Article

Occupancy maps of 208 chromatin-associated proteins in one human cell type

https://doi.org/10.1038/s41586-020-2023-4

Received: 4 October 2017

Accepted: 9 January 2020

Published online: 29 July 2020

Open access

Check for updates

E. Christopher Partridge[1,10], Surya B. Chhetri[1,2,9,10], Jeremy W. Prokop[1,3], Ryne C. Ramaker[1,4], Camden S. Jansen[5], Say-Tar Goh[6], Mark Mackiewicz[1], Kimberly M. Newberry[1], Laurel A. Brandsmeier[1], Sarah K. Meadows[1], C. Luke Messer[1], Andrew A. Hardigan[1,4], Candice J. Coppola[2], Emma C. Dean[1,7], Shan Jiang[5], Daniel Savic[8], Ali Mortazavi[5], Barbara J. Wold[6], Richard M. Myers[1]✉ & Eric M. Mendenhall[1,2]✉

Transcription factors are DNA-binding proteins that have key roles in gene regulation[1,2]. Genome-wide occupancy maps of transcriptional regulators are important for understanding gene regulation and its effects on diverse biological processes[3–6]. However, only a minority of the more than 1,600 transcription factors encoded in the human genome has been assayed. Here we present, as part of the ENCODE (Encyclopedia of DNA Elements) project, data and analyses from chromatin immunoprecipitation followed by high-throughput sequencing (ChIP–seq) experiments using the human HepG2 cell line for 208 chromatin-associated proteins (CAPs). These comprise 171 transcription factors and 37 transcriptional cofactors and chromatin regulator proteins, and represent nearly one-quarter of CAPs expressed in HepG2 cells. The binding profiles of these CAPs form major groups associated predominantly with promoters or enhancers, or with both. We confirm and expand the current catalogue of DNA sequence motifs for transcription factors, and describe motifs that correspond to other transcription factors that are co-enriched with the primary ChIP target. For example, FOX family motifs are enriched in ChIP–seq peaks of 37 other CAPs. We show that motif content and occupancy patterns can distinguish between promoters and enhancers. This catalogue reveals high-occupancy target regions at which many CAPs associate, although each contains motifs for only a minority of the numerous associated transcription factors. These analyses provide a more complete overview of the gene regulatory networks that define this cell type, and demonstrate the usefulness of the large-scale production efforts of the ENCODE Consortium.

There are an estimated 1,639 transcription factors (TFs) in the human genome[2], and up to 2,500 CAPs when we include transcriptional cofactors, RNA polymerase-associated proteins, histone-binding regulators, and chromatin-modifying enzymes[1,7]. A typical TF binds to a short DNA sequence motif, and, in vivo, some TFs exhibit additional chromosomal occupancy mediated by their interactions with other CAPs[8–10]. CAPs are vital for orchestrating cell type- and cell state-specific gene regulation, including the temporal coordination of gene expression in developmental processes, environmental responses, and disease states[3–6,11–13].

Identifying genomic regions with which a TF is physically associated, referred to as TF binding sites (TFBSs), is an important step towards understanding its biological roles. The most common genome-wide assay for identifying TFBSs is ChIP–seq[14–16]. In addition to highlighting potentially active regulatory DNA elements by direct measurement, ChIP–seq data can define DNA sequence motifs that can be used, often in conjunction with expression data and chromatin accessibility maps, to infer likely binding events in other cellular contexts without performing direct assays. Although motifs identified by ChIP–seq are often representative of direct binding, this is not always the case, as co-occurrence of other TFs could lead to the enrichment of their motifs. Furthermore, the ChIP–seq method identifies both protein–DNA and, indirectly, protein–protein interactions, such that indirect and even long-distance interactions (for example, looping of distal elements) can be captured as ChIP–seq enrichments.

A long-term goal is comprehensive mapping of all CAPs in all cell types, but a more immediate aspiration is to create a catalogue of all

[1]HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA. [2]Department of Biological Sciences, The University of Alabama in Huntsville, Huntsville, AL, USA. [3]Department of Pediatrics and Human Development, College of Human Medicine, Michigan State University, Grand Rapids, MI, USA. [4]Department of Genetics, University of Alabama at Birmingham, Birmingham, AL, USA. [5]Department of Developmental and Cell Biology, University of California Irvine, Irvine, CA, USA. [6]Division of Biology, California Institute of Technology, Pasadena, CA, USA. [7]Department of Pathology, University of Alabama at Birmingham, Birmingham, AL, USA. [8]Pharmaceutical Sciences Department, St Jude Children's Research Hospital, Memphis, TN, USA. [9]Present address: Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MA, USA. [10]These authors contributed equally: E. Christopher Partridge, Surya B. Chhetri. ✉e-mail: rmyers@hudsonalpha.org; eric.mendenhall@uah.edu

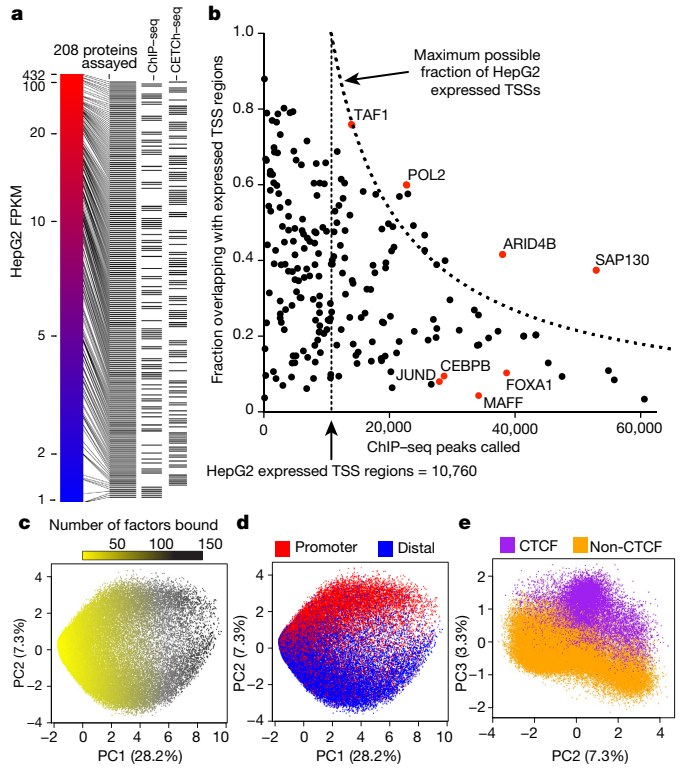


Fig. 1 | Overview and analysis of HepG2 data sets. a, The 208 chromatin-associated factors assayed in HepG2 cells, organized by expression (FPKM), and denoting whether the factors were assayed by ChIP–seq or CETCh–seq. **b**, Scatter plot of all 208 factors, showing broad distribution of fraction of called peaks at expressed TSSs (±3 kb from TSS) against total peak number; points beyond the maximum possible fraction are possible owing to multiple peaks at single TSS regions. **c**, Plot showing PCA of genomic segments (n = 282,105) with more than two factors bound, highlighting the separation on the basis of the number of factors bound. **d**, Same plot as in **c** showing promoter versus distal location. **e**, Same plot as in **c** showing PC2 versus PC3 and highlighting the presence of CTCF.

CAPs expressed in a single cell type. The resulting consolidation of hundreds of genome-wide maps for a single cellular context promises insights into CAP networks that are otherwise not possible. Such comprehensive data will also provide the backdrop for understanding large-scale functional element assays, and should improve the ability to infer TFBSs in other cell types that are less amenable to direct measurements.

Here we present an analysis of 208 CAP occupancy maps in the hepatocellular carcinoma cell line HepG2 performed as part of the ENCODE project, composed of 92 traditional ChIP–seq experiments with factor-specific antibodies and 116 CRISPR epitope tagging ChIP–seq (CETCh–seq) experiments[17,18]. Of all human CAPs, approximately 960 are expressed in HepG2 cells above a threshold RNA value of 1 FPKM (fragments per kilobase of transcript per million mapped reads), the lowest level at which we can routinely generate successful ChIP–seq and CETCh–seq results. This resource contains ChIP–seq and CETCh–seq maps for about 22% of these 960 CAPs, of which 171 are sequence-specific TFs and 37 are histone-binding or histone-modifying proteins, or other chromatin regulators or transcription cofactors (Fig. 1a, Supplementary Table 1). This large and unbiased sampling in one cell type allowed us to approach analysis from complementary directions, beginning with patterns of CAP occupancy and co-occupancy to find preferential associations with each other and with promoters, enhancers, or insulator functions, and in the other direction, working from genomic loci, sequence motifs, and epigenomic states to explain occupancy. These publicly available ENCODE occupancy data, together with the analyses and insights presented here, comprise a key resource for the scientific community.

- We analyse ChIP–seq and CETCh–seq maps for about 22% of TFs and other CAPs expressed in the human HepG2 cell line.
- We use clustering to classify major groups of CAPs, including those that are promoter- or enhancer-associated, or that are associated with both promoters and enhancers to a similar extent.
- Using this large amount of data, we demonstrate that DNA sequence motifs or ChIP–seq peak calls can distinguish between promoters and enhancers.
- We show that high-occupancy target (HOT) regions are driven by strong motifs for one or a few TFs and weaker, more degenerate motifs for many other CAPs.

CAPs segregate regulatory element states

As an initial analysis, we investigated how the binding of each of the 208 CAPs is distributed in the genome relative to known transcriptional promoters. We calculated the fraction of each of the called peaks of each CAP that was within 3 kilobases (±3 kb) of transcription start sites (TSSs), analysing only the TSSs of genes expressed (≥1 TPM (transcripts per kilobase million)) in HepG2 cells (Fig. 1b) and, separately, all annotated TSSs regardless of expression (Extended Data Fig. 1a). Individual CAPs exhibited variable proportions of promoter-associated peaks, independent of the number of peaks called in an experiment.

To further summarize the occupancy landscape, we merged all the called peaks from every experiment into non-overlapping 2-kb windows, limited to those windows in which two or more CAPs had a called peak, and performed principal component analysis (PCA) on these DNA segments, using the presence or absence of each CAP at each genomic segment. This analysis captured global patterns of ChIP–seq peaks, with principal component 1 (PC1) explaining about 28% of the variance and correlating strongly with the number of unique CAPs associated with a given genomic region (Fig. 1c). PC2 separates promoter-proximal from promoter-distal peaks, underscoring the relevance of promoters as a predictor of genomic state and CAP occupancy (Fig. 1d). Notably, the shape of this plot suggests that, as the number of CAPs associated at a locus increases, the promoter-proximal and promoter-distal regions lose separation along PC2. In addition, PC2 plotted against PC3 shows strong segregation based on occupancy of the factor CTCF (Fig. 1e), suggesting that discrete genomic demarcations are attributable to this factor, as expected given its insulator and loop-anchoring functions.

To assess the epigenomic context of each binding site, we used IDEAS (integrative and discriminative epigenome annotation system), a machine-learning method for biochemical mark-based genomic segmentation[19]. This IDEAS HepG2 epigenomic segmentation inferred 36 genomic states based on eight histone modifications, RNA polymerase ChIP–seq, CTCF ChIP–seq, and DNA accessibility data sets (DNase and formaldehyde-assisted isolation of regulatory elements (FAIRE)). Notably, IDEAS states for HepG2 cells were classified using mainly histone marks, augmented by only two chromatin-associated ChIP–seq maps included in our data set (CTCF and RNA polymerase). These segregate the anticipated major classes of correlations between epigenomic states in the IDEAS segmentation and CAP associations, such as enrichment of H3K4me3 at annotated promoters and H3K27ac at candidate active enhancers, as well as open chromatin status as assayed by DNA accessibility experiments, typical of TF-bound DNA. As expected, the resulting IDEAS states classified only a minority of the HepG2 genome as potential *cis*-regulatory elements (Extended Data Fig. 1b).

We calculated the relative IDEAS state enrichments of the peak calls for each CAP, and clustered the CAPs by these enrichments. The resulting matrix delineated several clear bins of genomic state associations, expanding and refining the previously noted preferential proximal versus distal genomic associations of CAPs[20]. Specifically, we found a subset of CAPs that are preferentially associated with promoters, another

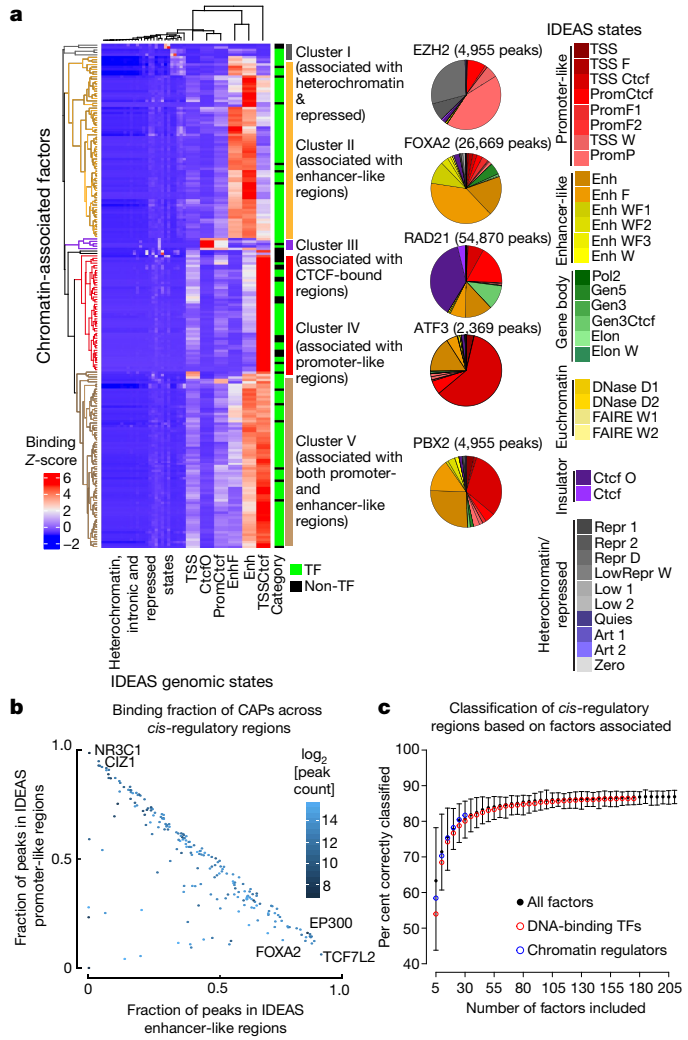


Fig. 2 | Landscape of factor binding to regulatory states. a, Unsupervised clustering of the 208 factors on the basis of binding enrichment at 36 IDEAS genome states and the 5 main clusters of factors, along with pie charts showing absolute binding fractions of an example of a factor from each cluster. **b**, Correlation plot showing the fraction of promoter (*y*-axis) or enhancer (*x*-axis) binding for all 208 factors, with points coloured by peak counts for each factor. **c**, Predictive ability of random forest classification of genomic regions as either enhancer or promoter on the basis of the number of factors used to train the algorithm; $n = 100$ iterations, lines from minimum to maximum with median indicated.

subset associated with candidate active enhancers, and a third group distributed across both proximal promoter regions and candidate active enhancers (Fig. 2a). We also found two smaller CAP-associated clusters: one associated with heterochromatin and repressed marks (including BMI1 and EZH2, both part of Polycomb repressive complexes), and one with CTCF regions (including CTCF and the known cohesin complex proteins RAD21 and SMC3; Fig. 2a, Supplementary Table 2). These categories contain members of different classes of CAPs, and point to distinct gene regulatory pathways. A PCA based on these IDEAS states also recapitulated these clusters (Extended Data Fig. 1c).

For roughly 40% of the CAPs assayed, most called peaks were in IDEAS promoter-like regions, while about 30% of CAPs were predominantly associated with IDEAS enhancer-like regions (Fig. 2b). Although these preferences are part of a continuous distribution, the unsupervised clustering using all IDEAS genomic states suggests that subsets of CAPS show strong localization preferences. We analysed whether the promoter-associated CAPs associated predominantly with CpG-island promoters by annotating promoter regions according to previous classifications for low, intermediate, and high CpG content[15,21]. The promoter-associated CAPs also cluster preferentially with promoters with high CpG content (Extended Data Fig. 2a, Supplementary Table 3). However, the GC content of motifs for CAPs in the promoter-associated cluster is not significantly different from that of CAPs associated with both promoters and enhancers, suggesting that motif GC content alone does not drive the clustering (Extended Data Fig. 2b).

The CAPs that associate with both promoters and enhancers do not have apparent bias in relation to the GC content of promoters. Previous publications have noted similarities between promoters and enhancers, ascribing enhancer activity to promoters, and transcription occurs directly at enhancers in the form of enhancer RNA (eRNA) and even as alternative promoters[22–24]. The subset of CAPs identified as associating with both promoters and enhancers may point to specific genomic loci or gene regulatory networks wherein the lines between promoters and enhancers are most blurred.

Because CAPs localize to specific genomic states, we were able to reproducibly train random forest models to predict the IDEAS state of a genomic region using binding information for only a small number of CAPs (Fig. 2c). The prediction method was successful when using a combination of TFs with chromatin regulators and other extended CAPs, but was also successful when trained only on direct DNA-binding TFs or only on non-TFs. Each approach required a subset of roughly any 30 CAPs to achieve approximately 80% accuracy.

CAP distribution in regulatory elements

Although the 208 CAPs do not represent a complete catalogue of all expressed CAPs in HepG2 cells, we investigated how much of the regulation in this cell line is captured by this partial compendium. We used IDEAS to define a set of 370,570 putative HepG2 *cis*-regulatory elements classified as promoters, 'strong' enhancers, or 'weak' enhancers, with merging of similar features within 100 base pairs (bp), resulting in a broad size distribution from 200 bp to 12–16 kb. We then calculated how many CAPs were associated in each region (Extended Data Fig. 1d). On average there were seven CAPs associated at any putative regulatory region. Approximately 67% of the regions did not contain any called peaks; however, the vast majority of these (about 85.5%) were classified as 'weak' or 'poised' enhancers by the IDEAS segmentation. Conversely, elements classified as promoters or 'strong' enhancers by IDEAS were enriched for occupancy by higher numbers of CAPs (Extended Data Fig. 1d). Of the IDEAS-determined active promoter-like regions, 61% contained a called peak for at least one CAP in this data set, and of the strong enhancer-like regions, 75% contained at least one called peak. Because most promoters and strong IDEAS-modelled enhancers had one or more CAPs associated, and these elements had an average of 15 and 18 unique associated CAPs per region, respectively, these data capture a substantial overview of the CAP regulatory network in HepG2 cells.

