LETTER

doi:10.1038/nature15545

Transcriptional regulators form diverse groups with context-dependent regulatory functions

Gerald Stampfel[1], Tomáš Kazmar[1], Olga Frank[1]†, Sebastian Wienerroither[1], Franziska Reiter[1] & Alexander Stark[1]

One of the most important questions in biology is how transcription factors (TFs) and cofactors control enhancer function and thus gene expression. Enhancer activation usually requires combinations of several TFs[1], indicating that TFs function synergistically and combinatorially[2,3]. However, while TF binding has been extensively studied, little is known about how combinations of TFs and cofactors control enhancer function once they are bound. It is typically unclear which TFs participate in combinatorial enhancer activation, whether different TFs form functionally distinct groups, or if certain TFs might substitute for each other in defined enhancer contexts. Here we assess the potential regulatory contributions of TFs and cofactors to combinatorial enhancer control with enhancer complementation assays. We recruited GAL4-DNA-binding-domain fusions of 812 *Drosophila* TFs and cofactors to 24 enhancer contexts and measured enhancer activities by 82,752 luciferase assays in S2 cells. Most factors were functional in at least one context, yet their contributions differed between contexts and varied from repression to activation (up to 289-fold) for individual factors. Based on functional similarities across contexts, we define 15 groups of TFs that differ in developmental functions and protein sequence features. Similar TFs can substitute for each other, enabling enhancer re-engineering by exchanging TF motifs, and TF–cofactor pairs cooperate during enhancer control and interact physically. Overall, we show that activators and repressors can have diverse regulatory functions that typically depend on the enhancer context. The systematic functional characterization of TFs and cofactors should further our understanding of combinatorial enhancer control and gene regulation.

We sought to characterize the potential regulatory contributions of different TFs to combinatorial enhancer control, that is, the regulatory functions of the TF proteins following DNA binding, regardless of their specific roles *in vivo*. We reasoned that such contributions could be best assessed using ectopic tethering assays in the context of DNA sequences that closely resemble active enhancers, ideally only lacking the input of a single TF. In particular, such a setup may allow the assessment of obligate combinatorial factors whose regulatory activities depend on partners and which would otherwise appear non-functional. We therefore developed enhancer complementation assays based on activator bypass experiments[4] used to test candidate TF or cofactor function in transcription control and to dissect promoters[5–10].

We mutated TF-binding-motif sequences within active enhancers to 'upstream activating sequence' (UAS) motifs for the GAL4 DNA-binding domain (GAL4-DBD), recruited 474 *Drosophila* TFs[11] via GAL4-DBD fusion proteins to the positions of the mutated motifs (enhancer context), and measured enhancer activities by luciferase assays in S2 cells, normalizing to GFP recruitment (Fig. 1a). Since expression and recruitment is standardized, the factors' regulatory functions can be assessed in a highly controllable manner independently of the factors' endogenous expression and DNA binding. Overall, the assays were highly reproducible: 75% of all data points had standard deviations (s.d.) $<10\%$, and 95% of the s.d. were $<18\%$ across four biological replicates.

We started with an enhancer that was highly active in *Drosophila* S2 cells and for which mutations of CGCG- or GATA-type motifs strongly reduced its activity[3] (Fig. 1b). We replaced either the CGCG- or the GATA-motifs with UAS motifs (Fig. 1a) and assessed which TFs restored enhancer function or repressed the remaining basal activities. Of the 474 TFs, 100 were activating ($\geq$1.5-fold compared to GFP; $P<0.05$ false discovery rate (FDR)-corrected for 474 tests) in the CGCG- and 84 in the GATA-context (Fig. 1c), including TFs that recognize the CGCG- and GATA-motifs, respectively (Extended Data Table 1). This compares to 77 TFs that activated on their own (that is, when recruited to UAS motifs outside an enhancer context), suggesting that TF function might be context-dependent. Indeed, 46 TFs activated the CGCG context at least 1.5-fold ($P<0.05$) more strongly than the GATA context, even though both contexts were derived from the same enhancer (Fig. 1d and Extended Data Fig. 1). To test if native untagged TFs recapitulate these results, we chose Ets at 21C (Ets21C), Deformed (Dfd), and Hairy/E(spl)-related with YRPW motif (Hey) that preferentially activated the CGCG context to different extents (Fig. 1d). We replaced the CGCG- and GATA-motifs with binding sites for these TFs and expressed the untagged TFs (Fig. 1e). This activated the mutant enhancers in a manner consistent with the results from GAL4-DBD-mediated recruitment: Ets21C and Dfd activated only the CGCG context, while Hey activated both (CGCG 1.3-fold more highly), confirming the similarity and context-dependency of these TFs' regulatory functions.

Intrigued by the context-dependency of some TFs even within a single enhancer, we decided to include more diverse regulatory contexts (Extended Data Fig. 1). We created 19 motif-mutant enhancer contexts for different types of TF motifs and different enhancers with broad, cell-type-specific[3], or hormone-inducible[12] activities. We also added five contexts consisting of UAS sites and core promoters specific towards developmental or housekeeping enhancers, respectively[13].

Nearly half of all TFs (42%) were activating and most (93%) of the remaining 276 TFs were repressing in at least one of the 24 contexts ($\geq$1.5-fold; $P<0.05$ FDR-corrected for 24×474 tests), suggesting that most TF-fusion proteins were functional. Many TFs had similar regulatory effects across the 24 contexts, suggesting that they might be functionally equivalent. We grouped all TFs into 15 clusters using unsupervised spectral clustering (Fig. 2a and Supplementary Table 1) and confirmed that these clusters are robust to bootstrapping and reproducible when using independent biological replicates (Extended Data Fig. 2). This revealed clusters of diverse regulatory functions (Fig. 2b and Extended Data Fig. 3), including cluster 8 with TFs that activated in most contexts (global activators) such as Antennapedia, Sox14 and Sox15, Clock (Clk), and Zelda, and clusters 3, 5 and 7 with global repressors (for example, Snail, Runt, Engrailed and Kruppel). These TFs seemed to dominate or override other regulatory cues, consistent with their ability to function in isolation. TFs of other clusters were only

[1]Research Institute of Molecular Pathology (IMP), Vienna Biocenter (VBC), Dr. Bohr-Gasse 7, 1030 Vienna, Austria. †Present address: Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstraße 108, 01307 Dresden, Germany.

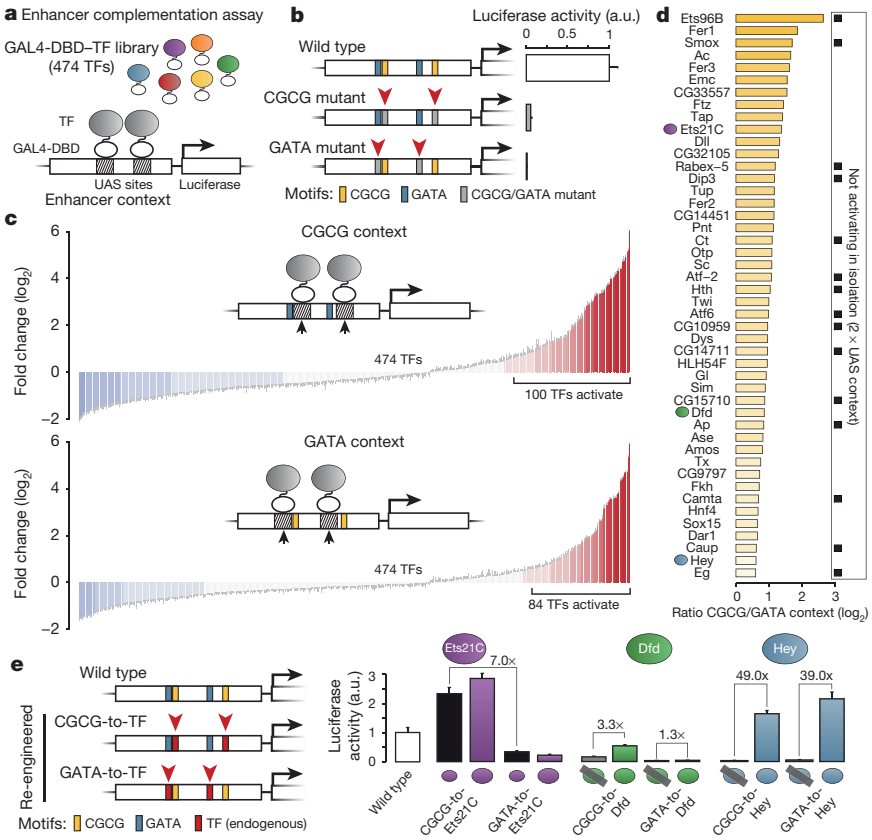


Figure 1 | Enhancer complementation assays for 474 TFs. a, Schematic overview. **b**, Activity of an enhancer and motif-mutant variants (data from ref. 3). a.u., arbitrary units. **c**, Enhancer complementation assays for CGCG- and GATA-contexts (normalized luciferase values for 474 TFs; red: activation, blue: repression). **d**, Preferential activation of CGCG- versus GATA-context (≥1.5-fold, FDR-corrected $P < 0.05$). Black boxes, TFs inactive on their own; colour, TFs tested in **e**. **e**, Validation of context-dependent TFs (details see text). Luciferase activities (firefly/*Renilla*) with (colours) or without (grey and black (*Ets21C* is expressed in S2 cells)) co-transfection of the respective untagged TF. Error bars show standard deviations ($n = 4$, biological replicates).

weakly active in the contexts tested, and, notably, the TFs of some clusters were context-dependent. For example, cluster 10 TFs preferentially activated the housekeeping core promoter and might constitute factors of a distinct transcriptional program[13], including Myb-interacting protein 120 (Mip120) and CG6813, the cluster's strongest activator (21-fold; $P = 4.4 \times 10^{-4}$). In contrast, cluster 1 TFs preferentially activated hormone-receptor contexts, that is, when recruited to ecdysone receptor (EcR)-binding sites in enhancers inducible by the insect steroid hormone ecdysone. Examples are Twist, Reversed polarity, Pointed, and other developmental TFs.

Intrigued by TFs that preferentially activated hormone-receptor contexts, we selected four such TFs from clusters 1 and 15, Ets96B, Helix loop helix protein 4C (HLH4C), Atonal (Ato), and Glass (Gl), and asked whether replacing the EcR motif with the motifs of these TFs would activate the enhancer in a TF-dependent but hormone-independent manner. This was indeed the case for all four TFs, and the effect was specific to the combination of motif and enhancer context (Fig. 2c), suggesting that these TFs might contribute regulatory functions equivalent to the activated EcR.

To assess the TF clusters independently of our approach, we asked whether they were enriched in Gene Ontology (GO) categories or in protein sequence features such as Pfam domains or short peptide motifs. Indeed, many such features were differentially distributed between the clusters ($P < 0.01$; empirical FDR = 0.1) and each cluster was enriched for at least one such feature (Fig. 2d and Extended Data Fig. 4; Supplementary Table 2). As expected, amino acid repeats known to mediate activation (for example, poly-glutamine[14]) or repression (for example, poly-alanine[15]) were enriched in activating clusters (1, 8) and repressing clusters (3, 5, 7), respectively. However, of several activating clusters, only cluster 1 was enriched in GO categories relating to development, suggesting a preferential use of cluster-1-type TFs during developmental gene regulation, which presumably relates to these TFs' dependence on partner TFs, enabling combinatorial control. Similarly, only repressing cluster 7 but not 3 or 13 was enriched in GO categories relating to Notch signalling, cardiocyte differentiation, or morphogenesis, suggesting that repression might occur through various means that are differently employed *in vivo*. Indeed, the three repressing clusters also differed in the enrichment of peptide motifs known to bind the co-repressors C-terminal binding protein (CtBP; cluster 7) or Sin3A (cluster 3), suggesting a functional association between the TFs in these clusters and the respective co-repressors (see below).

These results show that the different TF clusters, obtained solely based on the TFs' context-dependent regulatory functions, differ in several other aspects, which lends independent support to the clustering. It also suggests that the respective TFs are differentially employed *in vivo* (for example, during development), and that their distinct functions might arise through the recruitment of different types of cofactors (for example, CtBP versus Sin3A).

