Extensive binding of uncharacterized human transcription factors to genomic dark matter

Rozita Razavi[1*], Ali Fathi[1*], Isaac Yellan[1*], Alexander Brechalov[1*], Kaitlin U. Laverty[1,2], Arttu Jolma[1], Aldo Hernandez-Corchado[3], Hong Zheng[1], Ally W.H. Yang[1], Mihai Albu[1], Marjan Barazandeh[1], Chun Hu[1], Ilya E. Vorontsov[4], Zain M. Patel[1], The Codebook Consortium, Ivan V. Kulakovskiy[5], Philipp Bucher[6], Quaid Morris[2], Hamed S. Najafabadi[3,7], and Timothy R. Hughes[1**]

[1]Donnelly Centre and Department of Molecular Genetics, 160 College Street, Toronto, ON M5S 3E1, Canada
[2]Memorial Sloan Kettering Cancer Center, Rockefeller Research Laboratories, New York, NY 10065, USA
[3]Victor P. Dahdaleh Institute of Genomic Medicine, 740 Dr. Penfield Avenue, Room 7202, Montréal, Québec, H3A 0G1, Canada
[4]Vavilov Institute of General Genetics, Russian Academy of Sciences, 119991, Moscow, Russia
[5]Institute of Protein Research, Russian Academy of Sciences, 142290, Pushchino, Russia
[6]Swiss Institute of Bioinformatics, 1015, Lausanne, Switzerland
[7]Department of Human Genetics, McGill University, Montréal, Québec, H3A 0C7, Canada

*These authors contributed equally
** To whom correspondence should be addressed: t.hughes@utoronto.ca

The Codebook Consortium

Principal investigators (steering committee)

Philipp Bucher, Bart Deplancke, Oriol Fornes, Jan Grau, Ivo Grosse, Timothy R. Hughes, Arttu Jolma, Fedor A. Kolpakov, Ivan V. Kulakovskiy, Vsevolod J. Makeev

Analysis Centers:

University of Toronto (Data production and analysis): Mihai Albu, Marjan Barazandeh, Alexander Brechalov, Zhenfeng Deng, Ali Fathi, Arttu Jolma, Chun Hu, Timothy R. Hughes, Samuel A. Lambert, Kaitlin U. Laverty, Zain M. Patel, Sara E. Pour, Rozita Razavi, Mikhail Salnikov, Ally W.H. Yang, Isaac Yellan, Hong Zheng
Institute of Protein Research (Data analysis): Ivan V. Kulakovskiy, Georgy Meshcheryakov
EPFL, École polytechnique fédérale de Lausanne (Data production and analysis): Giovanna Ambrosini, Bart Deplancke, Antoni J. Gralak, Sachi Inukai, Judith F. Kribelbauer-Swietek
Martin Luther University Halle-Wittenberg (Data analysis): Jan Grau, Ivo Grosse, Marie-Luise Plescher
Sirius University of Science and Technology (Data analysis): Semyon Kolmykov, Fedor Kolpakov
Biosoft.Ru (Data analysis): Ivan Yevshin
Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University (Data analysis): Nikita Gryzunov, Ivan Kozin, Mikhail Nikonov, Vladimir Nozdrin, Arsenii Zinkevich
Institute of Organic Chemistry and Biochemistry (Data analysis): Katerina Faltejskova
Max Planck Institute of Biochemistry (Data analysis): Pavel Kravchenko
Swiss Institute for Bioinformatics (Data analysis): Philipp Bucher
University of British Columbia (Data analysis): Oriol Fornes
Vavilov Institute of General Genetics (Data analysis): Sergey Abramov, Alexandr Boytsov, Vasilii Kamenets, Vsevolod J. Makeev, Dmitry Penzar, Anton Vlasov, Ilya E. Vorontsov
McGill University (Data analysis): Aldo Hernandez-Corchado, Hamed S. Najafabadi
Memorial Sloan Kettering (Data production and analysis): Kaitlin U. Laverty, Quaid Morris
Cincinnati Children's Hospital (Data analysis): Xiaoting Chen, Matthew T. Weirauch

SUMMARY

Most of the human genome is thought to be non-functional, and includes large segments often referred to as "dark matter" DNA. The genome also encodes hundreds of putative and poorly characterized transcription factors (TFs). We determined genomic binding locations of 166 uncharacterized human TFs in living cells. Nearly half of them associated strongly with known regulatory regions such as promoters and enhancers, often at conserved motif matches and co-localizing with each other. Surprisingly, the other half often associated with genomic dark matter, at largely unique sites, via intrinsic sequence recognition. Dozens of these, which we term "Dark TFs", mainly bind within regions of closed chromatin. Dark TF binding sites are enriched for transposable elements, and are rarely under purifying selection. Some Dark TFs are KZNFs, which contain the repressive KRAB domain, but many are not: the Dark TFs also include known or potential pioneer TFs. Compiled literature information supports that the Dark TFs exert diverse functions ranging from early development to tumor suppression. Thus, our results sheds light on a large fraction of previously uncharacterized human TFs and their unappreciated activities within the dark matter genome.

KEYWORDS: Transcription factor (TF), ChIP-seq, SELEX, GHT-SELEX, PWM, Gene regulation, KRAB zinc finger protein, C2H2, Codebook

INTRODUCTION

Deciphering *cis* and *trans* gene regulation is a long-standing challenge in molecular biology and computational genomics. Transcription factors (TFs) are the sequence-specific DNA binding proteins that regulate gene expression, typically by associating with promoters and/or enhancers[1,2]. The human genome encodes over 1,600 apparent TFs, but hundreds of them have been identified as such only on the basis of conserved protein domain structures, and are otherwise poorly characterized, with no known DNA binding motif[3]. The function of much of the conserved (and presumably functional) noncoding DNA in human is also largely unknown, although at least some of it is involved gene regulation[4]. Indeed, phylogenetic footprinting, which identifies islands of conserved DNA sequence, is a long-established approach to identify functional regulatory elements[5,6], despite being limited by the frequent turnover of TF binding sites[7].

The functionality of most of the human genome is similarly ambiguous and uncertain, as most of it is both non-genic and unconserved[4,8]. Roughly half is composed of transposable elements (TEs), especially endogenous retroelements (EREs)[9]. Broadly, the human genome can be divided into gene-rich regions and gene deserts[10]; some of the latter contain transcriptional enhancers[11], while others appear to be dispensable[12]. Genome-wide chromatin contact maps also reveal two major genomic compartments that partially mirror the dichotomy in gene density: the “A” compartment is broadly associated with the presence of genes and accessible chromatin, while the “B” compartment is enriched for LINE-1 (L1) elements and constitutive heterochromatin[13]. The functions of the compartments and the mechanisms that create them are not well understood i.e. it is not clear whether they are defined by and/or recruit specific TFs, and how often (if at all) they recruit factors that contribute to regulation and/or structure of the genome.

It is known, however, that many human TFs bind to specific classes of TEs. TEs can rewire regulatory circuitry by introducing transcription factor binding sites, thus spawning novel cis-regulatory elements[14-16]. In some cases, these elements derive from the promoter of the TE (e.g. endogenous retrovirus long terminal repeats (ERV LTRs))[17,18]. Other cases may represent inadvertent matches to a host TF motif within a TE[19]. In addition, mammalian genomes encode a large family of KRAB-containing C2H2 zinc-finger (KZNF) TFs (~350 members in human), which evolves rapidly in parallel to ERE classes bound by its members[20]. KZNFs silence EREs via direct recruitment of KAP1/TRIM28, which associates physically with both readers (e.g. HP1/CBX proteins) and writers (SETDB1)[21,22] of the H3K9me3 mark that defines constitutive heterochromatin[23,22]. The KZNFs are also known for the potential to have very long binding sites, enabled by their long C2H2-zf domain arrays[24]. In general, the DNA binding preferences of C2H2-zf proteins, KRAB or otherwise, have proven difficult to characterize precisely, due to a lack of antibodies for ChIP-seq and low apparent functionality in biochemical assays; such that they are depleted from systematic studies of human TF motifs (e.g.[25]). The largest collections of binding data for C2H2-zf proteins have come from studies using ChIP-seq with epitope-tagged proteins in cultured cells[26,27]. These data reveal what families of EREs are bound by C2H2-zf proteins, but it

is difficult to accurately determine their precise sequence specificity (and thus the exact binding sites) because the repeat elements are related by common descent, which confounds motif discovery[28]. The lack of accurate knowledge of DNA sequence specificity of TFs complicates interpretation of ChIP-seq data, in general, because ChIP-seq readily detects indirect recruitment and nonspecific binding[29,30].

Regulatory DNA represents an expanding frontier in genetics, and it is critical that we gain a complete picture of human TF-DNA binding. As part of an international initiative termed the "Codebook consortium", aimed at obtaining binding motifs for all human TFs[31], we analyzed 315 uncharacterized human TFs by ChIP-seq in HEK293 cells, together with 58 controls. We evaluated the data in conjunction with other data from the Codebook project, which allowed base-level identification of direct binding sites. Previous ChIP-seq analyses have focused mainly on preferential association of TFs with promoters vs. enhancers[1], and indeed, many of the 217 TFs yielding reliable data in our study (i.e. reproducible and/or enriched for TF motif matches) bound predominantly and directly to such sites, consistent with conventional roles in gene regulation, and providing a likely explanation for the fact that their binding sites are frequently conserved. Surprisingly, however, roughly half of the uncharacterized Codebook TFs, including most KZNFs as well as other TF families, bound to apparently unique sites that are located in regions depleted from activating epigenetic marks. A subset of these TFs also bind mainly to closed chromatin; we refer to this subset as "Dark TFs". Multiple lines of evidence suggest diverse biochemical, cellular, and physiological functions of the Dark TFs, and by extension, the dark matter genome.

RESULTS

Generation of ChIP-seq data for hundreds of putative TFs

We surveyed the genomic binding sites of 314 poorly characterized, putative human TFs (the "Codebook" set, derived from Lambert 2018[32], and described in detail elsewhere[31]), and 58 previously characterized TFs as controls (selected from Isakova 2017[33] and Schmitges 2016[27]), using ChIP-seq in HEK293 cells (**Figure 1A**, **Table S1**). We used an inducible eGFP-tagged transgene system (**Figure 1B**) previously employed for ChIP-seq and to identify protein-protein interactions[27,34,35]. Using this system, we have shown that KZNFs bind to specific classes of retroelements, and that their binding sites are often depleted for open chromatin, indicating that the transgene system can readily assess binding to inactive or repressed regions of the genome[27,34]. The present study includes biological replicates performed by different experimentalists, such that the resulting dataset includes 678 ChIP-seq experiments for Codebook TFs, and 112 experiments for control TFs (**Table S1**). A full list of experiments is given in **Table S2**. Representative motifs obtained from control TFs of various families illustrate that the assay recovers known sequence-binding preferences, as expected (**Figure 1C**).

We used two criteria to determine which experiments were successful. First, the data were analyzed as part of a larger Codebook benchmarking effort, which is described in more detail in accompanying manuscripts[31,36]. The Codebook benchmarking included expert curation that relied mainly on obtaining similar motifs for the same TF from

different data types (ChIP-seq, Protein Binding Microarrays[37], SMiLE-seq[33], and several variants of HT-SELEX[38]) as evidence of direct, sequence-specific DNA binding. This Codebook motif benchmarking identified 130 Codebook TFs and 49 controls with "approved" ChIP-seq data, meaning that sequence-specific DNA binding is observed in ChIP-seq, and it is supported in almost all cases by *in vitro* experimental data.

