Discovering unknown human and mouse transcription factor binding sites and their characteristics from ChIP-seq data

Chun-Ping Yu[a], Chen-Hao Kuo[a], Chase W. Nelson[a,b], Chi-An Chen[a], Zhi Thong Soh[a], Jinn-Jy Lin[a], Ru-Xiu Hsiao[a], Chih-Yao Chang[a], and Wen-Hsiung Li[a,c,1]

[a]Biodiversity Research Center, Academia Sinica, 115 Taipei, Taiwan; [b]Institute for Comparative Genomics, American Museum of Natural History, New York, NY 10024; and [c]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637

Contributed by Wen-Hsiung Li, April 6, 2021 (sent for review December 30, 2020; reviewed by Han Liang and Zhongming Zhao)

Transcription factor binding sites (TFBSs) are essential for gene regulation, but the number of known TFBSs remains limited. We aimed to discover and characterize unknown TFBSs by developing a computational pipeline for analyzing ChIP-seq (chromatin immunoprecipitation followed by sequencing) data. Applying it to the latest ENCODE ChIP-seq data for human and mouse, we found that using the irreproducible discovery rate as a quality-control criterion resulted in many experiments being unnecessarily discarded. By contrast, the number of motif occurrences in ChIP-seq peak regions provides a highly effective criterion, which is reliable even if supported by only one experimental replicate. In total, we obtained 2,058 motifs from 1,089 experiments for 354 human TFs and 163 motifs from 101 experiments for 34 mouse TFs. Among these motifs, 487 have not previously been reported. Mapping the canonical motifs to the human genome reveals a high TFBS density ±2 kb around transcription start sites (TSSs) with a peak at −50 bp. On average, a promoter contains 5.7 TFBSs. However, 70% of TFBSs are in introns (41%) and intergenic regions (29%), whereas only 12% are in promoters (−1 kb to +100 bp from TSSs). Notably, some TFs (e.g., CTCF, JUN, JUNB, and NFE2) have motifs enriched in intergenic regions, including enhancers. We inferred 142 cobinding TF pairs and 186 (including 115 completely) tethered binding TF pairs, indicating frequent interactions between TFs and a higher frequency of tethered binding than cobinding. This study provides a large number of previously undocumented motifs and insights into the biological and genomic features of TFBSs.

ChIP-seq | transcription factor | binding site | promoter | position weight matrix

Transcription factor binding sites (TFBSs) are a central element of gene regulation. The collection of TFBSs bound by a TF is commonly represented by a position weight matrix (PWM) describing the relative likelihood of observing a given nucleotide at each position of the TFBS. Once determined, the PWM of a TF can be used to predict its target genes and the biological pathway(s) it affects, which together define the TF's function.

The importance of TFBSs is reflected by the many techniques that have been developed for their identification, including chromatin immunoprecipitation followed by sequencing (ChIP-seq) (1, 2), protein-binding microarray (3), systematic evolution of ligands by exponential enrichment (SELEX) (4), and DNA affinity purification sequencing (5). The availability of so many techniques notwithstanding, accurate determination of TFBSs is still not simple and "known" PWMs are still limited in number (6) and follow variable quality standards (7).

The preferred method for determining TFBSs has been ChIP-seq, an in vivo technique that can be used for genome-wide mapping of TFBSs and epigenetic marks. For humans, the ENCODE project has produced 1,621 "released" status experiments (accessed 5 October 2019) (8), representing the largest quantity of uniformly processed ChIP-seq data currently available. However, fewer than half of these experiments have previously been analyzed to infer PWMs, as a substantial volume of new ENCODE ChIP-seq data has accumulated since previous systematic analyses (9, 10). The present study analyzed all the currently available ENCODE human and mouse ChIP-seq data.

For the above purpose, we developed a computational pipeline that separately and simultaneously utilizes all available experiments and biosamples to infer PWMs for each TF studied (https://github.com/chpngyu/chip-seq-pipeline). As a resource for future research, our results are presented in a freely accessible database (dbTFBS: https://dbtfbs.cistro.me/), which also displays comparisons to PWMs available in other databases. After inferring PWMs, we mapped PWMs to the human genome to address questions such as the positional distribution of TFBSs in the human genome, the potential regulators of specific genes, and the clustering of TFBSs in promoters. Moreover, we addressed the issue of cooccurring motifs from different TF families within the same ChIP-seq data (6, 9). Our study provides not only many previously unreported motifs but also abundant data on the biological features of human TFBSs.

Results

Inferred PWMs in Human and Mouse. Our analysis of the ChIP-seq data was divided into two stages: 1) analysis using a set of

Significance

Transcription factor binding sites (TFBSs) are essential for gene regulation, but the majority of TFBSs remain unknown. To discover new TFBSs, we developed a computational pipeline to analyze human and mouse ChIP-seq (chromatin immunoprecipitation followed by sequencing) data. We found that the number of motif occurrences in ChIP-seq peaks is a highly effective quality-control criterion. Analyzing 1,089 human (101 mouse) TF experiments, we inferred 2,058 (163) motifs, including 487 previously unreported motifs, revealing predicted TFBS enrichment in promoters, particularly near transcription start sites. However, some TFs have TFBSs enriched in intergenic regions, including enhancers. We found frequent interactions between TFs, with a higher frequency of tethered binding than cobinding. This study provides many previously unreported TFBSs and describes their biological features.

Author contributions: C.-P.Y., C.-H.K., C.W.N., and W.-H.L. designed research; C.-P.Y., C.-H.K., C.W.N., C.-A.C., Z.T.S., R.-X.H., C.-Y.C., and W.-H.L. performed research; C.-P.Y., C.-H.K., C.W.N., C.-A.C., Z.T.S., R.-X.H., and C.-Y.C. analyzed data; and C.-P.Y., C.-H.K., C.W.N., C.-A.C., Z.T.S., C.-Y.C., and W.-H.L. wrote the paper.

Reviewers: H.L., The University of Texas MD Anderson Cancer Center; and Z.Z., The University of Texas Health Science Center at Houston.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: whli@uchicago.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2026754118/-/DCSupplemental.

Published May 11, 2021.

stringent rules and 2) recovery of data discarded in the first stage using a modified rule set (Fig. 1*A*). In the first stage, we restricted our analysis to ChIP-seq experiments having two or more replicates with an irreproducible discovery rate (IDR) (11, 12) score ≥1.3 [$-\log_{10}(0.05) = 1.3$; https://github.com/IDR/idr-py] and to replicates containing at least one PWM having both E <0.0001 and occurrences in >100 of the top 500 peaks (see *Methods*). For a TF with a single experiment, the replicates passing these two criteria were merged to infer PWMs. For a TF with multiple experiments, we used a nonparametric method to rank peaks from all experiments and selected the top 500 peaks to infer PWMs (*Methods*). After inferring the PWMs for a TF, we chose up to five PWMs that were supported by >100 peaks. We considered the PWM supported by the largest number of peaks to be the top (primary) PWM and the remaining to be secondary PWMs.

For human TFs, there were 1,621 ChIP-seq experiments in "released" status in the ENCODE project (accessed 5 October 2019). All experiments were analyzed except for CTCF, for which 129 of 311 experiments were sampled. A total of 1,439 experiments for 502 TFs in 49 TF families were analyzed, from which we obtained 1,829 PWMs for 346 TFs in 45 TF families from the 981 experiments (Dataset S1).