To assess the regulatory activities and the clustering in different cell types, we tested 171 TFs (9 to 17 TFs from each cluster) across six contexts in Kc167, BG3 and ovarian somatic cells derived from embryos, larvae and adult ovaries, respectively. These cell types differ increasingly from S2 cells in gene expression, enhancer activities, and the enhancers' motif signatures[3], yet TF activities were remarkably similar: all 18 pairwise comparisons had Pearson correlation coefficients (PCCs) ≥0.5 and 15 had PCCs ≥0.8 (all $P < 1 \times 10^{-3}$; Extended Data Fig. 5).

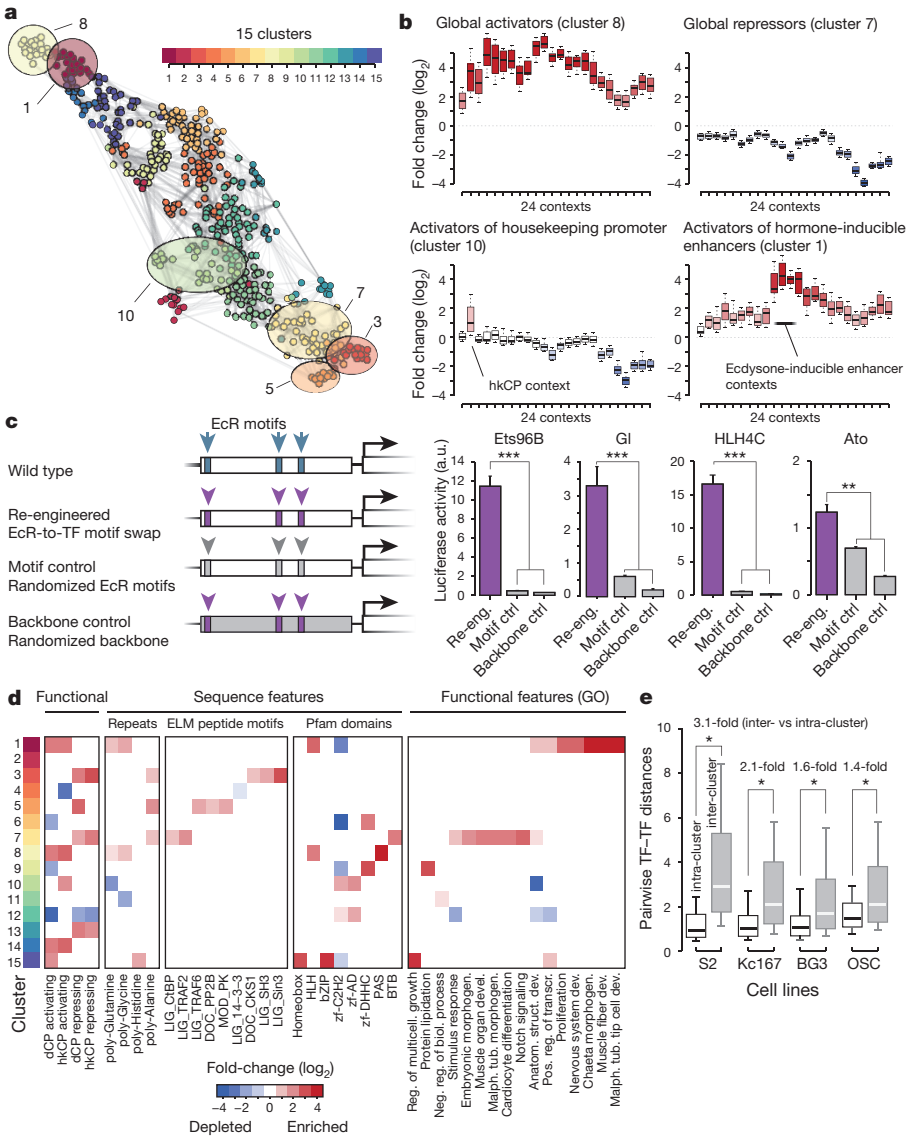


Figure 2 | TFs have diverse regulatory functions. a, 15 TF clusters (see text for highlighted clusters). **b**, Normalized luciferase values across 24 contexts (selected from Extended Data Fig. 3). **c**, Validation of hormone-context-preferential TFs (luciferase activities (firefly/*Renilla*) for re-engineered enhancers (re-eng.; purple) and controls (grey; see schematic)). Error bars show standard deviations ($n = 4$, biological replicates). $^{**}P < 1 \times 10^{-2}$; $^{***}P < 1 \times 10^{-3}$. **d**, Enrichments for TFs that activate or repress the 4×UAS-dCP or 4×UAS-hkCP contexts ('Functional'), protein-sequence features and GO-categories; significant ($P < 0.01$; empirical FDR = 0.1) enrichments, red; and depletions, blue; others, white; see Extended Data Fig. 4 and Supplementary Table 2 for details and all data. **e**, Pairwise distances between activity profiles in Kc167, BG3 and OSC cells support functional TF clusters (all empirical $P < 10^{-6}$, indicated with single asterisks).

Moreover, the TFs' activity profiles across the 6 contexts support the original clustering in each cell type: pairs of TFs from within original clusters were more similar than pairs across clusters (≥1.4-fold and $P < 1 \times 10^{-6}$ for all four cell lines; Fig. 2e). Intrigued by these results, we tested the ability of 107 *Drosophila* TFs (90) and cofactors (17) to activate transcription in human HeLa cells and found a good quantitative agreement across all factors (PCC = 0.74; Extended Data Fig. 6). These results suggest that many TFs function predominantly, but not completely, independently of cell type. We note that alternative splicing and post-translational modifications (for example, downstream of cellular signalling pathways) probably alter and diversify the regulatory functions of individual TFs.

The regulatory functions of TFs are generally mediated through transcriptional cofactors, which typically lack DNA-binding domains and are recruited to enhancers by TFs. To assess whether cofactor functions are similarly diverse or potentially more uniform, we cloned 338 putative cofactors from diverse protein families (Fig. 3a and Supplementary Table 3) and tested their activating and repressing functions in all 24 contexts using GAL4-DBD-mediated recruitment as for TFs (Fig. 3b).

Most cofactors (80%) were sufficient to activate or repress transcription in at least one context (≥1.5-fold; $P < 0.05$ after FDR correction for 24 × 338 tests), and the activities of well-studied factors matched their known functions (Fig. 3c–e): for example, P300 (also known as Nejire) strongly activated transcription in all contexts, as did the histone-methyltransferase Lost PHDs of Trr (Lpt) of the Set1/COMPASS-like complex, and the Mediator subunits MED15 and MED25, while the co-repressors CtBP, Sin3A and CoRest were strongly repressing in all contexts. Other cofactors had context-specific functions, including Chromator (Chro), TBP-associated factor 4 (Taf4) and Trithorax-related (Trr), which preferentially activated the housekeeping core promoter (Trr was even repressing in all other contexts). Chro is part of the non-specific lethal (NSL) complex which activates genes involved in cell proliferation and DNA replication[16] and Taf4 is important at TATA-less promoters[17]. Similar to the corresponding TFs above, these cofactors might be part of a dedicated housekeeping regulatory program[13].

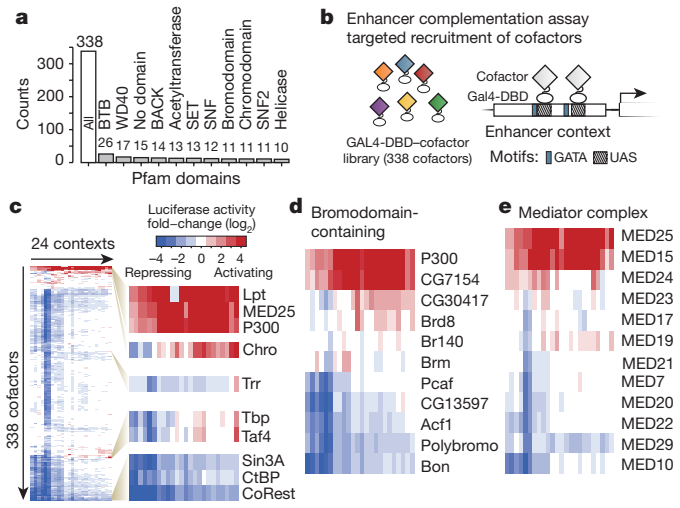


Figure 3 | Transcriptional cofactors can be sufficient for activation and repression and have context-dependent regulatory functions. **a**, Pfam-domain content of the 338 putative cofactors tested. **b**, Scheme depicting enhancer complementation assays for cofactors. **c**, Overview of the regulatory activities for cofactors across all 24 contexts (bi-clustered 338 × 24 heat map) and zoom for selected cofactors. **d**, **e**, Diverse regulatory activities for bromodomain-containing cofactors (**d**) and subunits of the Mediator complex (**e**) (see Extended Data Fig. 7 for additional complexes).

Notably, different members of a single complex or domain family frequently had different regulatory functions (Fig. 3d, e and Extended Data Fig. 7), cautioning against the transfer of annotations based on these grounds. For example, the activities of bromodomain-containing (BRD) cofactors (Fig. 3d) range from the strongly activating P300 and CG7154, the orthologue of human BRD7 and BRD9, to the context-dependent CG30417, and to Polybromo that was strongly repressing in most contexts, consistent with its previous implication in transcriptional repression[18]. Similarly, Mediator subunits MED15 and MED25 were strongly activating in most contexts, MED23 and MED24 were context-dependent, and MED29 was repressing, consistent with the function of human MED29[19] (Fig. 3e). This suggests that different TFs might interact with the Mediator complex through distinct subunits, or that complexes with variable composition and function might exist. Consistently, MED15 and MED25 interact directly with strong activators (for example, GAL4 and VP16[20,21]) and MED23 and MED24 are involved in signal-dependent and hormone-induced transcription, respectively[22,23].

The activating and repressing effects across 24 contexts were highly similar for many TFs and cofactors (Fig. 4a), which provided a means to infer functional associations (Fig. 4b). As expected, P300, Lpt, MED15 and MED25 were assigned to globally activating TFs (cluster 8) and context-dependent cofactors to context-dependent TFs (for example, Mip120 and Bsh to cluster 10 and Chro to cluster 14). Interestingly however, the globally repressing cofactors Sin3A and CtBP were assigned to different clusters of repressing TFs (cluster 3 versus 7), in agreement with the differential enrichment of peptide motifs involved in Sin3A and CtBP recruitment (Fig. 2d). Many of the assignments are consistent with known physical interactions, including the interaction between Chro and Pzg[24,25] or CtBP and Sna[26]. Indeed, the assignments were enriched for interactions reported in large-scale studies that used yeast two-hybrid assays[27] or co-affinity purification[25,28] (between 1.4- and 3.0-fold; all $P \leq 0.05$; Fig. 4c). In addition, the human orthologues of TF–cofactor pairs interacted 1.8-fold ($P = 0.025$) more frequently than expected[29]. These results suggest that the TF–cofactor assignments reflect functional associations and predict that the enhancer activation obtained by TF recruitment