Motif analysis reveals CAP associations

We assessed motif enrichment in peaks, and found many previously derived motifs for both direct and potentially indirect associations, as well as some potentially novel motifs. We derived a high-confidence set of 293 motifs called from 160 of the 171 putatively direct DNA-binding TFs in our data set[2]. We compared these motifs to the JASPAR databases[25,26] and to the Catalog of Inferred Sequence Binding Preferences (CIS-BP) database[8] to determine whether our de novo derived motifs matched previous findings from various in vivo and/or in vitro assays[27]. Overall, more than 80% of the 293 motifs had a similar motif in these databases (86% in CIS-BP build 1.02, 82% in JASPAR 2018, 81% in JASPAR 2016; Extended Data Fig. 3a–c). For 114 motifs derived from peaks for 89 unique TFs, the most similar motif in the database was

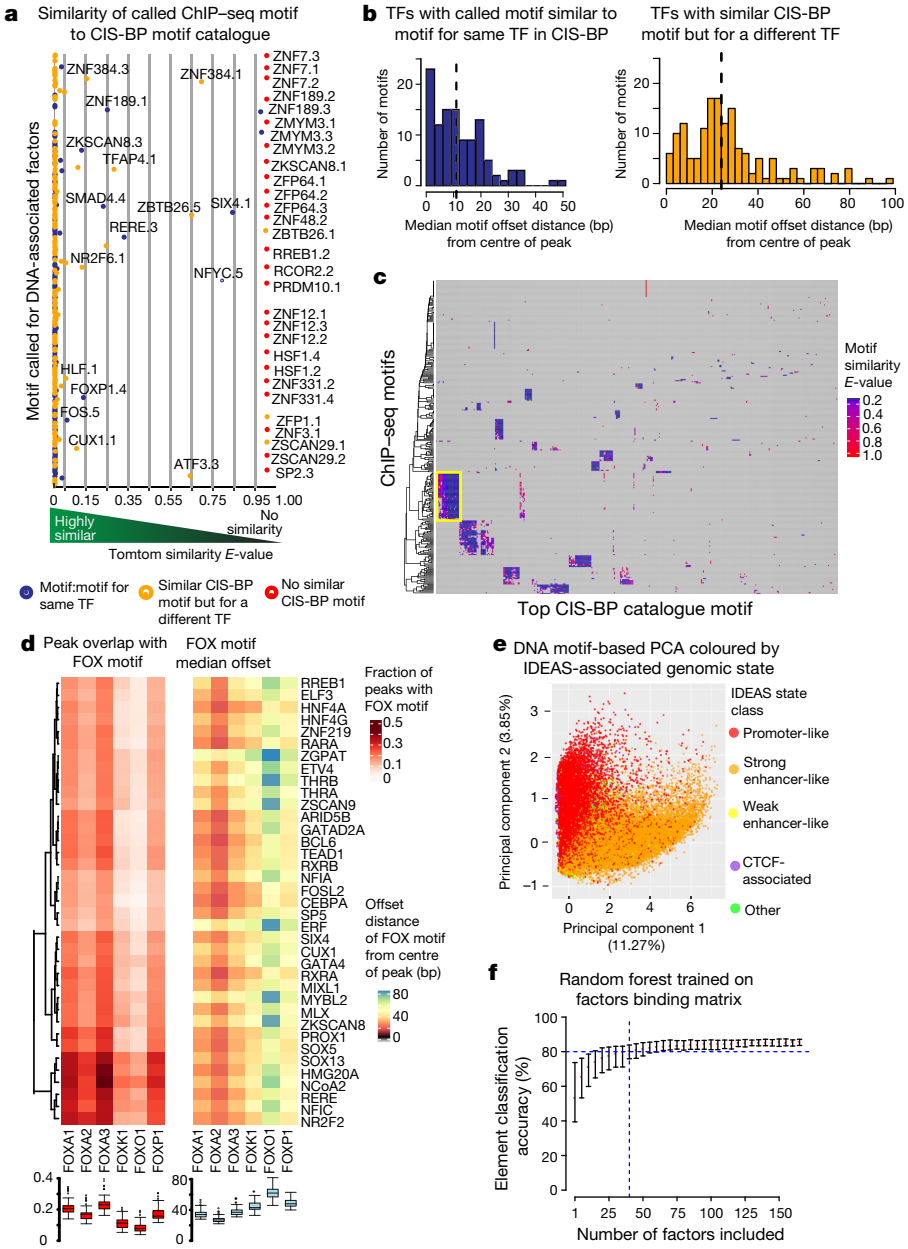


Fig. 3 | Motif identification and analysis. a, The 293 high-confidence motifs derived from analysis of the ChIP–seq data were quantitatively compared to all (human) motifs in the CIS-BP database and plotted according to similarity scores. Blue points represent motifs that matched the assayed factor, yellow points represent motifs that match a factor other than the one assayed, and red points represent motifs not similar to any in CIS-BP. **b**, Histograms showing the distance from the centre of the ChIP–seq peak for motifs that do (left) or do not (right) match the TF. **c**, Clustered heat map showing the similarity of all 293 significant motifs to 733 motifs from CIS-BP for the assayed factors. **d**, Further analysis of the cluster containing 37 factors that had FOX family motifs, showing the overlap of FOX TF binding in these peaks, as well as the median offset of the FOX motif from the centre of the ChIP–seq peaks. For box plots (bottom), $n = 37$ CAPs; boxes show middle quartiles, centre line shows median, whiskers show 1.5× interquartile range (IQR). **e**, PCA showing separation of motifs that fall in promoters versus those that fall in enhancers; $n = 408{,}382$ genomic elements. **f**, Prediction accuracy for calling whether an element is a promoter or enhancer on the basis of motifs that are present; $n = 100$ iterations, lines from minimum to maximum with median indicated.

annotated as the motif for the TF that was the target of the ChIP–seq or CETCh–seq assay, and we call these cases 'concordant' (Fig. 3a, Supplementary Table 4). There were 156 motifs derived from peak data for 99 TFs that were more similar to the database motif of a different TF, and we denote these as 'discordant'. We also observed 23 motifs derived from peaks of 14 TFs that were highly dissimilar to any motifs in the databases and may be previously undescribed motifs. Most of these were from zinc finger TFs, a large class of factors that has been virtually unassayed by endogenous ChIP–seq.

We note that concordant calls were sometimes problematic, specifically when the motif in a database originated from a previous ChIP–seq experiment. In some cases, these motifs probably do not represent the specific sequence recognized by the TF assayed, but are spurious calls from associated TFs that replicate across multiple ChIP–seq experiments. For example, two motifs for ATF3 matched an ATF3 ChIP–seq motif in CIS-BP, which qualifies these motifs as concordant, but they more closely resemble an E-box motif. We overruled the automatic concordant call for this case, and manually changed it to discordant. For Supplementary Table 4, we curated each called motif to clarify results from the matching algorithm, and included a column with this information.

Among the 163 discordant motifs, motifs representing pioneer TFs such as FOXA1 were enriched, and we hypothesize that these motifs

were called owing to their substantial co-occurrence with the assayed TFs. Previous studies have noted the enrichment in ChIP–seq data of sequences that do not appear to be binding motifs for assayed TFs, but rather are more similar to other TF motifs[28]. There are several potential explanations for why the ChIP–seq-derived motif would most closely match a motif previously annotated for another factor. Related TFs often recognize very similar sequence motifs; for example, the motif we derived for TEAD4 was very similar to the motif previously found for TEAD1[29]. There are also instances in which a CAP lacks a strong and specific DNA-binding domain and no motif would be expected unless the motif represents a frequent co-binding partner, a scenario we explore below with GATAD2A. A similar explanation involves a particular TF acting as an 'anchor' at a locus, and either through direct protein–protein interactions, or by inducing an open chromatin environment, behaving as a mechanism for localization of other proteins. A well-studied example of this highlighted in our data was the enrichment of the CTCF motif in RAD21 ChIP–seq data, as RAD21 lacks a DNA-binding domain but interacts with CTCF. It is difficult to determine confidently whether a discordant motif represents a key co-factor interaction or a commonly co-localized protein. When we called multiple, distinct, high-confidence motifs in a single ChIP–seq experiment, with one motif annotated in databases as the direct target of the assayed TF and another motif representing a different TF that we also assayed separately, the results of the secondary factor's ChIP–seq experiment suggested that both TFs are likely to be associated at these loci, as both experiments yielded called peaks at these loci.

Supporting our hypothesis that, in the discordant cases, the motif of the secondary TF was not a site of direct binding for the primary CAP, examination of the precise location of the motifs within peaks showed a significant difference (Kolmogorov–Smirnov test, $P = 2.481 \times 10^{-12}$); the direct matching motifs of the assayed TFs were closer to the centres of called peaks and the discordant motifs for other TFs were more offset, providing evidence for co-occurrence at these locations (Fig. 3b). Direct interaction and co-recruitment between these pairs of TFs could explain these observations, and numerous examples of such combinatory and cooperative activities between TF pairs have been reported[30]. We found no significant trend for secondary TF motifs in any factor clusters we identified by IDEAS state preferences or other methods, suggesting that no biases were introduced by contributions from particular genomic loci (Extended Data Fig. 3d). We also analysed the peak locations of the 23 novel motifs found with the 14 factors that were highly dissimilar to any motifs in CIS-BP, and the majority showed enrichment at the centres of peaks (Extended Data Fig. 3e, f), supporting the notion that these are previously undescribed motifs for direct DNA binding by these TFs.

To better understand discordant TF motif calls, we constructed a similarity heat map using all 293 high-confidence motifs from our data and motifs for each assayed TF annotated in the CIS-BP database ($n = 733$; Fig. 3c). This analysis clustered TFs both by similarity of their direct binding motifs (such as all Forkhead factors) and by co-occurrence with other motifs. We thereby identified TFs that associate at genomic loci near particular motifs, such as CTCF. Most obvious was a set of 37 CAPs for which a Forkhead motif was called, indicating the high prevalence of this motif in HepG2 cells at active enhancers and promoters, and the key role of TFs such as FOXA1 and FOXA2 in the gene regulatory network in these cells. We examined these cases using our ChIP–seq data from six FOX TFs (FOXA1, FOXA2, FOXA3, FOXK1, FOXO1, and FOXP1), testing how often each of these FOX TFs yielded called peaks with a FOX motif that overlapped with a peak for any of these 37 other CAPs, and we found that most of the 37 contained a FOX peak with a FOX motif in about 20% of their peaks, with FOXA1 and FOXA3 motifs being the most common (Fig. 3d).

We next examined the locations of the FOX motifs in the overlapping peaks and found that all were offset to varying degrees, always with a median distance of more than 20 bp from the centres of peaks (Fig. 3d). In addition, we examined all peaks called for each of the 37 CAPs and identified the fraction that contained a primary motif specific to the individual CAP (where known) along with a FOX motif, the fraction that contained only the primary motif, the fraction that contained only a FOX motif, and the fraction that contained neither motif (Extended Data Fig. 4a). For the 30 CAPs with a described motif, the majority of peaks did not contain a primary motif, a result that may indicate protein–protein interactions and/or looping events in these peaks. Furthermore, when we examined peak overlaps between these 37 TFs and the six FOX TFs, we observed varying associations and co-occupancy partners, including factor preferences for individual FOX TFs and a cluster of components of the nucleosome remodelling and histone deacetylase (NuRD) complex (Extended Data Fig. 4b–d).

Motif information alone was predictive of genomic segments, clearly showing segregation between IDEAS states in a PCA (Fig. 3e). A random forest algorithm trained only on motifs was able to predict IDEAS states almost as well as one trained on ChIP–seq peaks, achieving approximately 80% success with any roughly 40 motifs (Fig. 3f).

Known and novel CAP associations

TFs and chromatin regulatory proteins can interact with and recruit other CAPs through direct and indirect physical associations. Although the activity of a few key CAPs may be very important for cell-state-specific expression, it is likely that combinatorial events are necessary to fine-tune expression[31]. We found both known and novel associations by examining occupancy overlaps and trends in a variety of analyses.

To identify candidate co-occupancy events mediated by direct DNA binding or by indirect interactions, both of which produce peaks in ChIP–seq data, we performed several analyses. We used the PCA of the protein-bound genomic loci described above (in which genomic loci clustered according to the CAPs associated at each region; Fig. 1c–e), and generated a correlation matrix based on the cumulative PC distances (weighted by the proportion of variance explained by each component) between all CAPs. The resulting unsupervised clustering of respective pairwise distances highlighted punctate groups that represented both known and potentially novel complexes, including a group containing POL2 and TSS-associated chromatin-modifying enzymes and transcriptional cofactors, a group of cohesin complex members, a group of liver-specific factors (the tissue type from which HepG2 is derived), and a group containing the NuRD complex, among others (Fig. 4a).

To quantitatively analyse the overall data, we performed read count Spearman correlations between all 208 CAPs by calculating raw sequencing counts at every unique locus present in called peaks in any experiment (±50 bp from peak centre). The resulting correlation heat map also showed clusters of related CAPs as well as both known and potentially novel interactions (Extended Data Fig. 5, Supplementary Table 3). Network plots based on pairwise peak overlaps highlighted a number of known interactions, including CTCF–RAD21 and CEBPA–CEBPG networks, as well as CAPs that associate with a large number of other CAPs, usually chromatin regulatory proteins such as SAP130, GATAD2A, and ARID5B (Extended Data Fig. 6b). We examined the associations at the level of called motifs by finding the peaks in each experiment where a specific called motif was present, limiting the analysis to the 293 high-confidence motifs. Upon identification of the primary motif, we looked for associations between motifs 1–40 bp away (Extended Data Fig. 6a, Supplementary Table 3). This analysis revealed the TFs (and motifs) that were more likely to associate with the motif of any other particular TF. RAD21 was highly associated with CTCF motifs, as expected, and we also found several other known complexes as well as some novel associations. FOXA1 peaks with the canonical Forkhead motif were more likely to contain relatively few motifs for other factors, but many factors, such as HNF4A, HNF4G, and RXRB, were enriched for nearby FOXA1 motifs.

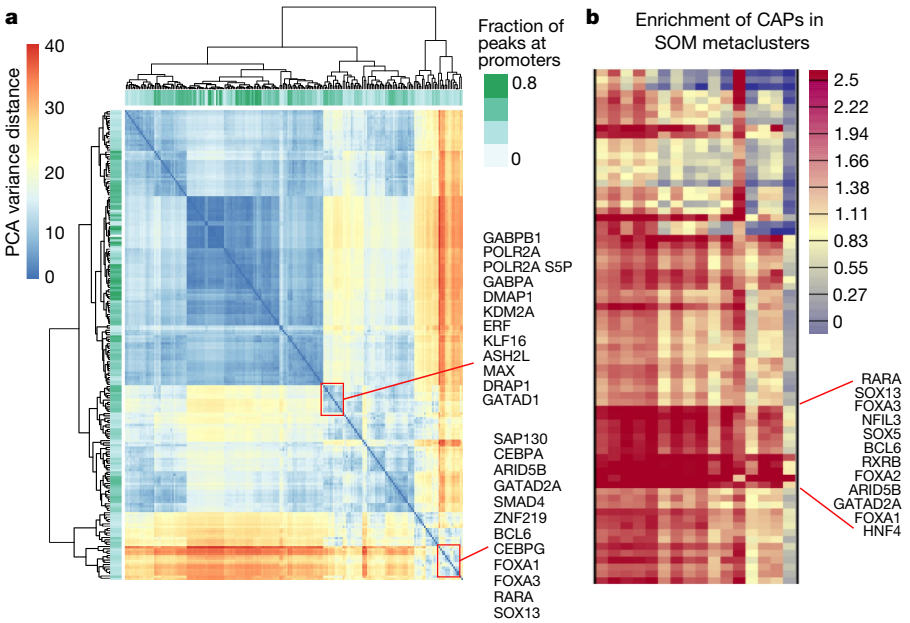


Fig. 4 | Co-localization of factors. a, Correlation matrix based on the cumulative principal component distances weighted by the proportion of variance explained by each component between all factors, derived from the PCA of all genomic loci with a peak containing at least two factors. **b**, SOM for a group of FOX TFs in HepG2 cells, with metaclusters showing major associations with specific factors.

To independently assess co-occupancy and provide an additional quantitative analysis, we trained a chromatin self-organizing map (SOM)[32] using all 208 CAPs with the SOMatic package[33]. We found key metaclusters around the key HepG2 TFs FOXA1/2 and HNF4A, in association with CAPs that are important for liver development, nucleosome remodelling (NuRD complex), and cohesin subunits (Fig. 4b, Extended Data Fig. 7a–f, Supplementary Notes).

The indirect motif, co-occupancy, and SOM analyses identified novel CAPs associated with GATAD2A, a core component of the NuRD complex. In GATAD2A CETCh–seq experiments, 53% of the GATAD2A peaks in HepG2 cells were annotated as active enhancers (Extended Data Fig. 8a), which was unexpected given the association of the NuRD complex with transcriptional repression and enhancer decommissioning[34–36]. GATAD2A has a very degenerate DNA-binding domain and is not predicted to bind DNA independently, and indeed the called GATAD2A motif matched FOXA3 (Fig. 5a). To assess co-localization in an additional, quantitative manner, we examined signal intensity[37] at shared and unique sites for GATAD2A and FOXA3 (Fig. 5b). Many of the unique sites showed signal above background, indicating a limitation of the conservative peak calls we used and adding support for extensive co-localization for these factors.

In our co-association analysis in HepG2 cells, we identified six CAPs that co-occurred with GATAD2A in discrete genomic regions (Fig. 5c). We analysed GATAD2A–FLAG protein immunoprecipitation by mass spectrometry and found that multiple components of the NuRD complex also co-immunoprecipitated with GATAD2A (Supplementary

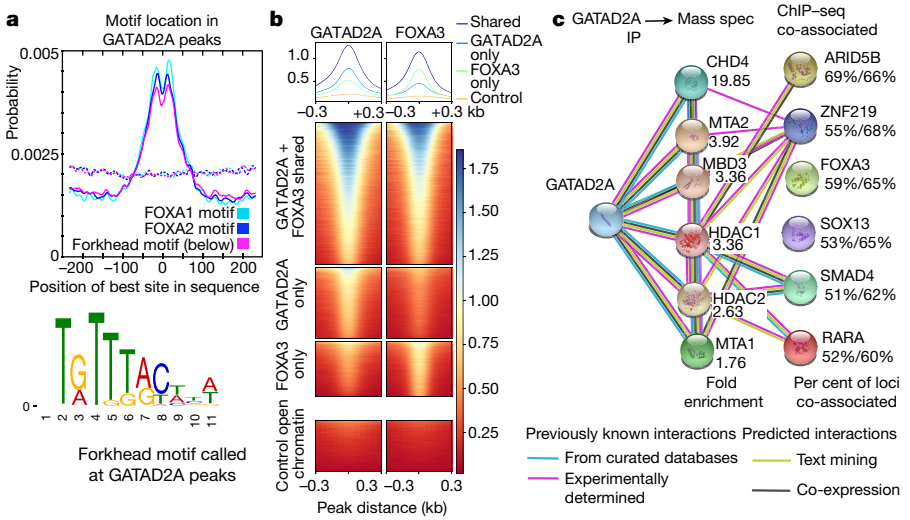


Fig. 5 | Analysis of GATAD2A co-localization. a, Presence of top motifs at GATA2DA-bound regions (top) and the top motif called at these peaks (bottom). **b**, Heat map showing signal intensity at shared and unique peaks for FOXA3 and GATAD2A. A set of random open chromatin regions is shown as a control. **c**, NuRD complex members and their identification through immunoprecipitation (IP)–mass spectrometry of GATAD2A immunoprecipitations, and through co-binding at GATAD2A-bound loci. Annotations from the String Database on protein interactions are shown as coloured lines connecting the proteins.

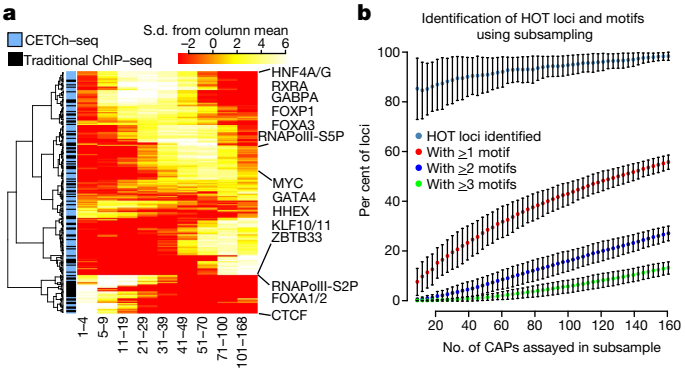


Fig. 6 | Association and motif trends in high CAP co-localization. a, CAP enrichment at loci with increasing number of factors bound. **b**, Subsampling plot showing the frequency of identification of motifs in HOT regions using increasing number of factors in permutations. Points represent median percentage of loci with one or more motifs (red), two or more motifs (dark blue), or three or more motifs (green) for CAPs bound at those regions; $n = 100$ iterations, lines from minimum to maximum.

Table 5). Of the GATAD2A-associated CAPs, ZNF219[38], SMAD4[39], and RARA[40] have previously been associated with the NuRD complex (Fig. 5c). We additionally identified ARID5B, SOX13, and FOXA3 (see above) as proteins that were associated with the known NuRD group, specifically at active enhancers where Forkhead binding sites were enriched (Fig. 5b, c). The classic NuRD complex has been suggested to function at enhancer regions associated with tissue-specific gene regulation[41], and our data confirm that the core NuRD component GATAD2A is recruited into these regions. Note that NuRD binding at these open and presumably active regions is thought to function through a NuRD complex that contains MBD3 and not MBD2, and our GATAD2A–FLAG immunoprecipitation–mass spectrometry data confirmed this, as MBD3 peptides but not MBD2 peptides immunoprecipitated with GATAD2A[42] (Supplementary Table 5).

We examined the expression of the genes nearest to peaks with both GATAD2A and FOXA3 association, as well as those with GATAD2A or FOXA3 binding but not both. All of these sites were near genes that were expressed at significantly higher levels than genes near random GC-matched sites (Extended Data Fig. 8b). Moreover, sites with both GATAD2A and FOXA3 peaks were near genes with significantly higher expression than those nearest sites with only GATAD2A or FOXA3 (Extended Data Fig. 8b). The genes nearest the GATAD2A–FOXA3 co-associated sites were enriched for liver biology gene ontology (GO) terms, including cholesterol metabolic processes and regulation of lipids, whereas FOXA3 sites without GATAD2A were near genes with additional liver biology GO terms, such as regulation of insulin and triglyceride biosynthesis, and GATAD2A sites without FOXA3 were enriched for negative regulation of sequence-specific DNA binding TFs (Extended Data Fig. 8c–e). Additional analyses indicated that there were strong associations between CAPs and important liver biology genes (Supplementary Notes, Supplementary Fig. 1).

CAPS in highly occupied regions

We examined how many factors were bound at putative HepG2 *cis*-regulatory elements by merging all peaks from all 208 CAP experiments, with a maximum merged size of 2 kb. This analysis yielded a total of 282,105 genomic sites with at least one associated CAP, with a maximum of 168 CAPs at one site. We investigated whether certain CAPs were more likely to co-occupy genomic loci with a high number of other CAPs, by performing hierarchical clustering of the degree of co-association for each CAP; this resulted in three distinct clusters (Fig. 6a). The first was a cluster of 33 proteins, including previously described key pioneer factors such as FOXA1 and FOXA2[43], which exhibit a low degree of co-occupancy with other CAPs at a relatively high proportion of their binding sites. The second cluster, comprised of 32 CAPs, displays frequent association at higher co-occupancy regions and is composed of CAPs already known to be recruited by, or to interact with, a large number of other CAPs, such as MYC and DNMT3B[44,45]. The third cluster contains the remaining CAPs, which exhibit an intermediate degree of co-occupancy, including key HepG2 TFs such as HNF4A and FOXA3.