As an example of our method, Fig. 1 shows our analysis of JUND (*JunD proto-oncogene*), a TF in the bZIP family. Fig. 1*B* shows the top five PWMs obtained from a JUND experiment using liver tissue (accession ENCSR837GTK). Among the top five PWMs in the top 500 peak regions, only one has >100 peaks ("Motif 1"). This PWM is found in >80% of the top 500 peaks, but in only 50% of the top 4,000 peaks (Fig. 1*C*), affirming the advantage of using only top-scoring peaks for inferring PWMs. As each of the remaining four PWMs is found in <100 of the top

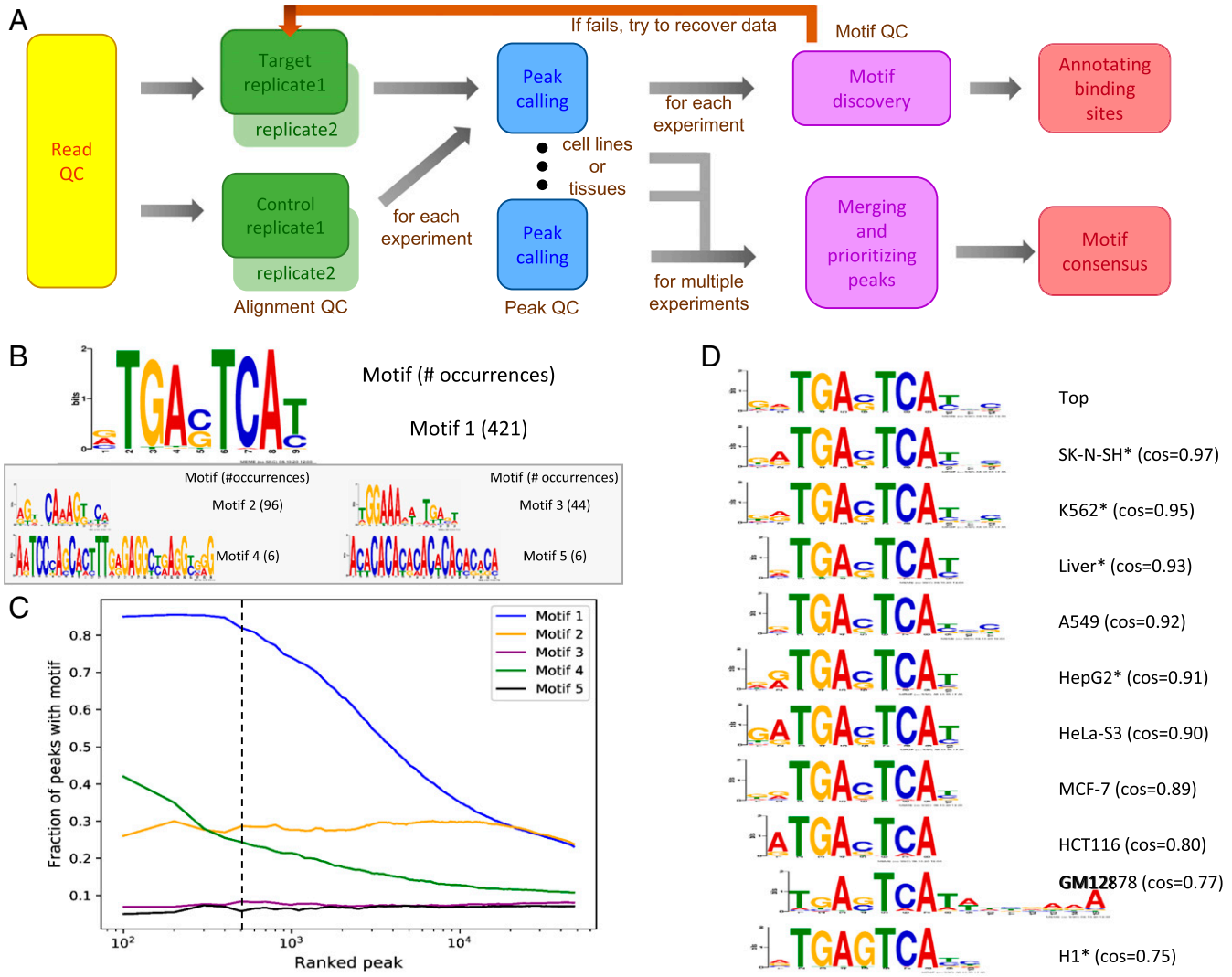

Fig. 1. Computational pipeline and method for selecting the top PWMs of a TF from individual experiments and the top PWM across multiple experiments, using JUND as an example. (*A*) Flowchart of the computational pipeline. The detailed criteria and parameters are given in *Methods* and *SI Appendix*, Fig. S1. (*B*) The top five PWMs from one experiment (liver tissue; accession ENCSR837GTK). The numbers of occurrences of each PWM in the top 500 peak regions are indicated on the right; one PWM (motif 1; top) has >100 occurrences. (*C*) Cumulative proportion of called peaks containing each of the top five PWMs. The ranked peaks are sorted by MACS2 score, with ranks closer to 1 corresponding to higher scores (x axis). For each PWM, the fraction of peaks (y axis) was calculated as the occurrence of a motif (FIMO $P < 0.0001$) in the top N peaks (x axis). The dashed vertical line denotes the peak with rank 500, with peaks to the left of the line used to infer PWMs (top 500 peaks). (*D*) The top PWM obtained from all 10 JUND biosamples is used as the reference for comparison with the top PWM of each individual biosample, each of which could contain one or more experiments. The PWM from a biosample with multiple experiments is indicated by an asterisk (*). For each PWM logo, positions with information content (IC) <0.3 bits were trimmed from both ends.

500 peaks (Fig. 1*C*), none of them is considered a PWM of JUND. In Fig. 1*D*, the first PWM ("Top") is the top PWM obtained by analyzing the data from all passing JUND experiments (15 of 17) across 10 biosamples. The remaining PWMs are those obtained from each of the 10 individual biosamples. Similarity between two PWMs is measured by the k-mer frequency cosine angle (hereafter "cos") between two PWMs, which is similar to the Pearson correlation coefficient (PCC) (see *Methods*). Eight of the 10 top PWMs from individual biosamples are highly similar to the top PWM among all samples, with cos ≥0.80. These eight PWMs all have a core motif sequence with the palindrome TGA[C/G]TCA. The two PWMs with cos < 0.80 are inferred from the GM12878 and H1 cell lines and show different binding specificities in the core sequence or flanking regions.

For the mouse, the ENCODE project included 193 experiments for 54 TFs in 14 TF families. We obtained 158 PWMs for 30 TFs in 12 TF families from 97 experiments (Dataset S2).

In the second stage of analysis, we tried to recover discarded data (*Methods*). We recovered each experiment that had at least one replicate with at least one PWM supported by >250 of the top 500 peaks. In this analysis, if merging increased the number of peaks supporting the top PWM, replicates of an experiment were merged irrespective of their IDR score. Otherwise, the replicate with the PWM supported by the largest number of peaks was selected to infer PWMs (*Methods*). In this effort, we recovered 108 experiments and obtained 229 additional PWMs for 56 human TFs (https://dbtfbs.cistro.me/?jump=Total) and 4 experiments and 5 additional PWMs for 4 mouse TFs (https://dbtfbs.cistro.me/?pwd=db-mmu&jump=Total). The recovered experiments are indicated as "Recovered" in the "Notes" column in our database.

Putting together the PWMs inferred from the two stages, we obtained a total of 2,058 PWMs for 354 human TFs and a total of 163 PWMs for 34 mouse TFs.

C2H2 TFs. Because C2H2 binding sites are often dominated by endogenous retroelements, we also conducted an RCADE analysis, as recommended by Najafabadi et al. (13) (see *Methods*). Our MEME-ChIP analysis yielded at least one PWM for each of 166 C2H2 TFs. Our RCADE analysis yielded 87 PWMs that are similar to the PWMs obtained from MEME-ChIP analysis: 5 cases with cos = 0.76, 0.77, 0.78, 0.78, and 0.79, respectively, with the remaining 82 cases having cos > 0.8 (https://dbtfbs.cistro.me/?jump=C2H2_ZF). The RCADE analysis failed to yield any acceptable PWM for 55 C2H2 TFs, which are indicated as "No results" in our database. The remaining 24 C2H2 TFs having PWMs from MEME-ChIP analysis exhibited cos <0.75 with the PWMs obtained from RCADE analysis. Thus, in 79 cases, RCADE failed to provide support for the PWMs obtained from MEME-ChIP analysis. These PWMs should be taken with caution (indicated as "uncertain" in our database). However, more than half of the cases yielded similar PWMs using MEME-ChIP (87) and RCADE (79), possibly because our analysis excluded Blacklist (repetitive) regions, which may have excluded most endogenous retroelements.

For the mouse, we obtained PWMs for five C2H2 TFs by MEME-ChIP and by RCADE, and in every case the PWMs obtained by the two methods are highly correlated (cos = 0.92).

Canonical Motifs and Cooccurring Motifs. The canonical motif of a TF refers to the sequence-specific DNA motif that is directly bound by the TF. Cooccurrence can be inferred when a motif found in the ChIP-seq data of one TF (TF1) is similar (cos ≥ 0.80) to the canonical motif of a second TF (TF2) from another family. We have developed a set of rules for classifying the PWMs inferred from an experiment as "canonical," "candidate canonical," "cooccurring," or "unannotated" (see *Methods*). Briefly, a PWM is classified as canonical if it is similar (cos ≥ 0.80) to a known canonical PWM of a TF in the same TF family, while it is classified as cooccurring if it is similar to a known canonical PWM of a TF that belongs to another family. If the top PWM is not similar to any known canonical PWM in any database or literature, it is likely a previously uncharacterized canonical PWM; however, we call it "candidate canonical" because it might be similar to an unknown motif of another TF. A secondary PWM is classified as "unannotated" if it is not similar to any known PWM.