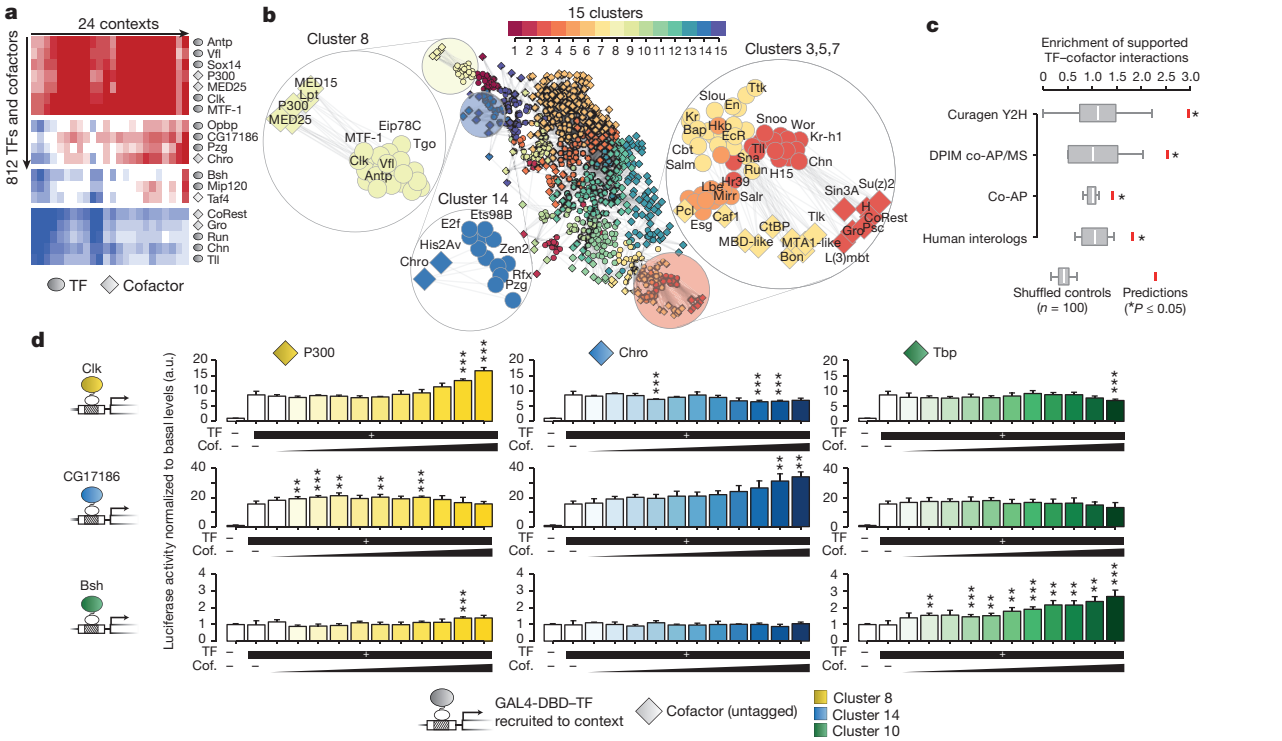


Figure 4 | TF–cofactor assignments through functional similarities. **a**, TFs (ovals) and cofactors (diamonds) show similar regulatory activities and localize next to one another in a bi-clustered 812 × 24 heat map (selection shown; see Supplementary Table 1 for all activities). Coloured as in Fig. 3c–e. **b**, TF clusters (as in Fig. 2a) and assigned cofactors. Highlighted are clusters 8, 14 and 3, 5, 7 and assigned cofactors. **c**, TF–cofactor assignments are enriched for physical interactions (red bars) compared to shuffled assignments, for which the box-plots indicate the 10th, 25th, 50th, 75th and 90th percentiles; *hypergeometric $P \leq 0.05$. **d**, Boosting of Clk (top), CG17186 (middle) and Bsh (bottom) induced enhancer activities by untagged P300 (left), Chro (centre) and Tbp (right). Error bars show standard deviations ($n = 4$, biological replicates). $^{**}P < 1 \times 10^{-2}$; $^{***}P < 1 \times 10^{-3}$.

should be boosted by the assigned but not by unrelated cofactors, even if the cofactors are not tethered via GAL4-DBD[30] (Fig. 4d). Indeed, in this experimental setup[30], the activation by Clk-recruitment (cluster 8) was further boosted by increasing amounts of untagged P300, but not by Chro or Tbp. In contrast, Chro specifically boosted CG17186 (cluster 14) and Tbp specifically boosted Bsh (cluster 10), consistent with the respective assignments.

Enhancer complementation assays provide a unique annotation and categorization of TFs and cofactors based on their regulatory functions, independent of the factors' endogenous roles and complementary to previous classifications through genetics, sequence comparisons, or genomics (for example, ChIP-seq). For many factors, including 266 putative TFs and cofactors ('CG' genes; Extended Data Fig. 8), our work provides the first functional characterization. All data are available at http://factors.starklab.org.

The existence of equivalence groups among TFs and cofactors with diverse context-dependent functions, even amid activators and repressors, has profound implications for our understanding of transcriptional gene regulation: while some enhancers might be controlled predominantly by individual activators, others may rely on specific combinations of distinct regulatory functions that are complementary and each insufficient for activation. It is therefore conceivable that different types of enhancers are controlled through non-overlapping sets of TFs and cofactors, enabling separate transcriptional programs even within individual cells (for example, ref. 13). The approach and categorization presented here provide a framework to dissect the molecular and biochemical nature of these functions and the mechanisms by which cooperativity at enhancers is established and transcriptional activation of target core promoters is achieved. Understanding these mechanisms will be crucial at a time when enhancer function and its control by TFs and cofactors are becoming increasingly central to our understanding of gene regulation in development and disease and the focus of novel therapeutic strategies.

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 9 January; accepted 2 September 2015.
Published online 9 November 2015.

1. Arnone, M. I. & Davidson, E. H. The hardwiring of development: organization and function of genomic regulatory systems. *Development* **124,** 1851–1864 (1997).
2. Han, K., Levine, M. S. & Manley, J. L. Synergistic activation and repression of transcription by *Drosophila* homeobox proteins. *Cell* **56,** 573–583 (1989).
3. Yáñez-Cuna, J. O. *et al.* Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. *Genome Res.* **24,** 1147–1156 (2014).
4. Ptashne, M. & Gann, A. Transcriptional activation by recruitment. *Nature* **386,** 569–577 (1997).
5. Cheng, J. X., Gandolfi, M. & Ptashne, M. Activation of the Gal1 gene of yeast by pairs of 'non-classical' activators. *Curr. Biol.* **14,** 1675–1679 (2004).
6. Keung, A. J., Bashor, C. J., Kiriakov, S., Collins, J. J. & Khalil, A. S. Using targeted chromatin regulators to engineer combinatorial and spatial transcriptional regulation. *Cell* **158,** 110–120 (2014).
7. Yuasa, Y. *et al. Drosophila* homeodomain protein REPO controls glial differentiation by cooperating with ETS and BTB transcription factors. *Development* **130,** 2419–2428 (2003).
8. Papadopoulos, D. K. *et al.* Functional synthetic *Antennapedia* genes and the dual roles of YPWM motif and linker size in transcriptional activation and repression. *Proc. Natl Acad. Sci. USA* **108,** 11959–11964 (2011).
9. Arnosti, D. N., Barolo, S., Levine, M. & Small, S. The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development* **122,** 205–214 (1996).
10. Brodu, V., Mugat, B., Fichelson, P., Lepesant, J. A. & Antoniewski, C. A UAS site substitution approach to the *in vivo* dissection of promoters: interplay between the GATAb activator and the AEF-1 repressor at a *Drosophila* ecdysone response unit. *Development* **128,** 2593–2602 (2001).
11. Hens, K. *et al.* Automated protein-DNA interaction screening of *Drosophila* regulatory elements. *Nature Methods* **8,** 1065–1070 (2011).
12. Shlyueva, D. *et al.* Hormone-responsive enhancer-activity maps reveal predictive motifs, indirect repression, and targeting of closed chromatin. *Mol. Cell* **54,** 180–192 (2014).
13. Zabidi, M. A. *et al.* Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518,** 556–559 (2015).
14. Gerber, H. P. *et al.* Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science* **263,** 808–811 (1994).
15. Galant, R. & Carroll, S. B. Evolution of a transcriptional repression domain in an insect Hox protein. *Nature* **415,** 910–913 (2002).
16. Lam, K. C. *et al.* The NSL complex regulates housekeeping genes in *Drosophila*. *PLoS Genet.* **8,** e1002736 (2012).
17. Wright, K. J., Marr, M. T. & Tjian, R. TAF4 nucleates a core subcomplex of TFIID and mediates activated transcription from a TATA-less promoter. *Proc. Natl Acad. Sci. USA* **103,** 12347–12352 (2006).
18. Martens, J. A. & Winston, F. Recent advances in understanding chromatin remodeling by Swi/Snf complexes. *Curr. Opin. Genet. Dev.* **13,** 136–142 (2003).
19. Wang, Y. *et al.* IXL, a new subunit of the mammalian Mediator complex, functions as a transcriptional suppressor. *Biochem. Biophys. Res. Commun.* **325,** 1330–1338 (2004).
20. Mittler, G. *et al.* A novel docking site on Mediator is critical for activation by VP16 in mammalian cells. *EMBO J.* **22,** 6494–6504 (2003).
21. Bryant, G. O. & Ptashne, M. Independent recruitment *in vivo* by Gal4 of two complexes required for transcription. *Mol. Cell* **11,** 1301–1309 (2003).
22. Ihry, R. J. & Bashirullah, A. Genetic control of specificity to steroid-triggered responses in *Drosophila*. *Genetics* **196,** 767–780 (2014).
23. Kim, T. W. *et al.* MED16 and MED23 of Mediator are coactivators of lipopolysaccharide- and heat-shock-induced transcriptional activators. *Proc. Natl Acad. Sci. USA* **101,** 12153–12158 (2004).
24. Gan, M., Moebus, S., Eggert, H. & Saumweber, H. The Chriz-Z4 complex recruits JIL-1 to polytene chromosomes, a requirement for interband-specific phosphorylation of H3S10. *J. Biosci.* **36,** 425–438 (2011).
25. Guruharsha, K. G. *et al.* A protein complex network of *Drosophila melanogaster*. *Cell* **147,** 690–703 (2011).
26. Nibu, Y., Zhang, H. & Levine, M. Interaction of short-range repressors with *Drosophila* CtBP in the embryo. *Science* **280,** 101–104 (1998).
27. Giot, L. *et al.* A protein interaction map of *Drosophila melanogaster*. *Science* **302,** 1727–1736 (2003).
28. Rhee, D. Y. *et al.* Transcription factor networks in *Drosophila melanogaster*. *Cell Rep.* **8,** 2031–2043 (2014).
29. Murali, T. *et al.* DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for *Drosophila*. *Nucleic Acids Res.* **39,** D736–D743 (2011).
30. Amelio, A. L. *et al.* A coactivator trap identifies NONO ($p54^{nrb}$) as a component of the cAMP-signaling pathway. *Proc. Natl Acad. Sci. USA* **104,** 20314–20319 (2007).

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

Acknowledgements We are grateful to K. Hens and B. Deplancke for sharing the TF entry clones, J. O. Yáñez-Cuna for help designing the enhancer contexts, and O. Bell, J. Brennecke, L. Cochella and S. Westermann for comments on the manuscript. We thank IMP/IMBA services, especially H. Scheuch, R. Heinen and Z. Dzupinkova, for technical support, and A. Posekany and A. Aszódi for advice on multiple-testing correction. Deep sequencing was performed at the CSF Next-Generation Sequencing Unit (http://csf.ac.at). The Stark group is supported by a European Research Council (ERC) Starting Grant (no. 242922) awarded to A.S., Boehringer Ingelheim GmbH, and the Austrian Research Promotion Agency (FFG).

Author Contributions G.S. and A.S. conceived the project. G.S. and O.F. cloned the cofactors and the GAL4–DBD fusions. G.S. and F.R. performed the luciferase assays in *Drosophila* cells and S.W. in HeLa cells. T.K. and G.S. conducted the bioinformatics analyses. G.S., T.K. and A.S. wrote the manuscript.

Additional Information All data are available at http://factors.starklab.org. The next-generation sequencing data have been deposited at the NCBI Sequence Read Archive (SRA) under the accession SRS806429. The cofactor Gateway entry clones and other plasmids are available from Addgene (http://www.addgene.org/Alexander_Stark/). 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 A.S. (stark@starklab.org).

METHODS

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

Cloning of N-terminal GAL4-DBD-tagged TF and cofactor library. For TFs, gateway-compatible entry clones (Invitrogen) containing the open reading frames (ORFs) lacking stop codons were obtained from ref. 11. *Drosophila Act5C*-promoter driven expression clones were created using the Gateway system. The TF ORFs were shuttled into the GAL4-DNA binding domain (DBD) containing destination vector pAGW-GAL4-DBD (cloned as described below) by mixing 100 ng of TF entry clone, 100 ng of pAGW-GAL4-DBD and 0.7 µl of LR clonase II enzyme mix (Invitrogen). The identities of all TF entry clones have been confirmed by Sanger sequencing using the primers 5′-CCCAGTCACGACGTTG-3′ and 5′-CACAGGAAACAGCTATG-3′. Note that we tested the full-length transcription factors, including their DBDs, as *trans*-activating and DNA-binding functions might not always reside in entirely separate protein domains. While this implies that the fusion proteins might bind via the TFs' DBDs in addition to the GAL4-DBD mediated recruitment, this does not influence the results of the assay: the assay itself measures transcriptional activation independently of where TF binding occurs and we expect that the TFs' DBDs have at most minor effects on binding strengths as the GAL4-DBD binds to DNA already very strongly.