Second, we identified experiments in which the peak overlaps of biological replicates exceeded what is expected at random (i.e. with TF identities permuted). ChIP-seq experiments that were classified as "approved" based on the motif similarity analysis described above displayed a higher overlap between TF replicates relative to mismatched TFs (median Kulczynski II coefficient of 0.57 vs. 0.034; **Figure S1A**; this statistic is a modified Jaccard value that compensates for class imbalance). A Kulczynski II coefficient threshold of 0.4 captures 78% of approved experiments with replicates, and 90% of controls, but eliminates 94.5% of mismatched experiments. Among proteins for which there was no "approved" experiment but for which there were biological ChIP-seq replicates, 36 putative TFs (and two controls) displayed peak overlaps between replicates that exceeded a Kulczynski II coefficient of 0.4. In these cases, the DNA binding may be indirect, i.e. these proteins may be DNA-associated chromatin factors, rather than TFs. Alternatively, they may recognize properties of the DNA sequence that are not captured by common motif models, or the constructs used may be inactive for direct DNA binding in HEK293 cells, but competent for association with chromatin. We included these 38 putative TFs in subsequent analyses, and we refer to the entire set of 217 successful proteins (130+49 "approved", and 36+2 with matching replicates; **Table S3**) as "TFs", for simplicity, although we caution that the subsets that are not "approved" may instead be chromatin factors.

We merged the peaks from TF replicates among the 489 ChIP-seq experiments deemed successful to produce a dataset for downstream analysis (**Table S3**). This dataset encompasses 217 proteins (166 Codebook putative TFs, and 51 controls), with a median of 12,681 peaks per protein (range 76-163,602) (using a MACS threshold of $P<10^{-10}$; see **Methods** for explanation of threshold choice)[39].

Overview of ChIP-seq data illustrates that half of the Codebook TFs bind genomic dark matter

To begin characterizing the ChIP-seq data, we surveyed for preferential association of the putative TFs with promoters and/or enhancers. **Figure 2A** shows the fraction of peaks from each protein that overlaps with protein-coding promoters (defined by RefSeq[40]) and enhancers (defined by HEK293 chromatin state[41], and corresponding mainly to H3K4me1 signal; see **Methods**). Indeed, many proteins are highly associated with promoters, and a smaller number with enhancers, although a large number of proteins did not associate with either promoters or enhancers. On average, the HEK293-derived enhancer set yielded higher overlaps than the larger, more universal "GeneHancer" set[42] (**Figure S2A**), indicating that this lower number of enhancer-favouring TFs (relative to promoter-favouring) is not due to incomplete enhancer annotations.

To gain further insight into the properties of the ChIP-seq binding sites, we compared the peak sets for each of the proteins to those of all other proteins in the dataset, and to a panel of genome annotations. **Figure 2B** (bottom) shows a symmetric heatmap of Jaccard similarity (intersection/overlap) between all 217 ChIP-seq datasets, providing an overview of the overlap between all pairs of TF peak sets. The heatmap at the top of **Figure 2B** shows the fraction of each corresponding TF peak set that overlaps with each type of genome annotation. The chromatin states were derived mainly from public-domain data for unperturbed HEK293 cells; we therefore expect them to reflect the state of the chromosomes prior to induction of the tagged TF transgene. This state could be involved in recruiting the TF, but it could also result from endogenous expression of the native, untagged protein, as most of the studied TFs are already expressed in HEK293 cells[27] (**Figure S3**).

Figure 2B reflects and expands upon trends observed in **Figure 2A**. The large bright square in the lower right quadrant of the bottom heatmap corresponds to TFs that associate primarily with open chromatin (ATAC-seq) and H3K4me3. These TFs also often associate with many promoters (median 6,140 coding gene promoters; see below), leading to high overlap between the peak sets. The observation that many TFs co-bind promoters and/or enhancers is prevalent in the literature (e.g.[1]); we note, however, that the Codebook proteins were considered uncharacterized TFs at the outset of this study, and therefore it appears that even well-known regulatory sequences often contain previously unidentified TF binding sites.

A second main feature of **Figure 2B** is the diagonal line in the upper left quadrant of the heatmap at the bottom. These are proteins for which there is very little overlap in peaks with any other Codebook TF. In addition, these peaks often do not overlap with any peak from any other protein in the Codebook ChIP-seq dataset (**Figure S1B**). The unique binding profiles are not explained by experimental error or random events: there is strong overlap between replicates of the same protein (**Figure S1C**), and these proteins often bind to the same unique sequences in ChIP-seq and *in vitro* (see below). The ChIP-seq peaks for these proteins also tend to be outside open chromatin, and outside of either promoters or apparent enhancers in HEK293 cells (**Figure 2B,** *top*). Instead, roughly half of these TFs' peak sets are either enriched for marks that characterize heterochromatin, or lack any of the diagnostic marks of promoter or enhancer activity (i.e. "empty" ChromHMM regions, **Figure S2B**). Peaks for a subset of these TFs are mainly associated with the Hi-C "B" compartment, and many associate with specific classes of repeats (**Figure 2B** and see below). Roughly half (50/94) of these TFs are KZNFs, which would be expected to display these properties[34] (bar in middle of **Figure 2B**).

In subsequent analyses, we sought to gain a better understanding of the properties and characteristics of TFs (and their binding sites) that represent the major patterns shown in **Figure 2**. For simplicity, we defined four mutually exclusive groups (see **Table S4** for TF labels). One group we named "Promoter binders" (55 proteins); for these proteins, more than 37% of peaks overlap with promoters (this threshold captures the most prominent features in **Figure 2A**)**.** Another group was designated "Enhancer binders" (9 proteins); for these proteins, >35% of peaks overlap with enhancers (this threshold

corresponds to the visual separation of data points on the vertical axis in **Figure 2A**). A third group we named “Dark TFs”, after the genomic dark matter (54 proteins); for these, most peaks lie within either the “empty”, “constitutive heterochromatin”, or “facultative heterochromatin” states (i.e. in HEK293 cells, they are outside of the states that represent promoters, enhancers, insulators, or gene bodies), and fewer than half of the peaks overlap with ATAC-seq peaks in unperturbed HEK293 cells. These thresholds exclude some TFs that may associate significantly with specific regions, but also bind many other locations; for example, the control TF YY1 bound 53% of all promoters in human, but it also had 32,046 additional binding sites outside promoters, which represent 77% of all YY1 peaks. Similarly, a subset of KZNFs were not classified as Dark TFs because they had many binding sites within open chromatin. The remaining ~40% of TFs we labeled as “Other” (i.e. not Promoter, Enhancer, or Dark TFs); they include 32 control TFs and 64 uncharacterized proteins which display a diversity of attributes and patterns in the data. The “Other” TFs thus present a rich landscape for further exploration, but we did not attempt to further subclassify them here.

Direct DNA binding by Codebook TFs to specific types of genomic elements

We next asked whether preferential association of the TFs with different types of genomic regions and chromatin states could be accounted for by intrinsic sequence recognition of the individual TF analyzed. We mainly compared the Promoter TFs and Dark TFs, which are large groups that contrast in many ways. ChIP-seq can detect both direct binding (i.e. the TF intrinsically recognizes the bound DNA sequences) and indirect binding (e.g. recruitment by another factor)[29]. ChIP-seq also readily detects non-specific DNA-binding (e.g. by histones), and is biased towards open chromatin since the sonication step preferentially releases these regions[30]. Therefore, to accurately identify direct binding sites in the ChIP-seq data, we used two independent sources of information that were available as part of the larger Codebook initiative. First, we employed data from a novel assay, GHT-SELEX (Genomic HT-SELEX; described in detail in [43]), which surveys binding of synthetic TFs to fragmented, purified, and unmodified genomic DNA *in vitro*; GHT-SELEX yields peaks that resemble those from ChIP-seq, but with greater resolution due to the smaller DNA fragment lengths (~64 bases). Second, for each TF, we obtained genomic matches to its DNA binding motif, modeled as Position Weight Matrices (PWMs) with an associated PWM score. PWM derivation and benchmarking are described in more detail in accompanying manuscripts[31,36].

To make a conservative assessment of direct binding, we considered a ChIP-seq peak to be bound directly by a TF if the peak overlapped with a GHT-SELEX peak for the same TF, and also contained at least one motif match. In addition, the significance thresholds for all three (ChIP-seq and GHT-SELEX peaks, and PWM hits treated as peaks) were adjusted to maximize the Jaccard value (intersection/union) between all three peak sets; we refer to these as “triple overlap” or TOP sites (see **Methods** for details). In this approach, false negatives will arise due to any experimental error or inaccuracy of motif models, as well as widespread non-specific DNA binding, which will tend to raise the thresholds for sequence-specific binding in these procedures. Thus, the number of direct binding sites obtained are underestimates. In addition, 37% of the

217 proteins lacked GHT-SELEX data, and/or did not have motifs; therefore, this analysis could be conducted only on 137 TFs (101 Codebook TFs and 36 controls; see **Table S4**). For these 137, the fraction of peaks that could be accounted for by direct binding ranged from 0.04% (for SP140, which binds a short motif composed mainly of a CG dinucleotide) to 65.7% (for ZNF728, which has a unique 21-base motif), with a median of 10%**.** The fraction of ChIP-seq peaks that are due to apparent direct binding (i.e. % of all ChIP-seq peaks that are TOP sites) is similar for the Promoter TFs (9.5%) and Dark TFs (9.3%) (**Figure 3A**), and the absolute number of direct binding sites is similarly high for a subset of both groups (**Figure 3B**). We conclude that there is no systematic difference between Promoter TFs and Dark TFs in direct binding characteristics, and that many of the observed TF binding sites are direct.

To ask whether the relative preference of TFs for different types of genomic regions and chromatin states in ChIP-seq is intrinsic, we examined the fraction of GHT-SELEX and ChIP-seq peaks for each TF that are found within genomic regions corresponding to each type of genome annotation. To avoid circularity, we used universal peak thresholds (i.e. the same cutoff across all experiments, see **Methods**) which lowers the overlap between GHT-SELEX and ChIP-seq peak sets. The fraction of intrinsic (i.e. GHT-SELEX) and cellular (i.e. ChIP-seq) sites for each TF that overlap with protein-coding gene promoters, repeat sequences (of any kind), and the combination of the "empty" and "heterochromatin" states are shown in **Figures 3C**, **D**, and **E,** respectively. In each case, there is preferential binding *in vitro* which corresponds to that observed in cells, with Promoter TFs having much higher intrinsic preference for promoter DNA, and Dark TFs having higher preference for repeats and empty/heterochromatin. We note that many Promoter and Enhancer TFs have a greater tendency to bind "empty" and "heterochromatin" state DNA *in vitro* than in cells, which could be due to a *bona fide* preference for open chromatin, functional binding at these loci in other cell types (but not HEK293), or preferential extraction of these proteins at open chromatin in ChIP-seq experiments.

The Promoter TFs also displayed intrinsic preference for the regions that overlap or are just upstream of transcription start sites (TSS), by multiple measures (**Figure 3F**), similar to that described for characterized TFs in a variety of genomes[44-46]. This observation is consistent with functional roles for these uncharacterized TFs in promoter definition, delineation of TSS location, and/or gene regulation.