As previously described[46–48], many regions in the genome are occupied by large numbers of CAPs in ChIP–seq assays (example shown in Extended Data Fig. 9a). There are several possible explanations for these HOT regions[49]. Some researchers have filtered all or the majority of these regions from analyses under the assumption that they are artefacts[50,51]. It is also possible that they are the result of stochastic shuffling of direct binding of many CAPs in a population of cells; when assayed across the millions of cells used for an individual ChIP–seq experiment, this could result in apparent co-localization of peaks for many CAPs that do not actually co-occupy at the same time in the same cell. The mechanisms that underlie this phenomenon might include indiscriminant recruitment driven by key CAPs or some unknown property of these regions of open chromatin, or by densely packed DNA sequence motifs. Another possible explanation is that three-dimensional genomic interactions, including enhancer looping and/or protein complexes, lead to ChIP–seq cross-linking of CAPs in close proximity.

We define HOT regions in these data ($n = 5{,}676$) as those sites with more than 70 CAPs (about one-third of all assayed CAPs) within a 2-kb region. Intersecting HOT regions with IDEAS segmentations revealed that more than 92% of HOT regions map to candidate promoter or strong enhancer-like states (42.25% and 49.88%, respectively). We determined using GREAT (genomic regions enrichment of annotations tool) analysis that promoter-localized HOT regions are associated with housekeeping genes and that distal HOT regions are near genes associated with liver-specific pathways (Extended Data Fig. 9b). In addition, the number of CAPs correlates with sequence conservation of the putative regulatory element and with the level of expression of the nearest gene (Extended Data Fig. 9c–e). While previous researchers have noted apparent general ChIP bias in favour of highly expressed genomic regions[51], we performed ChIP in untagged cells with an antibody raised against the epitope tag used in CETCh–seq experiments, normalizing for this background in peak calling, and the HOT regions continued to be strongly enriched (data not shown).

We computationally examined the general DNA motif structure of the HOT sites using two analyses. We first used a subsampling test to test whether motif information was gained as the numbers of CAPs assayed increased. We ran permutations of 12–162 CAPs and determined how often we could identify a HOT region as being bound by more than 33% of the CAPs in the subsample (Fig. 6b). More than 80% of the HOT loci were identified with only ten factors, and the curve approached 100% as the number of CAPs increased. We then investigated how often the motif for any associated CAP was found; fewer than 20% of sites had even a single motif identified with 40 or fewer CAPs. However, once more than 130 factors were included, over half the sites contained one or more identifiable motifs. While this analysis required only motif presence, we also found evidence of direct DNA–protein interactions using protein interaction quantification (PIQ)[52]—a computational tool that uses DNase-seq experiments and user-supplied motif sequences to identify direct TF binding sites. Using TF footprints identified in ENCODE HepG2 DNaseI hypersensitivity data by PIQ, we observed that the number of TF footprints was significantly positively correlated with the number of CAPs that had called peaks in a locus (Extended Data Fig. 10a–d). This observation was true at multiple PIQ purity (positive predictive value) thresholds and also when using TF footprints called in the same data set from JASPAR motifs. This is consistent with TF motif-driven architecture being a major characteristic of HOT regions. To determine whether CAP occupancy at highly bound regions is driven

by specific DNA motifs, we trained a support vector machine (SVM) on the sequences of 'HOT-motif' sites, a set of peaks with 50 or more co-localized motifs derived from the HOT sites ($n = 2{,}040$). We tested the predictive ability of the SVM as the number of TFs increased and found that predictions remained constant, rather than declining, further strengthening the notion that these sites are not artefacts (Extended Data Fig. 10e). The average precision recall area under curve (PR-AUC) scores for the SVM were about 0.74 for motif-level predictions and about 0.66 for peak-level predictions. These scores were substantially higher than expected, given the random sample of a positive set of 5,000 sites tested against 50,000 GC-matched null sequences as the negative set (Extended Data Fig. 10f). We also found, using the *k*-mers generated by the SVM, that there are 1–5 TFs at each site with very high motif scores, and about 25–50 TFs with degenerate or weaker motifs (Extended Data Fig. 10g); this was true for both HOT-motif sites and the broader HOT sites.

We investigated whether this observation was unique to HOT regions ($n = 5{,}676$) when compared to an equal number of enhancer regions (as defined by IDEAS segmentation) with only 2–10 associated CAPs, or to a null set of random enhancer elements with any number (0–208) of associated CAPs. Sites with 2–10 CAPs had substantially smaller numbers of both high-affinity and low-affinity TF motifs, and the random enhancers were essentially devoid of strong motifs (Extended Data Fig. 11a–g). The distribution of SVM scores in HOT sites was significantly higher than that of the SVM scores of sites with 2–10 associated CAPs (Kolmogorov–Smirnov test, $P = 5.966 \times 10^{-11}$), and both were significantly higher than that of the null set of random enhancer elements (Kolmogorov–Smirnov test, $P < 2.2 \times 10^{-16}$ for each), indicating that the information imparted by the DNA sequence of HOT sites exceeds that of other *cis*-regulatory elements (Extended Data Fig. 11h). Moreover, in HOT sites, the strongest-affinity TF at any individual peak varied across sites, indicating that many different CAPs are involved in regulation at these sites. Important liver TFs, such as FOXA3, HNF1A, and CEBPA, had the strongest putative motif affinity at many of these sites (Extended Data Fig. 11i). This supports the notion that HOT sites are driven by a few strong and specific TF–DNA interactions and non-specific recruitment of other factors, probably through both protein complexes and binding to degenerate motifs, and possibly linking together multiple distal genomic regions through CAP interactions. Thus, it is essential to generate complete CAP maps to determine the full complement of CAPs associated with each locus, which would not occur by analysis of functional motifs alone.

Discussion

This study introduces a data resource of occupancy maps for human transcription factors, transcriptional cofactors, histone-binding or histone-modifying proteins, and other chromatin regulators that illustrates the strengths of building towards a complete catalogue of CAP interactions in an individual cell type. At this intermediate stage of completeness, the aggregated data enabled us to identify known complexes and associations, and to identify putative novel associations. We also gained insights into gene regulatory principles, clearly showing the segregation of categories of CAPs associated with particular genomic states, including promoters and enhancers, and uncovering DNA sequence motifs at the majority of HOT regions that would have been impossible with fewer CAPs assayed.

The large number of CAPs assayed provided the capacity to identify and study regions of the genome associated with very high numbers of CAPs, compared with expectations from detailed work on specific enhancer complexes such as the interferon enhanceosome[53]. Multiple lines of evidence argue that, as a group, the regions at which high numbers of CAPs were detected are neither biological noise associated with general open chromatin nor ChIP–seq or CETCh–seq artefacts. HOT regions have been previously described as being depleted of TF motifs, but we suggest that this was likely to be because earlier analyses lacked a large enough sampling of key TFs with strong 'anchoring' motifs. We propose a model in which HOT regions are nucleated by anchoring DNA motifs and their cognate TFs. They would form a core, with which many other CAPs associate by presumed protein–protein interactions, protein–RNA interactions, and relatively weak DNA interactions at poorer sequence–motif matches. Extensive apparent co-occupancy at domains possessing few or no anchor motifs can potentially be explained when the ChIP assay captures, through assumed protein–protein fixation, non-adjacent DNA regions that associate with each other by looping interactions.

It is important to appreciate that the standard ChIP assay is performed on populations of large numbers of cells. Patterns of computational co-occupancy cannot discriminate between the simultaneous association of many CAPs in a single large molecular complex and diversified smaller complexes that are distributed at any given time across the cell population, with each containing a smaller number of secondary associations, which sum to give massive computational co-occupancy. We can, however, state that at individual known transcriptional enhancers with more than 70 CAPs, the ChIP signal for identified anchor factors was significantly higher in magnitude than at enhancers with fewer CAPs.

The results thus far argue that a fully comprehensive catalogue of all CAPs will help us to distinguish among these possibilities, which are not mutually exclusive. Completeness should also contribute to the identification of additional novel motifs, and, in the cases of indirect motifs found for TFs with known direct motifs, allow more accurate motif calling. In addition, a complete catalogue of CAPs in a single cell type will support the imputation of critical contacts in CAP networks for three-dimensional assembly of genomic enhancer–promoter organization that is not possible from a few individual CAP binding maps, as demonstrated by our findings regarding the NuRD complex. The ENCODE Project continues to produce additional occupancy maps and to expand cellular contexts for these assays. We anticipate more large-scale analyses such as this, and hope that the perspectives gained from these will inform more targeted research endeavours and generate meaningful hypotheses.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2023-4.

1. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet*. **10**, 252–263 (2009).
2. Lambert, S. A. et al. The human transcription factors. *Cell* **172**, 650–665 (2018).
3. Yosef, N. et al. Dynamic regulatory network controlling T_H17 cell differentiation. *Nature* **496**, 461–468 (2013).
4. Busskamp, V. et al. Rapid neurogenesis through transcriptional activation in human stem cells. *Mol. Syst. Biol*. **10**, 760 (2014).
5. Chen, X. et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106–1117 (2008).
6. Iwafuchi-Doi, M. & Zaret, K. S. Pioneer transcription factors in cell reprogramming. *Genes Dev*. **28**, 2679–2692 (2014).
7. Wingender, E., Schoeps, T. & Dönitz, J. TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res*. **41**, D165–D170 (2013).
8. Weirauch, M. T. et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
9. Cowper-Sal-lari, R. et al. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet*. **44**, 1191–1198 (2012).
10. Dror, I., Golan, T., Levy, C., Rohs, R. & Mandel-Gutfreund, Y. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res*. **25**, 1268–1280 (2015).
11. Dasen, J. S., Tice, B. C., Brenner-Morton, S. & Jessell, T. M. A Hox regulatory network establishes motor neuron pool identity and target-muscle connectivity. *Cell* **123**, 477–491 (2005).

12. Black, J. B. et al. Targeted epigenetic remodeling of endogenous loci by CRISPR/Cas9-based transcriptional activators directly converts fibroblasts to neuronal cells. *Cell Stem Cell* **19**, 406–414 (2016).
13. Visel, A. et al. ChIP–seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
14. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
15. Mikkelsen, T. S. et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
16. Robertson, G. et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657 (2007).
17. Savic, D. et al. CETCh–seq: CRISPR epitope tagging ChIP–seq of DNA-binding proteins. *Genome Res.* **25**, 1581–1589 (2015).
18. Partridge, E. C., Watkins, T. A. & Mendenhall, E. M. Every transcription factor deserves its map: scaling up epitope tagging of proteins to bypass antibody problems. *BioEssays* **38**, 801–811 (2016).
19. Zhang, Y., An, L., Yue, F. & Hardison, R. C. Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res.* **44**, 6721–6731 (2016).
20. Gerstein, M. B. et al. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012).
21. Mendenhall, E. M. et al. GC-rich sequence elements recruit PRC2 in mammalian ES cells. *PLoS Genet.* **6**, e1001244 (2010).
22. Kowalczyk, M. S. et al. Intragenic enhancers act as alternative promoters. *Mol. Cell* **45**, 447–458 (2012).
23. Dao, L. T. M. et al. Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat. Genet.* **49**, 1073–1081 (2017).
24. Andersson, R., Sandelin, A. & Danko, C. G. A unified architecture of transcriptional regulatory elements. *Trends Genet.* **31**, 426–433 (2015).
25. Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32**, D91–D94 (2004).
26. Mathelier, A. et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **44**, D110–D115 (2016).
27. Oliphant, A. R., Brandl, C. J. & Struhl, K. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Mol. Cell. Biol.* **9**, 2944–2949 (1989).
28. Worsley Hunt, R. & Wasserman, W. W. Non-targeted transcription factors motifs are a systemic component of ChIP–seq datasets. *Genome Biol.* **15**, 412 (2014).
29. Jolma, A. et al. DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).
30. Morgunova, E. & Taipale, J. Structural perspective of cooperative transcription factor binding. *Curr. Opin. Struct. Biol.* **47**, 1–8 (2017).
31. Wei, B. et al. A protein activity assay to measure global transcription factor activity reveals determinants of chromatin accessibility. *Nat. Biotechnol.* **36**, 521–529 (2018).
32. Mortazavi, A. et al. Integrating and mining the chromatin landscape of cell-type specificity using self-organizing maps. *Genome Res.* **23**, 2136–2148 (2013).
33. Longabaugh, W. J. R. et al. Bcl11b and combinatorial resolution of cell fate in the T-cell gene regulatory network. *Proc. Natl Acad. Sci. USA* **114**, 5800–5807 (2017).
34. Whyte, W. A. et al. Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. *Nature* **482**, 221–225 (2012).
35. Liang, Z. et al. A high-resolution map of transcriptional repression. *eLife* **6**, e22767 (2017).
36. Zhang, Y. et al. Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes Dev.* **13**, 1924–1935 (1999).
37. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
38. Huttlin, E. L. et al. The BioPlex Network: a systematic exploration of the human interactome. *Cell* **162**, 425–440 (2015).
39. Faherty, N. et al. Negative autoregulation of BMP dependent transcription by SIN3B splicing reveals a role for RBM39. *Sci. Rep.* **6**, 28210 (2016).
40. Choi, W. I. et al. Promyelocytic leukemia zinc finger-retinoic acid receptor α (PLZF-RARα), an oncogenic transcriptional repressor of cyclin-dependent kinase inhibitor 1A (p21WAF/CDKN1A) and tumor protein p53 (TP53) genes. *J. Biol. Chem.* **289**, 18641–18656 (2014).
41. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
42. Günther, K. et al. Differential roles for MBD2 and MBD3 at methylated CpG islands, active promoters and binding to exon sequences. *Nucleic Acids Res.* **41**, 3010–3021 (2013).
43. Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.* **25**, 2227–2241 (2011).
44. Conacci-Sorrell, M., McFerrin, L. & Eisenman, R. N. An overview of MYC and its interactome. *Cold Spring Harb. Perspect. Med.* **4**, a014357 (2014).
45. Hervouet, E., Vallette, F. M. & Cartron, P. F. Dnmt3/transcription factor interactions as crucial players in targeted DNA methylation. *Epigenetics* **4**, 487–499 (2009).
46. Boyle, A. P. et al. Comparative analysis of regulatory information and circuits across distant species. *Nature* **512**, 453–456 (2014).
47. Gerstein, M. B. et al. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330**, 1775–1787 (2010).
48. Moorman, C. et al. Hotspots of transcription factor colocalization in the genome of *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **103**, 12027–12032 (2006).
49. Wreczycka, K. et al. HOT or not: examining the basis of high-occupancy target regions. *Nucleic Acids Res.* **47**, 5735–5745 (2019).
50. Shin, H., Liu, T., Duan, X., Zhang, Y. & Liu, X. S. Computational methodology for ChIP–seq analysis. *Quant. Biol.* **1**, 54–70 (2013).
51. Teytelman, L., Thurtle, D. M., Rine, J. & van Oudenaarden, A. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc. Natl Acad. Sci. USA* **110**, 18602–18607 (2013).
52. Sherwood, R. I. et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.* **32**, 171–178 (2014).
53. Panne, D., Maniatis, T. & Harrison, S. C. An atomic model of the interferon-β enhanceosome. *Cell* **129**, 1111–1123 (2007).

Methods

ChIP–seq and CETCh–seq

All protocols for ChIP–seq and CETCh–seq have been previously published and are available at the ENCODE web portal (https://www.encodeproject.org/documents/). In brief, HepG2 cells were obtained from ATCC (HB-8065), confirmed by morphological observation, and tested for mycoplasma (ThermoFisher C7028). Pools of cells were grown separately to represent replicate experiments. Crosslinking of cells was performed with 1% formaldehyde for 10 min at room temperature and the chromatin was sheared using a Bioruptor Twin instrument (Diagenode). Antibody characterization standards are published on the ENCODE web portal and consist of a primary validation (western blot or immunoprecipitation–western blot) and a secondary validation (immunoprecipitation followed by mass spectrometry) for traditional antibody ChIP–seq. With CETCh–seq experiments, a molecular validation (PCR or Sanger sequencing confirmation of edited genes) in addition to one of the immunological validations (western blot, immunoprecipitation–western blot, or immunoprecipitation–mass spectrometry) is required for release. Raw fastq data were downloaded from the publicly available ENCODE Data Coordination Center, and aligned to the human reference genome (hg19) using the BWA-0.7.12 (Burrows Wheeler Aligner) alignment algorithm[54]. Post-alignment filtering steps were carried out using samtools-1.3[55] with MAPQ threshold of 30, and duplicate removal was performed using picard-tools-1.88 (http://broadinstitute.github.io/picard/). After filtering, each CAP's genome-wide binding sites (peak enrichment) were computed using phantompeakqualtools, implementing the SPP algorithm[56,57], with replicate consistency and peak ranking determined by irreproducible discovery rate (IDR) using the IDR-2.0.2 tool[56] to generate narrow peaks passing IDR cutoff 0.02 (soft-idr-threshold). ENCODE blacklisted regions (wgEncodeDacMapabilityConsensusExcludable.bed.gz, downloadable from the UCSC genome browser at https://genome.ucsc.edu/) were filtered out. In addition, we note that plasmids used to generate edited cells with epitope-tagged CAPs have been deposited to Addgene, the non-profit plasmid repository, and are available for researchers to tag particular CAPs in other cell lines of interest. We also note that the GC content of DNA has been reported as a source of bias in ChIP–seq data, leading to over-representation of TFBSs and false positive peak calls, which could confound subsequent analyses[58,59]. To address this concern, we performed ChIP–seq experiments in unedited cell lines using the FLAG antibody (Sigma F1804) that we use in CETCh–seq, and used these libraries as background for peak calling. In these experiments, the only variable is the edited cell line used as foreground, and most biases should be accounted for.

De novo sequence motif analysis

To identify enriched sequence motifs in the binding sites of CAPs, de novo sequence motif and motif enrichment analysis were performed using the MEME-ChIP[60] suite and the pipeline was built as previously described[61], on 500-bp regions centred on peak summits based on the hg19 reference genome fasta. The top five motifs per data set were reported from the top 500 peaks based on signal value, using 2× random/null sequence with matched size, GC content and repeat fraction as a background. Central motif enrichment analysis was performed using Centrimo[62], to infer the most centrally enriched motifs with de novo motifs generated from the pipeline against the 2× null sequence background.

Comparative motif analysis

De novo motifs generated from CAPs were filtered for high-confidence motifs, including only those that were highly significant and strongly enriched in binding sites, based on MEME $E < 1 \times 10^{-5}$, Centrimo $E < 1 \times 10^{-10}$ and Centrimo binwidth <150. High confidence motifs were then compared, and quantified for similarity against the previously derived or known motifs available in the CIS-BP build 1.02 and JASPAR 2016/2018 databases[8,25,26] using the Tomtom quantification tool[63]. Tomtom *E*-values <0.05 represent highly similar motifs, and >0.05 represent motifs with increasing magnitude of dissimilarity, or more distantly related motifs.

Gene expression

RNA-seq quantification data for 56 cell lines and 37 tissues were retrieved from the Human Protein Atlas (version 17, downloadable from https://www.proteinatlas.org/)[64], and used to identify 57 genes that were highly and specifically expressed in liver as compared to all other cell and tissue types, and also found in HepG2 cells with at least 10 TPM. On average, these 57 liver-specific genes were 151.21 times more highly expressed than in any other cell type.

IDEAS segmentation

IDEAS segmentation for six cell-types (HepG2, GM12878, H1hESC, HUVEC, HeLaS3, and K562) were collected from the Penn State Genome Browser (http://main.genome-browser.bx.psu.edu/). All promoter-like and enhancer-like regions identified in at least one of five other cell lines were merged using pybedtools[65,66] and these regions were filtered from the HepG2 segmentation. Significant enrichment of CAPs in the *cis*-regulatory regions was evaluated using Fisher's exact test (adjusted $P < 0.001$, BH FDR corrected) against random or null sequences with matched length, GC content and repeat fraction using null sequence python script from Kmer-SVM[67]. Heat maps were generated using the heatmap.2 function from R gplots package (https://cran.r-project.org/web/packages/gplots/).

GREAT analysis

Cis-regulatory associated highly CAP bound sites were binned into promoter-associated and enhancer-associated sites using IDEAS segmentation. To assess the biological function and relevance of these highly occupied sites, GREAT[68] analysis was performed to predict the function of these *cis*-regulatory regions (http://bejerano.stanford.edu/great/public/html/) by associating the genomic regions to genes from various ontologies such as GO molecular function, MSigDB and BioCyc pathway. The parameters used for GREAT analysis were Basal+extension (constitutive 5.0 kb upstream and 1.0 kb downstream, up to 50.0 kb max extension) for all enhancer-associated sites, and Basal+extension (constitutive 5.0 kb upstream and 1.0 kb downstream, up to 5.0 kb max extension) for all promoter-associated regions with whole-genome background. MSigDB pathway[69,70] was noted for genomic region enrichment analysis.

GERP analysis

Genomic evolutionary rate profiling (GERP) was performed to assess whether highly bound *cis*-regulatory sites, categorized into promoter or enhancer-associated, correlate with increased evolutionary constraints. A highly constrained elements bed file containing high-confidence regions (significant *P*) generated from per base GERP scores was retrieved from the Sidow laboratory at Stanford (http://mendel.stanford.edu/SidowLab/downloads/gerp/). The fraction of overlapping bases for each bin of the 'CAP bound category' (low to high) with highly constrained elements was computed using bedtools-2.26.0[66] and pandas-0.20.3, python2.7, further normalized by the fraction of 'highly constrained elements' overlapping per 100-bp region of CAP bound categories. In addition, the Kolmogorov–Smirnov test was performed to evaluate statistically significant differences in distribution between the highly bound (20+ CAP bound) and not highly bound regions (1–19 CAP bound sites) for both promoter- and enhancer-associated sites.