In total, we inferred 776 canonical, 113 candidate canonical, 455 cooccurring, and 374 unannotated PWMs for the 308 human TFs studied (Dataset S1). The number of canonical PWMs was larger than the number of TFs studied because different experiments for a TF often gave somewhat different PWMs. For a TF with multiple experiments, we assigned the top PWM obtained by our ranking method utilizing all experiments (*Methods*) as the canonical PWM. We also inferred 95 canonical, 40 cooccurring, and 40 unannotated PWMs for the 34 mouse TFs studied.

Unannotated Motifs. We found 487 human PWMs that have not been annotated before (Dataset S1), likely representing novel PWMs. Among these PWMs, 113 are primary PWMs and so are likely canonical. However, we classify them as "candidate canonical" because such a motif may be the canonical motif of another TF (e.g., an uncharacterized TF). For these candidate canonical motifs, we examined their 1) enrichment in promoters and/or enhancers and 2) evolutionary conservation during primate evolution (*Methods*). We required candidates to pass two tests: 1) twofold or greater enrichment in promoters or ≥1.5-fold enrichment in enhancers and 2) twofold or greater increase in conservation score as compared to intergenic regions. For instance, we inferred a novel candidate canonical PWM of ZBED5 in the BED ZF family that has a fivefold higher probability of being found in promoters and twofold higher conservation score as compared to intergenic regions (*SI Appendix*, Fig. S2). Overall, 61 (54%) of these 113 PWMs passed the two tests, providing evidence that they represent bona fide motifs.

The remaining 374 PWMs are secondary PWMs. Among them, 242 motifs were found in experiments with a successfully identified canonical motif. Thus, they are likely cooccurring motifs bound by unknown TFs. We conducted the above two tests (enrichment and conservation) on each of these PWMs and found that 132 (55%) of them passed the tests. The remaining 132 unannotated motifs cooccurred with noncanonical primary motifs, so that they themselves might be canonical. We found that 56 (42%) of these PWMs passed the above two tests and are likely functional motifs (Dataset S3).

To find support for the inferred novel motifs, we checked newly released ENCODE ChIP-seq data (released in July 2020 or later) from cell lines not used before. For 31 TFs with novel PWMs, we found 32 experiments (one TF with two experiments) from the newly released data. We could infer PWMs for 23 of the 31 TFs (Dataset S4). Among the 23 PWMs, 11 were similar (cos > 0.8) to the inferred novel PWMs and another had cos = 0.75, and these (11 + 1) PWMs were each supported by a large number of peaks (325 on average). Thus, these PWMs are likely functional. The remaining 11 TFs had a cos < 0.6 and the PWMs were supported by a lower number of peaks (~200 on average). These 11 PWMs may not be functional and should be taken with caution.

Cooccurrence of Motifs from Different TF Families. Motifs from different TF families may cooccur in the same ChIP-seq data. Cooccurrence may result from two situations, where TF1 is the TF being assayed and TF2 belongs to a different family (9): 1) cobinding, in which the two TFs (TF1 and TF2) tend to bind neighboring sites, and 2) tethered binding, in which TF1 binds to TF2 which, in turn, binds directly to DNA. Cobinding is inferred when the primary motif of TF1 is equally or more frequent than

GENETICS

that of TF2, indicating that each TF can bind separate DNA motifs (9). Cobinding is also inferred if the motif of TF2 is less frequent than the joint occurrence of the two motifs in the same peak regions (*Methods*). Tethered binding is inferred if the motif of TF2 is the primary motif, that is, it is more frequently found than that of TF1. The mathematical expressions for these criteria are presented in *Methods*.

Using the above rules and the known and inferred canonical motifs of TFs, we found 142 TF pairs that showed cobinding in one or more cell lines (Fig. 2 and Dataset S5). For example, USF2 (bHLH) tends to cobind with NFYA/C (an unknown TF) with consensus CCAAT and SP1/2 or MAZ with consensus GGG[C/A]GGG. Moreover, 186 TF pairs were identified as participating in tethered binding. Among the tethered binders, 130 TF pairs were completely tethered, showing only the non-canonical motif. For example, TCF12 was studied in eight cell lines, five of which were completely tethered by GATA, JUN, or FOXA1. In our data, FOXA1 showed cooccurrence with the largest number of TFs, including cobinding with 23 TFs and tethered binding with 15 TFs.

We found that CTCF has a canonical motif with the consensus sequence rsyGCCmyCTrsTGGyCr (r = A or G; s = G or C; y = C or T; m = A or C) and another motif with the consensus GGAACTGCAG that was only found in certain cell lines cooccurring with the canonical motif. Recently, Vierstra et al. (14) found that CTCF has a canonical and a secondary motif that is similar to the first motif above. It is noted that CTCF has multiple DNA-binding domains (DBDs), which could potentially bind to multiple motifs, as this example suggests. In such cases, a second motif may be mistaken as a cooccurring motif bound by an unknown TF.

Because the novel PWMs are dissimilar to known canonical PWMs and a TF may have secondary PWMs, we computed the positional distribution of the TF's novel PWMs relative to its canonical PWM (Dataset S6). A PWM with a short distance to the canonical PWM might be bound by another DBD of the same TF if it has multiple DBDs, that is, it can be a secondary PWM (*SI Appendix*, Fig. S3), while a novel PWM with a long distance to the canonical PWM might be bound by a TF with an unknown PWM.

In summary, among the 336 human TFs studied (not including the 18 TFs for which the PWMs inferred were uncertain), 42 showed cobinding, 84 showed tethered binding, 75 showed both cobinding and tethered binding, and 135 showed no interaction with other TFs (Dataset S5). In addition, we found that 37 (~26%) of the 142 cobinding pairs and 33 (~18%) of the 186 tethered binding pairs (Dataset S5) have protein–protein interaction data in BioGRID (15). Note that lack of BioGRID data could be due to either incompleteness of the database or lack of interaction.

PWMs with a Core Motif. We obtained canonical PWMs for 267 human TFs and 32 mouse TFs in 37 TF families; for simplicity, here "canonical" motifs include "candidate canonical" motifs. We define a motif (PWM) as a core motif of a TF family if it is shared (cos ≥ 0.8) by at least three canonical PWMs of the TF family. To find a core motif for a TF family, we clustered similar PWMs (cos ≥ 0.8) into groups. For each group with three or more members the core motif (PWM) was then constructed using the rules described in *Methods*, representing the consensus of the PWMs of the group. Among the 37 TF families studied, only 13 families have one or more groups with three or more PWMs available and thus have one or more core motifs (*SI Appendix*, Fig. S4). For example, for the 30 bHLH TFs studied, 19 share a core motif, 7 share another, and the remaining 4 TFs share none. For these 13 TF families, bHLH, bZIP, C2H2, and HMG/Sox have more than one core motif, implying that each of these TF families can be divided into TF subfamilies. In particular, bHLH and bZIP TFs form a homo/heterodimer that binds to palindromic sequences, e.g., core1 (E-box: CACGTG) and core2 (CAGCTG) of bHLH core motifs. The other nine families (E2F, ETS, Forkhead, GATA, Nuclear receptor, MADS box, SMAD, IRF, and Unknown) each has only one core motif. These core motifs are shown in our database.