For cofactors, we compiled a list of 338 cofactors based on several criteria. We included proteins containing Pfam domains typical for transcriptional cofactors (for example, HAT, HDAC, SET, Chromo, Bromo), proteins which are part of chromatin modifying or remodelling complexes or part of complexes associated with RNA polymerases (for example, SAGA, Polycomb, TFIID, Mediator), and *Drosophila* proteins which are homologues of mammalian chromatin-associated proteins (Supplementary Table 3). We amplified the cofactor ORFs from cDNA using oligonucleotides containing Gateway-compatible attB-sites (5′-GGGGACAAGTTTGTACAAAAAAGCAGGCTTC-3′ and 5′-GGGACCACTTTGTACAAGAAAGCTGGGTC-3′) for subsequent entry clone creation. The primer sequences have been chosen to be as close as possible to an annealing temperature of 60 °C which we calculated using the formula $T_{ann} = \frac{64.9 + 41 \times (cG + cC - 16.4)}{(cA + cT + cG + cC)}$ with cA, cT, cG and cC being the number of adenines, thymines, guanines and cytosines, respectively. The full list of resulting primer sequences (lacking the attB sequences) is listed in Supplementary Table 3; for 18 of the cofactors no primer sequences are available because we obtained these entry clones from ref. 11 categorized as TFs but manually re-categorized them as cofactors based on their annotation in FlyBase[31] or their protein domain content[32]. For cDNA generation, RNA was isolated from S2 cells and reverse transcribed as described in ref. 33. For PCR amplification, KOD and KOD XL DNA (Merck Millipore) and KAPA HiFi (KAPA) polymerases were used according to manufacturer's specifications. We created Gateway entry clones by mixing 1 µl of PCR reaction, 100 ng of pDONR221, and 1 µl of BP clonase II enzyme mix (Invitrogen). The identities and correctness of all entry clones have been ensured using Sanger and next-generation sequencing (see below) and we deposited them at Addgene (http://www.addgene.org/Alexander_Stark/). The cofactor ORFs were then shuttled to the *Drosophila Act5C*-promoter driven destination vector pAGW-GAL4-DBD as described for TFs.

Verification of cofactor clones by Sanger and next-generation sequencing. The insert flanks of all obtained cofactor entry clones have been Sanger-sequenced and automatically checked to cover the TSS and TTS of one of the isoforms annotated by FlyBase. All entry clones passing additional manual visual inspection using BLAT and the UCSC genome browser have been subjected to further verification by next-generation sequencing as follows. A pool of 100–300 entry clones corresponding to a total of 5 µg DNA solved in 50 µl TE buffer was sonicated (duty cycle, 20%; intensity, 5; 200 cycles per burst; time, 90 s) to 200–400 bp using a S220 Focused-ultrasonicator (Covaris) as described in ref. 11. The fragmented plasmid pool was then prepared for deep sequencing using the Illumina DNA Sample Prep kit and sequenced using a HiSeq2000 (Illumina) producing 50-nt reads. The resulting reads have been assembled and analysed using PrInSes-C[34]. All insert sequences not starting with ATG, containing a stop codon or a frameshift were immediately rejected. All sequences with less than five mutations leading to non-synonymous amino acid changes were immediately accepted. The remaining sequences were translated, aligned against the respective protein sequence, and manually decided. The next-generation sequencing reads have been deposited at the NCBI Sequence Read Archive (SRA) under the accession SRS806429; the PrInSes-C-generated full-length transcript sequences are available at http://factors.starklab.org and in Supplementary Data 1, and the cofactor Gateway entry clones from Addgene (http://www.addgene.org/Alexander_Stark/).

Cloning of destination vector pAGW-GAL4-DBD. We cloned a destination vector to conveniently create vectors expressing N-terminally V5- and GAL4-DBD-tagged TFs and cofactors under the control of the *Drosophila Act5C* promoter using the Gateway cloning system. pAGW-GAL4-DBD was cloned by amplifying the GAL4-DBD from pBPGUw[35] using one oligonucleotide containing the V5-tag (peptide sequence MGKPIPNPLLGLDST) 5′-TCTGATATCATGGGGAAGCCAATCCCTAATCCCCTTCTGGGACTCGACTCTACCGGCGGCTCTATGAAGCTACTGTCTTCTATCGAACA-3′ and the oligonucleotide 5′-TATACCGGTGGCCGCCGCCCGACGATACAGTCAACTGTCTTTGAC-3′. Amplification was performed using KOD Polymerase (Merck Millipore) according to the manufacturer's instructions. The resulting PCR product was digested using EcoRV and AgeI and ligated into pAGW (*Drosophila* Gateway Vector Collection), which was digested using the same enzymes, thereby replacing eGFP with V5-GAL4-DBD.

Cloning of luciferase reporter vectors. We created Gateway-compatible (Invitrogen) destination vectors to conveniently clone reporter vectors for different regulatory contexts based on firefly luciferase transcribed from a housekeeping core promoter (hkCP; promoter of ribosomal gene *RpS12*[13]) or a developmental core promoter (dCP; *Drosophila* synthetic core promoter (DSCP) derived from *Eve*[35]).

We created the destination vector attR_dCP_luc by digesting pGL4.26 (Promega) with FseI and BglII and ligating a fragment containing DSCP and luc+, thereby replacing the minimal promoter and luc2 with DSCP-luc+. We digested the resulting vector with KpnI and BglII and ligated a fragment containing the attR Gateway cassette, yielding attR_dCP_luc. We created two hkCP-driven destination vectors containing a Gateway cassette either upstream (attR_hkCP_luc) or downstream (hkCP_luc_attR) of the luciferase reporter gene by using the plasmid pGL3 (Promega) as a basis and replacing the SV40 promoter with the promoter of *RpS12* as described in ref. 13. The resulting vector was digested using either KpnI and BglII (to create attR_hkCP_luc) or AfeI (to create hkCP_luc_attR); in both cases, we amplified a Gateway attR cassette using oligonucleotides containing the respective restriction sites, and digested and ligated it into the digested plasmid.

All enhancers, motif mutant contexts and other motif or backbone mutant variants were either PCR amplified with primers containing attB Gateway sites or ordered as synthesized fragments (IDT), shuttled into entry clones using TOPO or BP Clonase II (both Invitrogen), and shuttled into the luciferase destination vectors using the LR clonase II enzyme mix (Invitrogen) by mixing 1 µl of PCR product or synthesized DNA solved in TE buffer, 100 ng of destination vector and 0.7 µl of LR clonase II enzyme mix (Invitrogen).

We used a modified version of pRL-TK (Promega) to normalize the firefly signal for transfection efficiency and cell number. Ubi-RL has been created by cloning a region upstream of the gene *Ubi-p63E* (chr3L: 3901760-3902637) upstream of the *Renilla* luciferase gene in reverse orientation using NheI and BglII.

***Drosophila* cell culture.** S2 cells, derived from embryos[36], were obtained from Life Technologies and grown in Schneider's *Drosophila* Medium (Life Technologies 21720-024) supplemented with 10% FBS (Sigma F7524) and 1% penicillin/streptomycin (Life Technologies 15140-122) grown in T75 flasks (ThermoScientific 156499) at 27 °C and passaged every 2–4 days. BG3 neuroblast-like cells, derived from larvae[37], were obtained from the *Drosophila* Genomics Resource Center (DGRC) and grown in Schneider's *Drosophila* Medium supplemented with 10% FBS, 1% penicillin/streptomycin, and 10 µg ml^{-1} Insulin (Sigma-Aldrich I1882) in T75 flasks at 27 °C and passaged every 3–4 days. Kc167 cells, derived from embryos[38], were obtained from DGRC and grown in M3/BPYE Medium containing 5% FBS and 1% penicillin/streptomycin in T75 flasks at 27 °C and passaged every 2–3 days. Ovarian somatic cells (OSCs), derived from adult ovaries[39], were obtained from the laboratory of J. Brennecke and grown in Shields and Sang M3 Insect Medium (Sigma-Aldrich S8398) supplemented with 10% FBS, 1% insulin, 1% glutathione, 1% fly extract, and 1% penicillin/streptomycin in T75 flasks at 27 °C and passaged every 2–3 days. All cell lines used are regularly checked for mycoplasma contamination.

Transfections of *Drosophila* cell lines. S2 cell transfections were performed using jetPEI (peqlab 13-101-40N). Four hours before transfection, 30,000 cells (30 µl of a 10^6 cells per ml suspension) were seeded in clear polystyrene 384-well plates (ThermoScientific 164688). For each transfection, we used 30 ng firefly luciferase reporter plasmid, 3 ng *Renilla* luciferase expressing plasmid Ubi-RL, and 3 ng GAL4-DBD-TF/cofactor or GAL4-DBD-GFP fusion protein expressing plasmid. Beforehand, we assayed the effects of using different amounts of GAL4-DBD fusion protein expressing plasmid and chose 3 ng (Extended Data Fig. 9). The DNA solution containing 36 ng DNA in 5 µl TE buffer was filled up to 15 µl using sterile 150 mM NaCl (polyplus) and prepared in 96-well plates. Transfection reagent (15 µl total: 13.95 µl 150 mM NaCl, 1.05 µl jetPEI) was added to each well of the 96-well plates and mixed rigorously. After 30 min incubation at 25 °C, cells were transfected in quadruplicates by transferring each transfection mix four times (6 µl each) to four adjacent wells of a 384-well plate containing the seeded cells. Luciferase assays were performed after 48 h of growth at 27 °C. Handling the transfection mixes and all subsequent pipetting steps have been performed using a Bravo Automated Liquid Handling Platform (Agilent). Kc167, BG3, and

OSC cell transfections were performed using jetPEI in the same way as described above for S2 cells with the exception of transfection reagent composition: 15 µl total containing 14.1 µl 150 mM NaCl and 0.9 µl jetPEI.

HeLa cell culture and transfections. Human HeLa cells (gift from the laboratory of J. M. Peters) were grown in DMEM medium (Gibco 52100-047) supplemented with 10% heat-inactivated FBS, 1% penicillin/streptomycin and 2 mM L-glutamine (Sigma G7513) in T75 flasks at 37 °C in an atmosphere of 95% air and 5% carbon dioxide. All cell lines used are regularly checked for mycoplasma contamination. We performed HeLa cell transfections using a self-prepared 1 mg ml^{-1} PEI (25,000 MW, Polysciences 23966) stock solution in PBS (pH adjusted to pH 4.5 and sterile filtered). On the day before transfection we seeded 30 µl of a suspension containing 4,000 HeLa cells in medium (DMEM, 10% FBS, penicillin/streptomycin) into each well of a 384-well plate. Three microlitres of a PEI/DMEM mix (0.24 µl PEI filled to a total of 4.5 µl using DMEM without FBS and penicillin/streptomycin and incubated at room temperature for 5 min) were added to 3 µl of a DNA/DMEM mix (44.5 ng firefly luciferase reporter vector, 4.45 ng TF expression vector (created using pAGW-CMV_ GAL4-DBD, see below) and 4.45 ng pRL-CMV vector for transfection normalization (Promega #E2261) in DMEM without FBS and penicillin/streptomycin. The resulting DNA/PEI mix in DMEM was incubated at room temperature for 30 min and subsequently added to the seeded cells. We performed cell lysis and luciferase assays using the Promega dual-luciferase reporter assay system (Promega E1910) according to the manual.