Distinct conservation patterns of Dark TF vs Promoter TF binding sites

To further query functionality of direct binding sites (i.e. TOPs), we examined conservation of the TOP sites, producing an estimate of whether each site is under purifying selection. In essence, for many TFs, the TOP sites in aggregate display conservation patterns that mimic the selectivity of each base position in the TF's PWM. **Figure 4A** shows a graphic demonstration: when TOP sites are aligned to the PWM hit, and displayed as heatmaps that show base-level conservation scores (here, phyloP[47]), there are often vertical blue lines. These lines represent positions in the PWM hits that are preferentially conserved across many TOP sites. Similar to previous observations made with well-characterized TFs[47], the positions with highest conservation often

correspond to tall letters in the sequence logo (i.e. high information content), indicating selection on the binding site to match the sequence preferences of the TF.

We developed three heuristics to discriminate conserved vs. unconserved TOP sites (see **Methods**). Two of them test for a relationship between the information content at each base position of the PWM and the conservation score, while the third tests for higher overall conservation at the PWM hit than in immediate flanking sequence. As shown in **Figure 4A**, and in similar diagrams for all 137 TFs for which these tests could be run (**Document S1**), these tests together detected sites that appear plausible by visual inspection (i.e. apparent conservation signal relative to flanks). We considered a site to be conserved if any of the three criteria were met, and at least one nucleotide in the PWM hit had an FDR-corrected PhyloP score $\geq$1. By these criteria, conservation of TOPs is observed for both Promoter TFs and Dark TFs (**Figure 4A**), but the fraction and absolute number of conserved TOP sites for Promoter TFs is much higher (**Figure 4B,C** and **Table S4**). This outcome suggests that many Promoter TF binding sites are functional, and that the corresponding TFs have conserved functions at promoters. An individual conserved TOP site (hereafter, “CTOP” site), for the Promoter TF ZNF407, is shown in **Figure 4D**; like many Promoter TF TOPs and CTOPs (**Figure 3F**), it overlaps with a transcription start site. CTOPs are also often found adjacent to other CTOPs (explored in greater detail in an accompanying manuscript[31]); an example of multiple sites for ZNF131 and YY1 is shown in **Figure 4E**.

Despite their lower numbers, there are still thousands of CTOPs for Dark TFs: in aggregate, the criteria used here yielded 6,086. They tend to be distant from promoters, or each other (e.g., 2,916 are > 1000 bp away from any other CTOP), and they tend to have lower PhyloP scores than CTOPs for Promoter TFs. The Dark TF ZBTB40 recognized nearly 1,000 CTOP sites, the vast majority of which correspond to remnants of *hAT/Charlie* DNA transposons (**Figure 4A,F**). Its most strongly conserved CTOP falls outside of a transposon, however, and instead is within the PRKACA 3’ UTR (**Figure 4G**), which may be relevant to its known function (see below). ZNF689, in contrast, is an example of Dark TF that has a much smaller number of CTOP sites, and is enriched for binding L1M5 elements across its TOPs (**Figure 4A**; example CTOP shown in **Figure 4H**). Overall, these analyses indicate that Dark TFs occupy a unique and expansive fraction of the genome, and thousands of their direct binding sites show indications of conserved function. The interactions of TFs with TEs, and the known and potential functions of these and other Dark TFs, are explored in the next sections.

Widespread and specific binding of Codebook TFs to transposable elements

We reasoned that the generally low conservation in direct binding sites for Dark TFs could be due to domination by TEs, which are typically under neutral selection. In addition, TEs are only present in a subset of species that have retained an ancestral insertion, limiting power to detect purifying selection. Indeed, 92.0% of the Dark TF binding sites overlap with repeats (aggregated TOPs vs. Repeatmasker track) (vs. 25.3% for Promoter TFs, 39.9% for Enhancer TFs, and 46.8% for Other TFs).

The combination of ChIP-seq, GHT-SELEX and Codebook PWM data enables us to circumvent previous challenges in analysis of repeat sequences, and to examine binding of TFs to TEs with unprecedented precision, including detection of direct, base-level binding. **Figure 5A** provides an overview of high-confidence ($P < 10^{-8}$, Fisher's Exact Test) interactions between the Codebook TFs and specific TE classes, with ChIP-seq and GHT-SELEX peak sets calculated separately (**Figure S4** shows an expanded version with all rows labelled). A first observation that emerges from this analysis is that Dark TFs are much more likely than Promoter TFs to significantly bind a specific class of TEs (36% vs. 8%, respectively), but binding to a specific TE class is not a universal or discriminating property of either the Dark TFs or KZNFs. Among the 42 TFs that passed the cutoff, 20 are Dark TFs, 22 are KZNFs, 14 are both, and 14 are neither.

A second observation is that the TE enrichments in the GHT-SELEX data are virtually identical to those in the ChIP-seq data (shown adjacent to each other in **Figure 5A**), illustrating that specific binding to these elements is an intrinsic property of individual TFs. To our knowledge, this is the first experimental demonstration that KZNFs independently possess sufficient sequence specificity to discriminate ERE subfamilies from the rest of the genome: previous motif models derived from ChIP-seq data were unable to specify individual elements as precisely[48]. A third observation is that TEs of all major classes (LINE, SINE, LTR/ERV, and DNA transposons) are recognized by specific TFs. Moreover, for all four major classes of TEs, there are cases in which greater than 10% of a TF's TOP sites overlap one type of TE, and are conserved (**Figure S5**), consistent with a function for the host genome

A fourth observation is that the encompassed TEs span a very wide age range, from human-specific AluY elements, to L2, L3, and MIR, which pre-date eutherian mammals. These associations can provide insight into the evolution and molecular function(s) of the TFs. For example, ZNF836 and ZNF841 (which are both Dark TFs and KZNFs) are paralogs that arose from a pre-simian duplication event[49] and bind to distinct subtypes of the closely related, simian-specific MaLR LTR elements. They bind distinct motifs that specify the differing base identities at homologous positions within the diverged LTR, suggesting neofunctionalization and retention to maintain silencing of both LTR subtypes (**Figure 5B**). There are also cases in which the TFs and TEs they bind are grossly mismatched in age. For example, ZNF286B is a human-specific duplicate of ZNF286A which has lost its KRAB domain[50,51], but its binding sites are enriched for LINE-3 (L3), an ancient element found across all mammals, suggestive of coincidental adaptation (**Figure 5A**). In contrast, ZNF362 and ZNF384 (both non-KZNFs) are products of a duplication ~429 MYA (the duplication is found across bony vertebrates), but the binding sites for both proteins are enriched for the much younger, primate-specific Alu elements, as well as poly-A repeats, consistent with their DNA binding motifs (**Figure 5C**). These proteins have the largest number of TOP sites within the Codebook dataset, and it is possible that the recently-expanded target range of these proteins is a coincidental liability, as rearrangements of both ZNF362 and ZNF384 genes (most commonly as fusions to activating TFs and cofactors) are found frequently in leukemia[52,53].

Four of the Promoter TFs bind to specific TEs in this analysis (**Figure 5A**), potentially providing direct links between TE insertions and regulation of host genes. One of them, ZNF676, is a KZNF that was previously shown to associate with LTR12, and to repress "transpochimeric" gene transcripts[54], which are generated during human early embryogenesis. The GHT-SELEX data and Codebook motif pinpoint its exact binding site in LTR12 (**Figure 5D**). ZNF676 may also have other roles at promoters: at the TOP site upstream of QSER1 (**Figure 5D**), the LTR12 element is in the opposite orientation from the gene. Another Promoter TF, JRK, preferentially associates with the DNA transposon Tigger15a. JRK is itself derived from a Tigger element, and the Tigger15a consensus sequence contains binding sites for JRK at its terminus[31]. Thus, Tigger15a may have simultaneously contributed both to the rise of JRK protein and a set of JRK binding sites that are still utilized; this hypothesis is supported by the taxonomic distribution of JRK and Tigger15a to therian mammals (dating to 160 MYA).

Older TFs tend to bind older DNA

In addition to insertions such as TEs, new TF binding sites can emerge from random mutations in pre-existing sequences. This mechanism is thought to be dominant for traditional enhancer-binding TFs[55]. To determine whether binding sites for Codebook TFs evolved from ancestral DNA, we estimated the age of each TOP site for each TF, gauged as that of the oldest ancestral genome that contains the entire site (i.e. a gapless alignment, even if the base identities are different) in the Zoonomia mammalian reconstructions[56] (**Figure 6A**). This is a simple heuristic, but we obtained a qualitatively similar conclusions using other approaches to estimate binding site age (**Figure S6**). As expected, TOP sites for the Dark TFs (most of which correspond to TEs) are estimated to be younger on average than those of Promoter TFs (median ages of 46 and 72 MYA, respectively (**Figure 6B**, **Table S4**), but there is a large overlap of age distributions between the two TF groups. Both groups contain TFs with binding sites at both extremes (i.e. very old or very young binding sites). Thus, average age of the binding site is not a discriminating characteristic of Promoter vs. Dark TFs.

We also estimated the ages of the TFs, using catalogued ortholog and paralog relationships[57] and species divergence times[58] (**Figure 6C**, **Table S4**). Overall, TFs in both classes tend to be older than the sites they bind: Promoter TFs have a median age of 429 MYA, while Dark TFs have a median age of 97 MYA, which is still older than a typical binding site even for Promoter TFs. These results are consistent with the established phenomenon of TF binding site turnover[55]. They are also consistent with previous observations with KZNFs, which concluded that the correlation between age of binding sites and age of the KZNF is weaker than expected if they evolve only to silence TEs[26]. Together with the retention of many KZNFs that bind extinct TEs, this finding supports the notion that KZNFs must frequently take on additional regulatory roles, e.g. in regulation of host genes.

Functions of Dark TFs

Finally, we examined existing literature and databases to survey known and potential functions for the Dark TFs, and related it to the data we collected (**Figure 7**, **Table S5**).

Most Dark TFs have apparent roles in repression of transcription. 35 out of 54 are KZNFs, and for 14 of them (and one non-KRAB TF, ZNF888), physical association with KAP1 has been verified[27,59-62]. The KZNFs may also have repressive functions beyond the recruitment of KAP1[27,59]. Five of these Dark TF KZNFs also interact with TRIM39, which itself interacts with numerous ubiquitin conjugating enzymes, H3K4 demethylase KDM1A, and dozens of other KZNFs[60,63]. One additional Dark TF KZNFs (and two other Dark TFs) interact with TRIM33, a member of the TIF1Y complex that specifically suppresses TGFβ-responsive genes by directly interacting with the histone subunits as well as E3 ubiquitin ligase[64].

Ten of the twelve non-KRAB Dark C2H2-zf proteins also appear to contribute to the formation and maintenance of heterochromatin, by association with chromatin proteins (CBX/HP1) directly, or via recruitment of other C2H2-zf proteins. One of them, ZNF518B, was identified as a partner of both H3K27 methylase EZH2 and H3K9/H3K27 methylase G9A, and to promote H3K9me2[65]. ZNF518B and ZNF280D both associate with multiple CBX/HP1 proteins[60,63]. ZNF518B binds many primate-specific L1 elements, but its most conserved binding sites are in its own promoter, suggesting a critical negative feedback mechanism (**Figure S7**). In another example, ZNF516 associates with the multifunctional CTBP1/KDM1A/RCOR1 corepressor complex, and its repressive function was shown in reporter assays[66]. Intriguingly, 21 of the 47 C2H2-zf proteins, including both KRAB and non-KRAB C2H2-zf proteins (as well as transposon-derived ZBED9) interact with other C2H2-zf proteins, often extensively[60], suggesting a potentially widespread role in organization of chromosome topology.