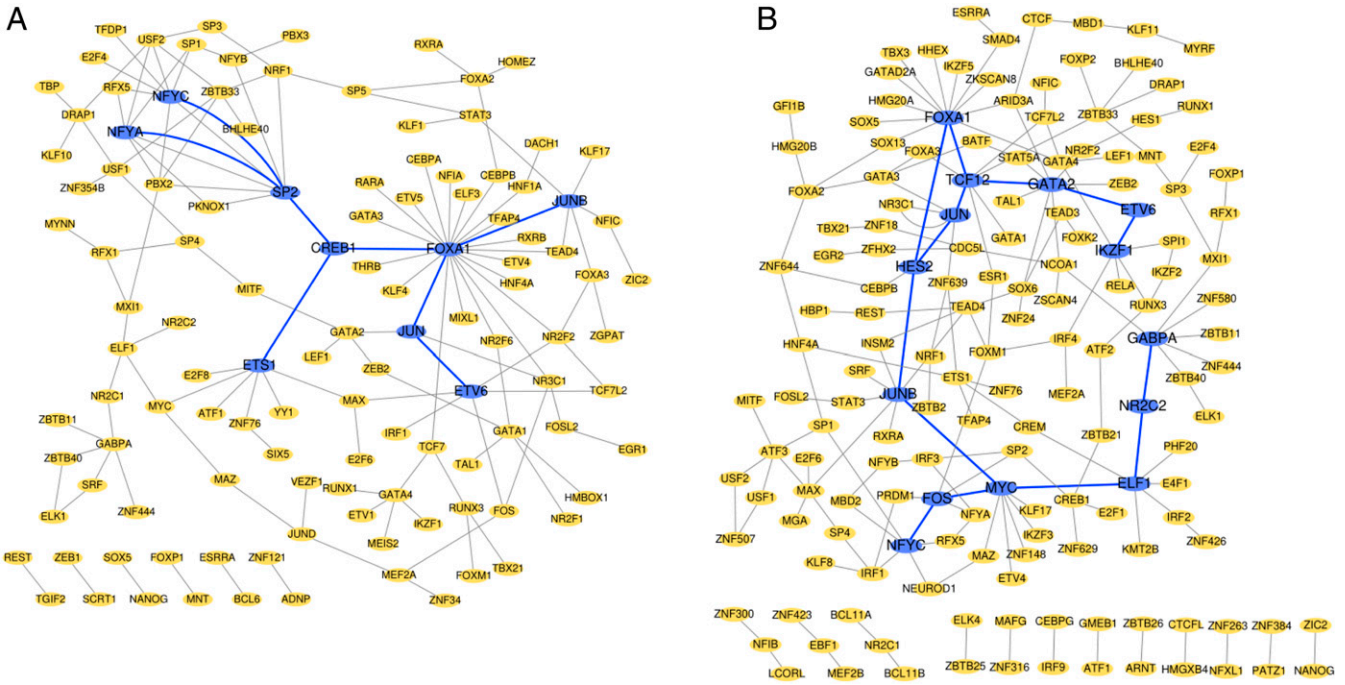

Fig. 2. Cooccurrence of PWMs between pairs of TF families. Each network is generated by Cytoscape (35). Each node is a TF. Hub TFs (degree >5) and bridge TFs between two hubs are highlighted in blue. For instance, TAL1 cobinds with GATA1, which cobinds with four TFs, including NR2F1, NR2F6, ZEB2, and HMBOX1. (*A*) Identified cobinding TF pairs. (*B*) Identified tethered binding TF pairs.

Spatial Distribution of TFBSs. It is useful to know not only a TF's motif(s), but also the locations of its potential binding sites (TFBSs) in the genome. To generate a list of candidate TFBSs, we used FIMO to locate significant matches ($P < 0.0001$) for each TF's canonical PWM within called peak regions across the human genome (all biosamples, not limited to the top 500 peaks; *Methods*). As a resource for future research, bedGraph and GFF format files are provided with the locations of these candidate TFBSs (Dataset S7), which can be displayed using a custom track in the UCSC Genome Browser. These data can be used to study the variation of TF binding among experiments (cell lines or tissues).

For a TF with only one experiment, we simply recorded the location of each PWM mapping to a peak region in the genome. For a TF with multiple experiments in the ENCODE project, we grouped the candidate TFBSs into clusters by concatenating overlapping appearances (≥1 bp overlap). The percentages of candidate TFBSs occurring in peak clusters are shown in Fig. 3.

To investigate variation in TFBSs among different samples of the same TF, we analyzed six TFs with the most abundant experiments in ENCODE (Table 1), where experiments utilizing the same biosample were merged and processed using our ranking method. For five of the six TFs, we found that ∼70% of the mapped TFBSs overlapped the same sites in at least two biosamples, while ∼30% of the sites were unique to one sample. However, about half of the unique sites were in intergenic regions. One exception is CTCF, for which ∼70% of candidate TFBSs occur in either intergenic or promoter regions, in accordance with its biological role of mediating intra- and interchromosomal contacts (16).

In promoter regions, defined as −1 kb to +100 bp relative to transcription start sites (TSSs), TFBSs have a sixfold or greater higher probability density than elsewhere (Fig. 3), with a maximum at −50 bp relative to the TSS. However, because promoters occupy only 1.2% of the human genome while intergenic regions and introns occupy 54% and 42%, only 12% of TFBSs are in promoters while 30% and 42% of TFBSs are in the intergenic regions and introns, respectively. Thus, both intergenic regions and introns contain many candidate TFBSs, likely because they contain enhancers.

We found that some TFBSs form long, continuous clusters in the genome wherein the motifs of one or more TFs directly overlap. These TFBSs can be tandem TFBS repeats of the same TF or combinations of different TFs, with length ≥500 bp (*SI Appendix*, Fig. S5).

Inferring the Biological Role of a TF. Based on the spatial distribution of TFBSs, we tested for the enrichment of PWM hits in specific genomic features, including promoters, exons, introns, and intergenic regions. To assess the positional preference of TFBSs for any specific TF, we selected candidate TFBSs derived from the most frequently used cell line, K562, and compared the proportion of TFBSs to the proportion of DNase I hypersensitive sites mapping to each genomic feature. DNase I sites were chosen to normalize for positional preference because they serve as a proxy for accessible chromatin; although they occupy only 4.4% of the human genome, 10% of them are in promoters (17), even though promoters occupy only 1.2% of the genome (18). Finally, we calculated the fold change (enrichment or depletion) of TFBSs relative to DNase I sites for each TF. We found that most TFs have predicted binding sites enriched in promoters, such as the YY1, MYC, ATF1, and E2F1 families, while some TFs have higher preference for intergenic regions, such as the CTCF, JUN, JUNB, and NFE2 families (Fig. 4 *A* and *B*). A TF that prefers promoters might be a core TF, while a TF that prefers intergenic regions might bind distal enhancers. For example, in the case of E2F1, >80% of binding sites are in promoters, 50% of which overlap TSSs. As another example, CTCF binds to border ranges, e.g., intergenic (34%) and intron (37%) regions, and is known to mediate long-range chromatin looping in promoter-distal regions. The binding preferences of the TFs are given in Dataset S8.

For TFs that prefer to bind intergenic regions, we also compared their TFBSs with known enhancer data from ENCODE (18) and FANTOM (19, 20). The enhancer data in ENCODE included four types of signatures: promoter-like signatures (PLSs), proximal enhancer-like signatures (pELSs), distal enhancer-like signatures (dELSs), and CTCF-only elements. The enhancers from FANTOM have 34.6% regions that overlap pELSs/dELSs in ENCODE. We found that the promoter-enriched TFs, such as YY1/YY2, ATF1, and E2F1, also have greater motif enrichment in PLSs; that the motifs of JUN, JUNB, and NFE2 are enriched in dELSs and in enhancer regions in FANTOM; and that KLF16 and ZNF654 tend to be associated with CTCF-only elements (Fig. 4*C*).

Comparison between Human and Mouse TFs. We found 25 TFs studied in both human and mouse (*SI Appendix*, Table S1). The sequence identity between orthologous human and mouse DBDs is ≥0.98 with two exceptions: 0.90 for MYC and 0.86 for NANOG. Similarly, the cos value between human and mouse PWMs is ≥0.94 with three exceptions: 0.87 for TCF3, 0.91 for TCF12, and 0.63 for NANOG. Thus, the canonical PWMs are generally well-conserved between human and mouse lineages. As the NANOG PWM has not been well conserved between human and mouse, the NANOG TF will not be included in the following discussion.

To infer the biological role of mouse TFs, we tested the enrichment of their TFBSs in five sets of enhancer-like regions as we did for human TFs (see *Methods*). The fold changes of the above 24 mouse TFs are significantly correlated with the fold changes of the 24 human TFs in dELSs (PCC = 0.78, $P = 7.6 \times 10^{-6}$), enhancers of FANTOM (PCC = 0.80, $P = 2.3 \times 10^{-6}$), and pELSs (PCC = 0. 72, $P = 8.6 \times 10^{-5}$). The correlations in PLSs and CTCF-only regions are both only PCC = ∼0.53 ($P = 8.4 \times 10^{-3}$), likely because most of the 24 TFs are not promoter- or CTCF-associated.

Database. We have constructed a database called "dbTFBS" (https://as0201821.github.io/dbTFBS/) to store the results of this study. The TFs studied are divided into TF families and each TF is given an entry. The first column shows the top PWM inferred for the TF; more than one entry is given if there are different top PWMs inferred from different experiments. The top PWM for an entry is compared with those from other studies and databases if available. Cooccurring motifs are classified into cobinding and tethered binding. The PWMs inferred for each experiment are also accessible.

