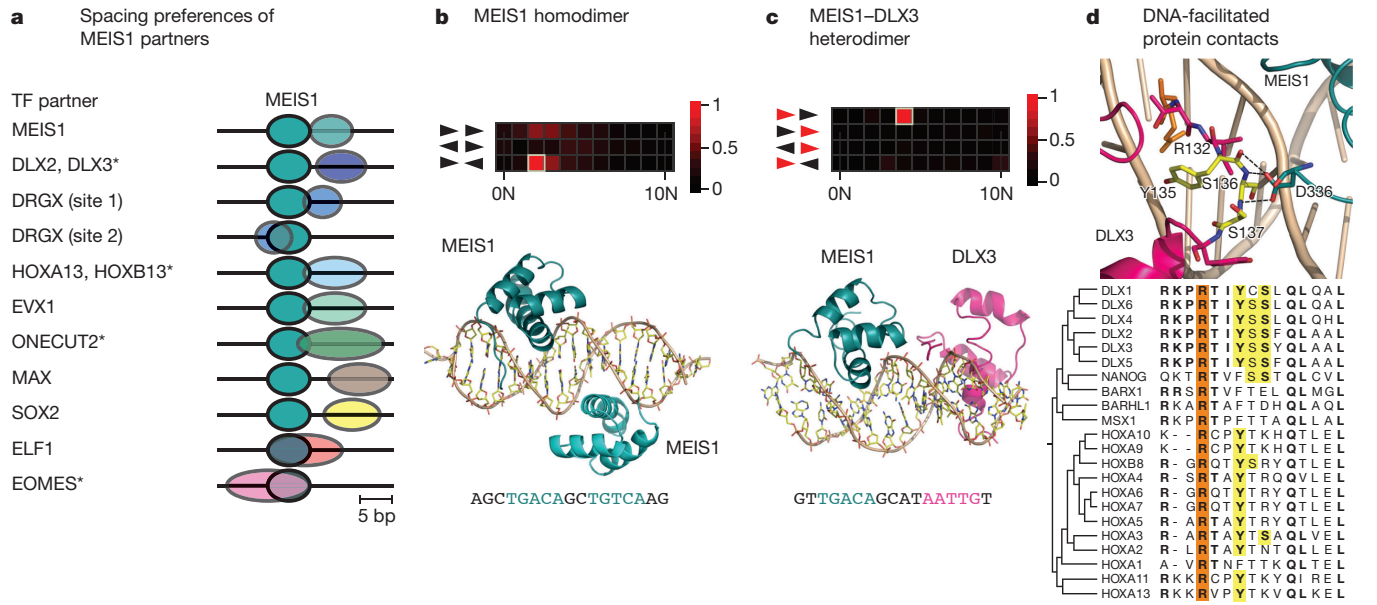


Figure 4 | Structural validation of TF–TF interactions. a, Positions of MEIS1 partner TFs in relation to the 7 bp MEIS1 motif[8] (cyan, orientation NTGACAN). Note that the partners bind to different positions, spanning a 26-bp region. Modelling suggests that in all pairs except MEIS1–ELF1, the proteins do not interact extensively (Supplementary Data Set 2). Asterisks indicate that the corresponding genomic motif matches are conserved in mammals (see Supplementary Table 2). **b**, Structure of MEIS1–MEIS1–DNA complex. Top, Heatmap based on HT-SELEX data[8] showing occurrence of MEIS1 subsequence TGACA in the orientations indicated (arrowheads, scale divided by highest count). The preferred spacing and orientation is indicated by yellow outline. Bottom, structure of two MEIS1 proteins bound to the preferred site. TGACA and its reverse complement sequence are in cyan. **c**, Structure of MEIS1–DLX3–DNA complex. Top, heatmap showing occurrence of MEIS1 and DLX3 5-mer subsequences TGACA (black arrowhead) and AATTG (red arrowhead), respectively. Bottom, crystal structure of MEIS1 and DLX3 bound to the preferred site. Note the narrowing of the DNA minor groove between the two proteins. **d**, DLX3 Arg132 (orange) inserts into the minor groove of DNA, and positions two adjacent serines and a tyrosine (yellow) so that an aspartate from MEIS1 hydrogen bonds (dotted lines) with DLX3 peptide backbone. Bottom, conservation of the residues in homeodomain proteins. Residues conserved in all human DLX proteins are in bold.

high false-negative rate, due to differences in dimers in different cell types, and the requirement that >50 motif occurrences are conserved to reach statistical significance. These results indicate that motifs identified in this study are biologically relevant.

Heterodimeric partners could also mask differences in binding specificities of individual TFs. For example, class I, II and III ETS factors ELK1, ELF1 and SPIB, respectively, prefer different 5′ flank sequences[8,21], but this difference is effectively masked by TEAD4 (Fig. 3e). This effect was rare; only two other similar cases were identified (Supplementary Data Set 1). Conversely, partners could be identified that revealed 'latent specificity'[14] of TFs, defined as binding of TFs to different heterodimeric sites, even when their primary specificities are indistinguishable. For example, ETV1, ETV4, ELK1 and ELK3 bind to similar monomeric sites, and also bound to similar heterodimeric sites with GCM1 (Supplementary Table 2). However, with HOXB2, ETV1 and ETV4 bound to one type of site, and ELK1 and ELK3 to another site (Fig. 3f). The ETVs are more closely related to each other than to the ELKs, suggesting that latent specificity evolves faster than primary specificity. Four other similar cases were identified (Supplementary Data Set 1).

To analyse the mechanisms of cooperativity, we studied all identified dimers that included the TALE-class homeodomain MEIS1. Twelve TFs from six TF families bound to diverse but specific positions at either side of MEIS1 (Fig. 4a). To understand the basis of such interactions, we solved the structure of MEIS1 bound to DNA alone, as a homodimer, and as a heterodimer with DLX3 using X-ray diffraction (3.5 Å, 1.6 Å and 3.5 Å resolutions, respectively). In the homodimer structure, the two monomers are on opposite sides of DNA and do not contact each other, indicating that the observed cooperativity is entirely mediated by DNA (Fig. 4b).

Interaction between MEIS1 and DLX3 in CAP-SELEX is much stronger than that observed for the MEIS1–MEIS1 dimer[8]. The proteins interact, but the contact surface is very small, covering only 2.0% of the solvent accessible surface of the dimer. However, the DNA between the proteins is significantly deformed, narrowing the minor groove (Fig. 4c and Extended Data Fig. 9). This facilitates interaction between the proteins, as Arg132 of DLX3 inserts into the minor groove, positioning the conserved amino acids Tyr135, Ser136 and Ser137 in such a way that the peptide backbone makes three hydrogen bond contacts with Asp336 of MEIS1 (Fig. 4d).

In summary, our sampling of a large number of TF–TF interactions revealed a much greater number of interactions than previously reported. Many novel DNA motifs were enriched in ChIP-seq TF clusters and conserved in mammalian genomes. Based on the fraction of pairs tested, we estimate that ~25,000 distinct TF pair specificities contribute to protein–DNA interactions in cells (Supplementary Table 4). The frequent interactions between TFs, together with nucleosome-mediated cooperativity[22,23] are consistent with the observation that TF binding in cells occurs in dense clusters. The clusters are also likely to be stabilized by TF co-factors, and complexes such as Mediator, cohesin and RNA polymerase II[19,24,25]. Such higher-order complexes, and complexes between TFs formed by domains that were not included in our constructs (Supplementary Table 1) are likely to further contribute to the cooperativity of TF binding *in vivo*.

Most of the observed interactions involve close packing of the individual TF's core motifs, and overlap between the motif flanks. The sequence between the core motifs commonly differs from that expected from the individual motifs. The composite sites would often be recognized by the individual TFs, but with relatively low affinity. Our findings are thus consistent with the general low affinity of sites bound *in vivo* in ChIP-seq experiments, and the fact that the conservation pattern of many regulatory elements extends beyond known TF binding sites. Taken together, our results show that cooperativity is an inherent feature of TF–DNA binding, and that DNA itself functions as an active

interacting partner, commonly facilitating the interactions between a wide range of TFs from diverse structural families.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 January; accepted 24 August 2015.
Published online 9 November 2015.

1. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Rev. Genet.* **15,** 272–286 (2014).
2. Levo, M. & Segal, E. In pursuit of design principles of regulatory sequences. *Nature Rev. Genet.* **15,** 453–468 (2014).
3. Slattery, M. *et al.* Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.* **39,** 381–399 (2014).
4. Rodda, D. J. *et al.* Transcriptional regulation of *Nanog* by OCT4 and SOX2. *J. Biol. Chem.* **280,** 24731–24737 (2005).
5. Panne, D., Maniatis, T. & Harrison, S. C. An atomic model of the interferon-beta enhanceosome. *Cell* **129,** 1111–1123 (2007).
6. De Val, S. *et al.* Combinatorial regulation of endothelial gene expression by Ets and Forkhead transcription factors. *Cell* **135,** 1053–1064 (2008).
7. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324,** 1720–1723 (2009).
8. Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152,** 327–339 (2013).
9. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature Rev. Genet.* **10,** 252–263 (2009).
10. Najafabadi, H. S. *et al.* C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nature Biotechnol.* **33,** 555–562 (2015).
11. Emery, P. *et al.* A consensus motif in the RFX DNA binding domain and binding domain mutants with altered specificity. *Mol. Cell. Biol.* **16,** 4486–4494 (1996).
12. Kurokawa, R. *et al.* Differential orientations of the DNA-binding domain and carboxy-terminal dimerization interface regulate binding site selection by nuclear receptor heterodimers. *Genes Dev.* **7,** 1423–1435 (1993).
13. Mohibullah, N., Donner, A., Ippolito, J. A. & Williams, T. SELEX and missing phosphate contact analyses reveal flexibility within the AP-2α protein: DNA binding complex. *Nucleic Acids Res.* **27,** 2760–2769, (1999).
14. Slattery, M. *et al.* Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* **147,** 1270–1282 (2011).
15. Grove, C. A. *et al.* A multiparameter network reveals extensive divergence between *C. elegans* bHLH transcription factors. *Cell* **138,** 314–327 (2009).
16. Nitta, K. R. *et al.* Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife* **4,** e04837 (2015).
17. Kim, S. *et al.* Probing allostery through DNA. *Science* **339,** 816–819 (2013).
18. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein–DNA interactions detected at single-nucleotide resolution. *Cell* **147,** 1408–1419 (2011).
19. Yan, J. *et al.* Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* **154,** 801–813 (2013).
20. Guturu, H., Doxey, A. C., Wenger, A. M. & Bejerano, G. Structure-aided prediction of mammalian transcription factor complexes in conserved non-coding elements. *Phil. Trans. R. Soc. Lond. B* **368,** 20130029 (2013).
21. Wei, G. H. *et al.* Genome-wide analysis of ETS-family DNA-binding *in vitro* and *in vivo*. *EMBO J.* **29,** 2147–2160 (2010).
22. Mirny, L. A. Nucleosome-mediated cooperativity between transcription factors. *Proc. Natl Acad. Sci. USA* **107,** 22534–22539 (2010).
23. Wasson, T. & Hartemink, A. J. An ensemble model of competitive multi-factor binding of the genome. *Genome Res.* **19,** 2101–2112 (2009).
24. Poss, Z. C., Ebmeier, C. C. & Taatjes, D. J. The Mediator complex and transcription regulation. *Crit. Rev. Biochem. Mol. Biol.* **48,** 575–608 (2013).
25. Kagey, M. H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467,** 430–435 (2010).
26. Nishizawa, M. & Nagata, S. cDNA clones encoding leucine-zipper proteins which interact with G-CSF gene promoter element 1-binding protein. *FEBS Lett.* **299,** 36–38 (1992).
27. Shen, W. F. *et al.* AbdB-like Hox proteins stabilize DNA binding by the Meis1 homeodomain proteins. *Mol. Cell. Biol.* **17,** 6448–6458 (1997).
28. Williams, D. C., Jr, Cai, M. & Clore, G. M. Molecular basis for synergistic transcriptional activation by Oct1 and Sox2 revealed from the solution structure of the 42-kDa Oct1·Sox2·*Hoxb1*–DNA ternary transcription factor complex. *J. Biol. Chem.* **279,** 1449–1457 (2004).
29. Cohen, S. X. *et al.* Structure of the GCM domain–DNA complex: a DNA-binding domain with a novel fold and mode of target site recognition. *EMBO J.* **22,** 1835–1845 (2003).
30. Mo, Y., Vaessen, B., Johnston, K. & Marmorstein, R. Structure of the Elk-1–DNA complex reveals how DNA-distal residues affect ETS domain recognition of DNA. *Nature Struct. Mol. Biol.* **7,** 292–297 (2000).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Yan, B. Schmierer, E. Kaasinen, C. Daub, E. Haapaniemi and Å. Kolterud for their review of the manuscript, the Karolinska Institutet protein science facility for protein purification, and S. Augsten, L. Hu and A. Zetterlund for technical assistance. This work was supported by Finnish Academy CoE in Cancer Genetics, Center for Innovative Medicine, Knut and Alice Wallenberg and Göran Gustafsson Foundations and Vetenskapsrådet.