Four additional Dark TFs have other potential roles in repression of transcription. Three of them are the paralogous nuclear speckle proteins SP100, SP140, and SP140L. Each contains a SAND domain, which we confirmed binds to unmethylated CG dinucleotides *in vitro*[67], and CG-containing motifs are enriched in their ChIP-seq peaks[36]. These proteins also contain PHD and BRD domains, which typically function as epigenetic readers[68]. In our ChIP-seq data, they are enriched at sites of H3K27me3 methylation (**Figure 7**, **Table S5**). SP140 is an exceptional TF among the Codebook data set; its ChIP-seq sites predominantly overlap with “GeneHancer” loci – i.e. these sites are catalogued as enhancers in other cell types, but not HEK293 enhancers (**Figure S2A**), suggesting that these loci may be actively silenced in HEK293. The fourth protein is SCML4, a polycomb group protein that was included in our study because it contains an AT hook, but we did not obtain evidence for its sequence-specific DNA binding. Thus, it may be more properly described as a chromatin protein. SCML4 is reported to associate with H3K4 demethylase KDM5C[69] as well as ubiquitination factors FBXO11 and UBR1[70].

Five of the non-KZNF Dark TFs may have roles other than repression. One of them, SOX2, is a well-known pioneer factor that can bind to motif matches within unmodified closed chromatin, but is inhibited to some extent by H3K9me3[71]. Indeed, in the ChIP-seq data reported here, most (66%) of its TOP sites are in “empty” chromatin, and only 5% overlap with ATAC-seq peaks in unperturbed HEK293 cells, consistent with its pioneer function. Less than 2% of SOX2 peaks overlap with heterochromatin (defined by ChromHMM mainly by H3K9me3 and H327me3), consistent with H3K9me3 being

refractory to SOX2 binding. Two additional Dark TFs may also represent pioneers: TPRX1 has recently been described as a master regulator in zygotic genome activation[72], while SALL3 controls the differentiation of hiPSCs into cardiomyocytes vs neural cells[73]. In contrast, two other Dark TFs have been described as impacting DNA metabolism. ZNF384, which we find binds many Alu and Poly-A repeats, as described above, is also known to bind Ku and recruit NHEJ factors to double-strand breaks[74]. ZNF146 binds L1 elements, and its depletion slows the replication fork[75].

The distinct binding sites and diversity of apparent effector mechanisms and cellular roles of the Dark TFs suggest that they may each regulate specific biological functions, and that they may also be multifunctional. Indeed, physiological consequences that have been reported for perturbation of the Dark TFs vary widely (**Figure 7**, right column; **Table S5** provides the values and sources), ranging from basic cellular processes to development. For example, the KZNF ZNF689, which we show above binds ~50 conserved sites enriched for L1M5 (**Figure 4G**), also binds promoters of various L1 subtypes, preventing genomic instability conferred by L1 retrotransposition[76]. ZBTB40, which we observe almost exclusively at DNA *hAT/Charlie* transposons (**Figure 4E**), and which is one of the oldest Dark TFs (**Figure 7**), was recently shown to bind telomeric dsDNA breaks and maintain telomeric length[77]. In mouse, Zbtb40 deficiency impacts spermatogenesis through disrupted telomeric lengthening and maintenance in spermatocytes[78]. *hAT/Charlie* transposons are enriched in telomeric regions of human DNA[79], suggesting that this function may be conserved. The most conserved ZBTB40 binding site, however, is within the 3'UTR of PRKACA (**Figure 4F**). PRKACA encodes the catalytic subunit α of protein kinase A, whose deficiency is associated with fertility defects in male mice and humans[80]. This site is also less than 1 kb from the TSS of the chromatin regulator SAMD1, which impacts sperm cells[81].

DISCUSSION

The Codebook ChIP-seq data provide cellular binding sites for 130 putative TFs, defined as previously lacking PWMs or other models of sequence specificity. It represents a valuable resource for studying TF function and evolution in the context of regulatory genomics. For a large majority of the proteins assayed, we have also now identified a binding motif which is supported by independent assays[36]. Thirty-six of the proteins did not produce a motif and may not be *bona fide* TFs. Their ChIP-seq profiles are nonetheless informative: enrichment of ChIP-seq peaks at different types of genomic features (e.g. promoters, repeats) or chromatin states, as well as co-occurrence with peaks for other proteins (e.g. TFs), can yield clues as to potential function.

Previous large-scale ChIP-seq analyses have mainly focused on the established roles of TFs in binding to promoters and enhancers (e.g.[1]). A major exception has been studies of KZNFs, which focus on binding to TEs, and specifically EREs[26,27,34]. The analysis scheme described here considers these models of TF function as hypotheses with equal weight. The known categories are clearly present, including preferences for promoters and enhancers, as well as the strong tendency for KZNFs to bind specific classes of EREs, and within constitutive heterochromatin. Overall, however, TF

558 behaviour with respect to chromatin states and genomic landmarks appears more
559 varied than a simple categorization scheme would imply. We did not systematically
560 explore the “Other” category, used here as a catch-all. Like the Dark TFs, it appears to
561 encompass proteins that satisfy some expectations of “pioneers”, given that they bind
562 both *in vitro* and *in vivo* to many regions that are labeled as inactive and/or closed
563 chromatin in HEK293 cells prior to induction of the TF. There are many intriguing TFs in
564 the “Other” category: one example is the non-KRAB protein ZSCAN2, which is involved
565 in spermatogenesis and fertility in mice (and perhaps human)[82]. We catalogued 183
566 CTOPs for ZSCAN2, and found that its binding sites are enriched for mammal-wide L3
567 elements.

568 Establishing functions, if any, for individual TF binding sites is a long-standing and
569 difficult problem in regulatory genomics. The level of binding site turnover observed on
570 evolutionary timescales requires that binding sites arise at random, many of which are
571 presumably irrelevant for gene regulation or reproductive fitness, at least initially. By this
572 reasoning, we expect that many biochemically verified, direct TF binding sites should be
573 non-functional, and indeed we find that, overall, most TOP sites are not conserved,
574 even for Promoter TFs. Phylogenetic footprinting does not discriminate between false
575 negatives due to binding site turnover or redundancy, and *bona fide* non-functional
576 sites, and therefore lack of conservation cannot be taken as lack of biological purpose.
577 Nonetheless, conserved TOP sites would seem most likely to yield interpretable results
578 in targeted laboratory studies. Sites overlapping TSS may be particularly fruitful, given
579 apparent constraint on both sequence and position of the binding site. More generally,
580 the Codebook TOP catalogue will provide a rich resource for future efforts in examining
581 genome function.

582 The low primary sequence conservation of Dark TF binding sites, especially relative to
583 those of Promoter TFs, could have several explanations. One possibility is that very few
584 of the binding sites are functional; in theory, only a single binding site that confers
585 modest selective advantage would be sufficient to drive retention of both the site and
586 the TF, with all other sites arising at random (and under neutral evolutionary pressure,
587 provided they are not detrimental). Another possibility is that the exact positioning of the
588 sites is not critical to their function, unlike Promoter TFs, which by definition must be
589 close to TSSs. Dark TF functions could simply require that binding sites are distributed
590 widely across non-functional DNA, and thus be highly redundant over large sequence
591 windows (e.g. TADs). Such functions would not preclude a small subset of sites being
592 co-opted for regulation of host genes, which would become constrained (i.e.
593 conserved). Regardless of what biochemical, cellular, and physiological functions are
594 revealed, the Dark TFs represent a new contribution to the decades-old odyssey into
595 the function and significance of the dark matter genome.

596

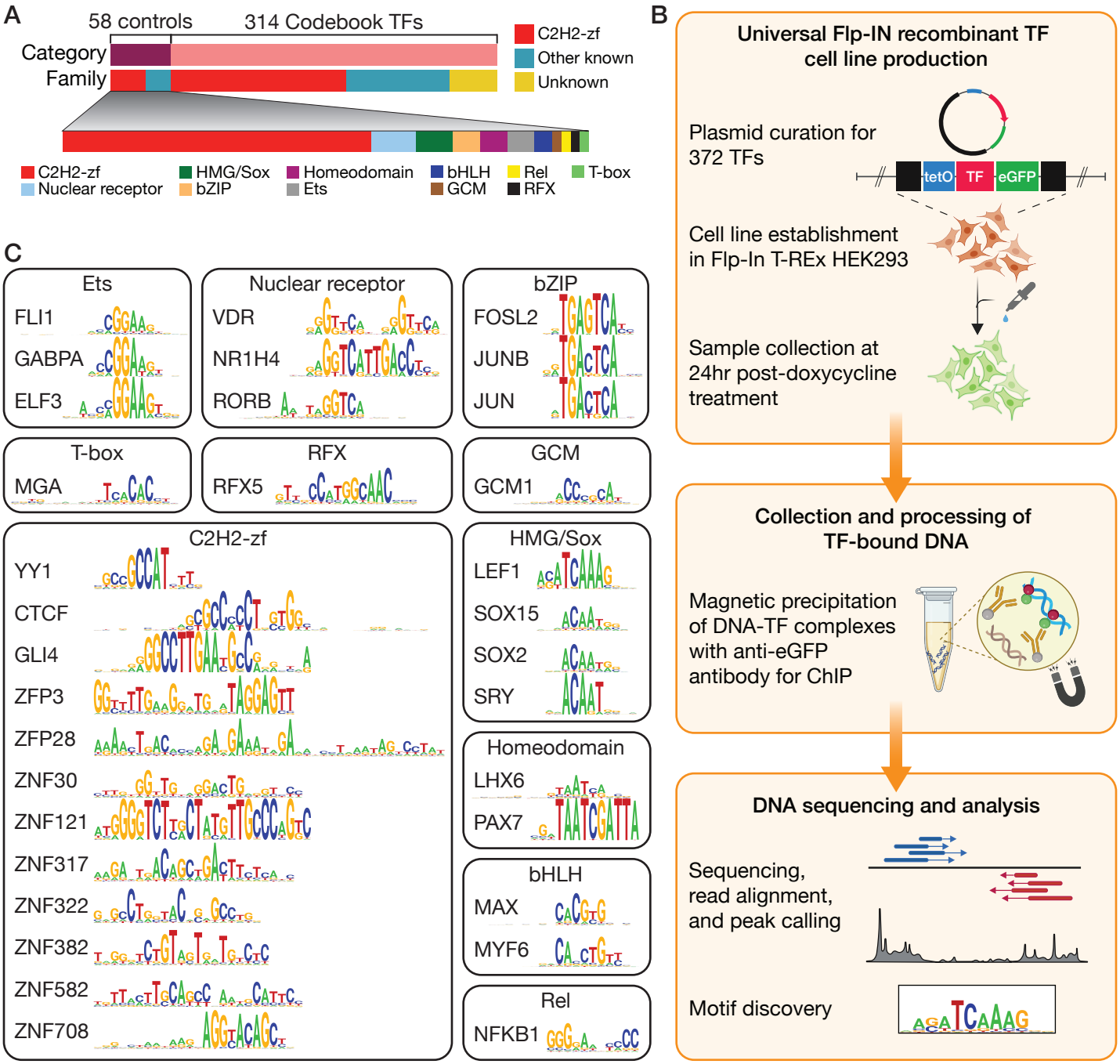
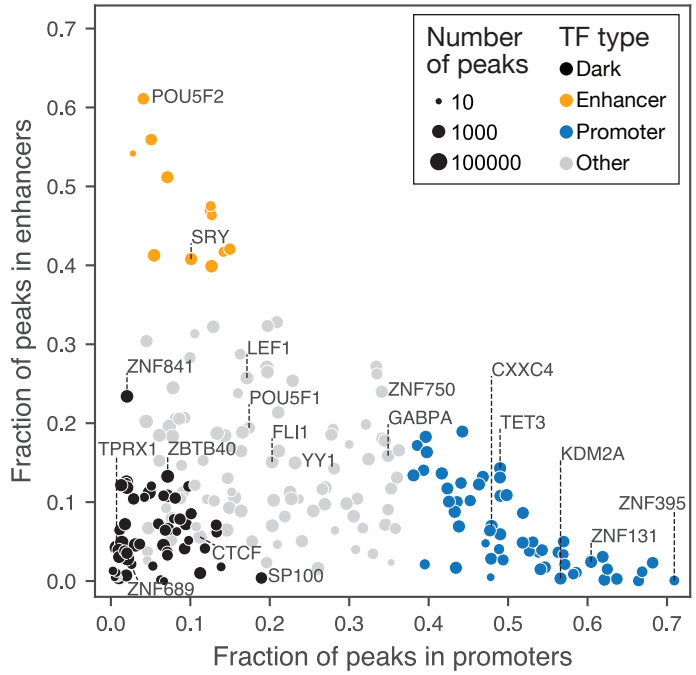