Co-binding analysis

Pairwise overlap of binding sites between each of the 208 CAPs was performed with 50 bp up- and downstream from the summit of peaks using

python-based pybedtools[65,66]. All other computations, and the pairwise peak overlap percentage for each CAP to build the pairwise matrix, were performed using pandas-0.20.3, python2.7 (Python Software Foundation) to construct network plots, using R igraph, implementing the Fruchterman Reingold algorithm. The interconnection between CAP shared binding sites for 208 CAPs was built with a minimum threshold of 75% or more overlap between any two CAPs. The sizes of vertices and nodes in the graph are representative of the number of connections each CAP has with its connected partner, while edges represent the degree of overlap between CAPs.

Co-binding was characterized by merging IDR-passing narrow peak files from 208 CAPs with the 'merge' function from the bedtools software package[71]. A minimum of 1 bp overlap was required and resultant peaks greater than 2 kb (~1%) were filtered from downstream analysis. Hierarchical clustering, using the Euclidean distance metric and Ward clustering method, of CAPs based on degree of co-binding was performed in R with the 'heatmap.2' function of the gplots package.

LS-GKM SVM analysis

At peak level, LS-GKM support vector machines (SVMs)[72] were trained on a random sample of up to 5,000 narrow peaks (using all peaks for those with fewer) as a positive set against 10× random/null sequence with matched size, GC-content and repeat fraction as a negative set. At motif level, LS-GKM support vector machines (SVMs)[72] were trained on a sample of 5,000 random motif sites found by FIMO (MEME-suite), extending ±15 bp, for all TFs (n = 171), as a positive set against the 10× random-null sequence with GC content and repeat fraction matched sequence as a negative set.

Null genomic sequences matched to observed binding events were obtained using the 'nullseq_generate.py' function available with the LS-GKM package. The fold number of sequences (−x) was set to ten and the random seed (−r) was set to 1. SVMs were trained using the 'gkmtrain' function with a k-mer length (−l) of 11, kernel function (−t) of 4, regularization parameter (−c) of 1, number of informative columns (−k) of 7, and maximum number of mismatches (−d) of 3. Precision-recall areas under the curve (PR-AUC) were calculated by obtaining the tenfold cross-validation results from 'gkmtrain' (after setting the −x flag to 10), and inputting the results into the 'pr.curve' function of the PRROC R package, resulting in mean PR-AUC of 0.66 at the peak level, and 0.74 at the motif level. Classifier values for all bound sequences were obtained using the 'gkmpredict' function, and HOT sites (n = 5,676) were scored with each CAP to assess their putative binding affinity at HOT regions, and percentile ranked to obtain the top 5% and bottom 75% k-mer compared to enhancers with 2–10 associated TFs (n = 5,676) and to random enhancers with any number of associated factors (0+) (n = 5,676).

Random forest and PCA analysis

PCA was performed on a CAP binding matrix composed of the presence or absence of motif in merged peaks as a binary matrix of loci, and implementing the python-based ML library scikit-learn Sklearn (0.19.0)[73]. Plots for motif-based analyses were generated using the R package ggplot2[74] and complex Heatmap[75]. A random forest classifier was trained on merged CAP binding matrices at both motif and peak level to predict *cis*-regulatory elements (promoter or enhancer, by IDEAS annotation) using the R package ranger[76], a faster implementation of random forest in R, and also tested using Sklearn 0.19.0. The median OOB (out-of-bag) error estimate was computed for 100 instances of randomly sampled (n = 1,000) loci iterations, to compute the element classification and misclassification accuracy using confusion matrix.

Immunoprecipitation with mass spectrometry

Whole-cell lysates of FLAG-tagged or unedited HepG2 cells (~20 million) were immunoprecipitated using a primary antibody raised against FLAG or the CAP, respectively. The immunoprecipitation fraction was loaded on a 12% TGX gel and separated with the Mini-PROTEAN Tetra Cell System (Bio-Rad). The whole lane was excised and sent to the University of Alabama at Birmingham Cancer Center Mass Spectrometry/Proteomics Shared Facility. The sample was analysed on a LTQ XL Linear Ion Trap Mass Spectrometer by liquid chromatography electrospray ionization with tandem mass spectrometry (LC–ESI–MS/MS). Peptides were identified using SEQUEST tandem mass spectral analysis with probability based matching at $P < 0.05$. SEQUEST results were reported with ProteinProphet protXML Viewer (TPP v4.4 JETSTREAM) and filtered for a minimum probability of 0.9. For ENCODE antibody characterization standards, all protein hits that met these criteria were reported, including common contaminants. Fold enrichment for each protein reported was determined using a custom script based on the FC-B score calculation[77]. Following ENCODE antibody characterization guidelines, the CAP must be in the top 20 enriched proteins identified by immunoprecipitation–MS, and the top CAP overall for release. For GATAD2A co-associated TFs, the peptides with minimum 0.9 probability were present in smaller quantities than those of GATAD2A.

TF footprints analysis

To identify TF footprints for comparison to ChIP–seq binding sites, we used PIQ[52]. ENCODE HepG2 DNase-seq raw FASTQs (paired-end 36 bp) of roughly equivalent size (accession numbers: ENCFF002EQ-G, -H, -I, -J, -M, -N, -O, -P) were downloaded from the ENCODE portal and processed using ENCODE DNase-seq standard pipeline (available at https://github.com/kundajelab/atac_dnase_pipelines) with flags: -species hg19 -nth 32 -memory 250G -dnase_seq -auto_detect_adapter -nreads 15000000 -ENCODE3. Processed BAM files were merged and used as input for PIQ TF footprinting using each TF's top motif position weight matrix (PWM). Next, identified TF footprints from every TF that met a specified PIQ purity (positive predictive value) were intersected with all identified ChIP–seq binding sites using BEDtools to correlate the number of unique TF footprints with the number of ChIP–seq factors identified at a given ChIP–seq binding site.

SOM analysis

The SOM was trained with the SOMatic package[33] using the previous chromatin analysis partitioning strategy[32] with modifications as described below. We calculated the RPKM of each data set's first replicate over each of the 951,022 genomic segments to build a training matrix. We used each data set's second replicate to build a separate scoring matrix. The training matrix was used to train five trial self-organizing maps with a toroid topology with size 40 × 60 units using 10 million time steps (~10 epochs) and selected the best, based on fitting error using the scoring matrix, for further analysis, and segments were assigned to their closest units based on the scoring matrix.

To properly fit the data, SOM units with similar profiles across experiments were grouped into metaclusters using SOMatic. In brief, metaclustering was performed using k-means clustering of the unit profiles to determine centroids for groups of units. Metaclusters were built around these centroids so that all of the units in a cluster remained connected. SOMatic's metaclustering function attempts all metacluster numbers within a range given and scores them on the basis of Akaike information criterion (AIC)[78]. The penalty term for this score is calculated using a parameter called the dimensionality, which is the number of independent dimensions in the data, which in this case are the individual cell subtypes. To estimate this number, we used a 60% cut on a hierarchical clustering done on the SOM unit vectors. For this work, the dimensionality was calculated to be 6. For metaclustering, all k between 50 and 250, with 64 trials, were tested and metacluster number 196 had the lowest AIC score and was chosen for further analysis.

To generate decision trees for these metaclusters, each of the segments in the training matrix was labelled with its final metacluster. For each metacluster, if the metacluster is of size n, n segments of other

clusters were chosen randomly, and this set of positive and negative examples was split, using 80% of the examples for training and 20% for scoring. The training data were fed through an R script using the rpart and rattle packages to create, score, prune, and re-score a tree for each metacluster. This entire process was repeated for 100 trials with only the tree with the highest accuracy drawn.

Reporting summary

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

Data availability

Data sets generated from this study are available at the ENCODE portal or at the Gene Expression Omnibus under accession number GSE104247. CETCh–seq reagents are available at https://www.addgene.org/crispr/tagging/.

Code availability

All code is available at https://github.com/chhetribsurya/PartridgeChhetri_etal.

54. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
55. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
56. Landt, S. G. et al. ChIP–seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res*. **22**, 1813–1831 (2012).
57. Kharchenko, P. V., Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP–seq experiments for DNA-binding proteins. *Nat. Biotechnol*. **26**, 1351–1359 (2008).
58. Worsley Hunt, R., Mathelier, A., Del Peso, L. & Wasserman, W. W. Improving analysis of transcription factor binding sites within ChIP–seq data based on topological motif enrichment. *BMC Genomics* **15**, 472 (2014).
59. Teng, M. & Irizarry, R. A. Accounting for GC-content bias reduces systematic errors and batch effects in ChIP–seq data. *Genome Res*. **27**, 1930–1938 (2017).
60. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011).
61. Ma, W., Noble, W. S. & Bailey, T. L. Motif-based analysis of large nucleotide data sets using MEME-ChIP. *Nat. Protocols* **9**, 1428–1450 (2014).
62. Bailey, T. L. & Machanick, P. Inferring direct DNA binding from ChIP–seq. *Nucleic Acids Res*. **40**, e128 (2012).
63. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol*. **8**, R24 (2007).
64. Uhlén, M. et al. Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
65. Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423–3424 (2011).
66. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
67. Fletez-Brant, C., Lee, D., McCallion, A. S. & Beer, M. A. kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. *Nucleic Acids Res*. **41**, W544–W556 (2013).
68. McLean, C. Y. et al. GREAT improves functional interpretation of *cis*-regulatory regions. *Nat. Biotechnol*. **28**, 495–501 (2010).
69. Liberzon, A. et al. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst*. **1**, 417–425 (2015).
70. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
71. Quinlan, A. R. BEDTools: the swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* **47**, 11.12.11–11.12.34 (2014).
72. Ghandi, M. et al. gkmSVM: an R package for gapped-kmer SVM. *Bioinformatics* **32**, 2205–2207 (2016).
73. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *JMLR* **12**, 2825–2830 (2011).
74. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, New York, 2016).
75. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
76. Wright, M. N. & Ziegler, A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw*. **77**, 1–17 (2017).
77. Mellacheruvu, D. et al. The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat. Methods* **10**, 730–736 (2013).
78. Akaike, H. Information theory and an extension of the maximum likelihood principle. *Intl Symp. Information Theory* 267–281 (1973).

Acknowledgements Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number U54HG006998 to R.M.M. and E.M.M. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was also supported by funds from The HudsonAlpha Institute for Biotechnology. We thank R. Nguyen, D. Moore, and M. McEown for their technical efforts in this study; B. S. Roberts and G. M. Cooper for comments; HudsonAlpha's Genomic Services Laboratory led by S. Levy for the high-throughput sequencing of much of the data used in this paper; and members of the ENCODE Consortium for public deposition of data generated by other Consortium groups.

Author contributions E.C.P., M.M., K.M.N., L.A.B., S.K.M., C.L.M., C.J.C., E.C.D., and D.S. developed the CETCh–seq method and performed ChIP–seq and CETCh–seq experiments and accompanying validations; S.B.C. performed peak calling and mapped TF binding sites; S.B.C. and E.C.P. performed motif analyses, gene expression analyses, IDEAS segmentation analyses, and co-association analyses; J.W.P. and S.B.C. performed GATAD2A analyses and experiments; M.M. performed immunoprecipitation–mass spectrometry analyses and managed the production of ChIP–seq and CETCh–seq experiments; C.S.J., S.J., and A.M. performed SOM analyses; S.B.C. and S.-T.G. performed conservation and co-association analyses; S.B.C., R.C.R., and A.A.H. performed LS-GKM SVM, random forest, PCA, and TF footprint analyses; E.C.P., S.B.C., B.J.W., R.M.M., and E.M.M. conceived and designed the study; R.M.M. and E.M.M. directed the study; E.C.P., S.B.C., and E.M.M. wrote the manuscript with assistance from all authors; and all authors read and approved the manuscript.

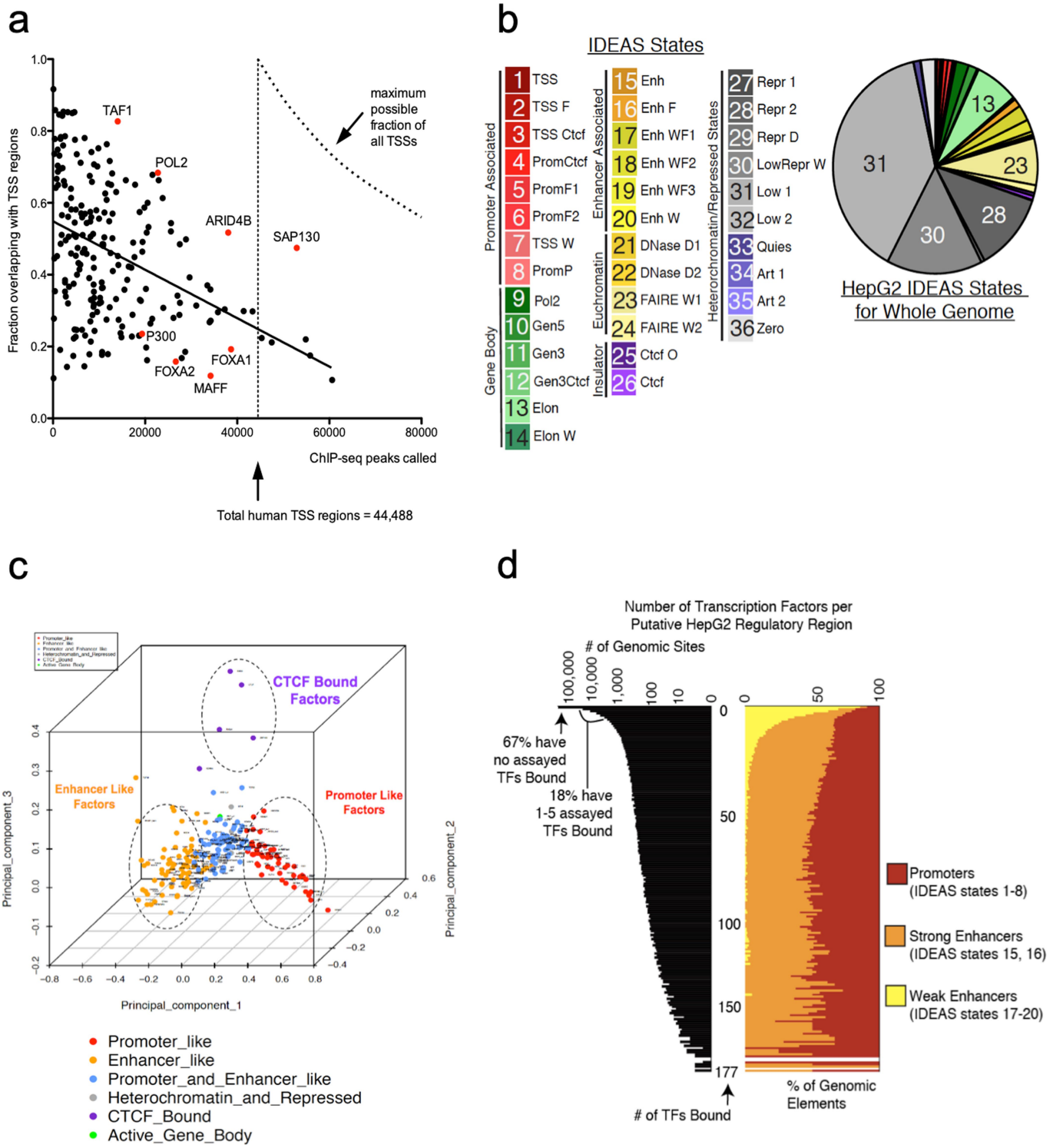
Competing interests The authors declare no competing interests.

Additional information

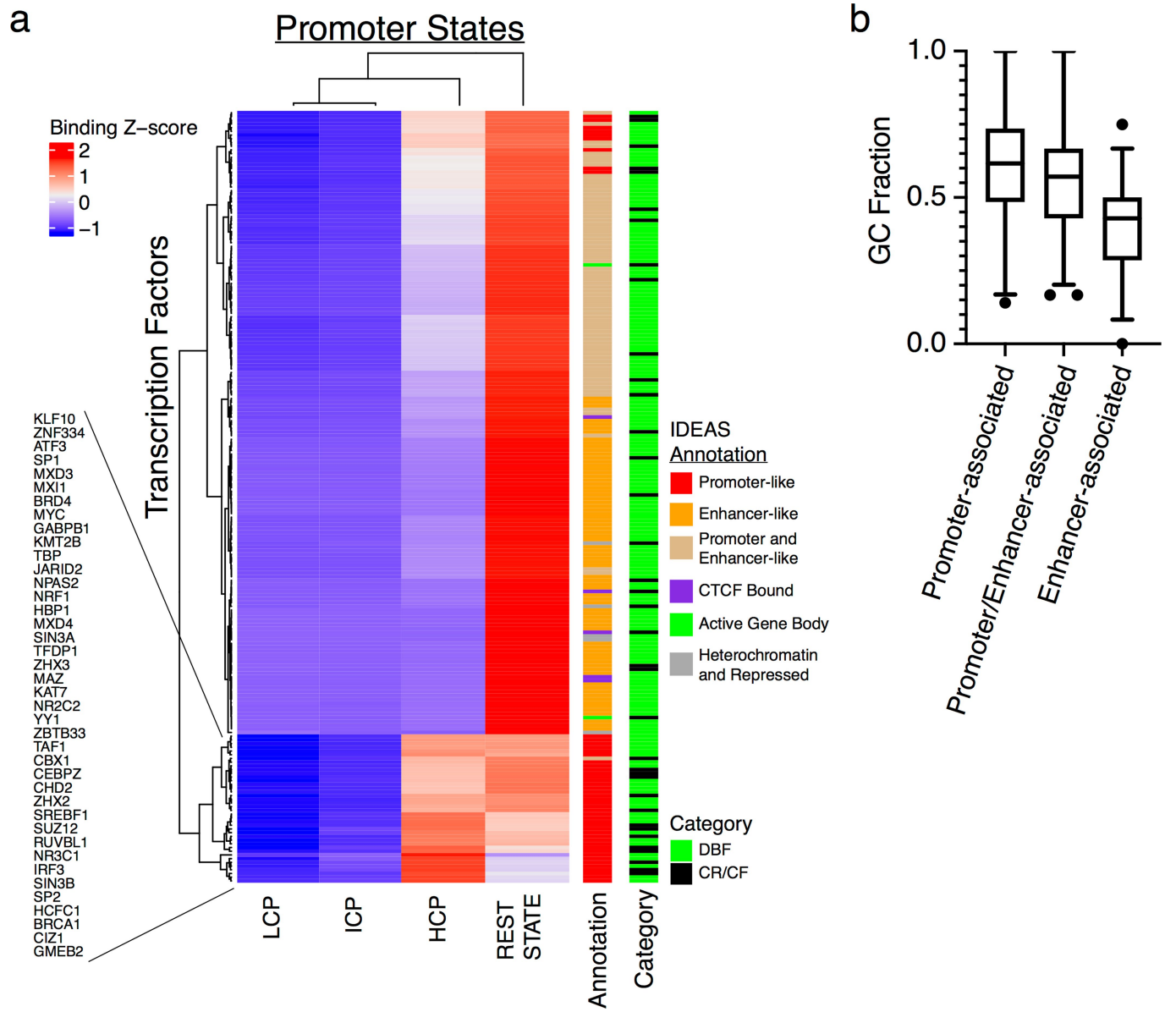
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2023-4.

Correspondence and requests for materials should be addressed to R.M.M. or E.M.M.

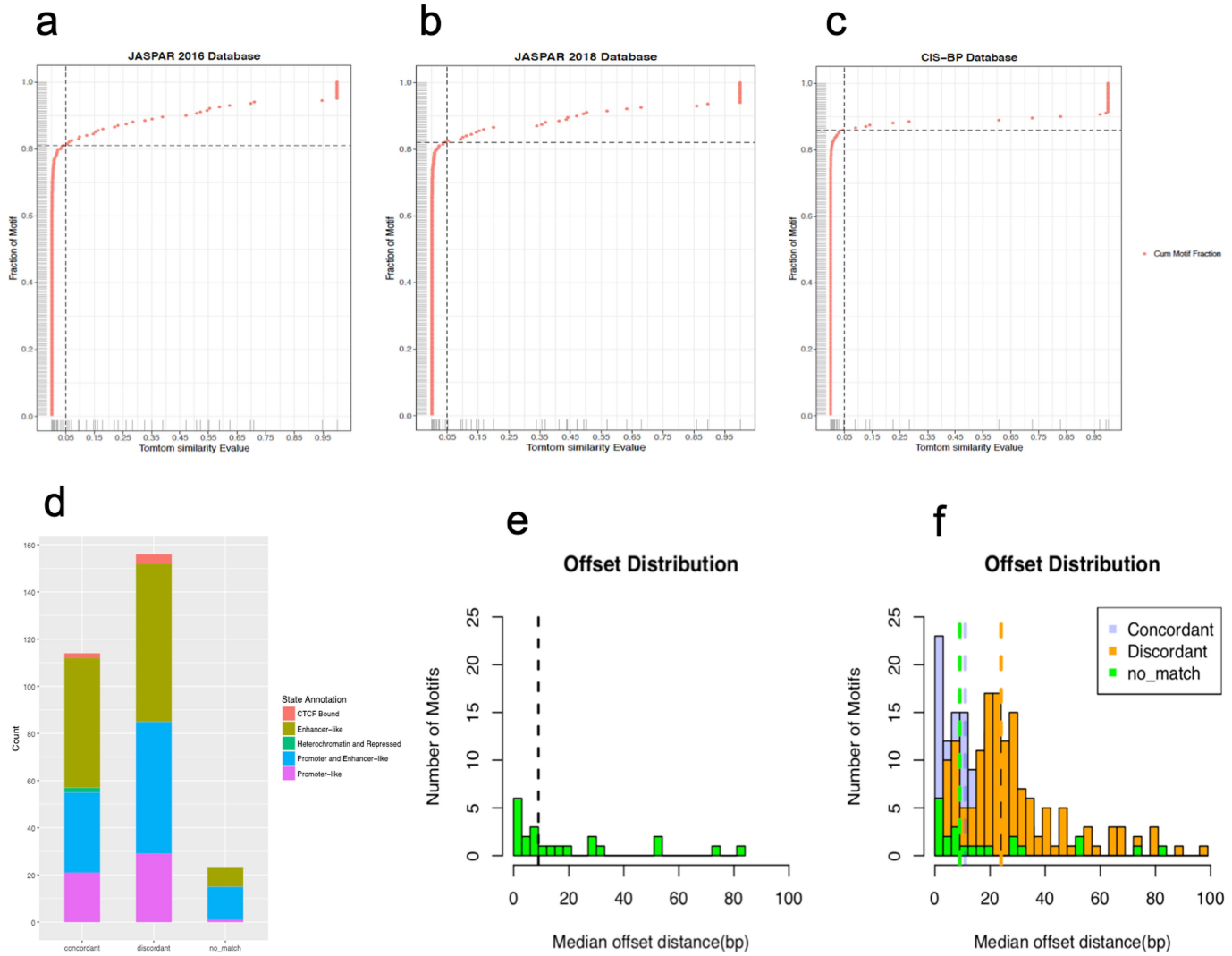
Reprints and permissions information is available at http://www.nature.com/reprints.