Author Contributions A.J. and J.T. designed the experiments. A.J. and Y.Y. performed CAP-SELEX, K.D. performed ChIP-exo, and M.T. the sequencing analyses. A.J., K.R.N., T.K., M.E. and J.T. wrote computer programs for the analyses. E.M. and A.P. performed X-ray crystallography, and E.M. solved the structures. A.J., K.R.N. and E.M. prepared illustrations and A.J., Y.Y. and J.T. wrote the article. All authors contributed to data analysis and reviewed the manuscript.

Additional Information Sequencing reads are deposited to European Nucleotide Archive (accession PRJEB7934). The atomic coordinates and diffraction data are deposited to Protein Data Bank (accession 4XRM, 5BNG and 4XRS). All computer programs and scripts used are either published or available upon request. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.T. (jussi.taipale@ki.se).

METHODS

Sequencing and data analysis. Unselected initial libraries and products of the third selection cycle were purified using a PCR-purification kit (Qiagen) and sequenced using Illumina HiSeq 2000 (multiplexed as in ref. 31; 55 bp single-read length). Raw sequence reads were demultiplexed, and initial quality control was performed using IniMotif[31]. Sequencing depth was set in such a way that on average each experiment would result in 250,000 sequence reads. Based on previous enrichment ratios, this should lead to more than 1,000 highly enriched seed subsequences to be detected (count statistics; Poisson distribution, 3.16% standard deviation; expected background for 10 bp seed 15 counts, P value 1.26×10^{-273} using winflat, expected 15, observed $\geq$1,000; Bonferroni corrected P value $< 1.32 \times 10^{-267}$). Average background-corrected count for the seed match at the indicated multinomial setting was 3,295 (Supplementary Table 2). All sequence data has been deposited to ENA (European Nucleotide Archive) under accession number PRJEB7934.

To identify overlapping composite sites, we used the AUTOSEED tool described in Nitta *et al.*[16]. This tool is based on identification of gapped and ungapped subsequences that represent a local maxima within a Huddinge distance[16] of 1; that is, they are more enriched than all subsequences that align with them with $k-1$ perfectly matching bases, where k is the length of the ungapped subsequences.

Cell culture and ChIP-exo. LoVo (source: ATCC, cat. no. CCL229TM), 293FT (source: Thermo Fisher Scientific, cat no. R700-07) and GP5d (source: ECACC, cat. no. 95090715) cells were cultured in DMEM supplemented with 10% fetal bovine serum (FBS) and antibiotics. Cells were obtained directly from the indicated source, and tested and found negative for mycoplasma contamination by immunofluorescence analysis after staining with 3.3 µg ml^{-1} bisBenzimide H 33342 trihydrochloride (Sigma cat no. B2261). All antibodies used in ChIP-exo experiments were ChIP-grade. In each experiment a non-specific IgG was used as a control. ChIP-exo was performed essentially as described in Rhee and Pugh[18] with modifications from ref. 32 using antibodies for HOXB13, MEIS1 and rabbit IgGs (Santa Cruz Biotechnology and Abcam cat. numbers sc-66923X, ab19867, and sc-2027, respectively). See Supplementary Table 1 for the sequences of the Illumina sequencing adapters. Sequence reads were mapped to the human reference genome (hg18), using bwa with default parameters. For peak-calling, we used GEM[33] with default parameters, and the genome size set to 2,700,000,000.

Construct design. TFs were selected to cover different structural classes and individual binding specificities. Thus, in the set, small TF families such as TEA and GCM are relatively overrepresented, and large families such as C_2H_2 zinc fingers and canonical homeodomains are underrepresented. The expression constructs contained the DNA-binding domain, and known dimerization and protein–protein interaction domains for TF families where such domains are known to be required for DNA binding. These included, for example leucine zipper domains of bZIP and bHLH proteins, pointed-domains of ETS factors, dimerization domains of nuclear receptors as well as short motifs such as 'YPWM' of the anterior homeodomains that are known to be involved in protein–protein interactions[34] (see Supplementary Table 1 for full sequences and removed and retained domains). Flanking sequences of 15 amino acids were also included on both sides to allow folding of the known protein domains, and to retain amino acids that are located close to DNA and could mediate interactions between closely packed TFs. We have previously shown that such constructs recover accurate binding specificities and homodimeric interactions by analysis of 125 pairs of such constructs and the corresponding full-length TF proteins[8,16].

Protein expression, purification and activity testing. Bacterial protein expression Gateway recipient vectors that incorporated a N-terminal Thioredoxin-6×His tag, with either a C-terminal streptavidin binding peptide (SBP) or 3×Flag tag were constructed using pETG20A-plasmid as a backbone. Inserts for protein expression were derived either by gene synthesis (Genscript; see Supplementary Table 1 for protein sequences and domains), or from previously published Gateway donor clones[8]. All proteins were expressed in the Rosetta 2(DE3)pLysS *E. coli* strain as fusion proteins using the auto-induction protocol described in Jolma *et al.*[8]. Proteins were purified using nickel affinity purification (GE Nickel sepharose Fast-Flow6, GE, Sweden) and stored at −20 °C in 50% glycerol, 150 mM NaCl, 250 mM imidazole, 15 mM Tris-Cl, pH 7.5.

Protein expression and purification from *E. coli* cells was performed as described in Jolma *et al.*[8]. Briefly, the expression system used Rosetta 2(DE3)pLysS strain of *E. coli* (Millipore) cultured in ZYP5052 autoinduction media, where the expression of proteins is induced upon consumption of the preferred carbon source (glucose). Transformed cells were first cultured in deep well 96-well plate (Thermo, AB0661) wells in TB-medium at 37 °C for overnight and then transferred to the auto-induction medium (1:40 dilution, see Vincentelli *et al.*[35]). When the cell density was between 2.0 and 3.0 optical density at 600 nm, the temperature was lowered to 17 °C, and the culture continued for 40 h. The cells were harvested by centrifugation (4,000 rpm for 15 min), and lysed by incubation with buffer A (300 mM NaCl in 50 mM Tris-Cl, pH 7.5) containing 10 mM imidazole, 0.5 mg ml^{-1} lysozyme (Sigma) and 1 mM PMSF (Sigma). The lysis was completed by a freeze–thaw cycle. DNase I and $MgSO_4$ were added to the thawed lysate at 10 µg ml^{-1} and 1 mM final concentration, respectively, and the lysates incubated with Ni-Sepharose 6 Fast Flow resin (GE Healthcare) and shaken for 45 min. The lysate was then transferred to a filter plate (Nunc, 278011, 20 µm pore size), and the beads washed two times each with 600 µl of 10 mM and 50 mM imidazole in buffer A using a vacuum manifold. The bound proteins were eluted from resin using 500 mM imidazole in buffer A. The expression of the purified proteins were checked by UV absorbance at 280 nm and SDS–PAGE electrophoresis (E-PAGE protein gels, Invitrogen) and Coomassie brilliant blue staining. 50% glycerol was added to the proteins before storage at −20 °C.

In most cases the activity of the proteins was assessed by HT-SELEX[8,31], and proteins that robustly enriched expected sequences were included in the CAP-SELEX process. As some TFs are only expected to bind to DNA as a heterodimer, we also included in CAP-SELEX some proteins that did not robustly enrich sequences in HT-SELEX. These included the known obligate heterodimers MYC, PBX1, PBX2 and PBX4. All included proteins are indicated in Supplementary Table 1. The HT-SELEX analysis yielded expected binding sites for most individual TFs, and in addition resulted in identification of novel motifs for TFs (see Supplementary Table 2). HT-SELEX was also used to validate some of the CAP-SELEX results with full length TFs. In these cases the pair of proteins were mixed together in an ~1:1 ratio before the further steps of the HT-SELEX (see Supplementary Table 1 for the details of the clones).

In most cases, a scaled-up culture (50 to 100 ml) was used to express more of the TF constructs for CAP-SELEX. The protocol used was similar to that used for the deep-well plate cultures, except that proportionally larger lysis and wash volumes were used, and that 1 ml Ni-Sepharose 6 Fast Flow gravity columns (GE Healthcare) were used as the affinity matrix.