Discussion

Computational Pipeline. Our computational pipeline for analyzing ChIP-seq data (*SI Appendix*, Fig. S1) is a combination of new and existing methods (9, 21–23). It has three key features. First, when merging reads from two replicates of an experiment, it requires both replicates to have one or more PWMs supported by >100 peaks in addition to the condition of IDR >1.3. The additional requirement is to ensure sufficient motif enrichment for inferring biologically meaningful PWMs. Second, rather than simply choosing one experiment or naïvely merging all read data from all available experiments, our procedure selects only the experiments that have passed the above two criteria, uses a ranking method to select the top 500 peaks from each selected experiment, merges the selected peaks from all selected experiments to form peak clusters, and finally selects the top 500 peak clusters to infer PWMs. This selective inclusion of top-quality peaks or peak clusters from all available experiments should increase PWM quality. Third, at the final stage of selecting PWMs, we choose the one supported by the largest number of peaks as the top (primary) PWM, instead of choosing the PWM with the smallest

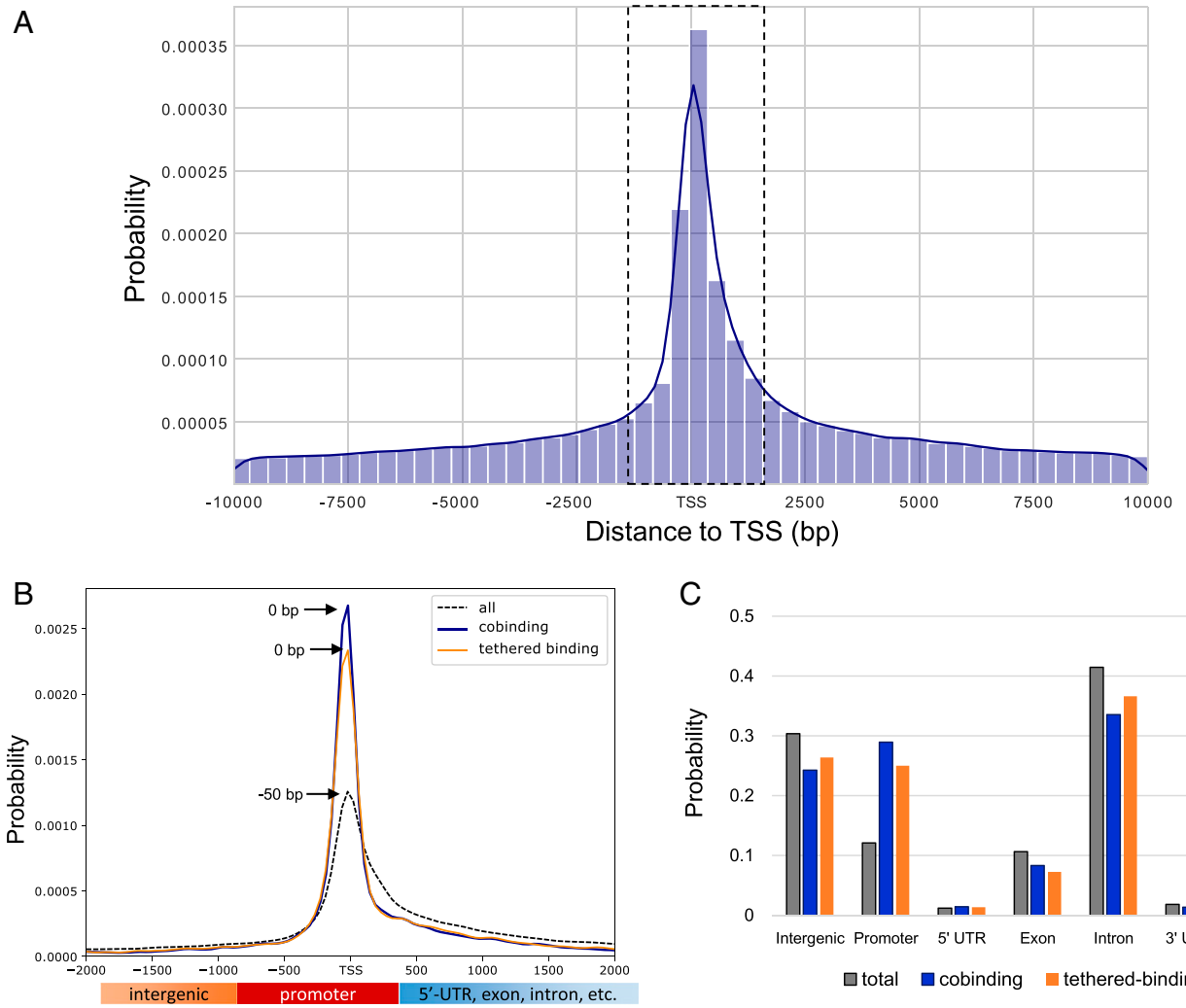

Fig. 3. Spatial distribution of human TFBSs. (*A*) Histogram showing the positional distribution of TFBSs using canonical PWMs inferred from all TFs that passed our criteria, within the region surrounding TSSs (±10 kb). The histogram is normalized and fitted by the Gaussian kernel density (line). The distribution in the dashed box is enlarged in *B*. (*B*) Positional distribution of mapped TFBSs within ±2 kb relative to TSSs. The peak (maximum) probabilities are indicated by arrows for all, cobinding, and tethered binding distributions, respectively. (*C*) The proportions of TFBSs occurring within eight different genomic features in the human genome. The proportions sum to 1.0 for each category of binding (color).

E-value, because a long candidate motif may have a low *E*-value even if supported by only a small number of peaks.

PWM Occurrences as the Criterion for Recovering Discarded Experiments. Reproducibility of peaks across replicates has been recommended as a criterion for judging the quality of an experiment (11, 12, 23). For this purpose the IDR has been developed, with IDR score >1.3 (IDR < 0.05) as a typical default threshold [i.e., $-\log_{10}(0.05) = 1.3$; https://github.com/IDR/idr-py]. However, in our ChIP-seq analysis this criterion resulted in unnecessarily discarding a large number of experiments. In contrast, we found that the number of peaks that support a PWM is a highly effective criterion for

Table 1. The distribution of mapped PWMs in intergenic and promoter regions

			Intergenic regions		Promoter (1 kb)		Promoter (500 bp)	
TF	TF family	No. of samples	No. of shared sites (%)	No. of unique sites (%)	No. of shared sites (%)	No. of unique sites (%)	No. of shared sites (%)	No. of unique sites (%)
SP1	C2H2	8	24,237 (51%)	23,016 (49%)	6,039 (70%)	2,554 (30%)	5,721 (72%)	2,207 (28%)
EGR1	C2H2	7	30,833 (42%)	42,035 (58%)	5,392 (66%)	2,763 (34%)	4,995 (68%)	2,310 (32%)
CTCF	C2H2	11	72,585 (71%)	29,823 (29%)	4,611 (71%)	1,909 (29%)	3,644 (72%)	1,428 (28%)
MAX	bHLH	20	32,713 (50%)	32,374 (50%)	3,416 (77%)	1,016 (23%)	2,614 (80%)	634 (20%)
MYC	bHLH	20	12,789 (56%)	10,099 (44%)	2,187 (79%)	598 (21%)	1,745 (81%)	413 (19%)
JUND	bZIP	17	56,444 (62%)	34,216 (38%)	1,767 (71%)	735 (29%)	1,164 (74%)	411 (26%)

The samples include cancer cell lines, primary cells, and tissues. Results are shown when defining the promoter of a gene as either the region 1 kb or 500 bp upstream of the TSS.