We created the Gateway destination vector pAGW-CMV_GAL4-DBD by replacing the *Drosophila Act5C* promoter in pAGW-GAL4-DBD with a region containing the CMV enhancer and the T7 promoter amplified from pRL-CMV using the primers 5′-CGACAGATCTTCAATATTGGCCATTAGCCATAT-3′ and 5′-GGTGGCTAGCCTATAGTGAGTCGTATTA-3′.

Luciferase assays. Dual-luciferase assays were performed using self-prepared substrate solutions (D-Luciferin and Coelenterazine have been obtained from GoldBio LUCK-250 and pjk-Gmbh 102111) and lysis buffer as described in ref. 40. For cell lysis, the supernatant was removed and 30 µl of lysis buffer added and incubated gently shaking for 30 min. Ten microlitres of the cell lysates were transferred to black 384-well plates for luminescence assays (Nunc MaxiSorp, Sigma-Aldrich P6491-1CS). All pipetting steps have been performed using a Bravo Automated Liquid Handling Platform (Agilent). Luminescence was measured after adding 20 µl of each substrate, for firefly and *Renilla* luciferase respectively, using a Biotek Synergy H1 plate reader coupled to a plate stacker.

Luciferase data analysis and plots. We normalized all firefly luciferase signals to the signal of *Renilla* luciferase to control for transfection efficiency and cell number (the relative luciferase signal). We then further normalized all relative luciferase signals for TF- and cofactor-GAL4-DBD transfections to relative luciferase signals obtained for GAL4-DBD-GFP transfections (fold-change over GFP). We assessed statistical significance by two-sided unpaired *t*-tests on the two sets of quadruplicate relative luciferase signals (GAL4-DBD-TF/COF versus GAL4-DBD-GFP). Throughout the paper, 'activation' was defined as a fold-change $\geq$1.5 ($P < 0.05$), and 'repression' was defined as a fold-change $\leq$1/1.5 ($P < 0.05$), both compared to the signal for GAL4-DBD-GFP. We corrected the *P* values for multiple testing using the Benjamini and Hochberg method as implemented in R (p.adjust with method 'BH' or its alias 'fdr'). All statistical calculations and graphical displays, if not stated otherwise, have been performed using version 2.15.3 of the R software suite[41].

TF cluster feature enrichment analysis. Enrichment analyses have been performed for each of the 15 clusters and for 6 types of features. To first obtain a coarse functional characterization of the clusters, we assessed the enrichments and depletions of TFs which are able to activate or repress a developmental (dCP) or housekeeping (hkCP) core promoter on their own ($\geq$1.5-fold activation or repression ($P < 0.05$), both compared to the signal for GAL4-DBD-GFP when tested on a context comprised of UAS sites upstream of a developmental core promoter (4×UAS-dCP) or a housekeeping core promoter hkCP (4×UAS upstream hkCP)). Homopolymeric amino acid repeat motifs have been *de novo* discovered using MEME[42] (version 4.8.1, *q*-value threshold of 1×10^{-5}) in TFs that activated or repressed on their own outside enhancer contexts (tested in the 4×UAS dCP context; $\geq$1.5-fold; $P < 0.05$). Pfam domain[32] signature matches in the *Drosophila* proteome have been generated using hmmer[43] (version 3.0b3, *e*-value threshold of 0.01). Eukaryotic Linear Motifs[44] (ELM; version 08/2014) were matched to the amino acid sequences of the tested TF protein isoforms, after masking the TFs' Pfam. Additionally, Gene Ontology[45] (GO) annotations, and gene expression patterns in the *Drosophila* embryo as annotated by ref. 46 (IMAGO) have been subjected to enrichment and depletion analyses.

To control for multiple testing, we empirically determined false-discovery rates (FDRs) for the different hypergeometric *P* values. For this, we repeated the feature enrichment analyses 1,000 times, each after randomly shuffling the TF-to-cluster assignments, and recorded the best (that is, most significant) *P* values. We then adjusted the original *P* values such that only 10% of the 1,000 random controls reached the *P* values of the original data (FDR < 10%). Following this protocol, we separately adjusted the FDR cut-off for each cluster (15) and feature type (ELM, MEME, Pfam, GO, IMAGO).

Validation with the TFs' endogenous motifs. To assess if tethering via the GAL4-DBD reflects the different TFs' regulatory functions when bound to their endogenous motifs, we selected two sets of TFs, three TFs that preferentially activated the CGCG- versus the GATA-context (Fig. 1e) and four TFs that preferentially activated the hormone-receptor contexts; Fig. 2c). We replaced each UAS site in the enhancer mutant contexts S2-1 CGCG, S2-1 GATA, and Nhe2 EcR[3,12] (which also corresponds to an endogenous TF motif in the wild-type enhancers, for example, the EcR motif for the hormone contexts) with a sequence corresponding to the consensus motif of the respective TF as reported in refs 47, 48. (Dfd: CTTAATGA, Hey: CAGCCGACACGTGCCCC, Ets21C: ATTTCCGGT, Ato: AACAGGTGG, Ets96B: ACCGGAAGTAC, Gl: ATTTCAAGAATA, HLH4C: AAAAACACCTGCGCC). The enhancer rescue constructs were synthesized by IDT, shuttled into the luciferase reporter vector attR_dCP_luc using the Gateway system and tested in luciferase assays in S2 cells exactly as described above.

TF–cofactor association assays. To assess potential functional associations of assigned TFs and cofactors, we followed the strategy from ref. 30, recruiting TFs via GAL4-DBD and providing untagged cofactors. For this, we chose contexts in which the different TFs (Clk of cluster 8, Bsh of cluster 10, and CG17186 of cluster 14) were active (4×UAS-dCP for Clk and 4×UAS-upstream-hkCP for Bsh and CG17186). We prepared DNA mixes to be transfected containing 29 ng firefly luciferase reporter plasmid, 3 ng *Renilla* luciferase expressing plasmid Ubi-RL, 1 ng (Bsh and CG17186) or 0.5 ng (Clk) of GAL4-DBD–TF fusion protein expressing plasmid and an increasing series of untagged cofactor expressing plasmid (0 ng, 0.003 ng, 0.006 ng, 0.012 ng, 0.023 ng, 0.047 ng, 0.094 ng, 0.188 ng, 0.375 ng, 0.75 ng, 1.5 ng, 3 ng). We kept the total amount of transfected plasmid DNA constant at 36 ng for all experiments using a GFP-expressing plasmid. To clone the expression plasmids for the untagged cofactors and GFP, we used the Gateway-compatible vector pAW (*Drosophila* Gateway Vector Collection). The remaining experimental procedure and analysis was performed as described above.

Transcription factor clustering, visualization, and assignment of cofactors to transcription factors. We clustered the 474 TFs based on the $\log_2$-transformed fold-change values (TF over GFP) from all 24 contexts. First, we standardized all contexts and constructed a *k*-nearest-neighbour graph ($k = 15$). We used the Euclidean distance as distance measure as it reflects both the variation of the enhancer activity profile across contexts and the effect sizes within each context; that is, it is able to discriminate between strong and weak activators and repressors even if they vary similarly across the 24 contexts. Next, we took a symmetrized ($A + A^T$) adjacency matrix of this graph and solved multiclass spectral clustering as described in ref. 49 and implemented in the Python package scikit-learn[50]. In order to decide about the number of clusters and to assess the clustering validity, we analysed the clustering stability upon bootstrapping the data set[51]. In order to visualize the data, we mapped the data onto a plane by a specialized nonlinear dimensionality reduction technique (t-SNE)[52]. The algorithm provides the visualization by mapping data points close in the original space to nearby locations in the plane, preserving the local structure. We extended the *k*-nearest-neighbour graph to include cofactors by comparing the $\log_2$-transformed fold-change values (cofactor over GFP) of cofactors and TFs ($k = 5$, Euclidean distance). The locations of the cofactors in the visualization were obtained from spring layout.

TF candidate recovery of enhancer mutants. We know that UAS sites in the enhancer mutant contexts most probably replace binding sites that are functional[3] but we do not know which TFs bind them *in vivo*. In order to check whether we recover these positive controls in the enhancer mutants, we took all the TFs expressed in S2 cells (RPKM > 1) (ref. 53) for which motifs are known[3]. We scanned the wild-type enhancer sequences (S2-1-wt, S2-2-wt, S2-3-wt, Ubi-1-wt, Ubi-2-wt, Ubi-3-wt) for motif matches with $P < 9.76 \times 10^{-4}$ (1/4,096) using an in-house motif-detection program. For each mutant context, we considered only those TFs for which any of its motif matches had at least 5 mutated base pairs. In the resulting set of TFs (Extended Data Table 1) there is at least one TF per each of the enhancer mutant contexts that activated the respective context when recruited via the GAL4-DBD ($\geq$1.5-fold activation compared to GFP; $P < 0.05$).

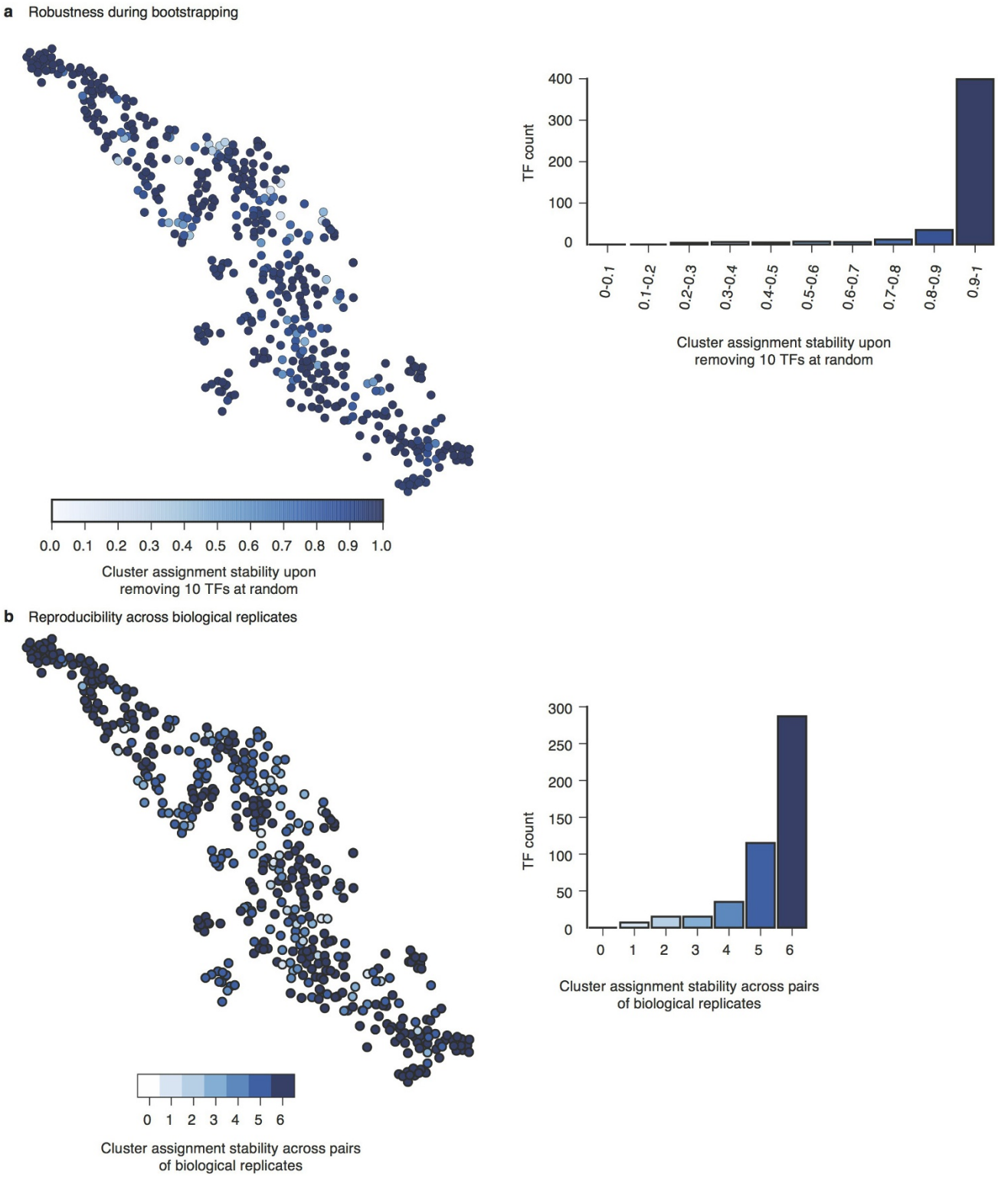
Cell type analysis—distances intra- versus inter-cluster. We tested a subset of the original 472 TFs in four different cell types (S2, Kc167, BG3 and OSC). This subset consists of 171 TFs covering all the 15 clusters by 9–17 TFs, including all the TFs mentioned in the main text. In each cell type, we computed Euclidean distances after standardizing the $\log_2$-transformed fold-change values in each context. Then we compared the distances of intra-cluster TF–TF pairs (both TFs belong to the same cluster) to inter-cluster TF–TF pairs (each of the TFs belongs to a different

cluster). In order to test whether the medians of these two groups of distances are significantly different, we determined empirical P values as follows. We randomly shuffled TF-to-cluster assignments 10^6 times and each time computed the medians of the distances for both groups. We mark the P values $P < 1 \times 10^{-6}$ as we never obtained a difference between the medians of intra- and inter-cluster distances as large as for the actual data for any of the cell types.