CAP-SELEX assay. Previous systematic efforts that have focused on heterodimerization between TFs have generally studied proteins that dimerize in the absence of DNA[36–38]. However, some cases of TFs that cannot dimerize in the absence of DNA, or only interact with each other indirectly through DNA have been described in the literature[39,40]. The CAP-SELEX process can capture both types of interactions, and is based on a combination of HT-SELEX[8,31] with tandem-affinity purification[41]. In this assay, a pair of TFs tagged with different affinity tags, SBP and 3×Flag, and a double stranded DNA ligand that contains a 40-bp randomized region are mixed together in individual wells of a 96-well plate in a buffer that mimics biological conditions in the nucleus[42], and the mixture is incubated for 30 min, after which the bound dsDNA ligands are separated from free ligands through consecutive affinity purification by first the SBP and then the 3×Flag tagged protein using affinity beads and an automated plate washer (Fig. 1a). Bound DNA is then amplified by PCR and sequenced, and the selection process repeated three times. Binding of the TFs is then revealed by enrichment of characteristic subsequences (see below).

The 40-bp random window corresponds to almost four complete turns of the DNA helix, allowing detection of interactions between two TFs that exclusively occupy 9 bp of sequence (see ref. 20) over two full helical turns. As with our previous HT-SELEX platform, the purified ligands are barcoded and directly compatible with multiplexed Illumina sequencing (for selection ligand sequences, please see Supplementary Table 1).

Of the proteins tested here, most (87%) were functionally validated by HT-SELEX (see Supplementary Table 1 for details). The low activity of some HT-SELEX validated proteins (Supplementary Table 1) was probably due to the fact that CAP-SELEX involves two consecutive affinity purifications, and is therefore more stringent than HT-SELEX.

For CAP-SELEX, 10–200 ng (see Supplementary Table 1) purified Flag- and SBP-tagged proteins were diluted into 25 µl volume of binding buffer (140 mM KCl, 5 mM NaCl, 2 mM $MgSO_4$, 3 µM $ZnSO_4$, 100 µM EGTA, 1 mM K_2HPO_4, in 20 mM HEPES, pH 7.0) containing approximately 10 µmol DNA selection ligands, and incubated for 20 min at room temperature. Subsequently, 0.2% BSA, 0.1% Tween 20 pre-blocked Streptavidin-coated magnetic Sepharose beads (1.25 µl; GE Healthcare Streptavidin Mag Sepharose) in two volumes of binding buffer were added, and the mixture incubated at room temperature for 2 h with vigorous shaking (800 r.p.m.; Edmund Bühler TiMix shaker). The beads were subsequently washed 5 times with binding buffer, using BioTek 405 CW plate washer with a magnetic platform. The protein–DNA complexes were eluted from the beads using 50 µl of 10 mM biotin (Sigma) in binding buffer. The eluate was transferred to a fresh plate containing M2 anti-Flag magnetic beads (1.25 µl; Sigma) in 50 µl of binding buffer, and shaken at 800 r.p.m. for 20 min at room temperature. The beads were washed ten times with binding buffer, suspended in 0.1% Tween 20, 0.5 mM EDTA, 10 mM Tris-Cl, pH 8.0, and transferred to a PCR plate. DNA was then eluted from the beads by

incubation at 95 °C for 10 min. A 9 µl aliquot of the bead suspension was transferred to a new PCR plate, and the DNA amplified by PCR (65 °C for 10 s, 72 °C for 36 s, 97 °C for 15 s for annealing, elongation and denaturation, respectively, for 25 cycles). A separate aliquot was analysed by qPCR (Roche LightCycler 480) to monitor progress of the experiment. Amplified selection ligands were then subjected to sequencing and new cycles of CAP-SELEX (up to 3 cycles total).

The input libraries and libraries selected for three cycles were then sequenced and analysed (see below). Cooperative complexes were initially detected from 5–12-bp long primary and secondary binding models generated by the previously described SELEX data-analysis tool IniMotif[31]. Initial results validated the method through confirmation, characterization and refinement of the binding specificity models for 12 previously known heterodimers (Fig. 1a and Extended Data Fig. 1).

Generation of PWMs. To detect heterodimeric sites that can occur in many different orientations and spacings, we used the *de novo* motif discovery algorithm Autoseed[16]. Autoseed finds gapped and ungapped subsequences that represent local maxima, that is, are more enriched than closely related subsequences. Control experiments established that such preferences were not observed in the input libraries.

Based on analysis of CAP-SELEX cycle 3, the sequencing was then extended to cover products from earlier selection cycles for samples that showed enrichment of motifs that were similar to the expected TF2 motif. We have previously shown that analysing SELEX data from early cycles allows recovery of low-affinity sites, and results in motifs that are very similar to those determined using competition assays[8,16].

For identification of co-operative sites where the individual TF motifs were spaced farther apart, we defined representative 6-mer sequences for each TF. For each experiment, we then identified cases where both representative 6-mers were found in the same sequence reads, and from these reads, counted the occurrence of each spacing and orientation combination. The expected distribution of spacings without any preferences is non-uniform due to the limited size of the randomized region but cannot explain the local maxima observed in the data (spacings that are preferred compared to both shorter and longer spacings). Given the size of both occupied sites m and a fixed spacing n between them, the number of ways the sites can be placed into the randomized region of size l is $c_n = l - 2m - n + 1$ ($0 \le n \le l - 2m$), or twice that if the order of sites is taken into account. As the expected count of spacing n is directly proportional to c_n, the count is a linearly decreasing function of n and thus has no maxima except the one at the boundary $n = 0$. Moreover, the expected relative difference of counts between consecutive spacings is $(c_n - 1 - c_n)/c_n = 1/c_n$ indicating that the expected differences are small in the range where spacing preferences are observed.

To display the preferred spacings and orientations of the two TFs, subsequences containing both consensus 6-mer sequences of the motifs with variable-length gap between them were counted, and the counts represented as a heatmap (for example, Fig. 1). For each orientation of the 6-mers, we identified the spacing with maximum counts. That spacing was considered to be preferred if the sum of the counts of it and its two neighbours was higher than 30% of the total count for all spacings, and less than 20% of the reads counted were derived from a single non-unique read. If one preferred case was identified for a pair of 6-mers, we then determined whether other spacings and orientations were also enriched. Up to five spacings and orientations were considered to be preferred if their respective counts were higher than 50% of the maximum count after mean normalization of all counts. Cases where both TF1 and TF2 6-mers were detected robustly, but no preferred orientation and spacing was detected were classified as 'weak or no co-operativity'. In case of experiments performed several times, interactions were called if they passed the thresholds in at least one case. In each case, control unselected ligands were also sequenced, to ensure that the oligonucleotide synthesis resulted in even distribution of mononucleotides. In addition, several ligands were sequenced very deeply (>10 million reads) to ensure that 6-mer subsequences were distributed at a similar frequency along the 40-bp random sequence window (Extended Data Fig. 4). For assessment of reproducibility, see Extended Data Fig. 3.

Subsequently, the enriched subsequences were used as seeds to generate position weight matrix (PWM) models for the complexes. We generated PWM models as described in Jolma *et al.*[8,31], using the seeds, selection cycles and multinomial setting indicated in Supplementary Table 2. The models were subsequently inspected to remove cases that could be also explained by homodimeric binding (for example, cases where two TFs with similar primary motifs were analysed). After identification of an enriched sequence, seed refinement was performed essentially as described in Jolma *et al.*[8]. In the majority of the cases, the PWM models were generated using selection cycle 3. The models were expert curated to separate different binding modes, and to remove cases where there was excessive positional bias or where the two proteins bound to very similar sites and thus we could not differentiate between homo- and heterodimeric binding. All seed, cycle and multinomial settings used are indicated on Supplementary Table 2, and the sequence reads have been deposited to ENA under accession (PRJEB7934).

Hydrogen bond contacts between TF amino-acids and DNA bases were identified using the program CONTACT that is included in the CCP4 software suite[43]. Hydrogen bond information was added to PWMs to generate contact annotated PWMs (pfmc). PWMs were aligned to each other by minimizing the sum of individual base-to-base comparison scores calculated as follows: Max (information content for PWM1 position n, information content for PWM2 position m) * (Manhattan distance between base frequencies of PWM1 position n and PWM2 position m). In regions where there was no overlap, the positions were compared to an equal frequency of all bases. Pairs of positions whose score was smaller than 0.25 are indicated by boxes in Supplementary Data Set 2. The same cut-off was used to count divergent base positions in Extended Data Fig. 6.

Enrichment of motifs at ChIP-seq TF clusters and prediction of ChIP-seq peaks. Interactions between TFs appear to be important *in vivo*, as recent large scale genome-wide location analyses of TFs in cultured cells have revealed that in a given cell type, TFs bind only to a subset of their potential target sites, and that the occupied target sites are located in high-density clusters, where many different TFs colocalize within a few hundred bp long regions[19,44,45]. Many of the occupied sites do not contain high-affinity sites for the analysed TF, suggesting that co-operative interactions allow binding of TFs to low-affinity sites[19,45,46]. Formation of TF clusters is probably at least in part the result of competition between TFs and nucleosomes, which indirectly results in increased occupancy of TFs close to each other, even when the TFs do not have direct cooperative interactions[22,23]. Another mechanism that could contribute to TF cluster formation is direct co-operativity between TFs. To test this, we analysed enrichment of monomer and heterodimer PWM matches at human LoVo colon cancer TF clusters[19] (Fig. 3a inset, Supplementary Tables 2 and 4) as described in Yan *et al.*[19], using a score cut-off for each motif that resulted in one match per 10 kb of genomic sequence. *P* values for the enrichment were calculated using winflat. Similarity of TF DBD amino-acid sequences was determined using PRANK[47]. Similarity of PWMs was determined using SSTAT[48] using the default parameters. Motif dendrogram was drawn by using Euclidean distance metric with average linkage with R package ape. Network shown in Fig. 3 was drawn based on the same distances. The dominating set of the models was generated essentially as described in Jolma *et al.*[5]. Briefly, we first generated a network where monomeric or homodimeric motifs from Jolma *et al.*[8], and heterodimer motifs from this study were connected to each other if they were similar. To determine how many novel specificities were identified in our study, we used the minimum dominating set of this network to identify a set of motifs that represents distinctly different specificities.

Sequence and barcode-logos were generated as described in Nitta *et al.*[16]. In barcode logos shown in Fig. 3 and Extended Data Figs 3 and 7, four bars are drawn that represent the frequency of the bases for each base position. Width of each bar is proportional to the frequency of the corresponding base (range 0 to 1), and both height and colour intensity of all the bars at a given position are proportional to the frequency of the most common base at that position (range 0.25 to 1). In the DNA base letter PWM logos shown, the height of each letter is directly proportional to the frequency of the indicated base at the indicated position.

Analysis of error rates in prediction of ChIP-seq/exo peaks using the dimeric PWMs was performed using a random forest classifier. For the random forest classifier, the R package randomForest, version 4.6–6, was used (http://cran.r-project.org/web/packages/randomForest/index.html). The classifier was trained on TF motif matches to discern peak summit regions from close by non-peak summit genomic positions. Stated accuracy estimates are based on out-of-bag error estimates.