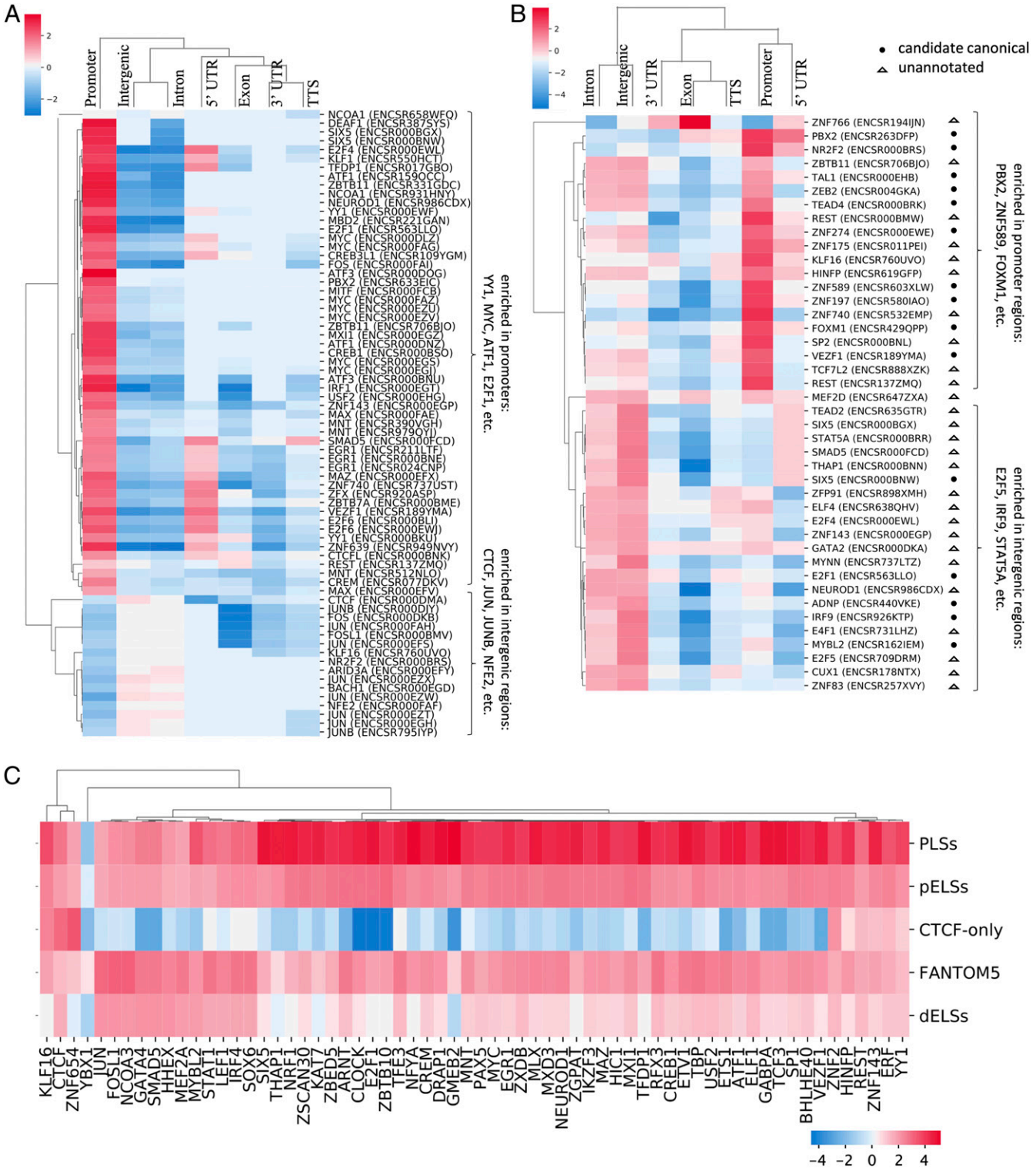

Fig. 4. Spatial distribution of TFBSs. (*A*) Fold enrichment of mapped TFBSs compared to DNase I sites in K562, shown for seven genomic features: promoter, intergenic, intron, exon, 5′ UTR (untranslated region), 3′ UTR, and presence near transcription termination sites (TTSs, from −100 bp to +1kb), annotated by HOMER. Fold change in comparison to DNase I hypersensitive sites is denoted by color, with red = positive and blue = negative. A positive fold change indicates that a TF has a higher proportion of its TFBSs mapping to the given genomic feature than the proportion of DNase I sites which map to the feature. The TF names and their accession numbers are provided, and some TFs whose TFBSs prefer promoters or intergenic regions are indicated on the right-hand side. (*B*) Heat map showing spatial preferences of TFBSs in K562 using unannotated motifs. (*C*) Spatial preferences of TFBSs in five enhancer-like regions, including four sets for ENCODE, i.e., PLSs, pELSs, dELSs, and CTCF-only elements, and one set of enhancers/enhancer-like regions from FANTOM5.

judging data quality and the reliability of an inferred PWM. We therefore proposed to use the number of PWM occurrences as a criterion to allow data recovery, when an experiment for a TF is discarded by a stringent IDR threshold. Specifically, in such cases, the IDR score was disregarded and the experiment was recovered if it yielded at least one PWM supported by >250 of the top 500 peaks (i.e., more than half of the top 500 peaks contain the inferred PWM).

To examine this criterion, let us consider the mouse data because the recovered PWMs can be compared to the corresponding human data. There are four recovered cases (*SI Appendix*, Fig. S6). First, for REST, a C2H2 TF, the data are initially discarded because the IDR score between the two replicates is 0. However, the top PWMs for the two replicates are supported by 417 and 387 peaks, respectively, and both PWMs show a high correlation with the human PWM (cos = 0.90 and 0.93). Moreover, merging the two replicates leads to a PWM that is supported by 418 peaks and has a correlation of 0.97 with the human PWM. This PWM is evidently credible. Second, for TCF3, a bHLH TF, the data are discarded because IDR score = 1.11 (<1.3). However, the top PWMs for the two replicates are supported by 495 and 499 peaks and show a good correlation with the orthologous human PWM (cos = 0.84 in both cases). Moreover, the top PWM for the merged data is supported by 500 peaks. Thus, the data indeed provide a reliable PWM. Third, for GATA2, a GATA TF, replicate 1 has only 12 peaks, which is obviously of poor quality. However, replicate 2 yields a PWM that is supported by 432 peaks and shows a correlation of 0.92 with the human PWM. Thus, even a single good-quality replicate can yield a reliable PWM. Fourth, for POU5F1, a homeodomain TF, the IDR score between the two replicates is 0. However, replicate 2 gives a PWM supported by 477 peaks. In this case, no ChIP-seq data are available for the orthologous human TF. However, the top PWM from replicate 2 has a good correlation with human NANOG (cos = 0.83), which is also a homeodomain TF. Thus, the PWM inferred from replicate 2 seems reliable. In addition to these four examples from the mouse, we have used this approach to recover 108 discarded human TF experiments. Clearly, the number of PWM occurrences is an effective criterion for judging the quality of an experiment and can be used to recover unnecessarily discarded data. This criterion also simplifies data analysis.

It is not clear why the IDR criterion sometimes does not perform well. This issue deserves further study. We suggest that the number of PWM occurrences is a better criterion for judging whether to merge replicates: In the absence of IDR support, merging is recommended only if it increases the number of peaks that support the top PWM. However, although the number of PWM occurrences alone seems to be sufficient to judge a replicate's quality, we still propose a two-stage approach for analyzing ChIP-seq data. PWMs inferred by a set of stringent rules would be more credible than relying on a single criterion. The second stage of analysis is needed only if the experiment is discarded and no PWM is obtained.

Interactions between TFs. Among the 336 human TFs for which we have obtained reliable PWMs, 201 showed cooccurring motifs, implying frequent interactions between TFs. As only 142 TF pairs showed cobinding while 186 pairs showed tethered binding, tethered binding occurs substantially more frequently than cobinding, as also observed by Wang et al. (9). It is interesting that 75 of the above 336 human TFs were found to be involved in both cobinding and tethered binding. Clearly, TF interaction is a complex phenomenon and its biological basis requires further exploration.

Methods

Data Collection. The primary ChIP-seq data (FASTQ files) for each target TF and its control(s) were downloaded from the ENCODE portal (8), including both single- and paired-end reads (accessed 5 October 2019). For humans, of the 1,639 TF genes cataloged by Lambert et al. (24), we identified 502 TFs having ENCODE data, representing 49 families and 1,621 released-status experiments (accessed 5 October 2019) (Dataset S9). The majority (61%) of human TFs were represented by a single experiment and a single control, each with two replicates. As CTCF was highly overrepresented in the ENCODE data (311 experiments), 129 experiments were somewhat randomly selected to represent a diversity of biosamples. For the mouse, we identified 54 TFs having ENCODE data, representing 14 TF families and 193 released-status experiments. To select controls, input DNA rather than immunoglobulin G (IgG) was used because 1) the IgG "mock" ChIP controls often had insufficient quantities of amplifiable DNA (22) and 2) the ENCODE ChIP-seq "blacklist" regions were defined using input DNA (https://www.nature.com/articles/s41598-019-45839-z) (25). Accession IDs and metadata are provided for human and mouse in Datasets S9 and S10, respectively.