Figure 1. Project overview. (**A**) Overview of the TF categories assayed in this study. (**B**) A schematic of the experimental pipeline for production of 372 inducible EGFP-labelled TF cell lines used in ChIP experiments and deriving TF binding sites. (**C**) Samples of representative motifs obtained from different families of control TFs.

A



B

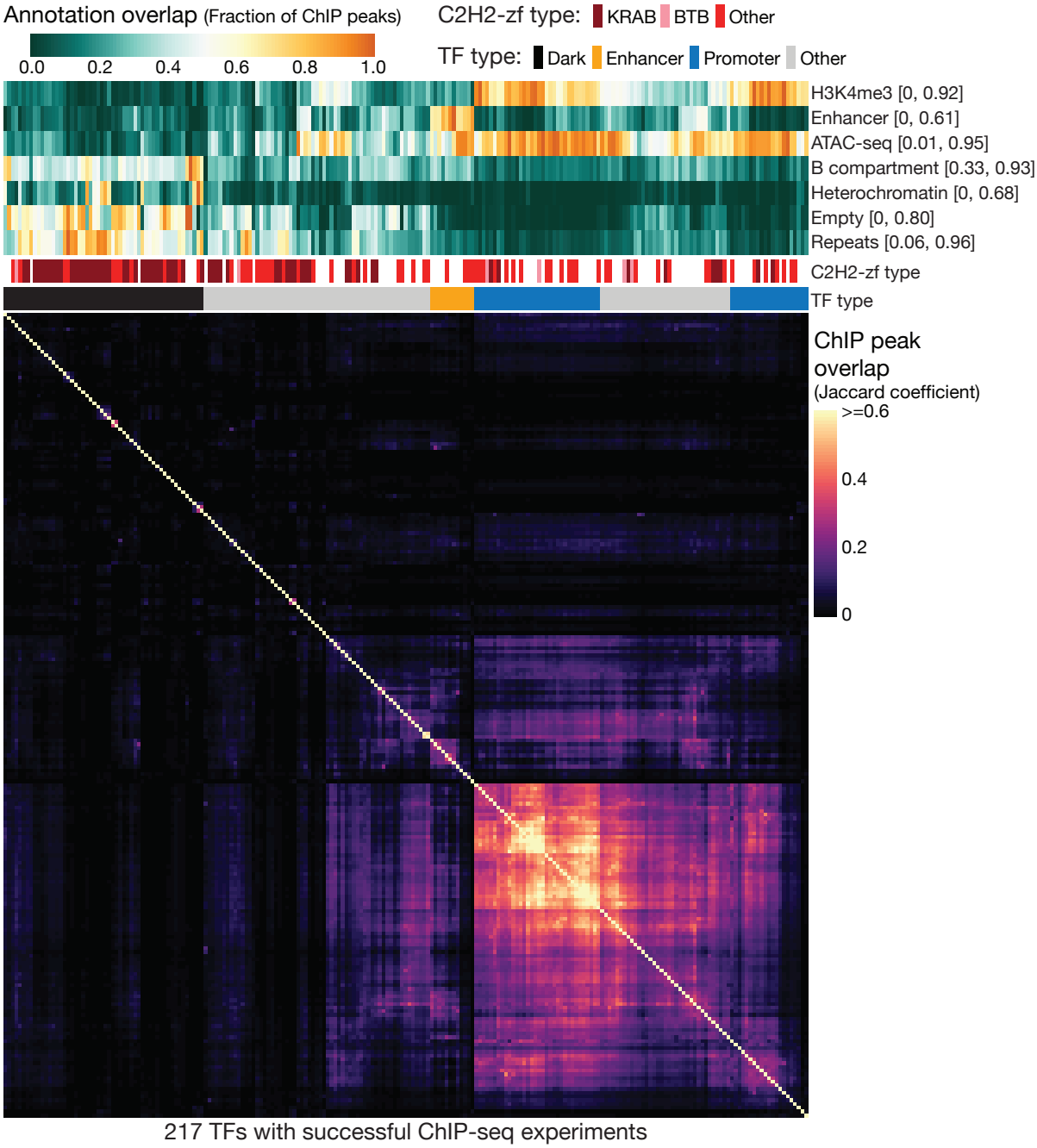


Figure 2. Overlapping *in vivo* binding sites of 217 TFs with each other and with various genomic regions. (**A**) Fraction of ChIP-seq peaks in protein-coding promoters (x-axis) and HEK293 enhancers (y-axis). Point sizes are proportional to the number of peaks for each TF (log scale). (**B**) *Bottom (square) heatmap*: Jaccard similarity coefficient between ChIP-seq peaks of all TF pairs. *Top heatmap*: Fraction of ChIP-seq peaks falling within genomic regions, as indicated, and other properties of the TFs. Fractions are scaled to fit in [min, max] range across the TFs for better visualization, as indicated in the right. TF ordering is determined by hierarchical clustering with Ward linkage and Euclidean distance, using the tracks 'H3K4me3', 'ATAC-seq', 'B compartment', 'Empty' + 'Heterochromatin', 'Repeats', 'CpG', 'Protein-coding promoters', 'H3K27ac' (the last three not shown), along with the one-hot encoded ‘TF type’ to aid in illustration.

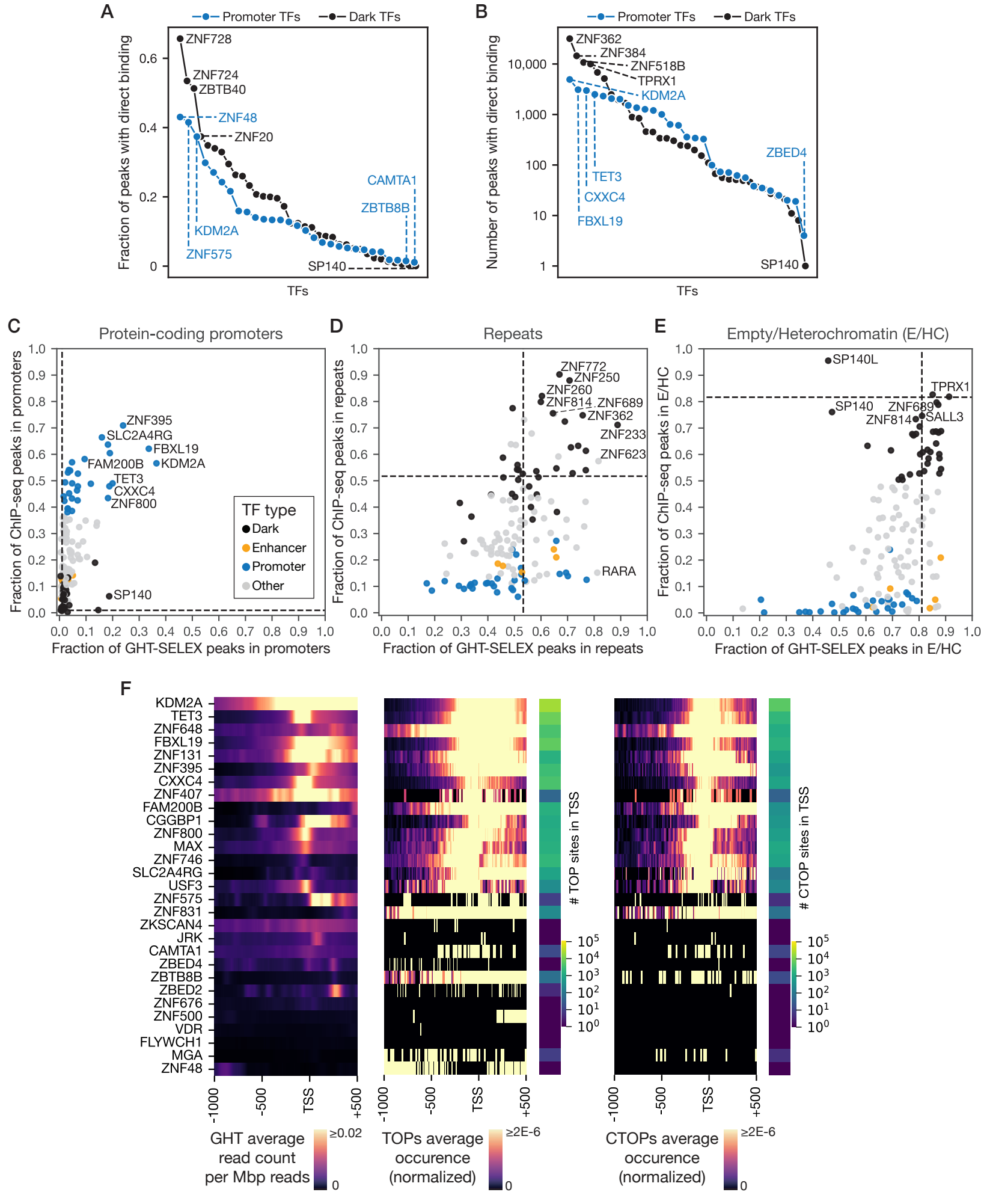


Figure 3. Characteristics of Promoter TFs, Enhancer TFs, and Dark TFs interaction with specific genomic sites. Fraction (**A**) and absolute number (**B**) of peaks with direct binding (i.e. TOP sites) for Promoter TFs and Dark TFs. TFs are sorted to compare distributions. The denominator for (A) is the total number of ChIP peaks at the same optimized threshold. (**C, D, E**) Fraction of GHT-SELEX (x-axis) and ChIP-seq (y-axis) peaks falling in the specified genomic regions (protein-coding promoters, repeats, and empty or heterochromatin), using the peaks at the universal threshold. Dashed lines show the expected fraction if peaks were distributed at random. **(F)** Density of GHT-SELEX signal (left), TOP sites (middle), and CTOP sites (right) by position relative to TSS of protein-coding promoters, for 29 Promoter TFs that have available GHT-SELEX data. Intensity of heatmaps for TOPs (middle) and CTOPs (right) have been normalized by the total number of PWM hits (of TOPs and CTOPs, respectively) in promoters (shown at the right of each heatmap).