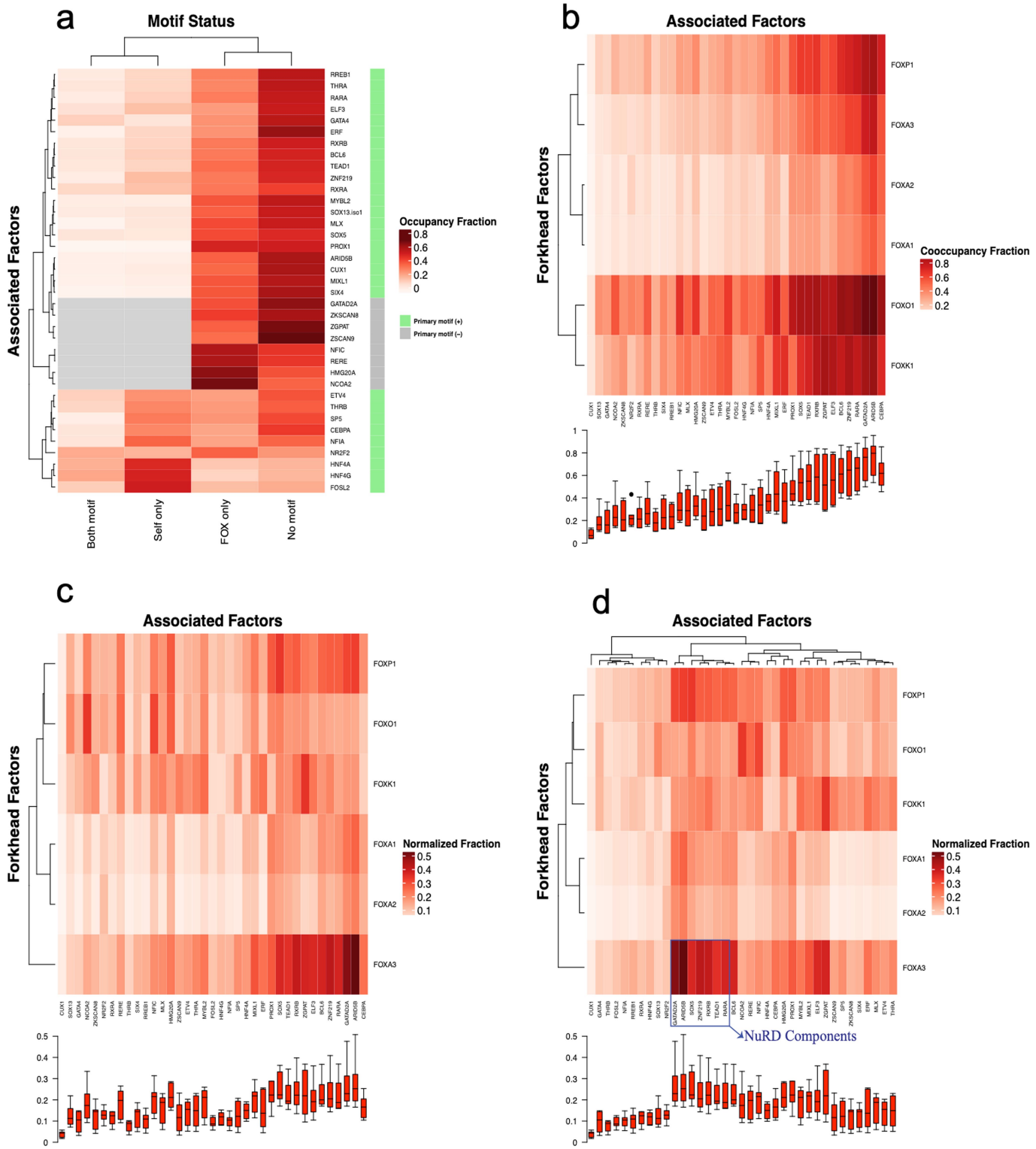
Extended Data Fig. 1 | CAP associations with annotated TSSs and IDEAS regions. a, The 208 ChIP–seq and CETCh–seq experiments plotted by number of peaks called in each experiment (*x*-axis) against fraction of peaks overlapping with any of 44,488 TSSs in the human genome (peaks ±3 kb from TSS). Selected individual CAPs are labelled. Solid line is linear regression through all points; dotted lines represent number of total TSS regions and maximum possible fraction of TSSs. **b**, IDEAS segmentation of HepG2 cell genome. Left, colour key for all IDEAS states; right, pie chart indicating fraction of HepG2 genome associated with each state. **c**, Clustering of 208 CAPs on the basis of chromatin state recapitulating the assigned cluster, with PC1 (63.50%), PC2 (16.51%) and PC3 (6.48%) variances explained. **d**, Left, distribution of regulatory regions by number of associated CAPs; right, distribution of horizontally matched sites by IDEAS state.



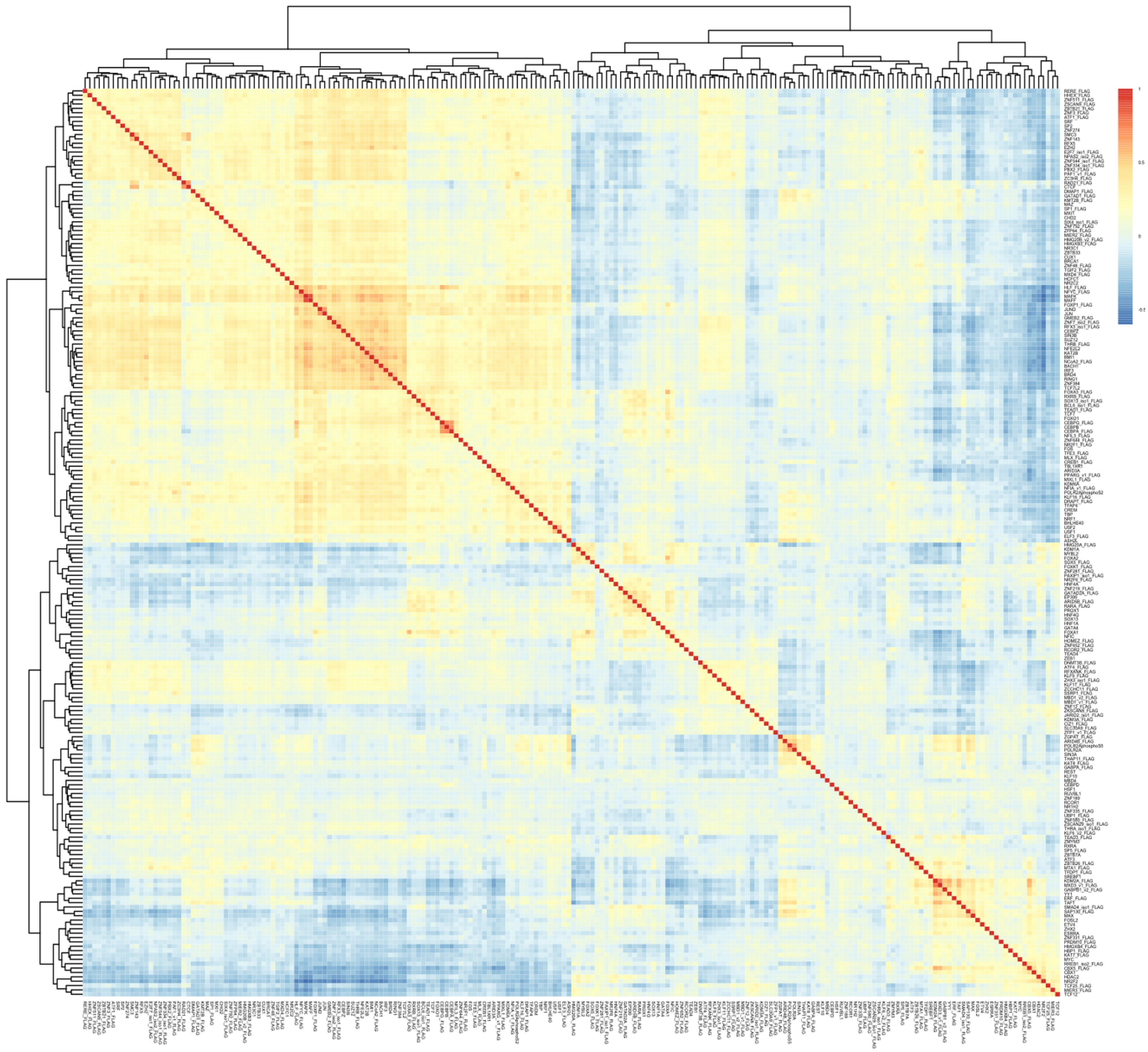
Extended Data Fig. 2 | CAP associations with varying CpG and GC content. **a**, Heat map and clustering of CAPs on the basis of association with low, intermediate, and high CpG content promoters (LCP, ICP, and HCP, respectively). All regions outside promoters are denoted as rest state. Annotation from Fig. 2a is shown, as are categories of direct DNA-binding factors (DBFs) and chromatin regulators or cofactors (CR/CF). **b**, Box plot of GC content of motifs for CAPs associating with promoters ($n = 26$), with both enhancers and promoters ($n = 45$), or with enhancers ($n = 55$). Centre line, median; boxes, 25th–75th percentiles; whiskers, 5th–95th percentiles.



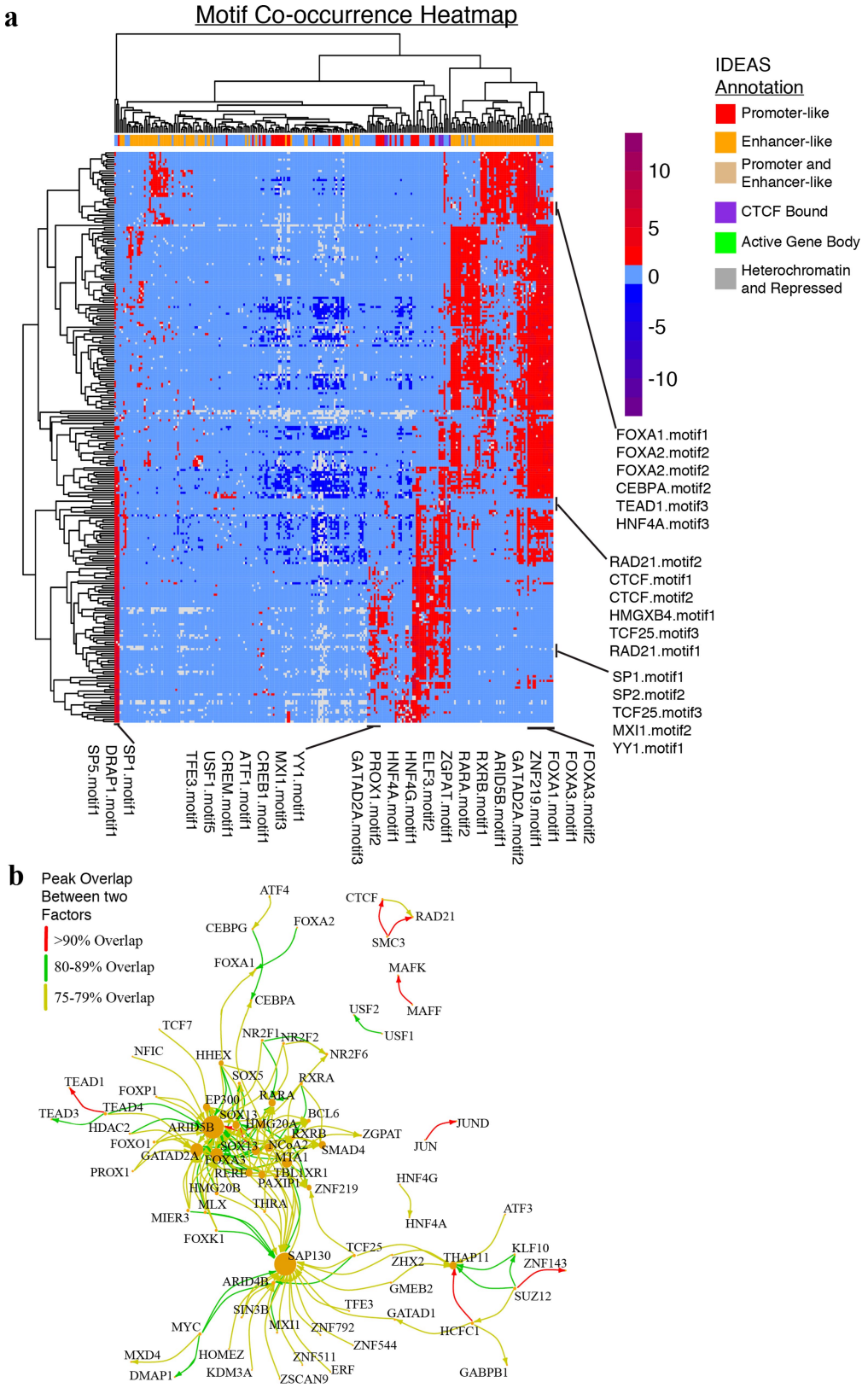
Extended Data Fig. 3 | Motif analysis. a, Cumulative fraction of called motifs in our data compared to motifs in the JASPAR 2016 vertebrate database as scored by Tomtom similarity *E*-value. **b**, Cumulative fraction of called motifs in our data compared to motifs in the JASPAR 2018 vertebrate database as scored by Tomtom similarity *E*-value. **c**, Cumulative fraction of called motifs in our data compared to motifs in the CIS-BP (build 1.02) *Homo sapiens* database as scored by Tomtom similarity *E*-value. **d**, Distribution of TF motifs by concordance (matching expected TF), discordance (matching different TF), and no match in the CIS-BP database. Stacked bar plots are coloured by main TF groups from previous unsupervised clustering. **e**, Distribution of TF motifs highly dissimilar to all motifs in CIS-BP (*y*-axis) and their median offset distance from the centre of peaks (*x*-axis). **f**, Stacked distribution of highly dissimilar motifs (no match; green) with similar (concordant; blue) and motif called for secondary factor (discordant; orange) and their median offset distances from the peak centre (*x*-axis).



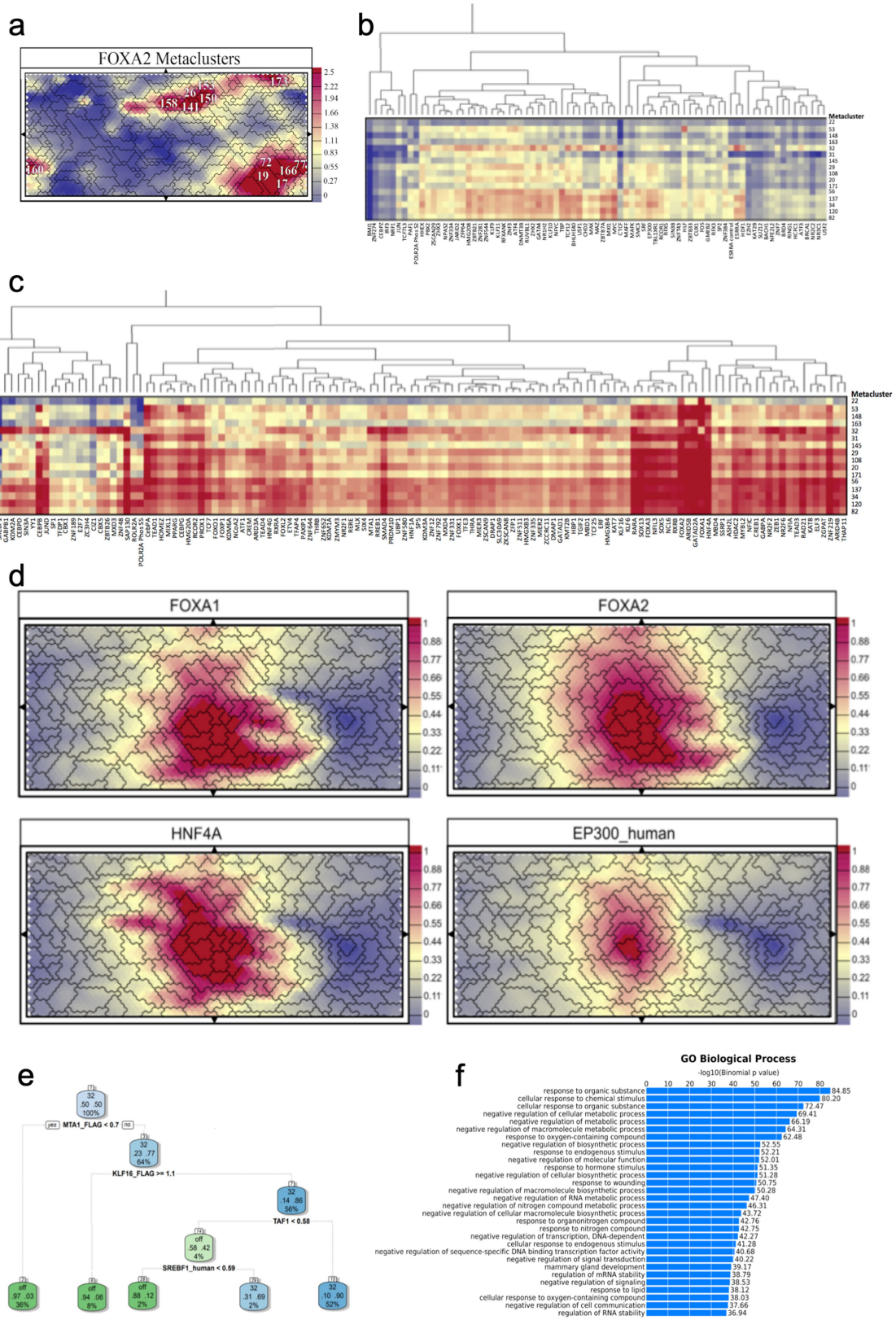
Extended Data Fig. 4 | CAPs associated with FOX TFs and motifs. a, Thirty-seven non-FOX TFs with a called Forkhead motif, with heat map denoting fraction of called peaks with both a primary (matched to specific TF) motif and a FOX motif, with a primary motif but not a FOX motif, with a FOX motif but no primary motif, and with neither a primary nor a FOX motif. The eight TFs with grey boxes do not have a known primary motif. **b**, Peak overlaps between the 37 TFs and 6 FOX factors for which we obtained ChIP–seq data; box plots represent distribution of all FOX overlaps for each of the 37 factors. **c**, Same as **b**, but normalized for peak counts of each of the 37 factors. **d**, Same as **c**, but clustered vertically, revealing NuRD component clustering. Box plots are vertically matched, $n = 6$ overlap measurements; boxes, middle quartiles; centre line, median; whiskers, 1.5 × IQR.