Read Quality-Control and Mapping. Read quality was assessed using FASTQC before and after read processing. Read trimming was performed using Trimmomatic (v0.39) (26) using the ILLUMINACLIP, LEADING:10, SLIDINGWINDOW:4:15, and MINLEN options. For MINLEN, 50 bp was used for read lengths >50 bp, 30 bp for read types ≤50 bp and >30 bp, and 25 bp for read types ≤30 bp. Trimmed reads not satisfying these length requirements were discarded. The remaining reads were aligned to the latest version of the human genome (GRCh38.p13) or mouse genome (GRCm38) using bowtie2 (v2.3.5) (27). A mappable rate of >70% was required for both replicates, and this criterion was met by 95% of experiments.

Peak Calling. For each experiment, peaks in read depth were determined using the callpeak function of MACS2 (v2.1.2) (28). Specifically, peaks were identified by comparing mapped reads from an experiment (MACS2 parameters: -t replicate1 replicate2) to its control (-c replicate1 replicate2). Peak summits, given by MACS2 summits.bed at 1-bp resolution, were called with q-value <0.05 and extended by 100 bp in both directions (total length 200 bp). Finally, sites overlapping blacklisted regions (including low-complexity repetitive regions) (25) were removed.

Motif Discovery and Merging Replicates of an Experiment. Motif discovery was performed using MEME-ChIP (v5.0.5) (29) to infer the top five PWMs from the top 500 peaks (200 bp each, ±100 bp from the peak summit), similar to the procedure of Wang et al. (9). An experiment was retained only if it had two or more replicates passing the following two quality-control criteria. First, each replicate was required to yield at least one PWM with occurrences in >100 peaks. Second, the top 500 peaks from each pair of replicates were required to have an IDR score, $-\log_{10}$(IDR), of ≥1.3 (11). Those experiments that failed to meet both criteria were discarded. However, as described below, if an experiment for a TF was discarded we allowed the data to be recovered for analysis if they met an alternative set of criteria.

For TFs with a single experiment passing the above criteria, replicates were merged, peak calling was repeated, and motif discovery was performed a second time using the merged-replicate peaks. In such cases, the PWM with the largest number of occurrences in called peaks was considered the top (primary) PWM. We did not rank motifs by *E*-value, as done in Wang et al. (9) and other studies, because we found that dubious (e.g., low-complexity) motifs could attain significant *E*-values due to length alone. For TFs with more than one experiment passing the above criteria, a procedure for ranking and selecting read peaks was used to derive a single top PWM for that TF (discussed in the next section).

Ranking and Selecting Read Peaks from Multiple Experiments. For those TFs with more than one experiment passing our two quality-control criteria, we obtained PWMs for each TF using a procedure modified from Satpathy et al. (30). We first describe the procedure for the case of a single biosample (e.g., a cell line) with multiple experiments (*SI Appendix*, Fig. S7).

First, an experiment using the biosample was retained only if it passed the above two quality-control criteria.

Second, for each retained experiment, the MACS2 scores ($s = -\log_{10}$[q-value]) of peaks were converted into score percentiles as rank(s)/total_no_peak, where rank(s) refers to sorted MACS2 scores in descending order and total_no_peak is the total number of called peaks. The highest score has the first rank and is assigned a value of 100% (i.e., 1.00) under this scheme, while the lowest score is constrained to a value of (100/total_no_peak)%.

Third, all peaks from all retained experiments were pooled and grouped using the bedtools cluster tool (31), so that overlapping (≥1 bp at termini) peak regions were merged into a single cluster. Each merged peak cluster was assigned a score equal to the sum of the score percentiles of all peaks

falling within that cluster, i.e., the rank-based scores from all experiments with a peak in the region.

Fourth, the summation scores of peak clusters were sorted in descending order to select the top 500 clusters. Within each cluster, we selected the peak with the highest score percentile as the representative peak, and the representative peaks from all 500 peak regions were used to infer the PWMs of the TF.

Fifth and last, we chose the PWM supported by the largest number of the representative peaks as the top PWM for the TF and also reported all other (at most four additional) PWMs with >100 peaks.

We now consider multiple biosamples.

First, for each biosample we analyzed the experiments following the first to third steps above to obtain the peak clusters and also the summation score of each peak cluster.

Second, the peak clusters from all biosamples were merged and ranked following the fourth step above to select new top 500 peak clusters to infer the PWMs of the TF.

Obtaining PWMs for C2H2 TFs Using RCADE. For each C2H2 TF experiment with ≥500 peaks, the sequences of the peaks and the protein sequence of the TF were uploaded to the RCADE webserver (http://rcade.ccbr.utoronto.ca/). Then, the computed PWMs were downloaded for comparison with those inferred using our pipeline.

Recovering Discarded Data. If an experiment for a TF was discarded by our selection criteria, we recovered data meeting the following alternative criteria:

1) Recover each experiment that has at least one replicate with at least one PWM supported by >250 of the top 500 peaks.
2) For a recovered experiment, merge the replicates irrespective of the IDR score, if merging increases the number of peaks supporting the top PWM. Otherwise, select only the replicate that has the PWM supported by the largest number of peaks.
3) Analyze each recovered experiment separately to infer PWMs.
4) Select the PWM supported by the largest number of peaks among all recovered experiments as the top PWM of the TF under study.

Identifying Canonical and Cooccurring Motifs. The canonical motif of a TF refers to the sequence-specific DNA motif that is directly bound by the TF. In this study, a cooccurring motif refers to a situation in which a motif found in the ChIP-seq data of one TF (TF1) is similar (cos ≥ 0.80) to the canonical motif of a second TF (TF2) that belongs to another TF family. We propose the following procedure to infer canonical and cooccurring motifs:

1) For each TF, determine groups of similar PWMs. Specifically, place all passing PWMs inferred by the ChIP-seq experiments for the TF into groups, where a PWM is added to a group if it is similar (cos ≥ 0.8) to at least one PWM already present in the group. Repeat until no new matches are found. Note that a PWM group may contain only one PWM.
2) If the TF has only one PWM group:
   a) if any member of the PWM group is similar to a known canonical motif of the same TF family, the group is considered to be a canonical motif of the TF. (If no similar known canonical motif is found [i.e., cos < 0.80], the condition is relaxed to "containing a submotif that is similar to a submotif of a known canonical motif.") The top PWM of the group is considered the candidate canonical motif.
   b) if no member of the PWM group is similar to any known canonical motif documented in any database or literature, the group likely represents a novel canonical motif of this or another TF. In our database such motifs are indicated as "candidate canonical."
   c) if an inferred PWM is similar to a canonical PWM of a TF in another TF family, it is considered a cooccurring motif.
3) If a TF has more than one PWM group, use the procedure in 2 to check if any group is canonical and/or contains cooccurring PWMs.

Reason suggests that the top PWM inferred from a TF ChIP-seq experiment is usually its canonical motif. However, in some cases the top PWM may be the canonical motif of another TF that belongs to a different TF family. Such a situation occurs when the TF under study binds to another TF rather than to DNA, i.e., tethered binding (see below). In general, one TF has one canonical motif. However, in some cases, a TF (e.g., CTCF) may have more than one canonical motif. This could be due to the existence of multiple independent motifs for the TF, or to a single large motif that is inferred as two or more PWMs during motif discovery.

Cobinding vs. Tethered Binding. As mentioned earlier, the cooccurrence of two motifs can be classified into 1) cobinding and 2) tethered binding (9). Wang et al. (9) have described rules for distinguishing between the two types of binding. Here we give the criteria in mathematical expressions. Let X = fraction of peaks containing the canonical motif only, Y = fraction of peaks containing only a noncanonical motif, and Z = fraction of peaks containing both motifs. We propose the following rules for identifying cobinding and tethered binding, where TF1 is the TF being assayed and TF2 belongs to a separate family: Cobinding is indicated when the primary motif of TF1 is equally or more frequent than that of TF2, due to each binding separate DNA motifs, and tethered binding is indicated when the primary motif of TF1 is inferred to be the primary motif of TF2, due to TF1 binding TF2 rather than a DNA motif. More specifically, the rules are as follows:

1) If Y > X and Y > Z, infer tethered binding. The condition Y > X requires that the noncanonical motif is the top (primary) motif, due to the TF binding a second TF rather than a DNA motif (9). The condition Y > Z requires that the noncanonical motif appears alone more often than together with the canonical motif.
2) If Y < X or Y < Z, infer cobinding. The condition Y < X requires that the canonical motif is the top motif and often occurs alone, while the condition Y < Z requires that the noncanonical motif appears more often with the canonical motif than alone.