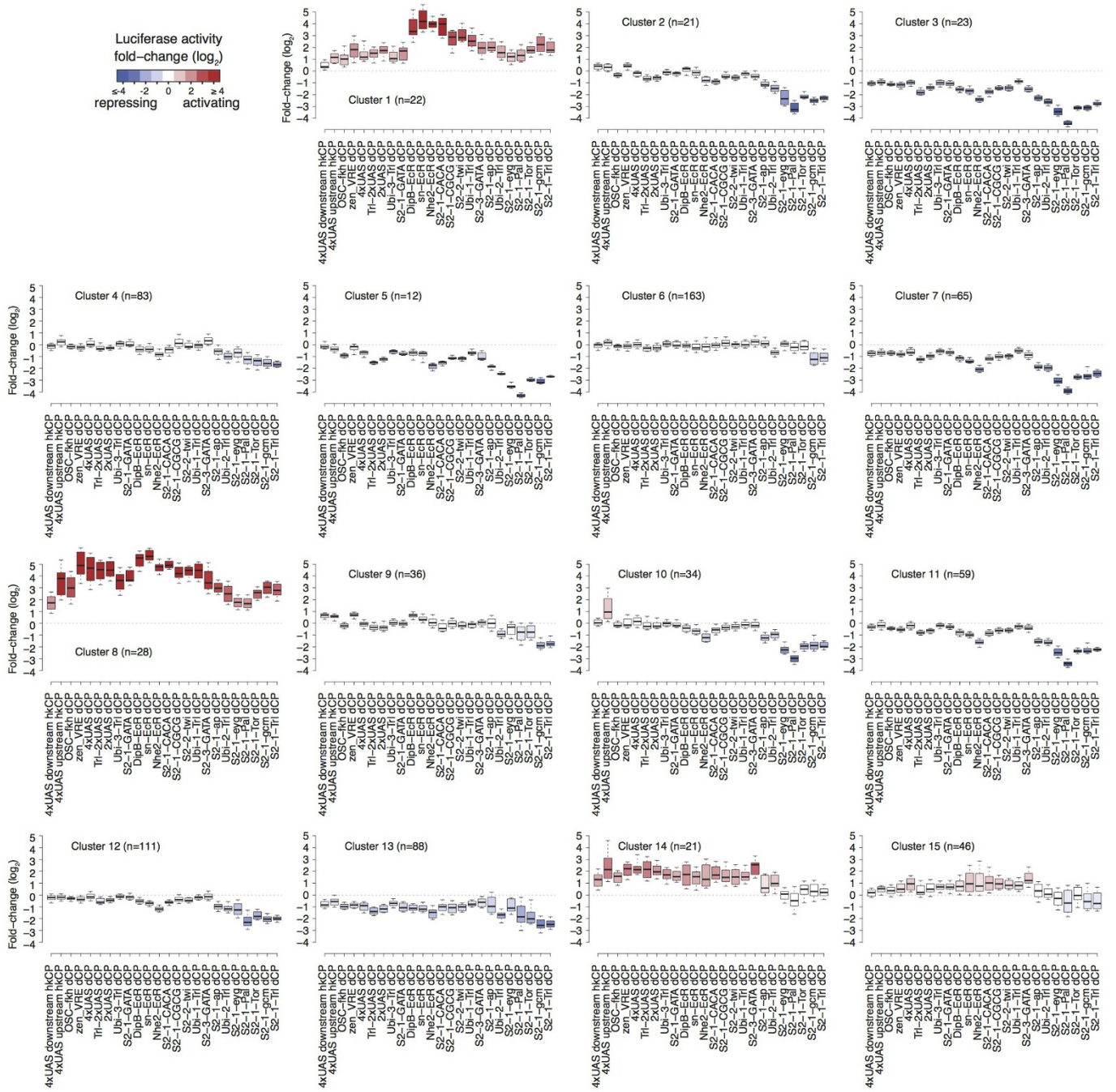
31. St Pierre, S. E., Ponting, L., Stefancsik, R. & McQuilton, P. FlyBase Consortium. FlyBase 102–advanced approaches to interrogating FlyBase. *Nucleic Acids Res.* **42,** D780–D788 (2014).
32. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42,** D222–D230 (2014).
33. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339,** 1074–1077 (2013).
34. Massouras, A., Decouttere, F., Hens, K. & Deplancke, B. WebPrInSeS: automated full-length clone sequence identification and verification using high-throughput sequencing data. *Nucleic Acids Res.* **38,** W378–W384 (2010).
35. Pfeiffer, B. D. *et al.* Tools for neuroanatomy and neurogenetics in *Drosophila*. *Proc. Natl Acad. Sci. USA* **105,** 9715–9720 (2008).
36. Schneider, I. The culture of cells from insects and ticks. I. Cultivation of dipteran cells *in vitro*. *Curr. Top. Microbiol. Immunol.* **55,** 1–12 (1971).
37. Ui, K. *et al.* Newly established cell lines from *Drosophila* larval CNS express neural specific characteristics. *In Vitro Cell. Dev. Biol. Anim.* **30,** 209–216 (1994).
38. Echalier, G. & Ohanessian, A. Isolation, in tissue culture, of *Drosophila melangaster* cell lines. *C. R. Acad. Sci. Hebd. Seances Acad. Sci. D* **268,** 1771–1773 (1969) [transl.].
39. Saito, K. *et al.* A regulatory circuit for *piwi* by the large Maf gene *traffic jam* in *Drosophila*. *Nature* **461,** 1296–1299 (2009).
40. Hampf, M. & Gossen, M. A protocol for combined *Photinus* and *Renilla* luciferase quantification compatible with protein assays. *Anal. Biochem.* **356,** 94–99 (2006).
41. R Development Core Team. *R: A language and environment for statistical computing.* (*R Foundation for Statistical Computing*, 2011).
42. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37,** W202–W208 (2009).
43. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14,** 755–763 (1998).
44. Dinkel, H. *et al.* ELM–the database of eukaryotic linear motifs. *Nucleic Acids Res.* **40,** D242–D251 (2012).
45. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25,** 25–29 (2000).
46. Tomancak, P. *et al.* Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.* **3,** (2002).
47. Zhu, L. J. *et al.* FlyFactorSurvey: a database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res.* **39,** D111–D117 (2011).
48. Stark, A. *et al.* Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450,** 219–232 (2007).
49. Yu, S. X. & Shi, J. Multiclass spectral clustering. In *ICCV*, 313–319 (IEEE, 2003).
50. Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *J. Machine Learn. Res.* **12** 2825–2830 (2011).
51. von Luxburg, U. Clustering stability: an overview. *Foundations and Trends in Machine Learning* **2,** 235–274 (2010).
52. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Machine Learn. Res.* **9,** 2579–2605 (2008).
53. The modENCODE Consortium *et al.* Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330,** 1787–1797 (2010).
54. Pfeiffer, B. D. *et al.* Refinement of tools for targeted gene expression in *Drosophila*. *Genetics* **186,** 735–755 (2010).

Context	Replaced motif	Scheme	Reference	Genomic coordinates	Length
S2-1 dCP ('GATA mutant context')	GATA		Yáñez-Cuna et al. 2014	chr2R:5326572-5327032	461bp
S2-1 dCP ('CGCG mutant context')	CGCG		Yáñez-Cuna et al. 2014	chr2R:5326572-5327032	461bp
2xUAS dCP	-		-	-	61bp
Nhe2-EcR dCP	EcR		Shlyueva et al. 2014	chr2L:21113350-21113776	439bp
DipB-EcR dCP	EcR		Shlyueva et al. 2014	chr3R:9616571-9616858	357bp
sn-EcR dCP	EcR		Shlyueva et al. 2014	chrX:7867729-7868227	514bp
4xUAS downstream hkCP	-		-	-	100bp
4xUAS upstream hkCP	-		-	-	100bp
4xUAS dCP	-		-	-	100bp
Trl-2xUAS dCP	-		-	-	73bp
zen_VRE dCP	dl		Jiang et al. 1993	chr3R:2581086-2581277	174bp
S2-1 dCP	eyg		Yáñez-Cuna et al. 2014	chr2R:5326572-5327032	461bp
S2-1 dCP	Trl		Yáñez-Cuna et al. 2014	chr2R:5326572-5327032	461bp
S2-1 dCP	Pal		Yáñez-Cuna et al. 2014	chr2R:5326572-5327032	461bp
S2-1 dCP	gcm		Yáñez-Cuna et al. 2014	chr2R:5326572-5327032	461bp
S2-1 dCP	Tor		Yáñez-Cuna et al. 2014	chr2R:5326572-5327032	461bp
S2-1 dCP	CACA		Yáñez-Cuna et al. 2014	chr2R:5326572-5327032	461bp
S2-1 dCP	ap		Yáñez-Cuna et al. 2014	chr2R:5326572-5327032	461bp
S2-2 dCP	twi		Yáñez-Cuna et al. 2014	chrX:4830533-4831008	476bp
S2-3 dCP	GATA		Yáñez-Cuna et al. 2014	chr3R:5262065-5262519	455bp
OSC dCP	fkh		Yáñez-Cuna et al. 2014	chr2L:19467959-19468425	467bp
Ubi-1 dCP	Trl		Yáñez-Cuna et al. 2014	chrX:1517186-1517657	470bp
Ubi-2 dCP	Trl		Yáñez-Cuna et al. 2014	chrX:6118311-6118795	485bp
Ubi-3 dCP	Trl		Yáñez-Cuna et al. 2014	chr3R:5376880-5377349	468bp