ChIP-seq binding site prediction error rate analysis was performed on data gathered from existing ChIP-seq experiments and from the HOXB13 ChIP-exo experiment from this study. The FASTA sequences of 1,001 bp genomic regions surrounding ChIP-exo peak summits for HOXB13 (from this work) and previously described ChIP-seq peak summits for ELF1 (ref. 21), ELK1 (ref. 49), HOXB13 (ref. 50) and two different MEIS1 experiments[51,52] were used as a positive set together with a negative set consisting of 1,001 bp FASTA sequences taken from 2,000 bp away from the peak summits. For each FASTA sequence in the collection, the score and relative position of the highest-scoring match to each motif in the PWM collection was recorded and used to train a randomForest classifier with 5,000 trees. To determine whether the dimer partners had predictive power, a classifier trained using the monomer motif of the relevant TF, and all its dimer motifs was compared to a classifier trained using the monomer motif and dimer motifs with the partner region of the motif reversed. Error rates were estimated using out-of-bag predictions.

Analysis of conservation of motif matches. To measure the conservation of genomic sites recognized by heterodimeric motifs we developed a procedure to test whether the heterodimeric motifs explained patterns of evolutionary conservation

observed in the human regulatory elements. To identify the potential conservation attributable to the specific TF–TF–DNA interactions described by the motifs, the genomic sites recognized by each heterodimeric motif were compared to sites recognized by artificial control motifs that represented different orientations of the two TFs but were otherwise comparable to the original motif, for example had the same width and information content. To obtain control motif sets containing the specificities of the individual TFs as embedded in the heterodimeric motifs, each heterodimeric motif was split into two partial motifs at all possible cut points with the restriction that both the 'left' and the 'right' sections were at least one-third of the width of the whole motif (rounded down). The control motifs were constructed by concatenating each of the partial motif pairs in all three alternative orientations ('right' followed by 'left', 'left' followed by the reverse complement of 'right', and the reverse complement of 'left' followed by 'right'). For example, a motif of length fifteen was split after 5–10 bases resulting in 18 control motifs.

Sites recognized by a heterodimeric motif and its control motifs were searched from the human genome constrained elements[53] (SiPhy pi 12-mers with 10% false discovery rate in reference genome hg19, regions shorter than 50 bp or overlapping exons or repeats according to Ensembl version 70 were removed resulting in total 41 Mb of sequence) using the program MOODS[54] with a loose cut-off (P value $<10^{-3}$ with flat background distribution) to obtain a large excess of putative binding sites for each motif. All found sites were merged into one list and 10,000 non-overlapping highest affinity sites selected for conservation analysis regardless of the motif identity (heterodimer or control).

Whether the evolutionary conservation of the high affinity sites was explained by the motifs was tested using program SiPhy[55] (version 0.5, task 16, seedMinScore 0) and multiz100way multiple alignments[56] of 99 vertebrate species to human (downloaded from UCSC genome browser, version hg19). A site was marked as being conserved according to the motif if its SiPhy score was positive meaning that the aligned bases at the site were better explained by the motif than by a neutral evolutionary model (hg19.100way.phastCons.mod obtained from UCSC genome browser).

The hypothesis that the heterodimeric motif sites were more likely to be conserved according to the motif than the sites of its control motifs was tested against the null hypothesis that there was no association between site conservation and motif identity using Fisher's exact test (one-sided). The P values given by the tests for individual heterodimeric motifs were corrected for multiple testing using Holm's method. This procedure detected genomic conservation for 149 out of 618 motifs (24%) at family-wise error rate <0.05 (including the previously known ETV2–FOXI1 motif[20] used as a positive control while developing the procedure).

Protein purification, crystallization and data collection. The DNA-binding domains of human MEIS1 (residues 277–339) and DLX3 (residues 122–193) were overexpressed as a thioredoxin-6His fusion protein in *E. coli* and isolated from the soluble cell lysate by affinity chromatography followed by gel-filtration chromatography as described in ref. 57. The DNA fragments used in crystallization were obtained from MWG as single strand oligonucleotides and annealed in 150 mM NaCl, 1 mM EDTA in 10 mM Tris-Cl, pH 7.5. The purified proteins were first mixed with solutions of annealed DNAs at a molar ratio of 1:1.2 and after incubation for 15–20 min on ice subjected to the crystallization trials. Crystallization experiments were carried out with an in-house developed crystal screening kit of different polyethylene glycols. The crystals of MEIS1–MEIS1–DNA complex were obtained in sitting drops at room temperature from 100 mM Tris-Cl (pH 8) solution containing 25.6% (w/v) PME (5000), 80 mM $MgCl_2$ and 10% PEG (400). The crystals of monodimer MEIS1–DNA complex were obtained from 100 mM HEPES (pH 7.09) solution containing 30% (w/v) PME(5000), 80 mM $MgCl_2$ and 10% PEG (400). Crystals of MEIS1–DLX3–DNA complex were obtained from 100 mM Tris-Cl (pH 7.5) containing 24% PEG (8000), 40 mM $MgCl_2$ and 5% butanol. All diffraction data for both complexes were collected at beam-line ID23-1 at the ESRF (Grenoble, France) using the reservoir solution as cryo-protectant. The data collection strategy was optimized with the program BEST[58]. The data were integrated with the program XDS[59] and scaled with SCALA[60]. Statistics of data collection are presented in Supplementary Table 5.

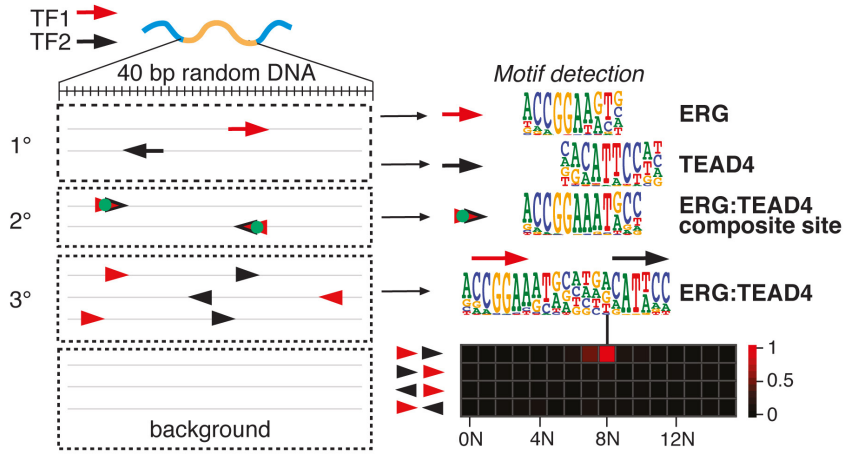
Structure determination and refinement. The structures of all three complexes were determined by molecular replacement using the program Phaser[61] in Phenix[62] with the structure of MEIS2 and DLX5 (PDB entries 3K2A and 2DJN, respectively) as a search model. The manual rebuilding of the model was done using COOT[63] combined with refinement with Phenix.refine using TLS option. The refinement statistics are presented in Supplementary Table 5. The atomic coordinates and diffraction data have been deposited to Protein Data Bank with the accession codes 4XRM, 5BNG and 4XRS, for MEIS1 homodimer, MEIS1 monomer and MEIS1–DLX3 heterodimer, respectively.

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

31. Jolma, A. *et al.* Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* **20,** 861–873 (2010).
32. Katainen, R. *et al.* CTCF/cohesion-binding sites are frequently mutated in cancer. *Nature Genetics* **47,** 818–821 (2015).
33. Guo, Y. *et al.* Discovering homotypic binding events at high spatial resolution. *Bioinformatics* **26,** 3028–3034 (2010).
34. Passner, J. M., Ryoo, H. D., Shen, L., Mann, R. S. & Aggarwal, A. K. Structure of a DNA-bound Ultrabithorax-Extradenticle homeodomain complex. *Nature* **397,** 714–719 (1999).
35. Vincentelli, R. *et al.* High-throughput protein expression screening and purification in Escherichia coli. *Methods* **55,** 65–72 (2011).
36. Keshava Prasad, T. S. *et al.* Human Protein Reference Database–2009 update. *Nucleic Acids Res.* **37,** D767–D772 (2009).
37. Ravasi, T. *et al.* An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140,** 744–752 (2010).
38. Newman, J. R. & Keating, A. E. Comprehensive identification of human bZIP interactions with coiled-coil arrays. *Science* **300,** 2097–2101 (2003).
39. Klemm, J. D. & Pabo, C. O. Oct-1 POU domain–DNA interactions: cooperative binding of isolated subdomains and effects of covalent linkage. *Genes Dev.* **10,** 27–36 (1996).
40. Panne, D., Maniatis, T. & Harrison, S. C. Crystal structure of ATF-2/c-Jun and IRF-3 bound to the interferon-β enhancer. *EMBO J.* **23,** 4384–4393 (2004).
41. Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnol.* **17,** 1030–1032 (1999).
42. Hallikas, O. *et al.* Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124,** 47–59 (2006).
43. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67,** 235–242 (2011).
44. Moorman, C. *et al.* Hotspots of transcription factor colocalization in the genome of *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **103,** 12027–12032 (2006).
45. Yip, K. Y. *et al.* Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* **13,** R48 (2012).
46. Meireles-Filho, A. C., Bardet, A. F., Yanez-Cuna, J. O., Stampfel, G. & Stark, A. *cis*-regulatory requirements for tissue-specific programs of the circadian clock. *Curr. Biol.* **24,** 1–10 (2014).
47. Löytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl Acad. Sci. USA* **102,** 10557–10562 (2005).
48. Pape, U. J., Rahmann, S. & Vingron, M. Natural similarity measures between position frequency matrices with an application to clustering. *Bioinformatics* **24,** 350–357 (2008).
49. Odrowaz, Z. & Sharrocks, A. D. The ETS transcription factors ELK1 and GABPA regulate different gene networks to control MCF10A breast epithelial cell migration. *PLoS ONE* **7,** e49892 (2012).
50. Huang, Q. *et al.* A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding. *Nature Genet.* **46,** 126–135, doi: 10.1038/ng.2862 (2014).
51. Huang, Y. *et al.* Identification and characterization of Hoxa9 binding sites in hematopoietic cells. *Blood* **119,** 388–398 (2012).
52. Penkov, D. *et al.* Analysis of the DNA-binding profile and function of TALE homeoproteins reveals their specialization and specific interactions with Hox genes/proteins. *Cell Reports* **3,** 1321–1333, (2013).
53. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–482 (2011).
54. Korhonen, J., Martinmaki, P., Pizzi, C., Rastas, P. & Ukkonen, E. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics* **25,** 3181–3182 (2009).
55. Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25,** i54–i62 (2009).
56. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14,** 708–715 (2004).
57. Savitsky, P. *et al.* High-throughput production of human proteins for crystallization: the SGC experience. *J. Struct. Biol.* **172,** 3–13 (2010).
58. Bourenkov, G. P. & Popov, A. N. A quantitative approach to data-collection strategies. *Acta Crystallogr. D* **62,** 58–64 (2006).
59. Kabsch, W. Xds. *Acta Crystallogr. D* **66,** 125–132 (2010).
60. Collaborative Computational Project Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
61. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
62. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
63. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).