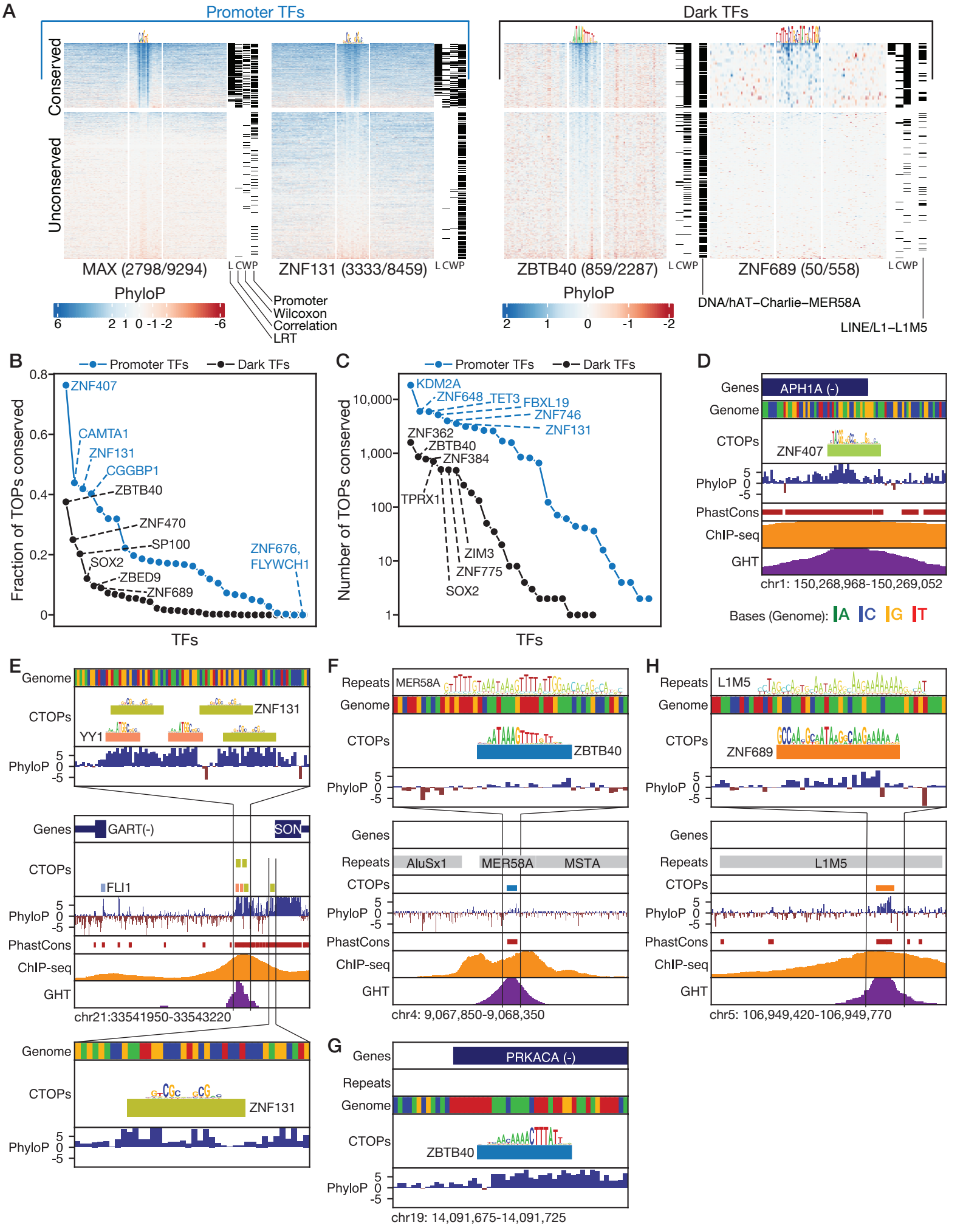


Figure 4. Conservation patterns of sequence-dependent TFs' target sites (TOPs). (**A**) Heatmaps of FDR-corrected phyloP scores across the TOP sites (rows), split into top and bottom segments that contain conserved and unconserved sites. Bars to the right indicate which tests of conservation are satisfied (Likelihood-ratio, Correlation, Wilcoxon), along with overlaps with promoters (P) and specific repeat families if applicable. 100 bp segments are shown with the PWM hit in the middle. Blue/positive phyloP indicates purifying selection, and red/negative phyloP values represent diversifying selection. (**B, C**) Fraction (**B**) and absolute number (**C**) of TOPs that are conserved, for Promoter TFs and Dark TFs, sorted to compare distributions. (**D, E, F, G, H**) Genome track displays of CTOP sites for ZNF407 (**D**), ZNF131 and YY1 (**E**), ZBTB40 at a *hAT/Charlie* (MER58A) element (**F**) and its most-conserved TOP (at the PRKACA promoter) (**G**), ZNF689 at an L1M5 element (**H**). The Dfam[103] repeat model sequence logo is also shown for MER58A (**F**) and L1M1 (**H**).

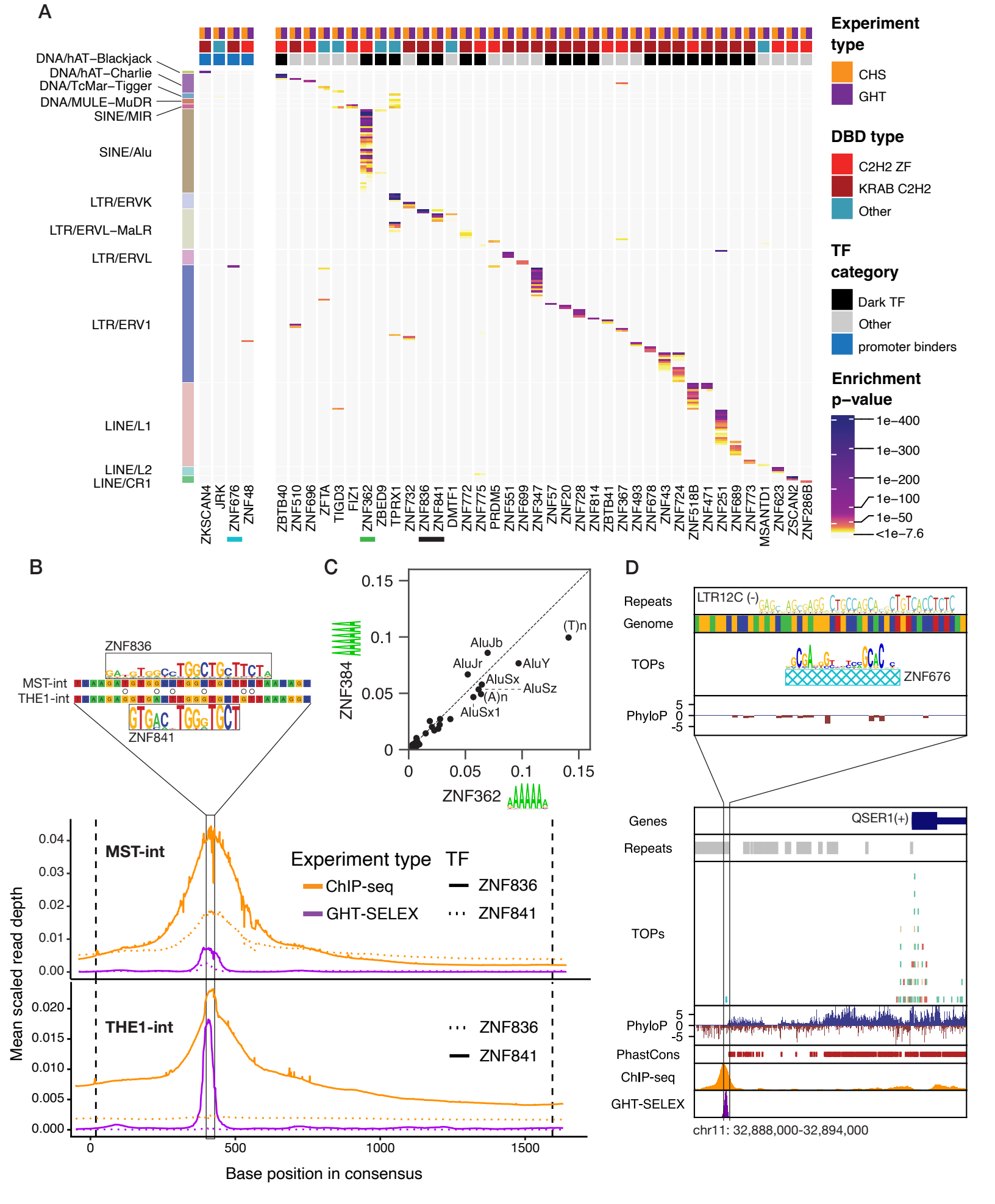
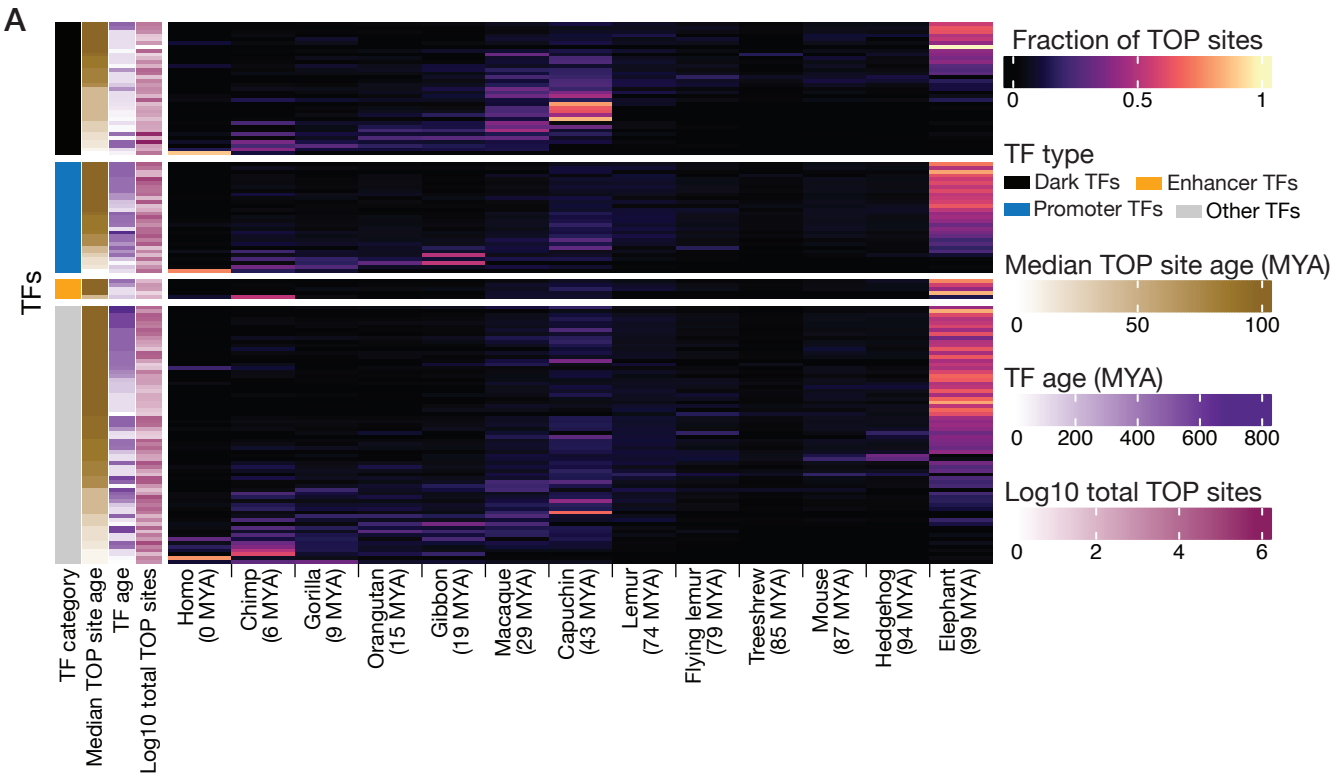
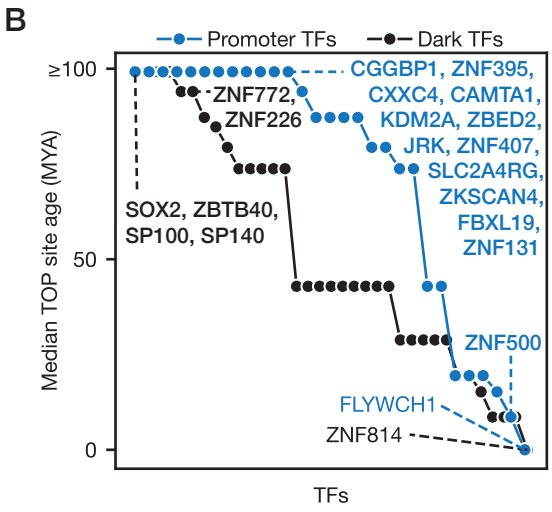