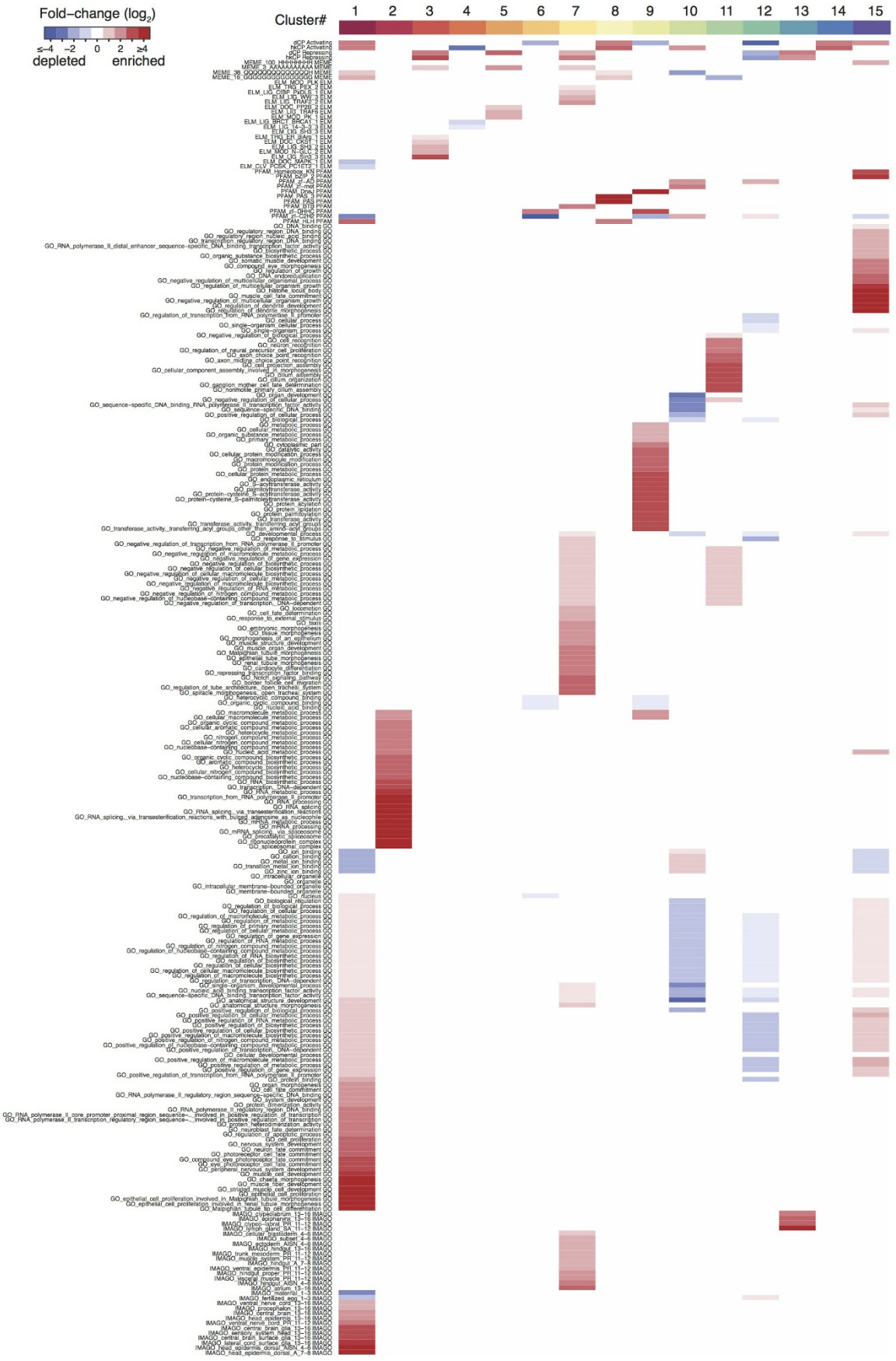
Extended Data Figure 1 | 24 regulatory contexts. Tested contexts included 19 motif-mutant enhancer contexts which were designed by replacing 43 occurrences of 15 different motif types in 11 previously characterized enhancers with broad ('Ubi-1' to 'Ubi-3') or cell type-specific ('S2-1' to 'S2-3': S2 cell-specific; 'OSC': ovarian somatic cell (OSC)-specific) activities (all from ref. 3) or hormone-inducible enhancers (from ref. 12.). We also designed five synthetic contexts consisting of UAS sites with or without Trl sites and a developmental core promoter (dCP; *Drosophila* synthetic core promoter (DSCP) derived from the transcription factor gene *Eve*[54]) or with a housekeeping core promoter (hkCP; derived from the ribosomal gene *RpS12*[13]). Shown are schemes of the luciferase reporter constructs used for the targeted recruitment of GAL4-DBD–TF/cofactor fusion proteins to UAS sites (the luciferase gene is not drawn to scale). Motif names denote the motifs (as named by refs 3, 12) that have been replaced by UAS sites (blue boxes) to create the enhancer context. Note that TF-to-motif assignments are not unique and typically several TFs can bind each of the motifs (see Extended Data Table 1).



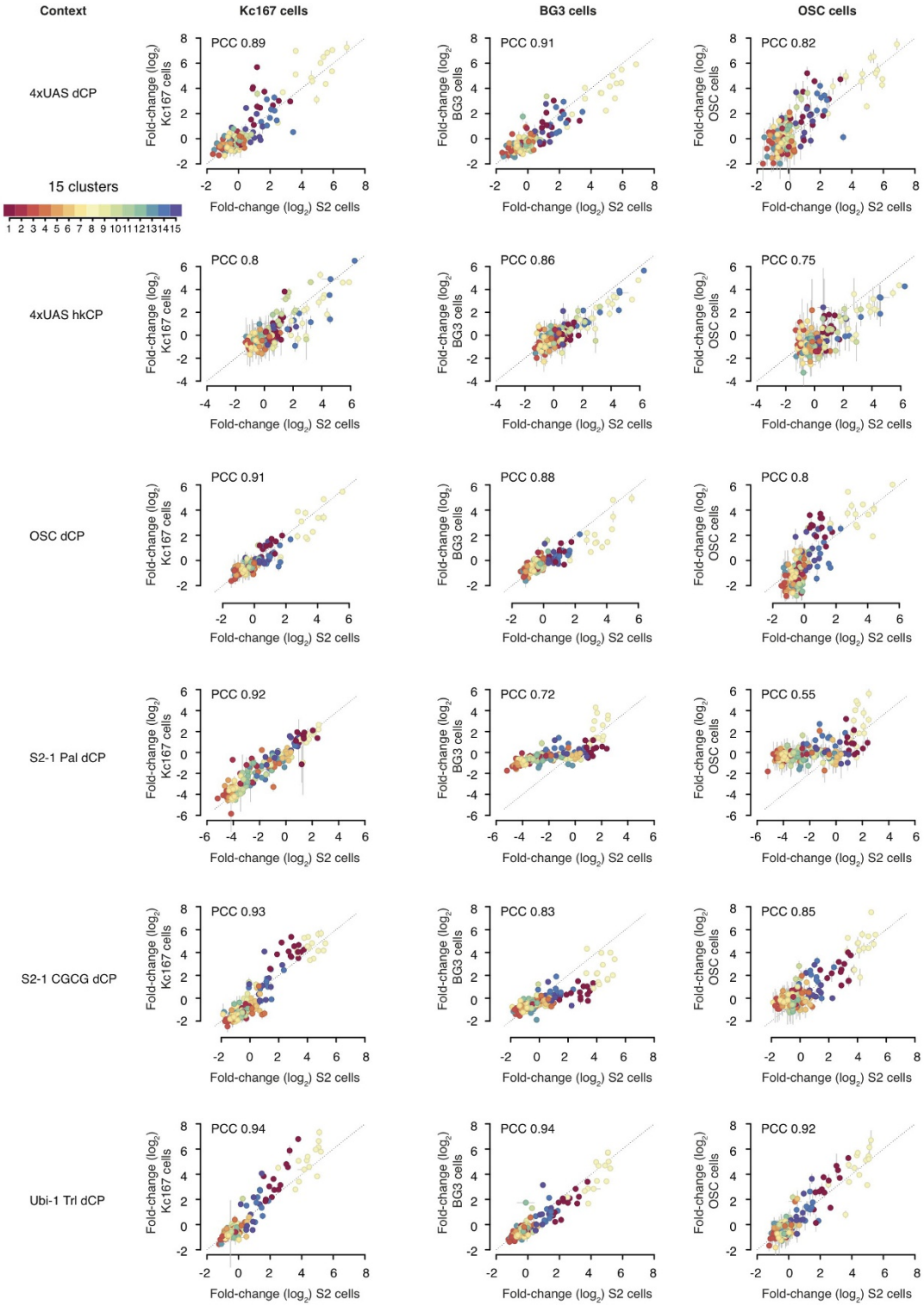
Extended Data Figure 2 | TF clustering is robust and reproducible. **a**, Cluster assignment is robust during bootstrapping (474 rounds of removing 10 randomly selected TFs). The cluster label stability denotes the fraction (out of 474 trials) a given TF was assigned to the same cluster as in the original clustering (node layout shown is identical to Fig. 2a). The vast majority of TFs were assigned to the same group in $\geq$90% of the cases (histogram). **b**, Cluster assignment for individual TFs is reproducible for biological replicates. We repeated the clustering six times, each time using only two out of the four biological replicates (six corresponds to all possibilities to choose two out of the four replicates). The cluster stability denotes the number of times (out of six) a given TF is assigned to the same cluster as in the original clustering. The majority of TFs were assigned to the same cluster, independent of which pair of biological replicates was used to generate the clustering (histogram).



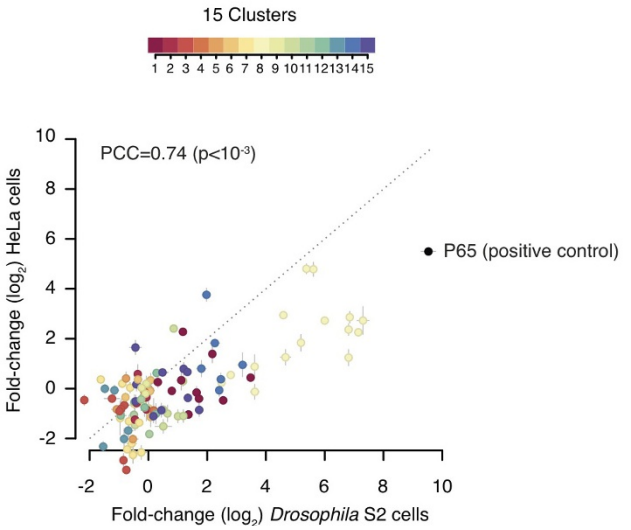
Extended Data Figure 3 | Cluster activity profiles. Normalized luciferase values for all TFs assigned to each of the 15 clusters across all 24 contexts. Shown are median and quartiles as boxes, and the tenth and ninetieth percentiles as whiskers for each of the 24 contexts. Boxes are coloured according to the median activity in each context (see colour legend).



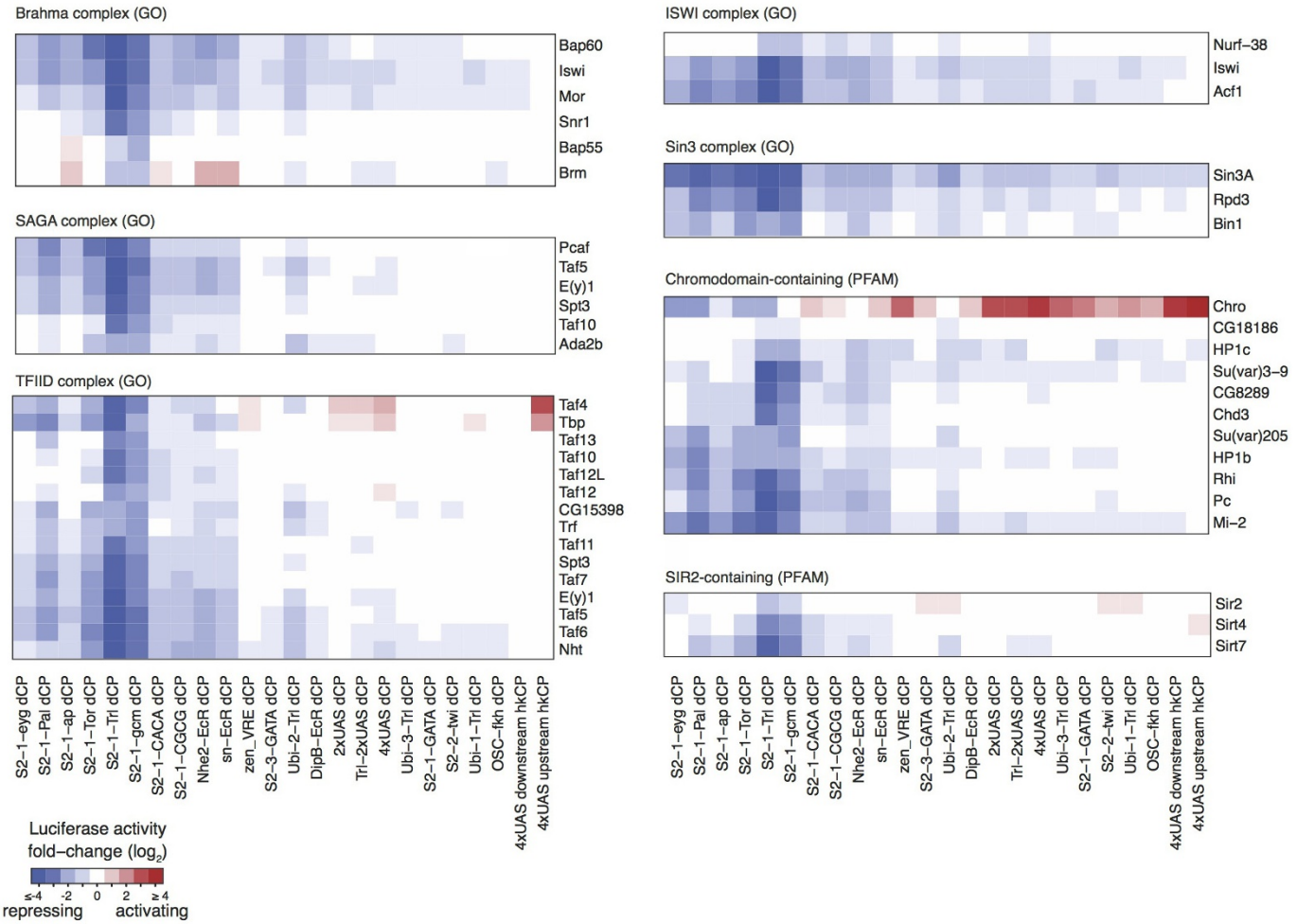
Extended Data Figure 4 | Complete feature enrichment analysis of 15 groups of TFs. Features analysed include eukaryotic linear motifs[44] (ELM), homopolymeric amino acid repeat motifs discovered using MEME[42], protein domains as annotated by Pfam[32], Gene Ontology[45] (GO), and gene expression patterns as annotated by IMAGO[46]. Red and blue shadings denote enrichments and depletions (log_2-transformed) with an empirical FDR of at most 10% per feature type and cluster (others are white).