64. Fitzsimmons, D. *et al.* Pax-5 (BSAP) recruits Ets proto-oncogene family proteins to form functional ternary complexes on a B-cell-specific promoter. *Genes Dev.* **10,** 2198–2211 (1996).
65. Kim, J. J. *et al.* Regulation of insulin-like growth factor binding protein-1 promoter activity by FKHR and HOXA10 in primate endometrial cells. *Biol. Reprod.* **68,** 24–30 (2003).
66. Vinson, C. R., Hai, T. & Boyd, S. M. Dimerization specificity of the leucine zipper-containing bZIP motif on DNA binding: prediction and rational design. *Genes Dev.* **7,** 1047–1058 (1993).
67. Williams, T. M., Williams, M. E. & Innis, J. W. Range of HOX/TALE superclass associations and protein domain requirements for HOXA13:MEIS interaction. *Dev. Biol.* **277,** 457–471 (2005).
68. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnol.* **30,** 271–277 (2012).
69. Raveh-Sadka, T. *et al.* Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nature Genet.* **44,** 743–750 (2012).
70. Hochschild, A. & Ptashne, M. Cooperative binding of λ repressors to sites separated by integral turns of the DNA helix. *Cell* **44,** 681–687 (1986).
71. Moretti, R. *et al.* Targeted chemical wedges reveal the role of allosteric DNA modulation in protein–DNA assembly. *ACS Chem. Biol.* **3,** 220–229 (2008).
72. Aggarwal, A. K., Rodgers, D. W., Drottar, M., Ptashne, M. & Harrison, S. C. Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. *Science* **242,** 899–907 (1988).
73. Jordan, S. R. & Pabo, C. O. Structure of the lambda complex at 2.5Å resolution: details of the repressor–operator interactions. *Science* **242,** 893–899 (1988).
74. Rohs, R. *et al.* Origins of specificity in protein–DNA recognition. *Annu. Rev. Biochem.* **79,** 233–269 (2010).

a *CAP-SELEX data analysis*



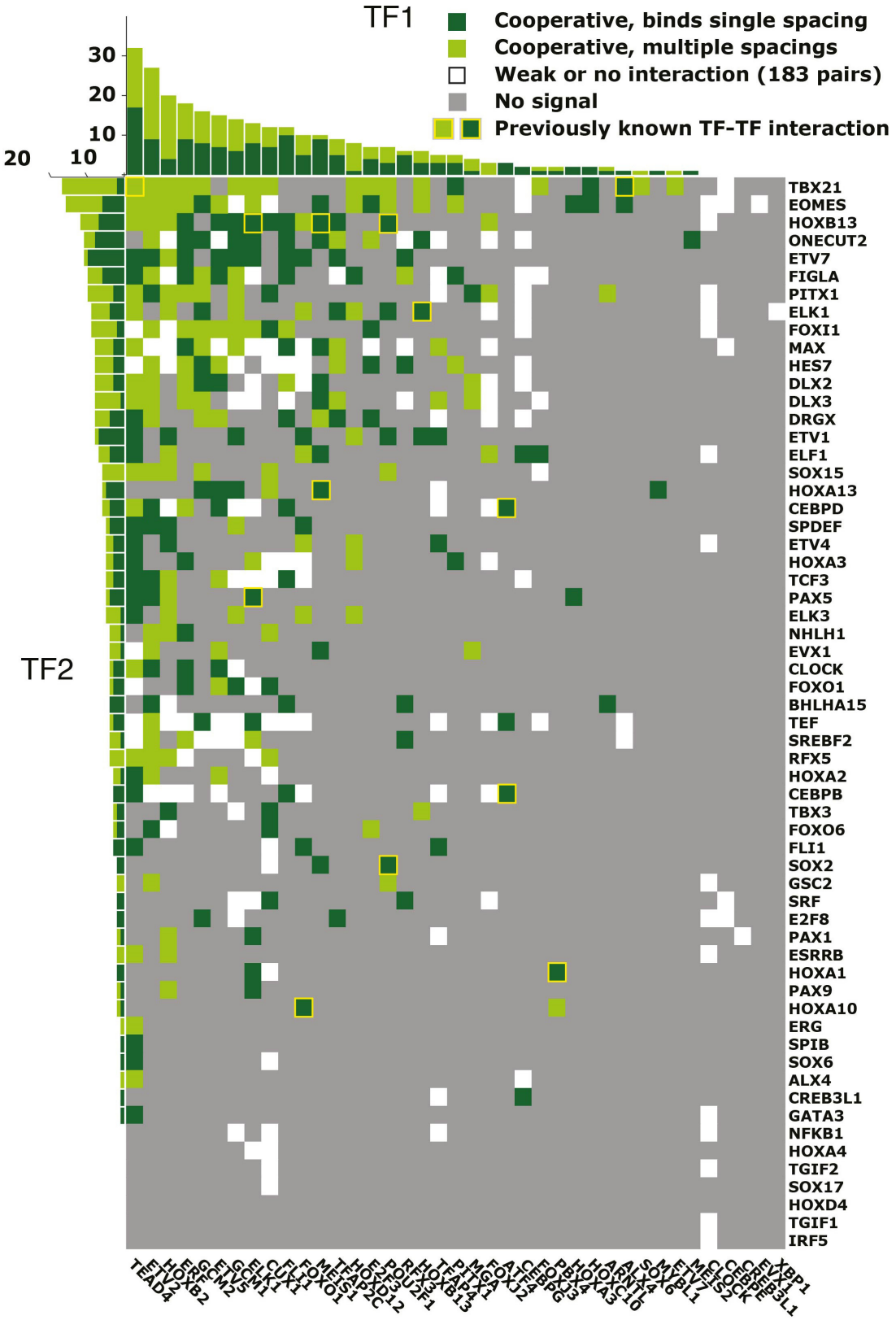
b *Comparison of CAP-SELEX PWMs to previous data*

TF pair	*CAP-SELEX model*	*Known consensus*	*Method*	*Ref.*
ATF4:CEBPG*		**CTGACGCAAT**	EMSA	**(26)**
ATF4:CEBPB		**CTGACGCAAT**	EMSA	**(66)**
HOXB13:MEIS1*		**GTCGTAAAACTGTCA**	EMSA	**(27)**
POU2F1:SOX2*		**CATTAGCATGACAAAGACA**	X-ray structure	**(28)**
ELK1:PAX5		**GCCACTGGAGCCCATCTCCGGCA**	EMSA	**(64)**

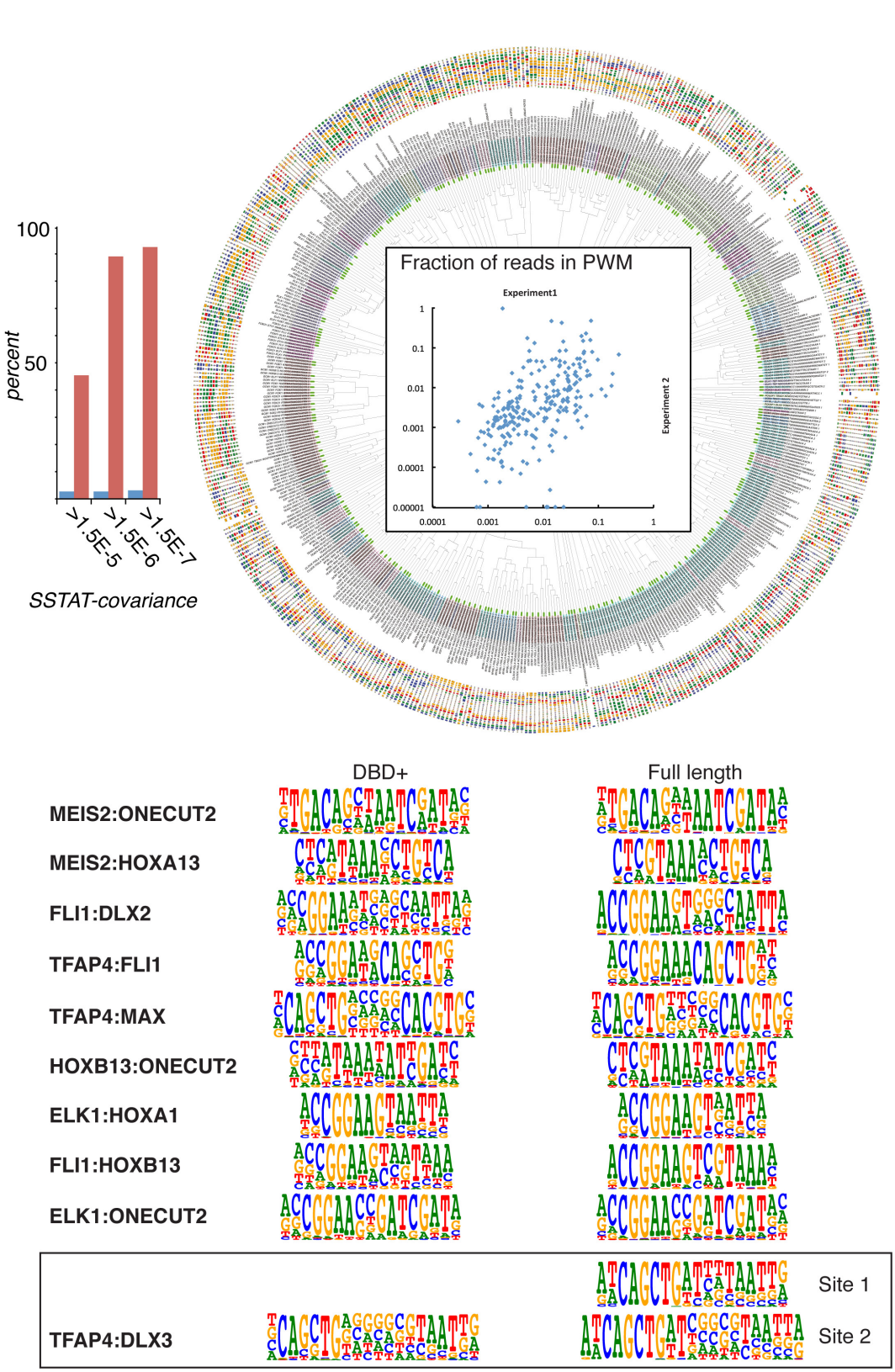
c *CAP-SELEX PWMs for TF pairs with a known protein-protein interaction*