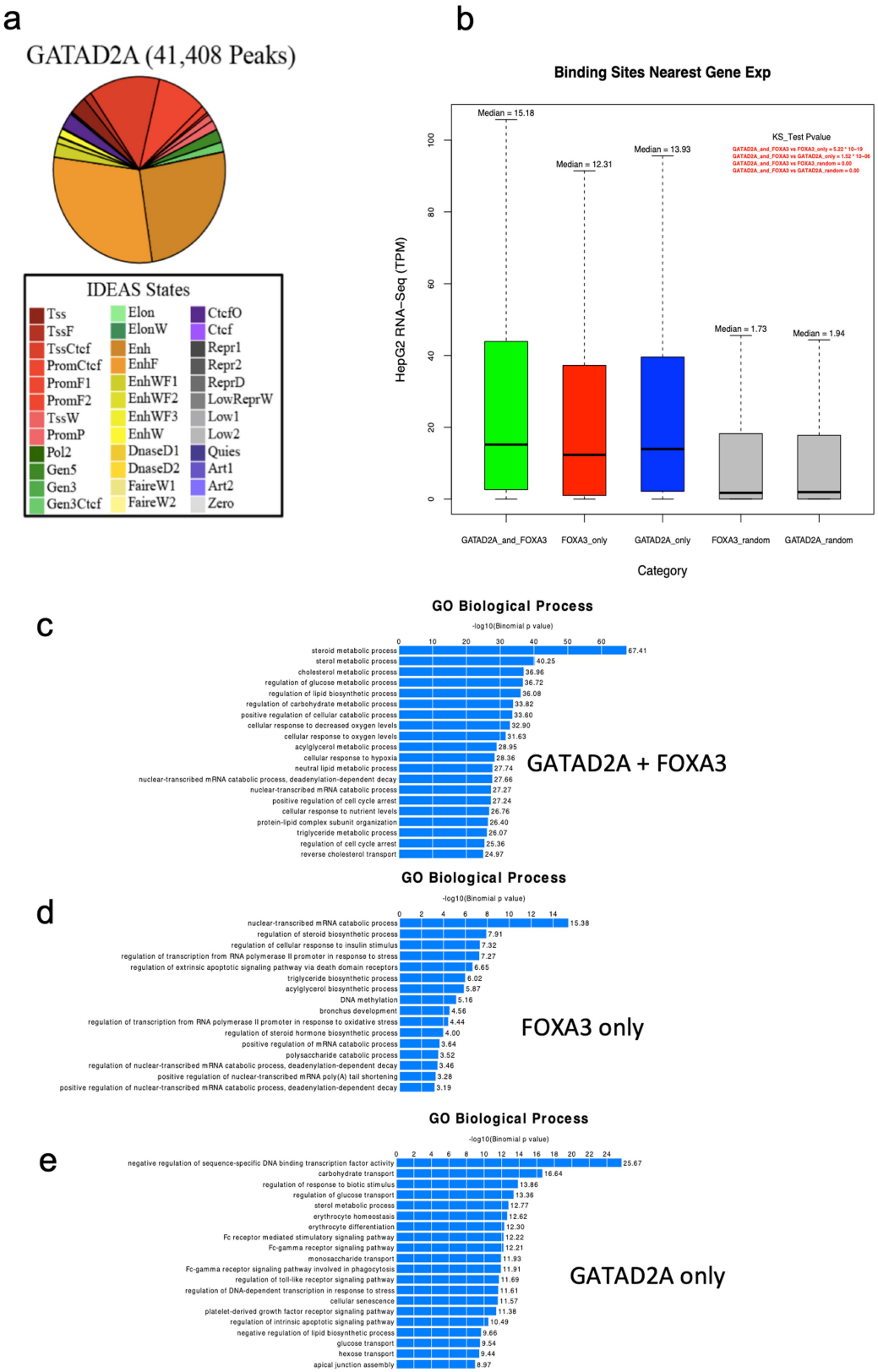
Extended Data Fig. 5 | Read count correlations between CAPs. Read count correlations between all 208 assayed CAPs, mean centred and squared, with unsupervised clustering.



Extended Data Fig. 6 | Motif and peak associations. a, Directional co-occurrence of motifs in ChIP–seq called peaks. **b**, Subset of network plot derived from peak overlaps between all factors, showing strong associations between a subset of factors.

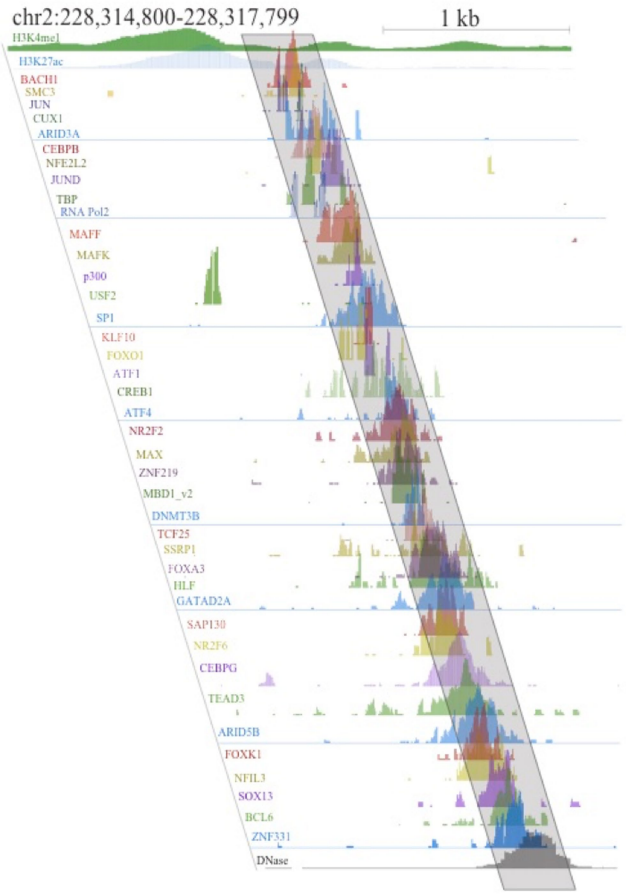


Extended Data Fig. 7 | Self-organizing maps. a, SOM showing FOXA2 metaclusters. **b**, Example heat map showing CAP enrichment in 16 key SOM metaclusters. **c**, Example heat map showing CAP enrichment in 16 key SOM metaclusters. **d**, SOMs for FOXA1, FOXA2, HNF4A, and EP300. **e**, Example decision tree showing the presence or absence of CAPs for metacluster 32. **f**, GREAT analysis of metacluster 32-assigned genes that are likely to be regulated in this metacluster, and GO term analysis for these genes; *P* represents sample frequency probability.



Extended Data Fig. 8 | GATAD2A analyses. a, GATAD2A genome-wide ChIP–seq binding in HepG2 cells annotated by IDEAS state. **b**, Box plots showing expression level (RNA-seq TPM) of genes nearest sites with both GATAD2A and FOXA3 ChIP–seq peaks (green), genes nearest sites with FOXA3 peaks but no GATAD2A peaks (red), genes nearest sites with GATAD2A peaks but no FOXA3 peaks (blue), and GC-matched null regions for each CAP (grey). Boxes, middle quartiles; centre line, median; whiskers, 1.5 × IQR; $n = 27{,}440$ binding sites (GATAD2A + FOXA3), $n = 10{,}658$ binding sites (FOXA3 only), $n = 13{,}706$ binding sites (GATAD2A only), $n = 37{,}073$ binding sites (FOXA3 null matched), $n = 40{,}441$ binding sites (GATAD2A null matched). **c**, GO enrichments for genes with both GATAD2A and FOXA3 peaks. **d**, GO enrichments for genes with FOXA3 peaks but no GATAD2A peaks. **e**, GO enrichments for genes with GATAD2A peaks but no FOXA3 peaks. GO *P* value represents sample frequency probability.

a

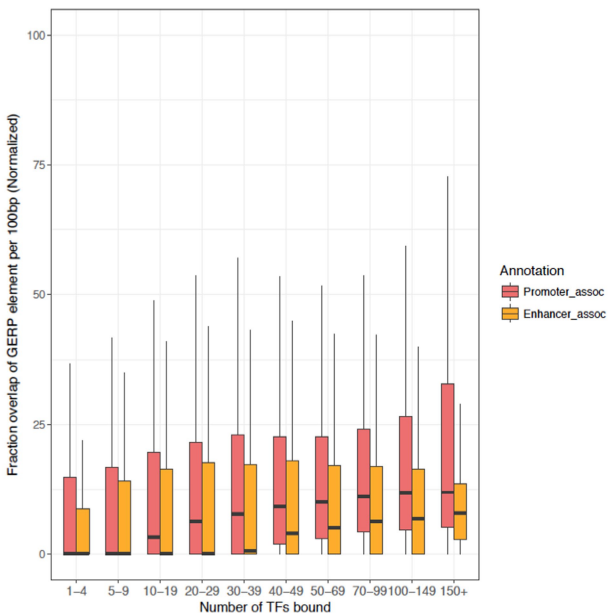


b

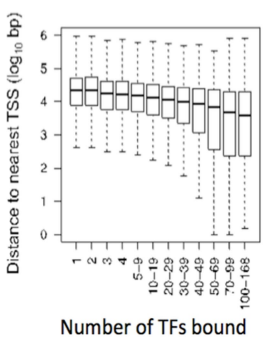
Highly Bound Regions at Promoters		
Top MSigDB Term	**Binomial P-Value**	**Fold Enrichment**
Metabolism of RNA	1.6e-147	50.5
Metabolism of Proteins	1.07e-146	34.0
Cell Cycle	2.29e-102	27.8
Cell Cycle, Mitotic	2.11e-90	30.0

Highly Bound Regions at Enhancers		
Top MSigDB Term	**Binomial P-Value**	**Fold Enrichment**
Metabolism of Lipids and Lipoproteins	2.42e-37	3.1
Fatty acid, triacylglycerol and ketone body metabolism	1.87e-24	4.0
FOXA2/FOXA3 TF networks	5.68e-17	7.6
HIF-1-alpha TF Network	8.86e-15	5.4

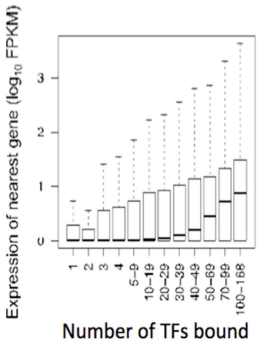
c



d

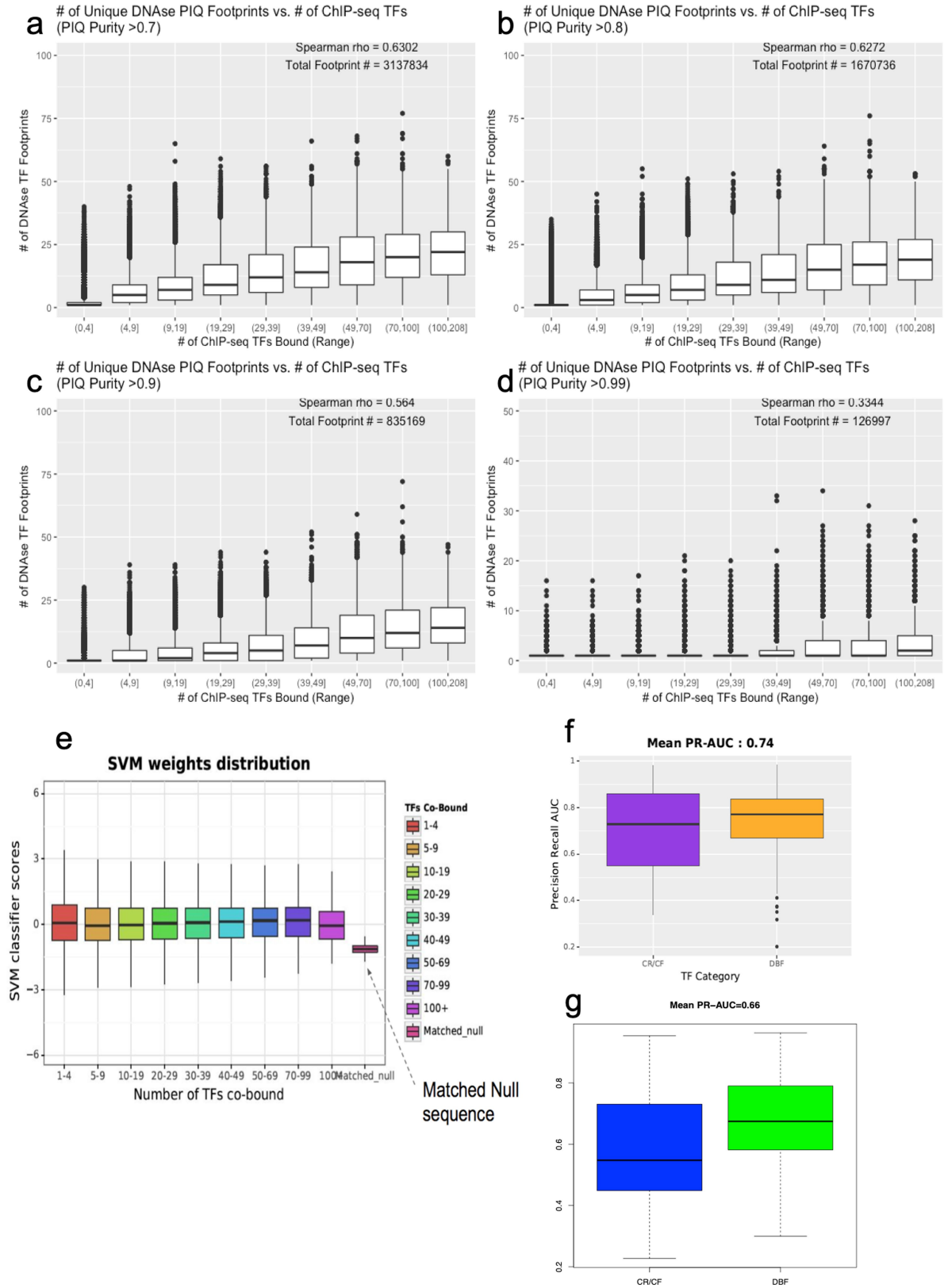


e



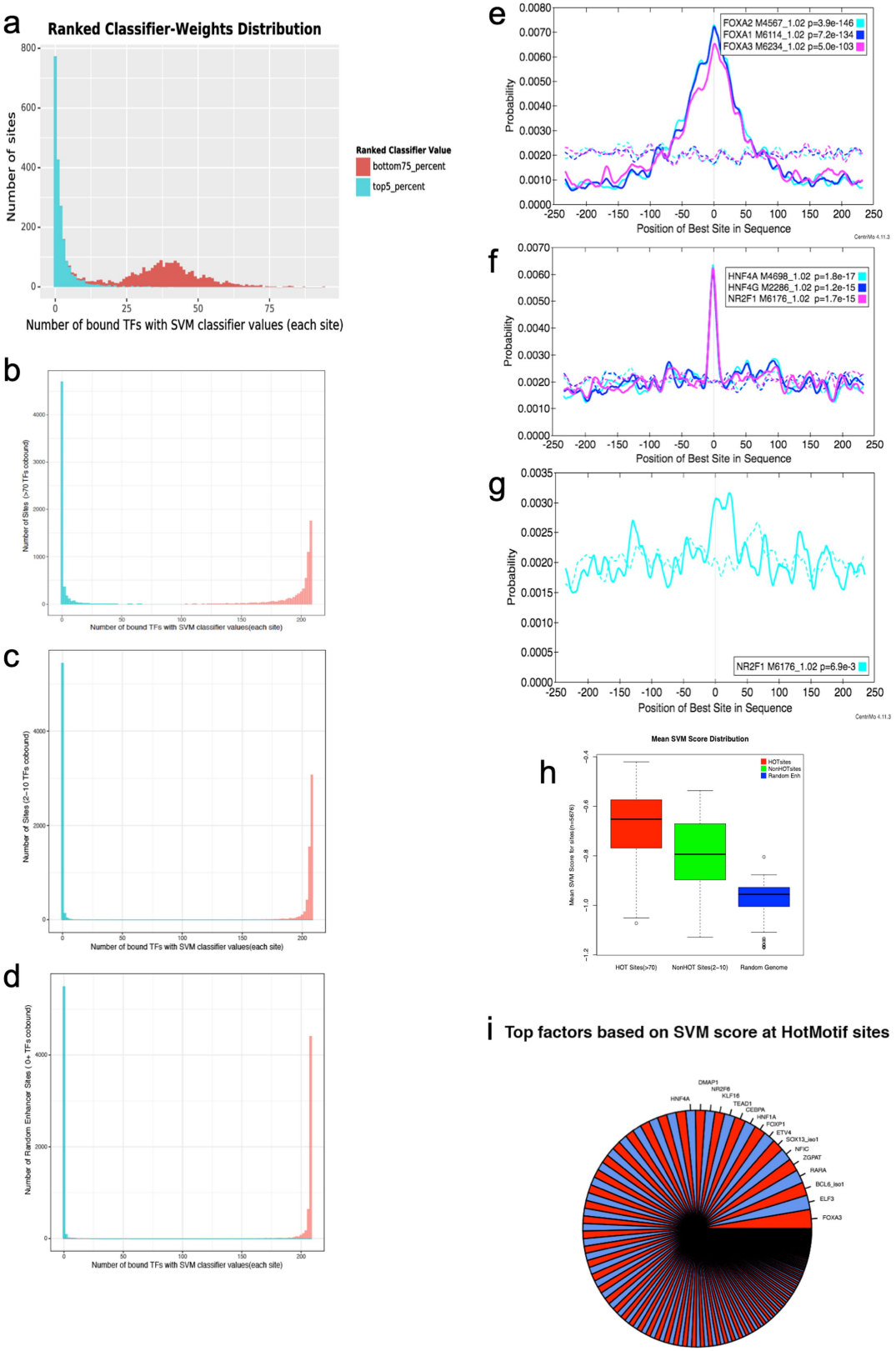
Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Extensive co-associations between CAPs. a, Example of genomic site with many associated CAPs. Each track shows aligned ChIP–seq reads, and is slightly offset to better show peaks for each experiment. **b**, Enrichment of biological pathways at HOT regions near enhancers or promoters; *P* represents sample frequency probability. **c**, Increasing numbers of CAPs bound at genomic sites correlate with increased evolutionary constraint as measured by GERP, showing incremental fraction overlap of highly constrained elements with CAP-associated sites for both promoter regions (red) and enhancer regions (orange). Boxes, quartiles; centre line, median; whiskers, 1.5 × IQR. **d**, Increasing numbers of CAPs bound at genomic sites (<2 kb in size) are associated with decreasing distance to nearest TSS; boxes, middle two quartiles; centre line, median; whiskers, 1.5 × IQR. **e**, Increasing numbers of CAPs bound at genomic sites (<2 kb in size) are associated with increasing expression of nearest gene; boxes, middle two quartiles; centre line, median; whiskers, 1.5 × IQR. **d**, **e**, Left to right: $n(1) = 124{,}074$, $n(2) = 59{,}407$, $n(3) = 19{,}661$, $n(4) = 12{,}433$, $n(5–9) = 23{,}517$, $n(10–19) = 14{,}757$, $n(20–29) = 7{,}077$, $n(30–39) = 4{,}703$, $n(40–49) = 3{,}542$, $n(50–69) = 5{,}061$, $n(70–99) = 4{,}655$, $n(>100) = 3{,}219$, total $n = 282{,}105$.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | PIQ and SVM analyses in CAP co-associated regions. **a**, Number of unique DNase PIQ footprints (*y*-axis) plotted by sites with varying numbers of associated CAPs (*x*-axis), for PIQ threshold >0.7. **b**, Number of unique DNase PIQ footprints (*y*-axis) plotted by sites with varying numbers of associated CAPs (*x*-axis), for PIQ threshold >0.8. **c**, Number of unique DNase PIQ footprints (*y*-axis) plotted by sites with varying numbers of associated CAPs (*x*-axis), for PIQ threshold >0.9. **d**, Number of unique DNase PIQ footprints (*y*-axis) plotted by sites with varying numbers of associated CAPs (*x*-axis), for PIQ threshold >0.99. **a**–**d**, Boxes, middle two quartiles; whiskers 1.5 × IQR; centre line, median; *n*(0–4) = 216,496, *n*(4–9) = 23,540, *n*(9–19) = 14,859, *n*(29–39) = 4,947, *n*(39–49) = 3,735, *n*(49–70) = 5,517, *n*(70–100) = 3,995, *n*(100–208) = 1,681. **e**, Distribution of SVM classifier scores (*y*-axis) for sites with varying numbers of associated CAPs (*x*-axis). The scores remain relatively constant across sites and are significantly higher than the scores of classifier values in matched null sites. Boxes, middle two quartiles; whiskers 1.5 × IQR; centre line, median; *n*(1–4) = 1,814,475 bins, *n*(5–9) = 643,997 bins, *n*(10–19) = 646,453 bins, *n*(20–29) = 330,795 bins, *n*(30–39) = 194,981 bins, *n*(40–49) = 118,622 bins, *n*(50–69) = 131,167 bins, *n*(70–99) = 57,819 bins, *n*(100+) = 3,545 bins, *n*(matched null) = 9,597,800 bins. **f**, SVM PR-AUC scores for non-TFs (chromatin regulators and cofactors; CR/CF) and for TFs at motif-level mean PR-AUC (0.74). **g**, SVM PR-AUC scores for non-TFs (chromatin regulators and cofactors) and for TFs at motif-level mean PR-AUC (0.66). **f**, **g**, Boxes, middle two quartiles; whiskers 1.5 × IQR; centre line, median; *n*(CR/CF) = 37, *n*(DBF) = 171.



Extended Data Fig. 11 | See next page for caption.