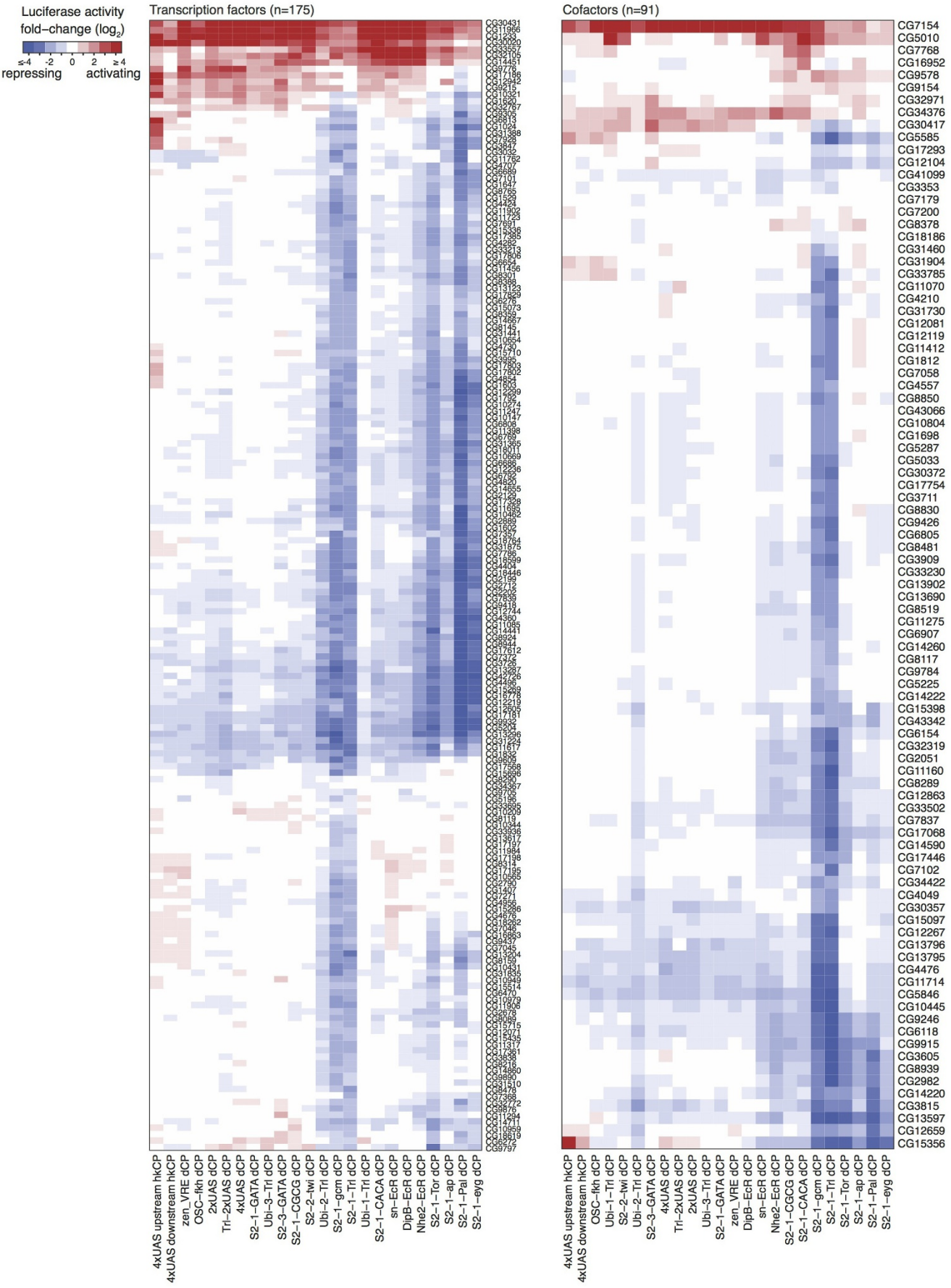
Extended Data Figure 5 | TFs behave consistently across *Drosophila* cell types. We tested 171 of the 474 TFs (36.1%) in 6 of the 24 contexts in Kc167, BG3 and OSC cell lines, which are derived from embryos, larvae and adult, respectively. Shown are normalized luciferase values and the Pearson correlation coefficients (PCC; $P < 1 \times 10^{-3}$ for all comparisons). We tested synthetic contexts containing an array of UAS sites upstream of a developmental and a housekeeping core promoter[13] (4×UAS dCP and hkCP), three contexts derived from cell-type-specific enhancers[3] (OSC dCP, S2-1 Pal dCP, S2-1 CGCG dCP), and one context derived from a broadly active enhancer[3] (Ubi-1 Trl dCP). The latter showed the highest similarities (PCCs of 0.94, 0.94 and 0.92 for Kc167, BG3 and OSC cells, respectively) while the lowest PCCs for the non-embryonic BG3 and OSC cells (0.72 for BG3 and 0.55 for OSC) were obtained for S2-1 Pal dCP, derived from an enhancer active only in the embryonic S2 and Kc167 cells, presumably because the corresponding wild-type enhancer sequence is inactive in larval and adult cells[3,12] such that combinatorial effects between the tethered TF and other enhancer-bound TFs may be less effective or lack entirely. Enhancer complementation presumes (and the results throughout this study confirm this presumption) that the regulatory functions of the tethered TFs are revealed (or altered) by other enhancer-bound factors; that is, factors that are bound to the enhancer in S2 cells (in which the corresponding enhancer is active[3]) but not in the other cell types (in which the enhancer is not active). This emphasizes the value of enhancer complementation for the study of regulatory activities and the importance of contexts derived from active enhancers. Error bars denote standard deviation ($n = 4$, biological replicates).



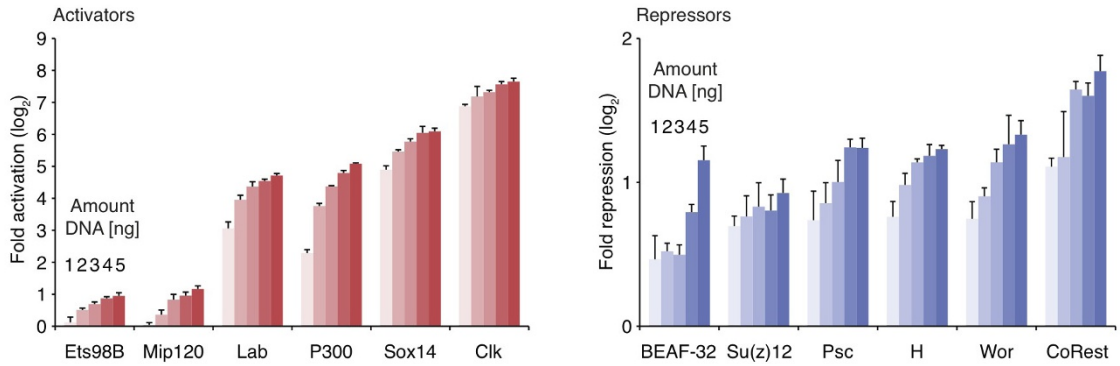
Extended Data Figure 6 | *Drosophila* TFs and cofactors retain their activating functions in human HeLa cells. We expressed GAL4-DBD fusion proteins for 107 of the 812 *Drosophila* factors (90 TFs and 17 cofactors) under the control of a constitutively active CMV promoter in human HeLa cells (see Methods). Shown are normalized luciferase values for the tested proteins recruited to the synthetic 4×UAS-dCP context. The values are remarkably similar quantitatively, with an overall Pearson correlation coefficient (PCC) of 0.74 ($P < 1 \times 10^{-3}$). The activation domain of the human TF P65 was used as a positive control. Error bars denote standard deviation ($n = 4$, biological replicates).



Extended Data Figure 7 | Regulatory activities of selected cofactor complexes or protein domain families. Heat maps of normalized luciferase values for sets of proteins annotated as being part of the same complex by Gene Ontology[45] (GO) or containing a chromodomain or SIR2 domain as annotated by Pfam[32].



Extended Data Figure 8 | Regulatory activities of uncharacterized TFs and cofactors. Heat maps of normalized luciferase values for all 'CG genes' among the tested TFs and cofactors, which activate or repress in at least one context ($\geq$1.5-fold compared to GFP; $P < 0.05$ FDR-corrected for 24×474 and 24×338 tests for TFs and cofactors, respectively).



Extended Data Figure 9 | Consistent effects of varying the amounts of plasmid DNA for TF and cofactor expression. The effects of using 1 ng, 2 ng, 3 ng, 4 ng and 5 ng of GAL4-DBD–TF/cofactor fusion protein expressing plasmids on luciferase assays in S2 cells suggest that reporter activity is robust to variation in TF levels. Shown are normalized luciferase values of GAL4-DBD N-terminally fused to six activating and six repressing TFs and cofactors of different strengths targeted to the synthetic 4×UAS-dCP context. The amount of plasmid expressing the GAL4-DBD–TF/cofactor fusion proteins was 3 ng for all factors throughout this study. Error bars denote standard deviation ($n = 4$, biological replicates).

Extended Data Table 1 | TF recovery analysis for S2 cell enhancer contexts

CONTEXT	TF	RPKM_S2-CELLS	FOLD ACTIVATION
S2-1-CACA dCP	Cf2	15.16	2.24
S2-1-CGCG dCP	vfl	13.54	27.04
S2-1-GATA dCP	Cf2	15.16	2.42
S2-1-GATA dCP	sd	109.42	1.53
S2-1-GATA dCP	sqz	12.39	1.68
S2-2-twi dCP	Cf2	15.16	1.81
S2-2-twi dCP	gl	24.29	5.55
S2-2-twi dCP	twi	92.36	5.82
S2-3-GATA dCP	ap	644.03	1.80
S2-3-GATA dCP	Atf6	17.04	6.60
S2-3-GATA dCP	lab	1.44	7.69
Ubi-1-Trl dCP	Atf6	17.04	2.75
Ubi-1-Trl dCP	Hsf	38.62	6.39
Ubi-1-Trl dCP	Sox14	17.55	29.68
Ubi-2-Trl dCP	Hsf	38.62	3.14
Ubi-3-Trl dCP	Hsf	38.62	5.00

TFs with known motifs that match the sequences we mutated to UAS sites in each of the different enhancer contexts are expressed in S2 cells (RPKM >1 (ref. 53)) and significantly activate the respective context when recruited via the GAL4-DBD ($\geq$1.5-fold activation compared to GFP; $P<0.05$).