Figure 5. Enrichment pattern of transposable elements in TFs' TOPs. (**A**) Heatmap of –log10 p-values for TFs (x-axis) that are enriched for binding specific TE families (y-axis). Labels show superfamily/family. (**B**) Binding of paralogous TFs, ZNF836 and ZNF841, to a homologous region in the two related LTR families, MSTA-int and THE1-int. Bottom plot shows the average ChIP-seq and GHT-SELEX signal (i.e. read count) across all the instances of MST-int and THE1-int aligned to their consensus. (**C**) Fraction of TOP sites in various repeat elements for two poly-A binding TFs ZNF362 and ZNF384. (**D**) An example of the Promoter TF ZNF676 binding site targeting an unconserved LTR12C sequence.





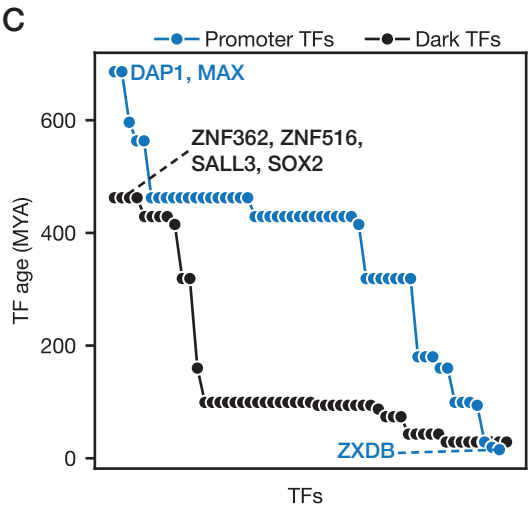


Figure 6. Age distribution of TOPs and their corresponding TFs. (**A**) Heatmap showing the fraction of TOP sites for each TF dating to different mammalian clades in the human lineage, along with information about the TF category, median age of TOP sites and TFs (million years ago, MYA), and log 10 of total TOP sites. (**B, C**) Sorted median age of the TOP sites (**B**) and the age of the TFs (**C**) are compared for Dark TFs and Promoter TFs.

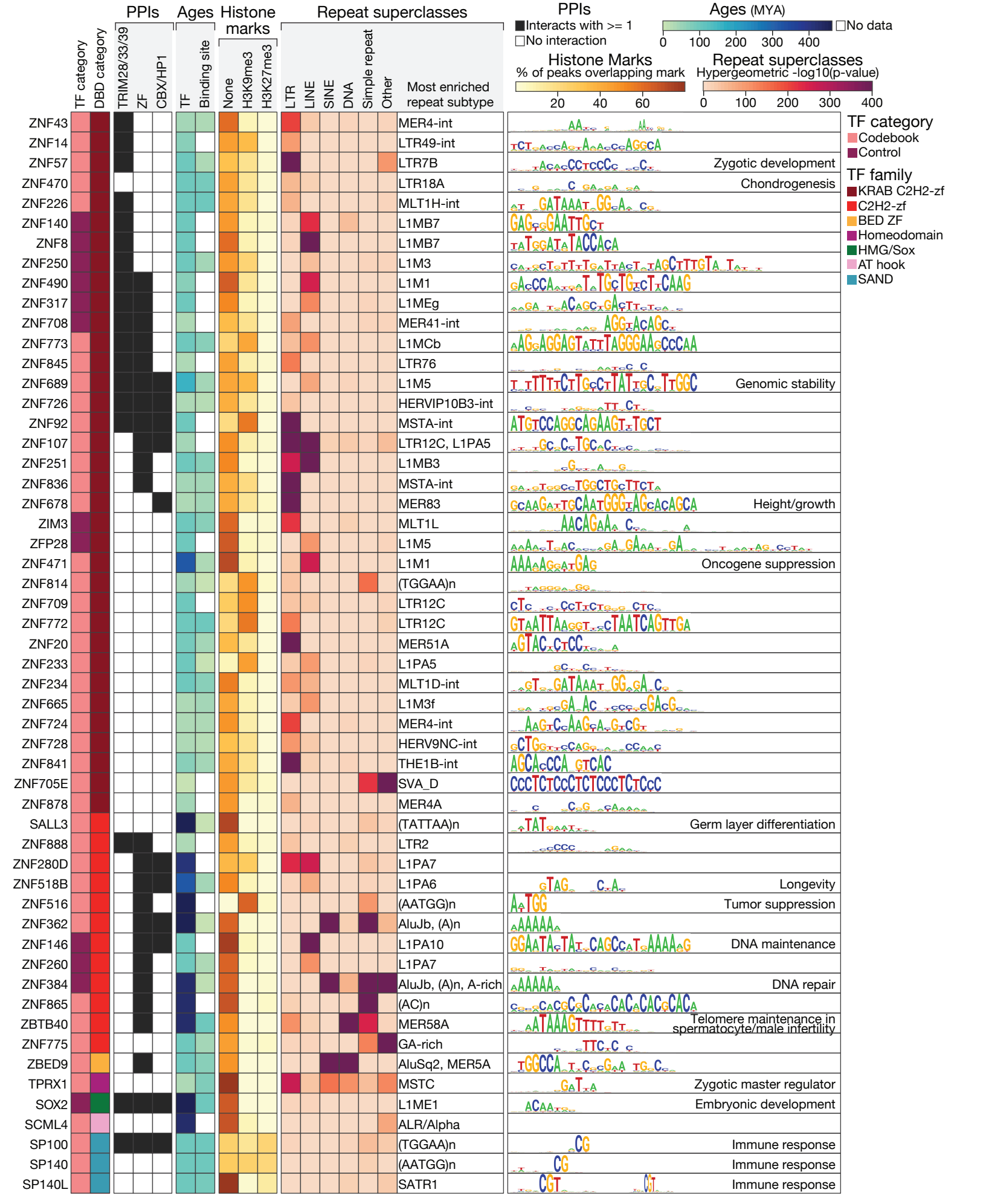


Figure 7. Consolidated functional information for Dark TFs. Compiled protein-protein interactions (PPIs)[63] mostly supported by two independent lines of support and grouped into three categories of TRIM28/33/39 interactions, zinc-finger (ZF) protein interactions, and CBX/HP1 interactions are shown at left. Median binding site age was calculated for TOP sites, only for the TFs with available GHT-SELEX data, shown along with the age of the TF. The fraction of ChIP-seq peaks (using the universal threshold) overlapping with H3K9me3 and H3K27me3 histone marks and with the ChromHMM "empty" state (None) are shown in the middle. For the repeat, in each superclass, the enrichment score (-log(p-value) hypergeometric test) for the most enriched repeat element within that superclass is plotted as a heatmap, and the most enriched repeat subtype across all the superfamilies is mentioned beside. The expert-curated sequence logos are displayed to the right (except for ZNF280D and SCML4 which did not produce any approved PWM), along with the corresponding phenotype for any TF with known biological function through literature review (in the same block).

METHODS

Plasmids. The Codebook project design is described elsewhere[31]. Putative TFs were those from Lambert et al.[3], with TFs we had already attempted as part of ENCODE removed. We attempted ChIP-seq analysis of all Codebook TFs. We designed full-length ORFs for synthesis (BioBasic.ca) and used conventional restriction cloning to insert them into Flp-In destination plasmid pTH13195 (a modified pDEST pcDNA5/FRT/TO-eGFP vector), which places ORFs under the control of a tet-on, CMV-driven promoter, with an N-terminally EGFP tag. We obtained the 58 controls independently[27,33] and cloned them into pTH13195 by the same process. See accompanying manuscript[31] for the sequences and other information about the inserts.

Cell line production. We used a previously established protocol for creating individual cell lines for each TF[34]. In brief, we cultured and maintained HEK293 Flp-In T-REx cells (Invitrogen) in Dulbecco's modified Eagle's medium with 10% fetal bovine serum and antibiotics. We created individual cell lines for each TF by flp recombination, in which individual destination plasmids were co-transfected into Flp-In T-REx 293 cells together with the pOG44 Flp recombinase expression plasmid using FuGene (Roche, 11814443001). We then selected cells for FRT site-specific recombination into the genome using selection media containing hygromycin (200μg/ml) for 1 to 4 weeks. For each TF cell line, we confirmed expression of EGFP by fluorescent microscopy after 24 hours of Doxycycline treatment (1ug/ml), at which point 10M-20M cells were used for downstream experiments.

Chromatin immunoprecipitation. We fixed ~20M cells on 15cm plates using 1% paraformaldehyde for 10 min on ice followed by 10 min quenching with 0.125 M glycine. We washed fixed cells twice with cold PBS, scrape collected, pelleted, flash froze, and stored the cells at -80C. Upon completion of cell collection for a panel of TFs, we thawed cell pellets on ice, lysed them as previously described[83], and sonicated them using a BioRuptor to shear chromatin. We then trapped the protein-DNA complexes on Dynabeads using an anti-EGFP antibody (Ab290, Abcam). Following wash, crosslink reversal, and elution, we assessed the size and concentration of DNA fragments with an Agilent bioanalyzer and Qubit, prior to sequencing.

Library preparation and sequencing. DNA library preparation and sequencing was performed at three different facilities (Memorial Sloan Kettering Cancer Center, SickKids Hospital in Toronto, and the Donnelly Centre at the University of Toronto) over a period of four years. The facilities prepared DNA libraries using the NEBNext® Ultra™ II DNA Library Prep Kit for Illumina. Samples were paired end sequenced (50-150bp) with a target depth of 20M reads per sample.