Before computing X, Y, and Z, we used FIMO to check whether a motif under study appears (hit with $P = 1 \times 10^{-4}$) in a peak region. However, as MEME-ChIP may include motif sites with a P value higher than the default value for FIMO, we increased the P value to the largest P value in the MEME-ChIP data if we found no presence of the motif under study. To reduce the chance of including randomly occurring motifs, the upper bound for the relaxed P value was set to 5×10^{-4}. As motif cooccurrence can differ by biosample (cell line or tissue), we analyzed each biosample separately to detect cooccurring motifs.

Inferring Core Motifs. As described above, we categorized TF families into groups based on similar motifs. For each group with three or more members, we inferred the core motif as follows. First, the motifs within each group were aligned. Second, the average frequency of each of the four nucleotides was computed at each position to obtain the consensus motif (PWM) for the group. Third, the consensus motif was end-trimmed if the terminal position's information content (IC) was <0.3 bits. This was repeated until the positions at both ends had IC ≥0.3 bits. Each group was then regarded as a subfamily, and each trimmed final motif was regarded as its core motif.

Spatial Distribution of TFBSs. To study the distribution of TFBSs in the genome, we first mapped the inferred PWMs for each TF onto the genome as follows. For a TF with a single experiment, we selected the canonical PWM of the TF and mapped it to all the peak regions using FIMO. For a TF with multiple experiments, we used the ranking method to infer PWMs and selected one canonical PWM as described above, and then mapped it to all the peak regions. If a PWM sequence was detected in a peak region ($P < 0.0001$), the evolutionary conservation of the putative TFBS was assessed among primates. We calculated an average nucleotide conservation score using PhastCons30, which included 27 primates and 3 mammals. We required the score to be >50%, i.e., the PWM sequence was found in at least 14 species. For a TF with multiple experiments in one or more biosamples, the peak clusters identified by the ranking method were used instead. The distance of a putative TFBS (center position) with respect to the closest TSS was calculated using bedtools and the genomic regions of the TFBSs were annotated by annotatePeaks.pl in HOMER (32).

TFBSs separated by ≤100 bp were clustered using bedtools (cluster -d 100). The density of TFBSs in a cluster per kilobase was calculated by (N/L)*1,000, where N is the number of TFBSs in a cluster and L is the length of the cluster.

TFBSs of TFs Preferring Enhancer Regions. To infer the biological role of a TF that prefers intergenic regions, we compared the putative TFBSs of a TF with enhancers/enhancer-like regions from ENCODE and FANTOM5. For the ENCODE data, we downloaded the annotated regions from the database of The Registry of Candidate cis-Regulatory Elements (18). The annotated regions were divided into four categories: PLSs, pELSs, dELSs, and CTCF-only elements. The other enhancer data were downloaded from FANTOM5 (19, 20). The two databases included enhancer-like data for human (hg38) and mouse (mm10). Each set of the enhancer regions was compared with the set of TFBSs of a TF to estimate the ratio of the observed number of overlapping

peaks to the expected number for nonenhancer sites and *P* values using mergePeak.pl in HOMER (32).

Motif Comparisons. To assess the similarity between two PWMs, the method of k-mer frequency vector was used (33). Based on the authors' suggestion, the optimal k-mer length was set to 4 and the similarity measurement was chosen to be the cosine angle. The PCC and cosine angle (cos) metrics were both used to assess the similarity of two PWMs by Xu and Su (33). However, PCC transforms observed values by centering the mean value, which may result in negative values, while cos takes the absolute observed value without any shift. The study found that cos outperformed PCC and also two other metrics, Euclidean and Kullback–Leibler distances. Thus, we adopted cos as the metric for comparing PWMs in this study. The PWM databases compared in this study include Factorbook (34), SELEX in Cis-BP Build 2.00, and JASPAR Core eighth version 2020.

Functional Motif Inference. The PWMs were mapped to ChIP-seq peak regions using FIMO with $P < 1 \times 10^{-4}$. For analyzing promoter enrichment, the locations of PWM appearances were annotated by HOMER and the PWMs with $\log_2$ ratio >2 in promoter regions (−1 kb to +100 bp related to TSS) were classified as "promoter-enriched" PWMs.

In addition, we calculated the evolutionary conservation as described in TFBS prediction and compared with the average score in intergenic regions (score = 0.10). Motifs with odds ratios >2 were classified as enriched.

Data and Computational Pipeline Availability. All data and scripts used to generate results are freely available on GitHub at https://github.com/chpngyu/chip-seq-pipeline and https://as0201821.github.io/dbTFBS/. The commands used for data download are documented in the supplementary script at https://github.com/chpngyu/chip-seq-pipeline. The commands for quality-control, read trimming, read mapping, and peak calling are documented in https://github.com/chpngyu/chip-seq-pipeline. The commands for calculating IDR scores between replicates and cos values between PWMs are documented in https://github.com/chpngyu/chip-seq-pipeline. All other study data are included in the article and/or supporting information.

ACKNOWLEDGMENTS. We thank Frank Hsu, Federico Giorgi, Naoki Osato, Jeff Vierstra, and Meng Wang for helpful comments. This study was supported by Academia Sinica (AS-SUMMIT-109) and by Ministry of Science and Technology Taiwan (MOST 107-2311-B-001-016-MY3).

1. D. S. Johnson, A. Mortazavi, R. M. Myers, B. Wold, Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
2. G. Robertson *et al*., Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657 (2007).
3. M. F. Berger *et al*., Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* **24**, 1429–1435 (2006).
4. A. R. Oliphant, C. J. Brandl, K. Struhl, 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).
5. R. C. O'Malley *et al*., Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell* **165**, 1280–1292 (2016).
6. E. C. Partridge *et al*., Occupancy maps of 208 chromatin-associated proteins in one human cell type. *Nature* **583**, 720–728 (2020).
7. D. Mercatelli, L. Scalambra, L. Triboli, F. Ray, F. M. Giorgi, Gene regulatory network inference resources: A practical overview. *Biochim. Biophys. Acta. Gene Regul. Mech.* **1863**, 194430 (2020).
8. C. A. Davis *et al*., The encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
9. J. Wang *et al*., Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **22**, 1798–1812 (2012).
10. P. Kheradpour, M. Kellis, Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* **42**, 2976–2987 (2014).
11. Q. H. Li, J. B. Brown, H. Y. Huang, P. J. Bickel, Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, 1752–1779 (2011).
12. Y. Chen *et al*., Systematic evaluation of factors influencing ChIP-seq fidelity. *Nat. Methods* **9**, 609–614 (2012).
13. H. S. Najafabadi, M. Albu, T. R. Hughes, Identification of C2H2-ZF binding preferences from ChIP-seq data using RCADE. *Bioinformatics* **31**, 2879–2881 (2015).
14. J. Vierstra *et al*., Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020).
15. R. Oughtred *et al*., The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* **30**, 187–200 (2021).
16. J. E. Phillips, V. G. Corces, CTCF: Master weaver of the genome. *Cell* **137**, 1194–1211 (2009).
17. A. P. Boyle *et al*., High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
18. J. E. Moore *et al*.; ENCODE Project Consortium, Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
19. R. Andersson *et al*., An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
20. J. A. Ramilowski *et al*., Functional annotation of human long noncoding RNAs via molecular phenotyping. *Genome Res.* **30**, 1060–1072 (2020).
21. A. F. Bardet, Q. He, J. Zeitlinger, A. Stark, A computational pipeline for comparative ChIP-seq analyses. *Nat. Protoc.* **7**, 45–61 (2011).
22. S. G. Landt *et al*., ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
23. R. Nakato, K. Shirahige, Recent advances in ChIP-seq analysis: From quality management to whole-genome annotation. *Brief. Bioinform.* **18**, 279–290 (2017).
24. S. A. Lambert *et al*., The human transcription factors. *Cell* **175**, 598–599 (2018).
25. H. M. Amemiya, A. Kundaje, A. P. Boyle, The ENCODE blacklist: Identification of problematic regions of the genome. *Sci. Rep.* **9**, 9354 (2019).
26. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
27. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
28. Y. Zhang *et al*., Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9**, R137 (2008).
29. P. Machanick, T. L. Bailey, MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011).
30. A. T. Satpathy *et al*., Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).
31. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
32. S. Heinz *et al*., Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
33. M. Xu, Z. Su, A novel alignment-free method for comparing transcription factor binding site motifs. *PLoS One* **5**, e8797 (2010).
34. J. Wang *et al*., Factorbook.org: A wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res.* **41**, D171–D176 (2013).
35. P. Shannon *et al*., Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).