Extended Data Fig. 11 | SVM and motif analyses in HOT sites. a, Number of sites (*y*-axis) by measured number of TFs (*x*-axis) with classifier values in the top 5% of all classifier values (blue) or with classifier values in the bottom 75% of all classifier values (red) in highly bound regions, based on SVM scores of factor peaks associated with highly bound regions. **b**, Number of sites (*y*-axis) by measured number of TFs (*x*-axis) with classifier values in the top 5% of all classifier values (blue) or with classifier values in the bottom 75% of all classifier values (red), in HOT sites with >70 associated TFs. **c**, Number of sites (*y*-axis) by measured number of TFs (*x*-axis) with classifier values in the top 5% of all classifier values (blue) or with classifier values in the bottom 75% of all classifier values (red), in sites with 2–10 associated TFs. **d**, Number of sites (*y*-axis) by measured number of TFs (*x*-axis) with classifier values in the top 5% of all classifier values (blue) or with classifier values in the bottom 75% of all classifier values (red), in a random set of enhancers with any number of associated TFs (0+). **e**, Degree of motif enrichment in highly bound regions for all HepG2-expressed TFs with available motifs ($n = 365$) for top three motifs enriched in highly bound sites with 50+ CAPs (highest $P = 3.9 \times 10^{-146}$). **f**, Degree of motif enrichment in highly bound regions for all HepG2-expressed TFs with available motifs ($n = 365$) for top three motifs in enhancers with 2–10 CAPs (highest $P = 1.8 \times 10^{-17}$). **g**, Degree of motif enrichment in highly bound regions for all HepG2-expressed TFs with available motifs ($n = 365$) for top motif in random genome enhancers with 0+ CAPs (highest $P = 6.9 \times 10^{-3}$). **h**, Distribution of all SVM scores (*y*-axis) for HOT sites with >70 associated CAPs (red), for sites with 2–10 associated CAPs (green), and for random enhancer sites with 0+ CAPs (blue). **i**, Pie chart showing fraction of HOT sites in which each TF has the highest SVM classifier value, indicating the strongest motif present.

nature research

Corresponding author(s): Eric M. Mendenhall and Richard M. Myers

Last updated by author(s): Nov 20, 2019

Reporting Summary

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

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	Python v3.5 R v3.5 Shell Env (unix, bash) boost C++ libraries bwa v0.7.12 bowtie2 v2.1.0 spp v1.10.1 idr v2.0.2 meme v4.11.4 centrimo v4.11.4 tomtom v4.11.4 samtools v1.3 bedtools2 v2.20.0 fimo v4.11.4 phantompeakqualtools v2.0 picard-tools v1.88 trim_galore v0.3.7 cutadapt v1.16 fastqc v0.10.1 deeptools v3.0.0 kmersvm; gkmSVM v2.0 pybedtools v0.7.10 pandas v0.20.3

numpy v1.14.0
scipy v0.19.1
scikit-learn v0.19.0
ggplot2 v3.1.0; gplots v3.0.1
dplyr v0.7.5; gtools v3.0.1
ranger v0.10.1
ComplexHeatmap v1.18.0
circlize v0.4.3
igraph v1.2.1
GraphPad Prism 8 for macOS v.8.3.0
R v.3.3.2
All code available at https://github.com/chhetribsurya/PartridgeChhetri_etal

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

Data

Policy information about availability of data

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

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

All data are available at the ENCODE portal (encodeproject.org) or at Gene Expression Omnibus under accession number GSE104247.

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	No statistical methods were used to predetermine sample size.
Data exclusions	No data were excluded.
Replication	Replicate structure of ChIP/CETCh-seq experiments are described in the manuscript. Traditional antibody ChIP-seq experiments were replicated in separate grow-ups of cells, IPed separately; these are biological replicates and technical replicates for growth, crosslinking, IP, and sequencing library construction. CETCh-seq experiments were replicated at the point of nucleofection of CRISPR components, where cells are split in two equal amounts directly after nucleofection; since these represent separate pools of pre-edited cells, these are biological replicates and technical replicates for growth, crosslinking, IP, and sequencing library construction. Final data is composed of IDR-passed reproducible reads from both experimental replicates.
Randomization	The experiments were not randomized.
Blinding	The investigators were not blinded to allocation during experiments and outcome assessment.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

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

Methods

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

Antibodies

Antibodies used

BACH1,sc-14700,E1503,Santa Cruz Biotech,5 ug per 2e7 cells;
CUX1,sc-6327,E0709,Santa Cruz Biotech,5 ug per 2e7 cells;
SIN3B,sc-13145,B2802,Santa Cruz Biotech,5 ug per 2e7 cells;
KAT2B,3378S,1,Cell Signaling,5 ug per 2e7 cells;
POLR2AphosphoS2,ab5095,GR32890-1,Abcam,5 ug per 2e7 cells;
RFX5,200-401-194,14562,Rockland,5 ug per 2e7 cells;
SMC3,ab9263,963667,Abcam,5 ug per 2e7 cells;
FOS,sc-7202,K0810,Santa Cruz Biotech,5 ug per 2e7 cells;
HCFC1,NB100-68209,A1,Novus,5 ug per 2e7 cells;
MXI1,AF4185,ZIJ0107031,RD Systems,5 ug per 2e7 cells;
TBL1XR1,ab24550,GR340121,Abcam,5 ug per 2e7 cells;
ZNF384,HPA004051,A57874,Sigma,5 ug per 2e7 cells;
ZNF143,16618-1-AP,8059,Proteintech,5 ug per 2e7 cells;
JUN,sc-1694,C2206,Santa Cruz Biotech,5 ug per 2e7 cells;
RCOR1,sc-30189,C0806,Santa Cruz Biotech,5 ug per 2e7 cells;
CHD2,ab68301,762356,Abcam,5 ug per 2e7 cells;
SUZ12,3737BF,4,Cell Signaling,5 ug per 2e7 cells;
IRF3,sc-9082,I0908,Santa Cruz Biotech,5 ug per 2e7 cells;
BHLHE40,NB100-1800,A1,Novus,5 ug per 2e7 cells;
ARID3A,NB100-279,A1,Novus,5 ug per 2e7 cells;
BRCA1,A300-000A,2,Bethyl Labs,5 ug per 2e7 cells;
NFE2L2,sc-13032,A1711,Santa Cruz Biotech,5 ug per 2e7 cells;
JUND,sc-74,unknown,Santa Cruz Biotech,5 ug per 2e7 cells;
CEBPZ,SAB2100398,QC8343,Sigma,5 ug per 2e7 cells;
TBP,ab62126,unknown,Sigma,5 ug per 2e7 cells;
POLR2A,MMS-126R,14861301,Covance,5 ug per 2e7 cells;
MAFF,M8194,125K4837,Sigma,5 ug per 2e7 cells;
HSF1,sc-9144,unknown,Santa Cruz Biotech,5 ug per 2e7 cells;
SREBF1,sc-8984,10211,Santa Cruz Biotech,5 ug per 2e7 cells;
MAFK,ab50322,904274,Abcam,5 ug per 2e7 cells;
MAZ,ab85725,GR41711-2,Abcam,5 ug per 2e7 cells;
NR3C1,sc-1002,I0310,Santa Cruz Biotech,5 ug per 2e7 cells;
ESRRA,sc-66882,A0809,Santa Cruz Biotech,5 ug per 2e7 cells;
CEBPB,sc-150,I1010,Santa Cruz Biotech,5 ug per 2e7 cells;
BMI1,unknown,unknown,unknown,unknown;
RING1,unknown,unknown,unknown,unknown;
KDM6A,unknown,unknown,unknown,unknown;
BRD4,A301-985A50,3,Bethyl Labs,5 ug per 2e7 cells;
EZH2,39875,27210001,Active Motif,5 ug per 2e7 cells;
ASH2L,A300-489A,2,Bethyl Labs,5 ug per 2e7 cells;
KDM1A,A300-215A,1,Bethyl Labs,5 ug per 2e7 cells;
NR2C2,TR4,unknown,James Engel,5 ug per 2e7 cells;
ZNF274,H00010782-A01,060729QCS1,Abnova,5 ug per 2e7 cells;
TCF7L2,2569,2,Cell Signaling,5 ug per 2e7 cells;
MYC,sc-764,H0107,Santa Cruz Biotech,5 ug per 2e7 cells;
TAF1,sc-735,K0905,Santa Cruz Biotech,5 ug per 2e7 cells;
USF1,sc-229,A2109,Santa Cruz Biotech,5 ug per 2e7 cells;
SIN3A,sc-994,F1005,Santa Cruz Biotech,5 ug per 2e7 cells;
FOSL2,sc-604,unknown,Santa Cruz Biotech,5 ug per 2e7 cells;
RXRA,sc-553 ,C1811,Santa Cruz Biotech,5 ug per 2e7 cells;
TCF12,sc-357,F2305,Santa Cruz Biotech,5 ug per 2e7 cells;
POLR2AphosphoS5,ab5408,648628,Abcam,5 ug per 2e7 cells;
HNF4A,sc-8987,G1309,Santa Cruz Biotech,5 ug per 2e7 cells;
FOXA1,sc-6553,H1209,Santa Cruz Biotech,5 ug per 2e7 cells;
YY1,sc-281,B1010,Santa Cruz Biotech,5 ug per 2e7 cells;
ATF3,sc-188,E1410,Santa Cruz Biotech,5 ug per 2e7 cells;
SRF,sc-335,F3006,Santa Cruz Biotech,5 ug per 2e7 cells;
CTCF,sc-5916,F2906,Santa Cruz Biotech,5 ug per 2e7 cells;

HDAC2,sc-6296,G0307,Santa Cruz Biotech,5 ug per 2e7 cells;
HNF4G,sc-6558,I299,Santa Cruz Biotech,5 ug per 2e7 cells;
EP300,sc-585,E2010,Santa Cruz Biotech,5 ug per 2e7 cells;
ZBTB33,sc-23871,I0308,Santa Cruz Biotech,5 ug per 2e7 cells;
CEBPD,sc-636,C1010,Santa Cruz Biotech,5 ug per 2e7 cells;
REST,Custom,20081022MA,Anderson lab,5 ug per 2e7 cells;
SP2,sc-643,K1803,Santa Cruz Biotech,5 ug per 2e7 cells;
ZBTB7A,sc-34508,H2406,Santa Cruz Biotech,5 ug per 2e7 cells;
NFIC,sc-81335,L0808,Santa Cruz Biotech,5 ug per 2e7 cells;
MYBL2,sc-724,D2109,Santa Cruz Biotech,5 ug per 2e7 cells;
MBD4,sc-271530,H1210,Santa Cruz Biotech,5 ug per 2e7 cells;
TEAD4,sc-101184,A1811,Santa Cruz Biotech,5 ug per 2e7 cells;
NR2F2,sc-271940,I1410,Santa Cruz Biotech,5 ug per 2e7 cells;
MAX,sc-197,J0809,Santa Cruz Biotech,5 ug per 2e7 cells;
ZEB1,sc-25388,D2010,Santa Cruz Biotech,5 ug per 2e7 cells;
FOXA2,AM39828,1720001,Active Motif,5 ug per 2e7 cells;
NR1H2,61178,29111001,Active Motif,5 ug per 2e7 cells;
TFAP4,WH0007023M3,07040-7A10,Sigma,5 ug per 2e7 cells;
ZMYM3,JH39.2.2F10,20130506-RAP,CDI,5 ug per 2e7 cells;
ZHX2,GTX112232,40107,GeneTex,5 ug per 2e7 cells;
ZNF189,GTX117129,40730,GeneTex,5 ug per 2e7 cells;
RUVBL1,JH39.2.1A1,20130711.YRH,CDI,5 ug per 2e7 cells;
PROX1,61092,14511001,Active Motif,5 ug per 2e7 cells;
SOX13,WH0009580M1,10061-3E8,Sigma,5 ug per 2e7 cells;
TCF7,WH0006932M1,11181-1D2,Sigma,5 ug per 2e7 cells;
ETV4,GTX114393,40184,Genetex,5 ug per 2e7 cells;
HNF1A,GTX113850,40135,Genetex,5 ug per 2e7 cells;
GATA4,39894,26310001,Active Motif,5 ug per 2e7 cells;
CBX1,39980,11213003,Active Motif,5 ug per 2e7 cells;
CREM,WH0001390M2,11056-3B5,Sigma,5 ug per 2e7 cells;
NRF1,R157.1.3H3,20140422-DNF,CDI,5 ug per 2e7 cells;
GABPA_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
RAD21_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
USF2,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
KLF10,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
FOXO1,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ATF1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
CREB1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ATF4_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ZNF3_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
HHEX_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
PBX2_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ZNF219_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
MBD1_v1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
MBD1_v2_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
DNMT3B_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
TCF25_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
SSRP1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
TGIF2_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
HLF_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
HBP1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
KDM3A_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
FOXP1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
SLC30A9_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ZNF644_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
HOMEZ_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
RERE_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
SAP130_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
KLF11_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
KMT2B_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
NR2F6_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ARID4B_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
GATAD1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ZNF792_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ZNF652_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
GATAD2A_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
NCoA2_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
TEAD1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
NFYC_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
CEBPG_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
KLF9_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
DRAP1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
MLX_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ZNF511_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
MIXL1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;

ZSCAN9_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
NR2F1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
TFE3_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
KAT8_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
RXRB_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
SOX5_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
KLF16_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
KLF6_v2_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
THAP11_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
FOXA3_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ELF3_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ZBTB26_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
TEAD3_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
GABPB1_v2_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ERF_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
KAT7_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
MXD3_v1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ZNF580_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
CIZ1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
MIER3_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
HMGXB4_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ZGPAT_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
RARA_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ARID5B_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
MXD4_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
CEBPA_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ZFP1_v1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
NFIL3_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
SP5_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
TFDP1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
RFXANK_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
DMAP1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
THRB_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
PPARG_v1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
HMG20B_v2_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
PAF1_v1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
MIER2_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
NFIA_v1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
RCOR2_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
GMEB2_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ZKSCAN8_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
HMG20A_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ZNF48_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
UBP1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
MTA1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ZFP64_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
FOXK1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
RFX3_iso1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ZNF7_iso2_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
SOX13_iso1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
SMAD4_iso1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
BCL6_iso1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ZNF331_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
THRA_iso1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
SIX4_iso1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ZHX3_iso1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ZNF544_iso1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ZNF334_iso1_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ZNF281_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
NPAS2_iso2_FLAG,F1804,SLBK1346V,Sigma,5 ug per 2e7 cells;
ZSCAN29_iso1_FLAG,F1804,SLBN5629V,Sigma,5 ug per 2e7 cells;
E2F7_iso1_FLAG,F1804,SLBN5629V,Sigma,5 ug per 2e7 cells;
PRDM10_FLAG,F1804,SLBN5629V,Sigma,5 ug per 2e7 cells;
KDM2A_FLAG,F1804,SLBN5629V,Sigma,5 ug per 2e7 cells;
PAXIP1_iso1_FLAG,F1804,SLBN5629V,Sigma,5 ug per 2e7 cells;
JARID2_iso1_FLAG,F1804,SLBN5629V,Sigma,5 ug per 2e7 cells;
RREB1_iso2_FLAG,F1804,SLBN5629V,Sigma,5 ug per 2e7 cells;
ZBTB21_FLAG,F1804,SLBN5629V,Sigma,5 ug per 2e7 cells;
ZC3H4_FLAG,F1804,SLBN5629V,Sigma,5 ug per 2e7 cells;
SP1_FLAG,F1804,SLBN5629V,Sigma,5 ug per 2e7 cells;
CBX5_FLAG,F1804,SLBN5629V,Sigma,5 ug per 2e7 cells;
ZNF12_FLAG,F1804,SLBN5629V,Sigma,5 ug per 2e7 cells;
ZNF335_FLAG,F1804,SLBN5629V,Sigma,5 ug per 2e7 cells;
HMGXB3_FLAG,F1804,SLBN5629V,Sigma,5 ug per 2e7 cells;

ZCCHC11_FLAG,F1804,SLBN5629V,Sigma,5 ug per 2e7 cells;

Validation

Primary characterization by Western blot or Immunoprecipitation/blot
For every TF ChIP-seq antibody, ENCODE data producers first perform an immunoblot characterization. This can be either a standard Western blot, or an immunoprecipitation followed by a Western blot ("IP Western"). If the blot results do not meet the parameters and thresholds given below, then Primary Characterization Method 2 (IP mass spec) is performed. In the latter case, the failed (or partially failed) immunoblot that preceded IP/mass spec is included in the report so that researchers and data users can independently evaluate the data for each antibody.
Immunoblot parameters:
a. ENCODE developed a set of working parameters and thresholds to identify antibodies with a high likelihood of being specific for the target factor. The parameters allow for modest variation in gel migration characteristics and in band number to accommodate known behaviors of typical nuclear factors. Current acceptable parameters are that the major band is within 20% of the size predicted by the size of the coding region and corresponds to >50% of all bands on the gel (excluding the antibody bands in the case of an immunoprecipitation). If the western or IP-western results meet these criteria, we consider the antibody to meet expectations for the primary characterization. The immunoblot results (which must include appropriate size markers) are submitted as evidence for each cell type or tissue tested. For IP-westerns, a control IgG precipitation is also performed and analyzed on the same gel.
b. If the antibody fails to pass the immunoblot tests because the bands observed are too numerous, or too far from the predicted migration behavior, it can be “rescued” by a secondary characterization that supports the conclusion that the band(s) detected correspond to the correct protein (e.g. all bands are reduced by treatment with a specific siRNA to that protein; see secondary characterizations).
c. If the antibody passes the immunoblot tests, a further characterization is required to support the successful immunoblot. This can be Primary Characterization Method 2 (IP mass spec) or any one of the Secondary Characterization methods in IB.

Primary Characterization Method 2: Immunoprecipitation followed by mass spectrometry ("IP mass spec")
If the immunoblot characterization data was not successful (ranging from no bands to patterns that do not meet the thresholds given above), then Mass spec of an immunoprecipitation can be performed. The failed or ambiguous immunoblot is, however, shown as part of the antibody characterization dataset. Because the IP/mass spec assay provides explicit evidence about the identity of the TF detected, it can also be used in lieu of Secondary methods after a successful Immunoblot (see flowchart above).
For TF mass spec, a cell or nuclear extract is immunoprecipitated with the same antibody used to perform ChIP-seq. That IP is then fractionated on a denaturing polyacrylamide gel, and the fractions are prepared and analyzed by mass spec as described below.
What is reported for IP mass spec:
• IP-western blot of gel image with outline of gel slices submitted for mass spec.
• All peptides (with peptide counts) from all immunoreactive bands.
• Fold enrichment of all peptides in the immunoreactive bands vs either mock IP or a set of proteins that have been immunoprecipitated from the same cell type using a collection of other antibodies from the same host species (the list of proteins used as the set of IP contaminants list must be provided).
• Indication as to which proteins above the target protein on the ranked list (ranked by fold enrichment) are TFs and which TFs are members of the same TF family as the target protein.
IP mass spec requirements to be considered fully validated for ENCODE data:
• The target protein should be enriched in the IP when compared to a mock IP or to a set of proteins that have been immunoprecipitated from the same cell type using a collection of other antibodies from the same host species.
• The target should be in the top 25 ranked proteins and the top most-enriched TF (by fold enrichment) in the immunoreactive band, unless the higher ranked TFs are known interacting partners of the target TF and/or a known interacting partner of one of the other higher ranked TFs that is a known partner of the target TF. Evidence for interaction can come from publications or refer to records in interaction databases such as BioGRID, or other sources..
• The target should be the top ranked member of that family of TFs (exceptions will be allowed if a publication is provided that demonstrates that a higher ranked family member is known to dimerize with the target protein).
• In situations for which the target protein has 0 peptides in the mock IP, a ranking by enrichment can not be performed. In that case, the following criteria are used:
• If the target TF is the top TF as ranked by number of detected peptides, then the antibody passes this characterization method.
• If the target TF is not the top TF but the TFs having more counts have previously been documented to be in the same complex and/or interact directly with the target TF, then the antibody passes this characterization method.
• If the target TF is not the top TF but the non-target TFs (having a greater number of detected peptides) were detected using mass spec analysis of two different 4 antibodies to the target TF, then the antibody passes this characterization method (with the assumption that the other TFs are bona fide interacting TFs that have not yet been documented in the literature).
• If the target TF is not the top TF and the TFs having more counts have never been linked to the target TF then this antibody is flagged, with the explanation that enrichment could not be determined due to the lack of detected peptides in the IgG and that

no published data exists linking the target to the non-target TFs.

• If an antibody doesn't meet these characteristics, the antibody characterization document can be submitted for consideration as a special request (see Note 3).

Additional situations for Primary Characterization
a. Guidelines for using the same lot number of a previously characterized antibody in a new cell type
If a specific lot number for an antibody has previously passed characterization in another cell type, and if the banding pattern on the immunoblot or immunoprecipitation is the same in the new cell type as in the characterized cell type, then no further characterization is needed for the antibody in that new cell type. If the banding pattern is different in the new cell type, a secondary characterization is performed in the new cell type. Exceptions to this guideline will be considered for studies of human tissues (due to the fact that it is often hard to obtain sufficient tissue for the antibody characterization and a ChIP-seq experiment). If an antibody has passed characterization criteria in 2 different human cell lines and/or tissues, it does not have to be characterized in each tissue type.
b. Guidelines for using a different lot number of a previously characterized antibody
For the first time that a new lot number is used for a previously-characterized antibody, a Primary Characterization method (immunoblot or IP mass spec) is performed with one of the same cell types used to characterize the previous lot number plus the cell type for which ChIPseq data will be deposited for the new lot number; the ENCODE antibody accession number of the specific previously characterized lot that should be used for comparison is indicated. If the patterns for the new lot number are the same in the previously characterized cell type and in the cell type for which ChIP-seq data will be deposited as shown in the characterization of the original lot number of that antibody, then no further characterization is required. If the banding patterns are different, a secondary characterization is performed. Exceptions to this guideline will be considered for analysis of tissues with antibodies that have been well-characterized and used extensively by the field (e.g a monoclonal antibody to RNAPII). In this case, if a previous lot number of an antibody has passed characterization criteria in 2 different human cell lines and/or tissues, the new lot number does not have to be characterized in each tissue type.
Other primary characterization methods. If other methods not specified above are used for primary characterization of an antibody, the antibody characterization document is submitted as a special request and is so annotated and flagged.