TF pair	*CAP-SELEX model*	*Method*	
ALX4:TBX21		Mammalian two hybrid	**(37)**
TEAD4:TBX21		Mammalian two hybrid	**(37)**
POU2F1:HOXB13		Mammalian two hybrid	**(37)**
HOXB13:ELK1		Mammalian two hybrid	**(37)**
ATF4:CEBPD		Protein microarray	**(38)**
FOXO1:HOXA10		co-immunoprecipitation	**(65)**
HOXA13:MEIS1		Yeast two hybrid, co-immunoprecipitation	**(67)**

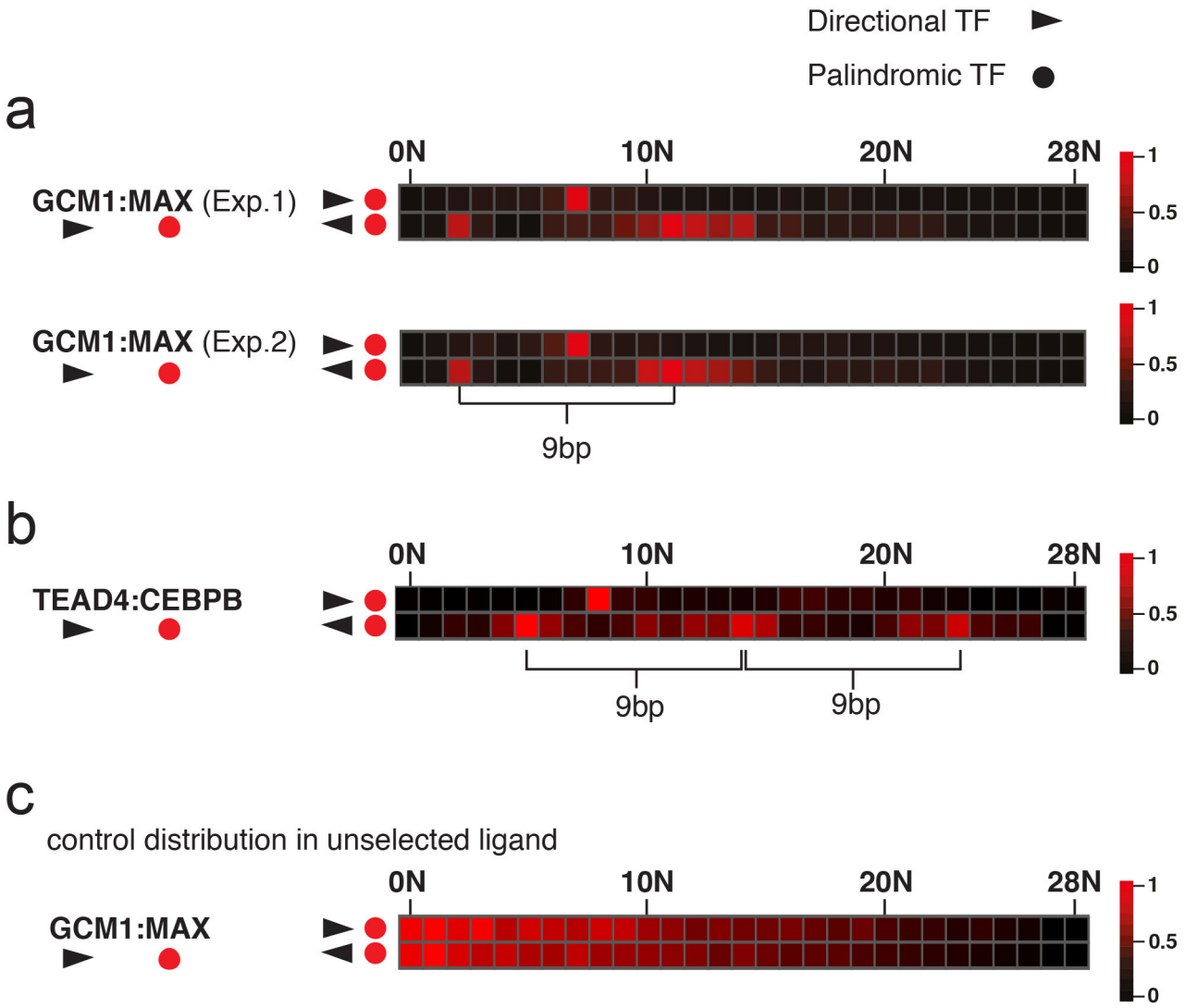
Extended Data Figure 1 | CAP-SELEX data analysis and comparison to previous data. a, Flowchart of CAP-SELEX data analysis. Left, a library of selection ligands with random sequences (yellow) is incubated with TFs. After CAP-SELEX, enriched individual TF motifs (1°; arrows) and composite motifs that are not simply combinations of the individual motifs (2°; green dots) are detected from the reads. To detect preferential spacings and orientations of the TF pair (3°), co-occurrence of the indicative 6-mer subsequences (arrowheads) are counted from the reads. The subsequences are then used to generate seeds for the PWM models (right). Heatmap (bottom right; scale divided by highest observed count) shows frequency of occurrence of the two 6-mers (CCGGAA, red arrowhead; CATTCC, black arrowhead) in all possible spacings (columns) and orientations (rows). Note that the 6-mer based approach cannot model the composite site, but identifies a strong case of cooperativity where the ERG 6-mer CCGGAA is followed by the TEAD4 6-mer CATTCC site with an 8 bp gap. Logo of the PWM for this site is also shown. **b**, Comparison between CAP-SELEX PWMs and previously characterized specificities for the indicated TF pairs. This method has been used previously and its references are also indicated. CAP-SELEX models also shown in Fig. 1 are indicated by asterisks. Note that four out of five of the CAP-SELEX models are similar to the previously identified consensus sequences. The exception is ELK1–PAX5 consensus, that matches poorly both the CAP-SELEX motif and individual motifs for ELK1 and PAX5 (not shown). **c**, CAP-SELEX PWMs for TF pairs known to interact at protein level. Method used to identify the protein–protein interaction and its reference are also shown[6,26–28,37,38,64–67].



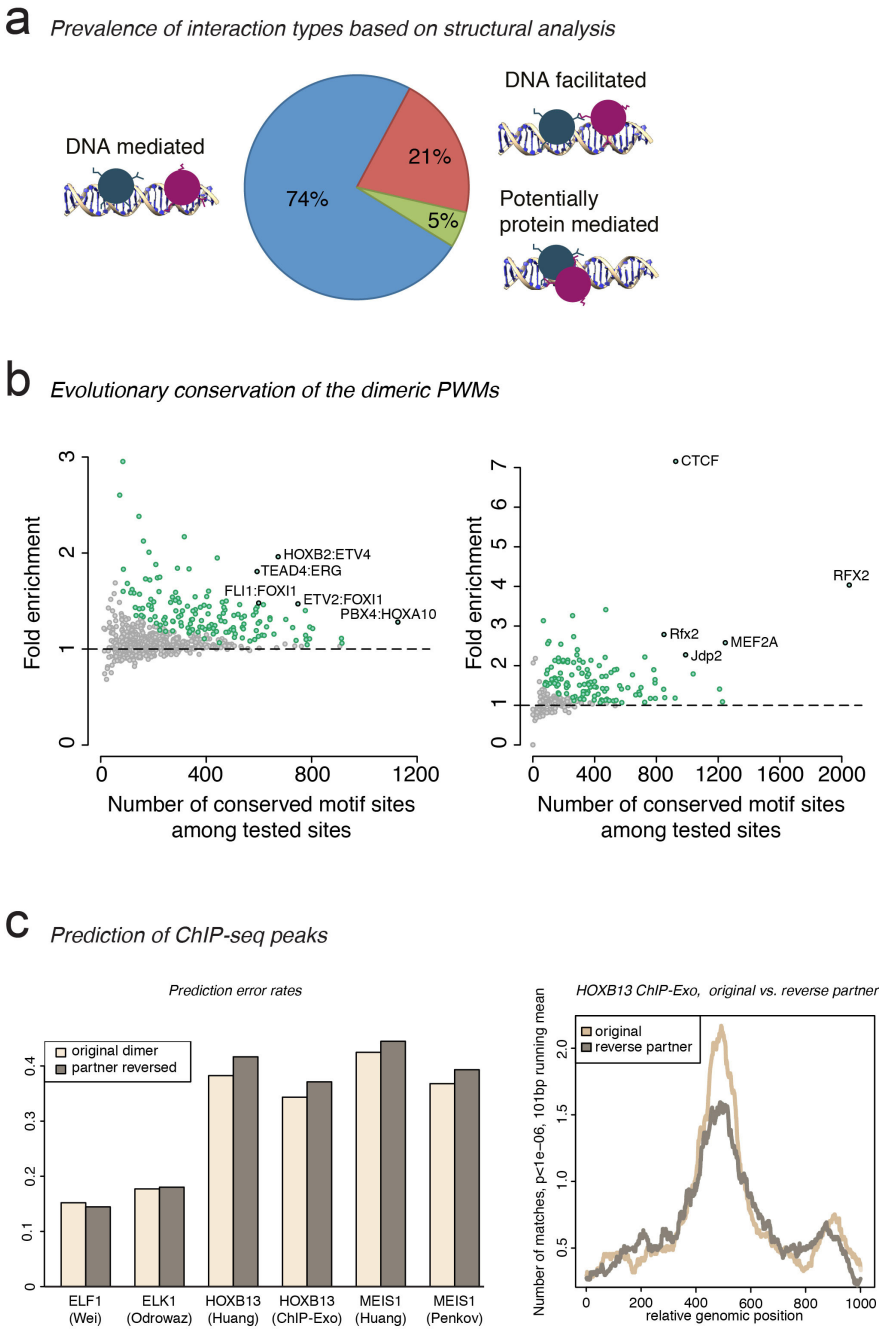
Extended Data Figure 2 | Pairwise interaction matrix between TFs. Columns indicate TF1 proteins, and rows TF2 proteins, subjected to the first and second affinity purifications, respectively. Pairs of TFs with a single spacing and orientation preference are indicated in dark green, and pairs with multiple preferred configurations in light green. White boxes indicate pairs that displayed weak or no interaction, and grey boxes cases where robust preference data was not recovered. Previously known interacting TF-pairs are indicated by a yellow outline (see Extended Data Fig. 1). Histograms show the counts for the interactions for each TF. Only TFs for which at least one clear interaction or independent binding was identified are included. The importance of including DNA in the interaction assay is highlighted by the fact that only four and five of the interactions detected are among those observed between 762 human or 877 mouse TF pairs identified using protein–protein interaction assays[37], or compiled from literature[36], respectively.



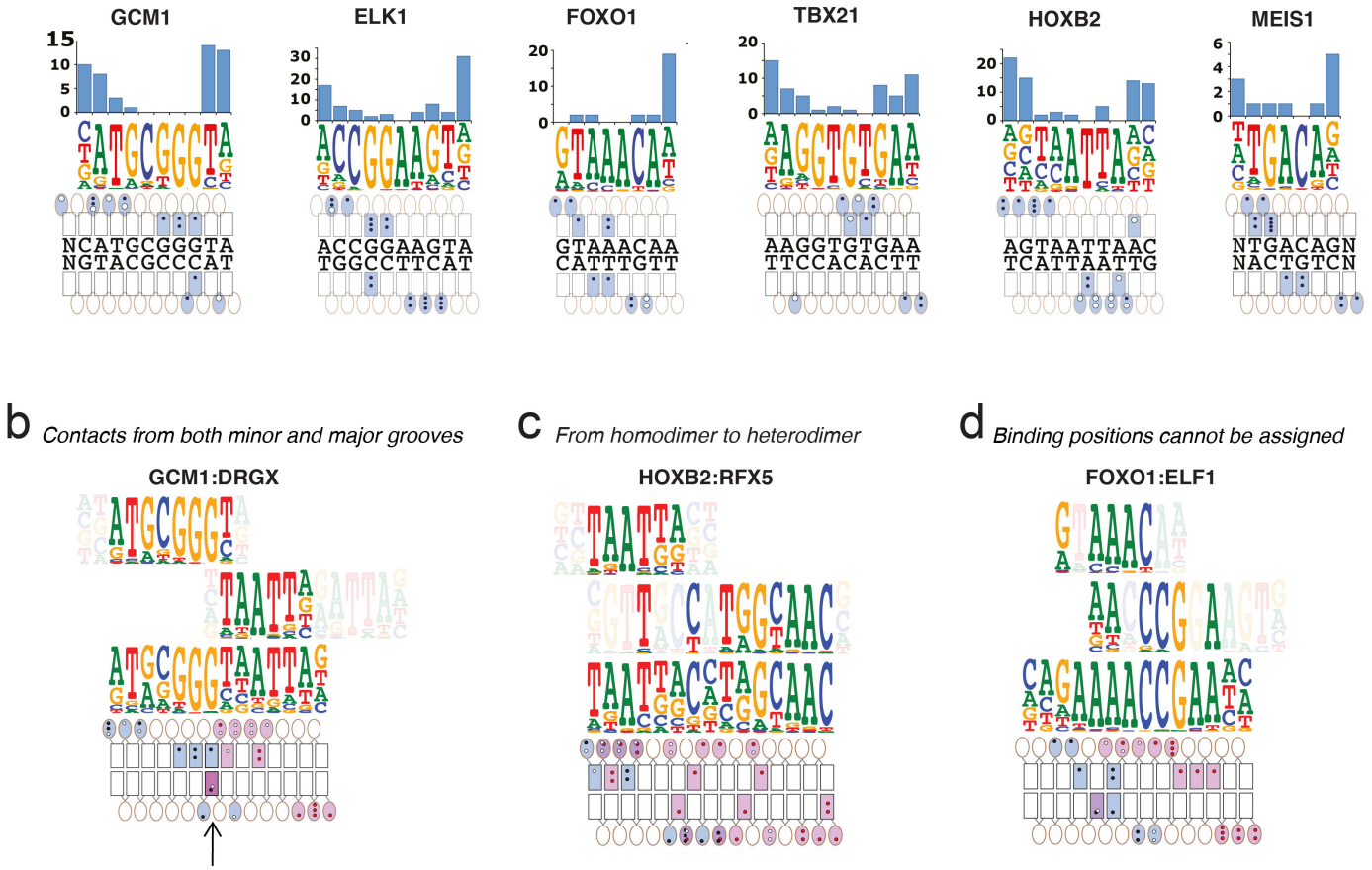
Extended Data Figure 3 | CAP-SELEX reproducibility. a, Replicate analysis of more than two hundred of the generated PWMs. The same seeds that had been used to generate PWMs for the primary experiments were used to seed new PWMs from the replicates. Left, red bars on the left show the percentage of the PWM pairs that are similar at the indicated cut-offs (measured as SSTAT covariance[8,48]). The highest threshold is the same used for identifying the dominating set of PWMs. Blue bars indicate fraction of all replicate PWMs that are similar using the same cut-off. Right, dendrogram and barcode logos of all PWM pairs. Plot in the middle shows fraction of reads included in the same models in replicate 1 and 2. **b**, Validation of the CAP-SELEX analysis using shortened TF constructs (DBD+) by HT-SELEX using full-length protein mixtures (full length). Note that the same orientation and spacing is preferred in all but one of the cases. In one case (bottom), full-length proteins show the highest preference to a different spacing than that observed in CAP-SELEX; even in this case, the second-most preferred spacing is the one identified using CAP-SELEX.



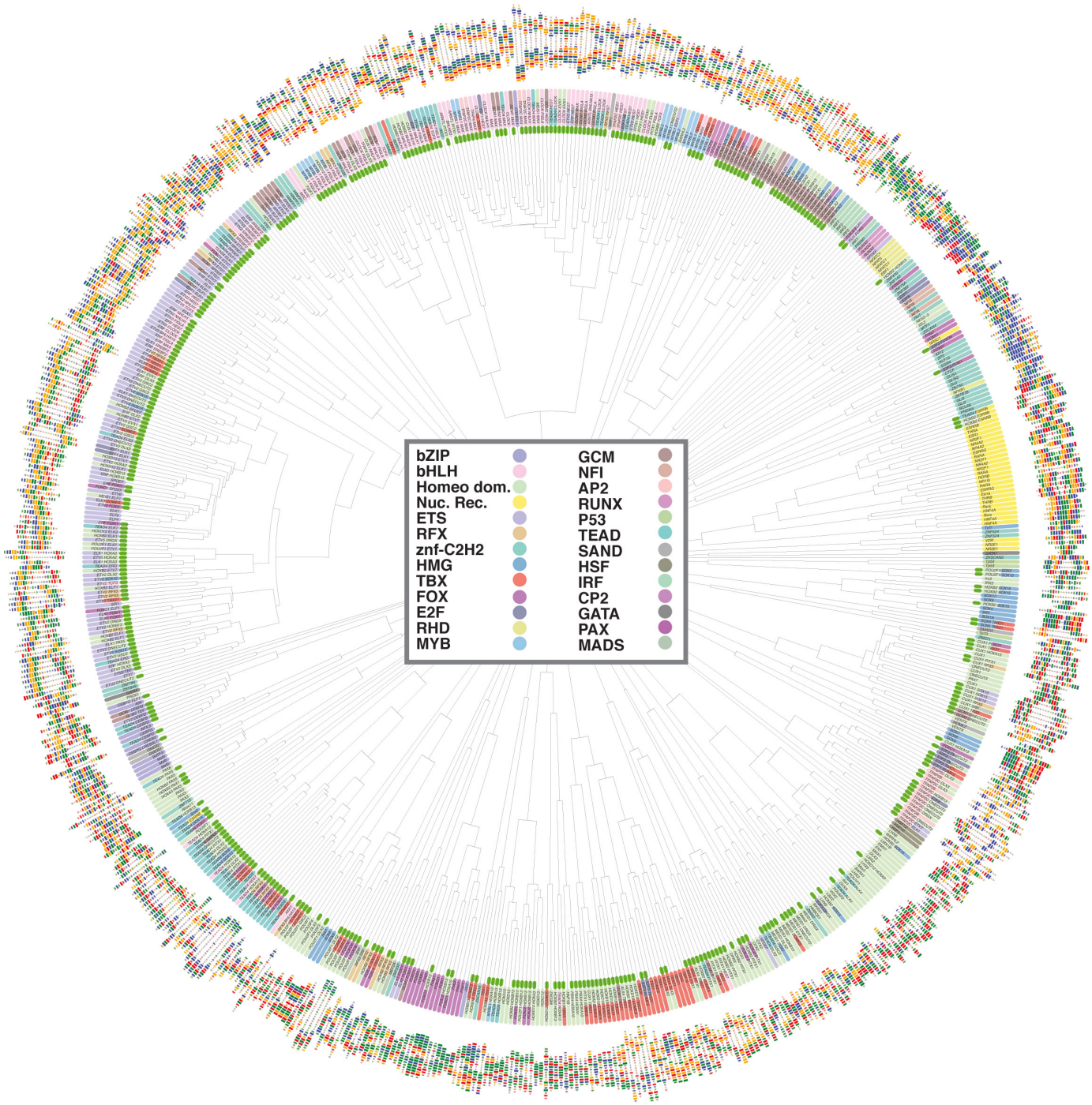
Extended Data Figure 4 | Long-range cooperativity. Many experiments where TFs bound sites that were relatively far apart showed preferential binding to sites that are separated by approximately nine to ten bases. Heatmap (maximum count set to 1) representations showing frequency of occurrence of the representative 6-mers for TF pairs in all possible spacings (columns) and orientations (rows). **a**, Replicate experiment of GCM1 (black arrowhead) and MAX (red ball) pair show very similar preference for cooperatively bound representative 6-mers (see Supplementary Table 1). While one of the orientations shows preference for a single spacing, the second has two preferentially recognized regions separated by ~9 bp. **b**, TEAD4–CEBPB pair shows a similar ~9 bp separation between three regions of preferred spacings (brackets). **c**, Very deep sequencing of the unselected input ligand does not show the same preference, instead counts decrease linearly as a function of gap length (due to decreasing number of available positions in the 40N random sequence). The mode of cooperativity seen in **a** and **b** appears similar to that reported by Kim *et al.*[17]. In addition to high-affinity sites, lower affinity spacings and orientations between TF pairs could be employed in fine-tuned transcriptional responses (see refs. 68, 69).