IB. Secondary Characterization Methods. These methods are used to support and clarify the Immunoblot data. In particular, they aim to verify that a band or bands observed on the prior blot correspond to the intended TF. At least one successful Secondary Characterization (or alternatively IP/Mass spec as shown in figure 1 above) is required to support a successful Western or IP/Western.
Secondary Characterization Method 1: siRNA or shRNA against the mRNA of the target protein
For siRNA or shRNA knockdown characterization, the band(s) detected by the antibody on a western blot should be reduced by at least 50% of the control signal. These methods are especially intended to address instances where the Western or IP Western data give multiple bands and unpredicted migration patterns. The sequence or vendor and catalog number of the oligonucleotide(s) reagent should be provided. A control knockdown should also be performed. Cell types will be labeled and size markers should be included on the immunoblot. A brief description of the transfection protocol will also be provided.
Secondary Characterization Method 2: ChIP-seq data from a previously characterized antibody
If ChIP-seq data for a different lot number of a previously characterized antibody or a previously characterized, but different, antibody for a given transcriptional regulator is available, this ChIPseq data can be used to evaluate a new antibody or new lot number. The ChIP-seq data from the new antibody or new lot number are compared to the previous ChIP-seq using IDR. If the two datasets pass the ENCODE IDR cuts-offs for narrow peak ChIP-seq reproducibility (see below for current IDR standards), then the secondary characterization of the new antibody/lot number is scored as successful. For data submission, the specific antibody lot and ChIP-seq data used for the comparison are identified by their ENCODE antibody and experiment accessions, respectively. In a similar way, ChIP-seq data obtained using an endogenous epitope-tagged version of the target protein can be used for comparison.
Secondary Characterization Method 3: Expression patterns of an epitope-tagged transcription factor
Especially useful for TFs that are resistant to knockdown using shRNA or siRNAs (e.g. very stable proteins) is a secondary characterization method that involves comparison to overexpressed or endogenously epitope-tagged TF proteins. In this case, the primary characterization of the TF antibody must first show the appropriate specificity on the western or IP-western. Then, two side-by-side immunoblots can be performed using control cells and cells expressing the tagged-factor. The first immunoblot employs the antibody to the tag to show the position of the exogenous factor band(s) and the second immunoblot employs the antibody to the endogenous factor to show that the band(s) in the control and ectopically expressing cases correspond.

Secondary Characterization Method 4: Motif analysis
Motif enrichment for antibody characterization requires pre-existing information about the DNA

sequence to which the factor binds. Enrichment of a known motif for a target TF in a ChIP experiment is evidence that the antibody does in fact recognize the target TF.
Motif enrichment can be used as a validation method for antibodies that meet the following criteria:
i. The antibody under consideration binds a sequence-specific transcription factor
ii. The DNA motif sequence bound by the transcription factor has been previously wellcharacterized by either in vitro or in vivo experiments
iii. The antibody is raised to a unique region of the transcription factor (in relation to other TFs in the same family)
Motif analysis can be performed using high-quality peaks (0.01 IDR cut off) from the ChIP experiment. Proper use of motif enrichment analysis for antibody validation should include metrics indicative of:
i. Global Enrichment z-score: Enrichment of the motif sequence in the ChIP peak over shuffled randomized motifs of the same sequence composition
ii. Positional Bias z-score: A measure of the distance of the motif to the peak center
iii. Peak Rank Bias z-score: A measure of the distribution of the motif in peaks ranked by ChIP intensity
The mean of these three z-scores is used in computing the final enrichment rank among 282 motif groups, as well as the “accept probability”. The “accept probability” is a combined metric that measures confidence in the antibody under investigation being of high quality for ChIP experiments. An accept probability greater that 0.6 is the current criteria for accepting an antibody as passing secondary characterization by motif enrichment (see note 4)..
The Characterization report where Motif enrichment is used for antibody secondary validations includes:
i. The ENCODE DCC file identifiers for the peaks files (.bed files) used in the analysis
ii. A brief description of the analysis method and a reference to the standards documents
iii. The accept probability score from the motif analysis pipeline
iv. The identified motif (PWM) and its enrichment rank
v. The positional bias score as well as the peak rank score
Motif analysis cannot be used when:
i. The transcription factor does not bind in a sequence-specific manner
ii. There is no information for the DNA motif bound by the TF
iii. When it has been shown that the TF bind to DNA indirectly by interacting with other proteins that directly bind DNA

Because transcription factors are recruited by multiple mechanisms, failure of a data set to meet the motif enrichment criteria does not indicate poor antibody quality or poor data quality. Such antibodies can be validated using other Secondary Characterization methods.
Additional notes on methods for antibody characterization:
1. These methods refer to characterization of antibodies that recognize endogenously expressed proteins. The requirements for characterization of antibodies that recognize epitope-tagged proteins are described elsewhere.
2. Current IDR standards for a narrow-peak ChIP-seq dataset are: Rescue Ratio RR_new = | Np U Nt | / | Np ^ Nt | Self consistency ratio SR_new = | N1 U N2 | / | N1 ^ N2 | where ^ = intersection (common) of 2 peak sets U = union (merge) of 2 peaks sets
If (R_new > 2) AND (SR_new > 2) then the replicates are proclaimed to have low reproducibility (failed) and flagged with -1 quality score If (RR_new > 2) OR (SR_new > 2) but not both, then the replicates are proclaimed to have moderate reproducibility (passed) and flagged with a 0 quality score If (RR_new <= 2) AND (SR_new <= 2) then the replicates are proclaimed to have high reproducibility (passed) and flagged with +1 quality score.
3. Scientists within and outside ENCODE have learned over time that some antibodies that perform well in ChIP assays nevertheless fail to pass the conventional tests that comprise primary and secondary analyses. Therefore, exceptions to the basic characterization can be considered for such cases. The antibody characterization review committee together with the ENCODE Production PIs, will consider these on a case by case basis. Datasets using such reagents, referred to as “exempt” antibodies, will be flagged in the ENCODE data.

Guidelines for ENCODE Epitope-tagged transcription factor ChIP-seq
ENCODE uses a variety of methods to characterize tagged TFs in ChIP-seq experiments, and these methods are categorized as being either genomic characterizations (to ensure the correct locus of interest was tagged properly), or immunological characterizations (to ensure the antibody recognizes the epitope-tagged protein) Typically, one form of the experiments listed under part A (Genomic DNA Characterization) and one form of the experiments listed under part B (Immunocharacterization) is used for a given TF.

A. Genomic DNA characterization (A-1 or A-2 should be performed)
The experimental design relies on correct integration of the epitope tag sequence into genomic DNA of the recipient cell line. One of the following genomic characterizations is performed:
A-1. PCR analysis
PCR is used to verify the presence of the intended integrated sequence at the intended site of integration. PCR primers are designed such that the amplification product is generated only if the epitope tag is integrated correctly in the genomic DNA. In this design, one primer is selected to anneal outside the region used for the homology-directed repair (the mechanism used for integration), and one primer is located inside the tag sequence.
A-2. DNA sequencing of integrated tag segment

Genomic DNA is used to show epitope-tag integration at the designed target site. Sanger or next-generation DNA sequencing of genomic DNA showing correct integration of the tag sequence is performed for this determination.

What is reported for Genomic Characterizations:
A gel image of the PCR reaction products with a DNA sizing ladder. A negative control sample (amplification from wild-type DNA) should be included if available. The expected size should be indicated, along with the PCR primer sequences and thermocycling conditions used to generate the products. For sequencing data, an electropherogram (Sanger sequence trace) or genome browser screenshot with an indication of the integration region within the wild-type genomic DNA.
Genomic Characterization requirements to be considered fully validated for ENCODE data:
Ideally for PCR and sequencing data, results from both replicates should be represented. If, however, only one replicate is present or passes genomic validation, then a passing grade can be assigned if both replicates passed IDR from ChIP-seq.

B. Immunocharacterization (B-1 or B2 should be performed)
The epitope-tagged ChIP-seq experiment relies on a well-characterized antibody raised against the epitope tag. Immunological characterization of the antibody in each parental target cell population or type, prior to introduction of the tag, is performed. This characterization is used to detect any significant off-target ChIP signals due to cross-reactivity of the antibody with proteins other than the designed tagged protein. Epitope-tagged cell immunocharacterization is done by performing one of the methods below (B-1 or B-2).
B-1. Immunoblot (Western blot) or Immunoprecipitation blot (IP-Western blot)
It is preferred that the antibody used for the blots is the same one as used in the ChIP-seq experiment. However, it is recognized that antibodies differ in their ability to detect denatured and native proteins. Therefore, if necessary, another antibody raised against the epitope tag can be used for the Western blot. A band (or bands) corresponding to predicted migration for the epitope tagged protein (or multiple forms, if they are predicted) should be visible when comparing the epitope-tagged cell line versus the “wild-type” cell line. The background control for immunocharacterization is the “wild type” cell line without a tag integration event. This control experiment is performed at least once for each parental cell line that is used.
What is reported for Immunoblot or Immunoprecipitation blot (IP-Western blot):
An image of the blot/gel showing affinity of the antibody for the epitope-tagged protein from either cell lysates (immunoblot) or immunoprecipitated proteins from cell lysates (immunoprecipitation blot). A protein sizing ladder should be included as well as a description of the blotting method and conditions for immunostaining. For immunoprecipitation blots, the antibodies used for both immunoprecipitation and visualization should be indicated. The expected size of the tagged target protein should be indicated as well as other bands that might correspond to either lower size degradation products or putative post-translational modifications.
Immunoblot or Immunoprecipitation blot requirements to be considered fully validated for ENCODE data:
The protein band of interest must be within 20% of the size predicted by the coding region. If the Western blot or IP-Western blot result meets this criteria, we consider the engineered cell line to meet expectations. If however, protein sizes do not match expected sizes which include the tag, then Western blots with native antibodies from commercial vendors can be used for compliance if the sizes are equivalent. Protein modifications and degradation products are known to complicate the sizing and intensity of bands, therefore, all instances must be thoroughly explained in the corresponding captions so that users of the data are made aware.
B-2. Immunoprecipitation followed by mass spectrometry
A cell or nuclear extract from cells expressing the tagged protein is immunoprecipitated with the same antibody used to perform ChIP-seq. These characterizations should be performed using the same lot number of antibody as used in the reported ChIP-seq experiments. The IP product is then fractionated on a denaturing polyacrylamide gel, and the fractions are prepared and analyzed by mass spec as described below.
What is reported for IP mass spec:
An IP-Western blot gel image with an outline of gel slices that were submitted for mass spec should be reported. If, however, the entire IP was used for the mass spec analysis, a Western blot or IP-Western blot image is not required. A list of all peptides (with peptide counts) from all immunoreactive bands should be presented in tabular format. Fold enrichment of all the peptides in the immunoreactive bands vs either mock IP or a set of proteins that have been immunoprecipitated from the same cell type using a collection of other antibodies from the same host species (the list of proteins used as the set of IP contaminants list must be provided) should also be determined.
IP mass spec requirements to be considered fully validated for ENCODE data:
The target protein should be enriched within the top 20 ranked proteins in the IP when compared to a mock IP or to a set of proteins that have been immunoprecipitated from the same cell type using a collection of other antibodies from the same host species. Ideally, the target TF would represent the highest ranking TF within this enrichment. If it is not however, then the production lab should indicate potential complexes or interacting partners (if known) that have co-immunoprecipitated with their target TF or provide an appropriate audit if the ChIP-seq data is deemed of high quality. In situations involving mock IPs for which the target protein has 0 peptides in the mock IP, a ranking by enrichment cannot be performed. In this case, the following criteria are considered for validation: the target TF is the top TF present as ranked by

the number of detected peptides or, the target TF is not the top TF ranked by peptide counts but is documented to be in a complex or have interactions with the other TFs having more counts. For situations where the target TF is not the top TF and there are no documented instances of interactions with other TFs having more counts, then an audit is assigned with the explanation that the enrichment could not be determined due to the lack of detected peptides in the IgG control and that no published data exists linking the target to the non-target TFs.

EXCEPTIONS
We realize that, in some cases, situations may arise in which antibodies or tagged factor lines do not pass the above standards, but the data producers feel that the datasets should be made available to users. Often there is data from other sources that support a ChIP-seq dataset that has not passed both A and B standards. Examples include the same epitope tagging reagents having passed in another cell type, or a high overlap of peaks to an antibody based dataset in the same cell type, or a highly similar motif found to one previously published for that factor. Therefore, exceptions to these characterization standards are considered for special cases. The antibody characterization review committee of the ENCODE Consortium will consider each special request. If an exception is granted, the datasets using these "exempt" antibodies will be flagged in the ENCODE datasets.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HepG2 - ATCC - HB-8065
Authentication	Phenotypic characterization.
Mycoplasma contamination	Cells are routinely tested for Mycoplasma contamination. All tests were negative.
Commonly misidentified lines (See ICLAC register)	HepG2 is not listed as being commonly misidentified.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	encodeproject.org GEO: GSE104247
Files in database submission	GSM2797484 ARID3A GSM2797485 ARID4B GSM2797486 ARID5B GSM2797487 ASH2L GSM2797488 ATF1 GSM2797489 ATF3 GSM2797490 ATF4 GSM2797491 BACH1 GSM2797492 BCL6_iso1 GSM2797493 BHLHE40 GSM2797494 BMI1 GSM2797495 BRCA1 GSM2797496 BRD4 GSM2797497 CBX1 GSM2797498 CBX5 GSM2797499 CEBPA GSM2797500 CEBPB GSM2797501 CEBPD GSM2797502 CEBPG GSM2797503 CEBPZ GSM2797504 CHD2 GSM2797505 CIZ1 GSM2797506 CREB1 GSM2797507 CREM GSM2797508 CTCF GSM2797509 CUX1 GSM2797510 DMAP1 GSM2797511 DNMT3B GSM2797512 DRAP1 GSM2797513 E2F7_iso1

GSM2797514	ELF3
GSM2797515	EP300
GSM2797516	ERF
GSM2797517	ESRRA
GSM2797518	ETV4
GSM2797519	EZH2
GSM2797520	FOS
GSM2797521	FOSL2
GSM2797522	FOXA1
GSM2797523	FOXA2
GSM2797524	FOXA3
GSM2797525	FOXK1
GSM2797526	FOXO1
GSM2797527	FOXP1
GSM2797528	GABPA
GSM2797529	GABPB1_v2
GSM2797530	GATA4
GSM2797531	GATAD1
GSM2797532	GATAD2A
GSM2797533	GMEB2
GSM2797534	HBP1
GSM2797535	HCFC1
GSM2797536	HDAC2
GSM2797537	HHEX
GSM2797538	HLF
GSM2797539	HMG20A
GSM2797540	HMG20B_v2
GSM2797541	HMGXB3
GSM2797542	HMGXB4
GSM2797543	HNF1A
GSM2797544	HNF4A
GSM2797545	HNF4G
GSM2797546	HOMEZ
GSM2797547	HSF1
GSM2797548	IRF3
GSM2797549	JARID2_iso1
GSM2797550	JUND
GSM2797551	JUN
GSM2797552	KAT2B
GSM2797553	KAT7
GSM2797554	KAT8
GSM2797555	KDM1A
GSM2797556	KDM2A
GSM2797557	KDM3A
GSM2797558	KDM6A
GSM2797559	KLF10
GSM2797560	KLF11
GSM2797561	KLF16
GSM2797562	KLF6_v2
GSM2797563	KLF9
GSM2797564	KMT2B
GSM2797565	MAFF
GSM2797566	MAFK
GSM2797567	MAX
GSM2797568	MAZ
GSM2797569	MBD1_v1
GSM2797570	MBD1_v2
GSM2797571	MBD4
GSM2797572	MIER2
GSM2797573	MIER3
GSM2797574	MIXL1
GSM2797575	MLX
GSM2797576	MTA1
GSM2797577	MXD3_v1
GSM2797578	MXD4
GSM2797579	MXI1
GSM2797580	MYBL2
GSM2797581	MYC
GSM2797582	NCoA2
GSM2797583	NFE2L2
GSM2797584	NFIA_v1
GSM2797585	NFIC
GSM2797586	NFIL3
GSM2797587	NFYC
GSM2797588	NPAS2_iso2

GSM2797589 NR1H2
GSM2797590 NR2C2
GSM2797591 NR2F1
GSM2797592 NR2F2
GSM2797593 NR2F6
GSM2797594 NR3C1
GSM2797595 NRF1
GSM2797596 PAF1_v1
GSM2797597 PAXIP1_iso1
GSM2797598 PBX2
GSM2797599 POLR2A
GSM2797600 POLR2AphosphoS2
GSM2797601 POLR2AphosphoS5
GSM2797602 PPARG_v1
GSM2797603 PRDM10
GSM2797604 PROX1
GSM2797605 RAD21
GSM2797606 RARA
GSM2797607 RCOR1
GSM2797608 RCOR2
GSM2797609 RERE
GSM2797610 REST
GSM2797611 RFX3_iso1
GSM2797612 RFX5
GSM2797613 RFXANK
GSM2797614 RING1
GSM2797615 RREB1_iso2
GSM2797616 RUVBL1
GSM2797617 RXRA
GSM2797618 RXRB
GSM2797619 SAP130
GSM2797620 SIN3A
GSM2797621 SIN3B
GSM2797622 SIX4_iso1
GSM2797623 SLC30A9
GSM2797624 SMAD4_iso1
GSM2797625 SMC3
GSM2797626 SOX13
GSM2797627 SOX13_iso1
GSM2797628 SOX5
GSM2797629 SP1
GSM2797630 SP2
GSM2797631 SP5
GSM2797632 SREBF1
GSM2797633 SRF
GSM2797634 SSRP1
GSM2797635 SUZ12
GSM2797636 TAF1
GSM2797637 TBL1XR1
GSM2797638 TBP
GSM2797639 TCF12
GSM2797640 TCF25
GSM2797641 TCF7
GSM2797642 TCF7L2
GSM2797643 TEAD1
GSM2797644 TEAD3
GSM2797645 TEAD4
GSM2797646 TFAP4
GSM2797647 TFDP1
GSM2797648 TFE3
GSM2797649 TGIF2
GSM2797650 THAP11
GSM2797651 THRA_iso1
GSM2797652 THRB
GSM2797653 UBP1
GSM2797654 USF1
GSM2797655 USF2
GSM2797656 YY1
GSM2797657 ZBTB21
GSM2797658 ZBTB26
GSM2797659 ZBTB33
GSM2797660 ZBTB7A
GSM2797661 ZC3H4
GSM2797662 ZCCHC11
GSM2797663 ZEB1

GSM2797664 ZFP1_v1
GSM2797665 ZFP64
GSM2797666 ZGPAT
GSM2797667 ZHX2
GSM2797668 ZHX3_iso1
GSM2797669 ZKSCAN8
GSM2797670 ZMYM3
GSM2797671 ZNF12
GSM2797672 ZNF143
GSM2797673 ZNF189
GSM2797674 ZNF219
GSM2797675 ZNF274
GSM2797676 ZNF281
GSM2797677 ZNF331
GSM2797678 ZNF334_iso1
GSM2797679 ZNF335
GSM2797680 ZNF384
GSM2797681 ZNF3
GSM2797682 ZNF48
GSM2797683 ZNF511
GSM2797684 ZNF544_iso1
GSM2797685 ZNF580
GSM2797686 ZNF644
GSM2797687 ZNF652
GSM2797688 ZNF792
GSM2797689 ZNF7_iso2
GSM2797690 ZSCAN29_iso1
GSM2797691 ZSCAN9
GSM2797692 Input 1
GSM2797693 Input 2
GSM2797694 Input 3
GSM2797695 Input 4
GSM2797696 Input 5
GSM2797697 Input 6
GSM2797698 Input 7
GSM2797699 Input 8
GSM2797700 Input 9
GSM2797701 Input 10
GSM2797702 Input 11
GSM2797703 Input 12
GSM2797704 Input 13
GSM2797705 Input 14
GSM2797706 Input 15
GSM2797707 Input 16
GSM2797708 Input 17
GSM2797709 Input 18
GSM2797710 Input 19
GSM2797711 Input 20
GSM2797712 Input 21
GSM2797713 Input 22
GSM2797714 Input 23
GSM2797715 Input 24
GSM2797716 Input 25
GSM2797717 Input 26
GSM2797718 Input 27
GSM2797719 Input 28
GSM2797720 Input 29
GSM2797721 Input 30
GSM2797722 Input 31
GSM2797723 Input 32
GSM2797724 Input 33
GSM2797725 Input 34
GSM2797726 Input 35
GSM2797727 Input 36
GSM2797728 Input 37
GSM2797729 Input 38
GSM2797730 Input 39
GSM2797731 Input 40

Genome browser session
(e.g. UCSC)

no longer applicable

Methodology

Replicates

Duplicate experiments as described above and on ENCODE portal.

Sequencing depth	Each experiment >20M reads, single end 50, single end 75, single end 100, paired end 100. Details listed on ENCODE portal.
Antibodies	Listed above and on ENCODE portal.
Peak calling parameters	All settings described on ENCODE portal.
Data quality	All validation and QC are described on the ENCODE portal.
Software	Software listed above, described in the methods section of the manuscript, and on the ENCODE portal.