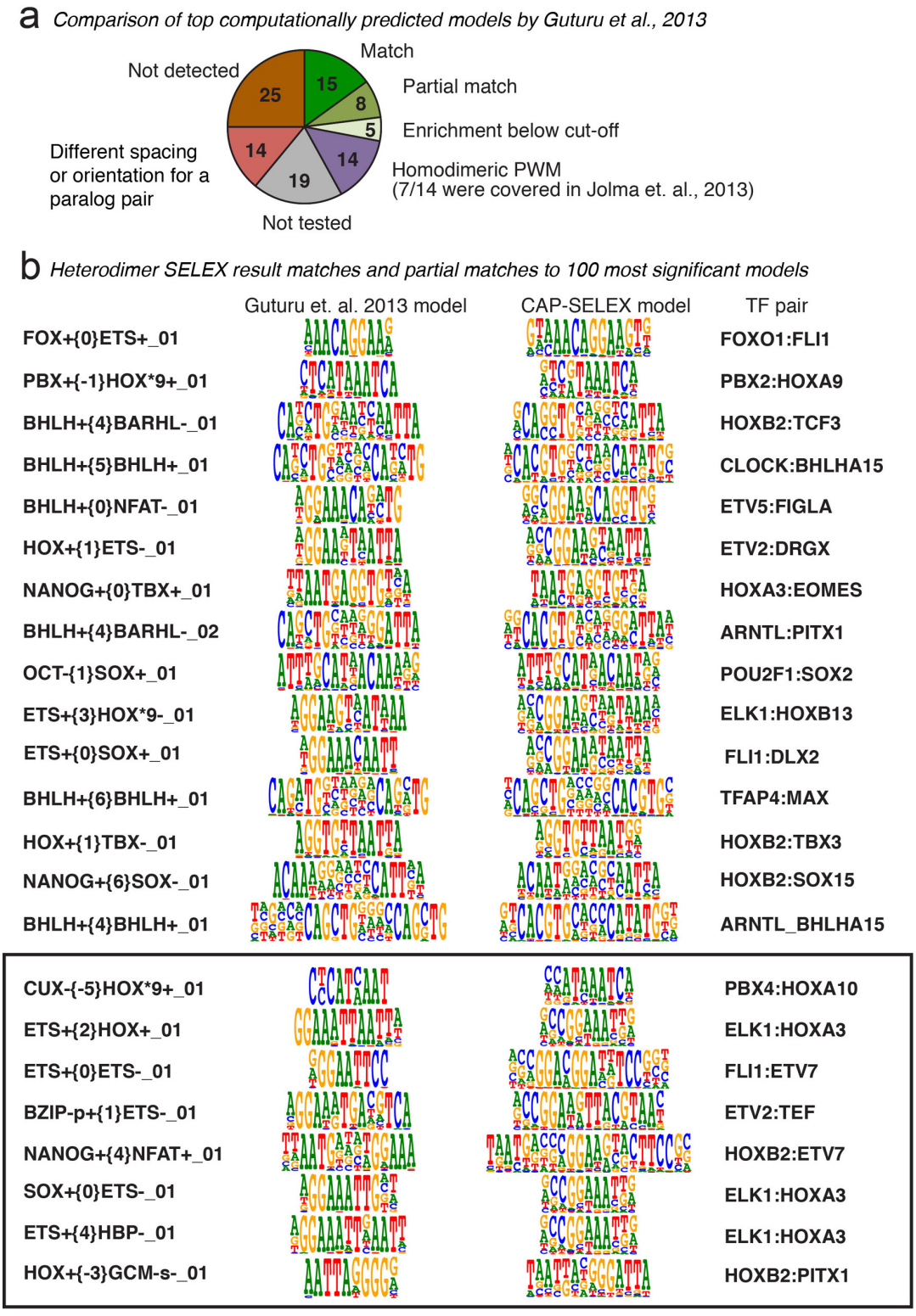
Extended Data Figure 5 | CAP-SELEX motifs are conserved and enhance prediction of *in vivo* peaks. a, Pie chart showing the frequency of DNA-mediated, DNA-facilitated and potentially protein–protein interaction mediated heterodimers in the CAP-SELEX data set. Cooperativity between TFs can result from direct contacts between the proteins (protein-mediated), DNA-facilitated protein contacts (DNA-facilitated) or arise indirectly from DNA-mediated interactions[17,34,39,40,70,71]. The last type of cooperativity is caused by the DBD binding-induced changes in DNA shape, and do not involve other domains or direct contact between the proteins[17,39,40]. The dimers were classified to DNA-mediated, DNA-facilitated and potentially protein–protein interaction mediated classes manually, based on structural models shown in Supplementary Data Set 2 . **b**, Conservation of the genomic sites recognized by the CAP-SELEX identified heterodimeric motifs (left) compared to monomeric and homodimeric sites identified by HT-SELEX (right, motifs from ref. 8). For each motif, ten thousand non-overlapping highest affinity sites within human constrained non-coding regions recognized by the motif or one its control motifs (see Methods for details) were selected and their conservation was tested. The fold enrichment (*y* axis), that is, the fraction of conserved sites among the motif sites divided by the fraction of conserved sites among the control motif sites, is shown as a function of the number of conserved motif sites among the top ten thousand sites (*x* axis). The motifs that are significantly conserved (multiple testing adjusted *P* value <0.05) are marked green. Five motifs with lowest *P* values are also indicated. Note that ~50% of the HT-SELEX and ~25% of the CAP-SELEX motifs are conserved above the significance threshold. **c**, Inclusion of heterodimeric motifs improves prediction of ChIP-seq peaks. Left, the error rate of prediction of ChIP-seq peak positions using either the monomer motifs and CAP-SELEX dimers (light grey), or monomer motifs and control motifs where the partner of the indicated TF is reversed but not complemented (dark grey) are shown. Note that inclusion of the correct heterodimeric motifs decreases the prediction error rate in the cases of HOXB13 and MEIS1. The relatively modest effect is likely due to the fact that only a subset of heterodimers were identified in our study, and that ChIP-seq peak positions are also influenced by other factors such as nucleosome binding and chromatin structure. Right, number of PWM matches as a function of distance from HOXB13 ChIP-exo peaks. Note that using the original heterodimer motifs clearly outperforms the control motifs.



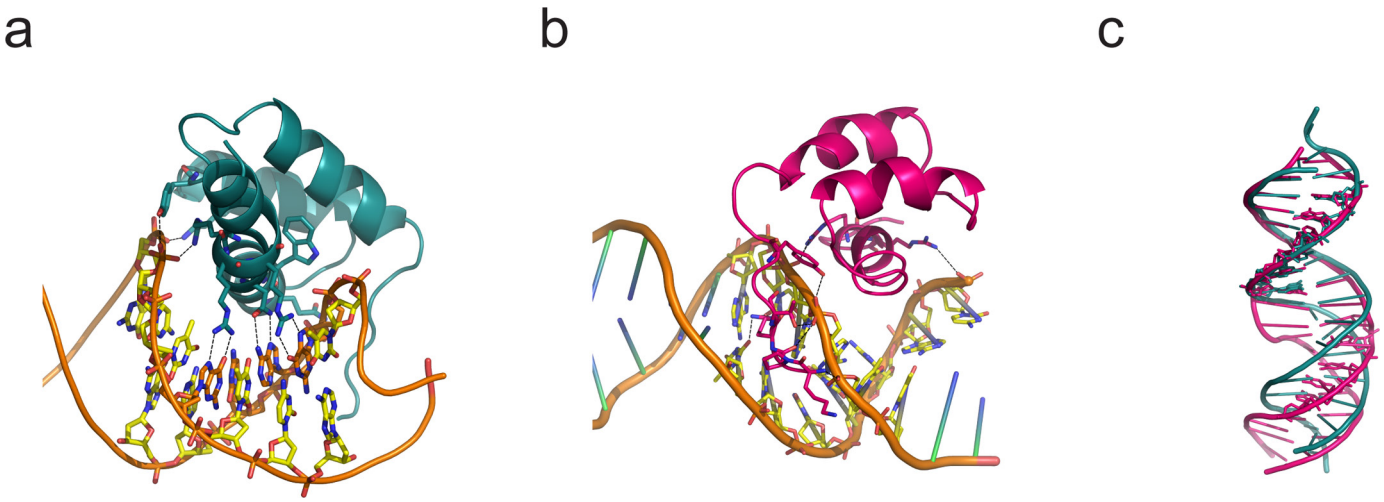
Extended Data Figure 6 | Heterodimers where the individual TF core recognition sites appear to overlap. a, Composite site formation alters specificity at bases flanking the core TF site. TFs often directly read specific 'core' sequence motifs via hydrogen bonding to DNA bases. The sequences flanking this core are commonly read indirectly, through contacts to the sugar and phosphate backbone of DNA[72–74]. The backbone contacts specify a preferred DNA conformation, which then leads to a preference of a sequence that is optimal for stacking interactions between consecutive base pairs (reviewed in ref. 74). Figure shows summary of base positions whose specificity is affected in all composite sites identified in this study for the indicated TFs. Note that the bases comprising the core motif that is recognized by direct hydrogen bonds to the DNA bases are not commonly affected by heterodimer formation. In contrast, specificity at flanking positions that are recognized by contacts to the sugar or phosphate backbone of DNA are commonly altered by binding of the heterodimer partner. Hydrogen bonds contacts were determined based on the indicated (refs 29 and 30) or homologous TF structures (see Supplementary Table 3). **b**, A base (arrow) can be contacted both from the side of the major groove (black dot; G contacted by GCM1) and the minor groove (white dot; C contacted by DRGX homeodomain). **c**, A TF that can bind to a homodimeric site appears instead to bind as a heterodimer. A composite site is shown where HOXB2 appears to form a heterodimer with a monomer of RFX5. **d**, In some cases, the binding positions of the TFs cannot be unambigiously assigned based on the composite recognition sequence. In **a**, the annotation of hydrogen bond contacts is as described in main Fig. 2; in **b**–**d**, the major groove contacts of the left and right TFs are indicated in black and red dots, respectively.



Extended Data Figure 7 | Specificities of individual TFs and heterodimer pairs. Dendrogram shows motif similarities between the representative heterodimer and monomer motifs. Heterodimer models are indicated by green bars. Barcode logos for each factor are also shown. Centre of dendrogram shows the colour key for the TF families.



Extended Data Figure 8 | Comparison of CAP-SELEX models to models inferred from conserved genomic sequences. a, Motifs that are very similar to the CAP-SELEX motifs are enriched and conserved. A previous study by Guturu *et al.*[20] made structural models for pairs of TFs to identify sterically possible configurations and predict sites that could be bound by such complexes. Enrichment of those target sites were then quantified in evolutionarily conserved noncoding regions over nonconserved control regions to infer putative target sites for cooperatively binding TFs. Pie chart shows comparison of top 100 most significant target sites predicted[20] to all heterodimeric PWMs generated in this study. 15 PWMs showed clear similarity to our heterodimeric PWMs (upper right, dark green slice), 8 were partially similar (green) and further 5 had enrichment for the site but under the threshold used in our study. We did not detect 25 motifs, and for 14 potential pairs, we identified a different spacing and orientation. This result is expected as we did not test all potential TF–TF pairs, and many TFs that bind to similar monomer sites prefer different dimer spacings and orientations. Finally, of the 100 Guturu *et al.*[20] top motifs, 33 were not analysed in our study (14 were homodimeric and no possible pair was tested for 19; for example, three of the pairs were predicted for pairs with a SMAD TF, and no SMAD TFs were tested in our study). **b**, Comparison of the 15 (top) and 8 (bottom, boxed) matching and partially matching PWMs, respectively.



Extended Data Figure 9 | Detailed view of MEIS1 and MEIS1–DLX3 structures. a, Contacts between MEIS1 (cyan) and DNA. **b**, Contacts between DLX3 (magenta) and DNA. **c**, Comparison of the DNA structures in MEIS1 homodimer (cyan) and MEIS1–DLX3 heterodimer (magenta). Note that the DNA bound to the heterodimer is more distorted.