ChIP-seq data processing steps. *Read mapping*: We mapped raw ChIP-seq reads to the human genome build hg38 with *bowtie2*[84] (options: *--very-sensitive*, and *--no-unal*). We used Samtools[85,86] (options: *-q 30*, and *-F 1548*) to remove reads that were unmapped, failed platform/vendor quality checks, were PCR duplicates, or had a mapping quality <30. *Peak calling*: We created sample-specific background models following a procedure established previously[27], with minor modifications. Specifically, for

each pull-down experiment, we pooled reads from different control experiments together in a manner that maximizes the similarity of the pooled dataset and the background signal observed in that specific pull-down experiment, while ensuring high coverage. To do so, we first identified genome regions that show high background signal in at least one control experiment, by performing peak calling directly on control experiments with MACS2 (options: p-value < 0.001, and *--nomodel*)[39,87,88]. We pooled these "background hotspots" from all control experiments, and merged those whose summits were within 50 bp of each other, to create a unified set of hotspots. Then, in each control experiment, we calculated the number of reads overlapping each of the hotspots from the unified set, resulting in a read count matrix (with hotspots as rows and control experiments as columns). Similarly, for each pull-down experiment, we calculated the number of reads overlapping each of the hotspots, which we then used as the response variable in a non-negative Poisson regression, with the matrix described above as the set of independent variables. This regression results in a set of non-negative coefficients, representing a weighted mix of the control experiments that reconstructs the read count profile of the pull-down experiment across the hotspot regions as closely as possible. We then pooled the BAM files from the control experiments, by sampling a number of reads from each file that is proportional to this experiment-specific coefficient, to create a pulldown-specific background file, which we subsequently used for peak calling on the pull-down dataset using MACS2[39,87,88].

ChIP peak replicate analysis and merging. For each TF with one or more replicate, we calculated the Kulczynski II similarity metric for each pair of replicates (**Figure S1**). We used the Kulczynski II metric in place of Jaccard as it is less affected by the uneven size of the peak sets. We additionally calculated the Kulczynski II similarity metric for each pair of mismatched replicates (i.e., with TF identities permuted). Based on the distributions of "approved" experiment replicates and mismatch replicates, we defined a Kulczynski II threshold of 0.4 as the separating value for those two distributions (**Figure S1**). For TFs with "not approved" experiments (i.e., two ChIP-seq experiments did produce a reliable motif) we retained 36 (plus two controls) that achieved a Kulczynski II value >0.4 for inclusion in downstream analyses.
To generate a single peak set for each transcription factor, we merged the peak data from all successful experiments by merging overlapping peaks from one or more replicates using BEDTools[89] merge to generate new, wider peaks, with the sum of component peak -log(p-values) assigned as the new peak score, and center of mass of summits as the new peak summit. By default, we employed a peak cutoff of $P<10^{-10}$ (MACS2). Modulation of thresholds is described below.

ATAC-seq experiment and data analysis. We performed ATAC-seq in HEK293 cells, in four replicates, as described[90]. Briefly, 50,000 viable HEK293 cells were pelleted (500 RCF at 4°C for 5 min). After removing the supernatant, the cells were lysed in 50 µl of cold ATAC–resuspension buffer (RSB) containing 0.1% NP40, 0.1% Tween 20, and 0.01% digitonin by pipetting up and down three times followed by 3 min incubation on ice. The lysate was then washed out with 1 ml of cold ATAC-RSB containing 0.1% Tween 20 and the nuclei were pelleted at 500 RCF for 10 min at 4°C. 50µl of transposition mixture (25 µl of 2× TD buffer, 2.5 µl of transposase, 16.5 µl of PBS, 0.5 µl of 1% digitonin, 0.5 µl of 10% Tween 20, and 5 µl of H_2O) was added to each pellet,

mixed well by pipetting up and down, and incubated for 30 min at 37C. Tagment DNA TDE1 Enzyme and Buffer Kit (Illumina) was used for this step. The tagmented DNA was then purified with DNA Clean and Concentrator kit (Zymo Research) in 21 ul of elution buffer. DNA amplification and barcoding were performed using Nextera DNA Library Prep kit (Illumina) and NEB barcoding oligos. Subsequent sequencing was performed at the Donnelly Center sequencing facility using 100bp paired end sequencing at 60M reads per sample. Adapter sequences were first trimmed using *cutadapt*. The resulting reads were then mapped to the human genome (hg38) using *bowtie2*[84], followed by the creation of BAM files using *samtools*[85,86] *view*, and sorting with *samtools sort*. Peak calling was performed on the sorted BAM reads by running *macs2*[39,87,88] *callpeak* with the options *-f BAMPE, -g hs, -B*, and *-q 0.01*. Finally, to generate a single peak file for HEK293 open chromatin, all the peak sets were merged using *bedtools*[89] *merge*.

Chromatin state analysis. We obtained chromatin states by training a ChromHMM[41] with ten states (see **Figure S2B**) on marks H3K4me1, H3H4me3, H3K36me3, H3K27ac, H3K9me3, and K3K27me3, collected in HEK293 cells by ENCODE[91], plus ATAC-seq and CTCF peaks from HEK293 cells generated as part of this project. We also employed promoter regions derived from the GENCODE annotation (release 44)[92] for our analyses (-1000 to +500). For Hi-C B compartment annotations, we labeled genomic regions with a Hi-C first eigenvector value less than 0.4 in ENCODE data for HAP1 cells[93], comprising 65% of the genome.

Overlap of ChIP-seq peaks between all pairs of TFs. Jaccard similarity is taken as O / (N1+N2-O) where O is the number of intersecting peaks and N1 and N2 are the size of each set. We utilized BEDTools[89] to calculate overlaps. To prevent miscounting of the cases in which one peak in one set overlaps with multiple peaks in another set, we used the average of overlapping peaks (O = (O1+O2)/2 where O1 is the number of peaks in set 1 overlapping with any peak in set 2 and vice versa) to calculate the intersection in Jaccard. The same methodology was used to calculate the overlap of ChIP-seq peaks with the chromatin tracks.

Selection of universal ChIP-seq and GHT-SELEX thresholds. We calculated the Jaccard similarity from all 137 pairs of TFs with ChIP-seq and GHT-SELEX data, using the merged ChIP-seq peaks. We performed a grid search for all TFs simultaneously, sampling ChIP-seq P-value and GHT-SELEX cutoffs (determined by selecting different "knee" values[94] in the graph of sorted enrichment coefficients[43]), to identify a pair of thresholds that maximize median Jaccard. Two ChIP-seq cutoff (10^{-10} and 10^{-20}) yielded an almost identical maximum; we chose 10^{-10} as it includes a larger number of peaks. A corresponding knee threshold of 30 emerged for the GHT-SELEX knee-based cutoff.

Derivation of TOP sites. To define binding sites supported by ChIP-seq, GHT-SELEX, and PWM hits, we optimized the cut-offs of all three to maximize the overlap between all three data types. We first sorted the peaks based on their statistical scores, i.e., merged p-values for ChIP-seq peaks, cycle enrichment coefficient for GHT-SELEX peaks (see accompanying manuscript[43]), and sum-of-affinities for clusters of PWM hits with a p-value < 0.001 (from MOODS[95]), merged with neighboring hits in the case of having a distance less than 200 bp. Then, for different values of N, we took the top N peaks and

calculated the overlap (measured as the Jaccard index; intersection of all three divided by the union of all) using the top N ChIP-seq peaks, top N GHT-SELEX peaks, and top N merged PWM hits. The N that maximizes the Jaccard overlap was taken as the optimized threshold, and the overlap of all three sets at this threshold (N) is referred to as triple overlap or "TOP" sites.

Analysis of purifying selection and classification as conserved and unconserved binding sites. We extracted phyloP scores[47] for each PWM hit, and for flanking regions of equal length (for a total of 100 bp including the PWM hit and its flanks) from the 241 eutherian mammal Zoonomia alignment[96] using DeepTools[97]. We excluded PWM hits overlapping with ENCODE Blacklist sites[98] or protein coding sequences, due to the skew in phyloP scores caused by codons. All phyloP scores reported here are FDR-corrected. We conducted three tests to classify PWM hits as 'conserved' or 'unconserved':

1, LRT (Likelihood-Ratio Test)*:* This test scores the likelihood that the phyloP scores are driven by the PWM information content (IC) at each base position in the PWM. For each TF, we created a scoring model that represents the relationship between the phyloP scores at a PWM hit, and the information content at each base position of the PWM. This model is an l x 1 vector, where l is the length of the motif. To derive this vector, we first took the correlation of phyloP scores at each base position within the PWM hit to the IC at that position, for each PWM hit in the TOP dataset. We then selected the 100 PWM hits with the highest correlation and calculated the standard deviation (σ) of the phyloP score at each position of these 100 PWM hits. If a position has an invariant phyloP score (i.e. $\sigma = 0$), the σ at this position was replaced with a 1. As a null model, an IC value of 0 was assumed at each position, and the same σ values as the phyloP model. The LRT statistic for each PWM hit m was then taken according to the equation:

Equation 1: $$L(m) = -2 \times \left(\sum_{i=1}^{l} \frac{{P_i}^2}{\sigma_i^2} - \sum_{i=1}^{l} \frac{(P_i - I_i \times C)^2}{\sigma_i^2} \right)$$

Where P_i represents the phyloP score of position i in PWM hits m, σ_i is the standard deviation of the phyloP model at position i, and I_i is the IC of the PWM at position i. The IC value is first multiplied by a coefficient C, which is the linear regression coefficient describing the relationship between the position-wise phyloP means of all of a TF's TOPs and position-wise PWM IC. It therefore has units phyloP/bits and converts I_i to a phyloP score. Based on manual inspection of phyloP patterns across TOPs at different test statistic thresholds, we selected a threshold of L < -10 to be considered "conserved" according to this test, which manual inspection indicated is conservative.

2, Correlation Test: For each TF, we permuted the position-wise IC of the PWM using the *permute* R package, up to a maximum of 1,000 unique permutations (not every PWM has 1,000 unique permutations). We then took the Pearson correlation of each of these permuted PWM IC vectors using the phyloP scores of 1,500 randomly selected PWM hits from the unfiltered PWM scan results (or fewer, if there are <1,500 total hits). This resulted in a maximum of 1,500,000 correlations per TF, dependent upon the number of unique PWM IC permutations and number of PWM hits. We used these correlation values as a null distribution, converted to Z scores, and determined a

threshold correlation value corresponding to an alpha of 0.05; this threshold was chosen manually. PWM hits with a Pearson correlation to the unpermuted motif IC values greater than this threshold were considered to have passed this test.

3, Wilcoxon Test*:* For each TOP site, we performed two Wilcoxon tests, one comparing values in the PWM hits to those in the 25bp downstream flank, and the same for the 25bp upstream flank. All p-values were FDR corrected, and an FDR-corrected p-value less than a threshold of 0.1 for both flanks was considered a positive.

A TOP PWM hit was considered conserved if it passed one of the three conservation tests above, and had at least one site with an FDR-corrected phyloP score > 1 (i.e. corresponding to an FDR-corrected p-value < 0.1).

Determination of binding site ages. We used halLiftover[99] to map TOP PWM hits to syntenic loci in all other genomes in the Zoonomia 241-mammal alignment[96], including 9 reconstructed genomes ancestral to human, and calculated the alignment's % identity to the human sequence. We then assigned an age using multiple criteria. For the method used in the main text, we identified the oldest ancestral genome with a gapless alignment of any % identity to the human PWM hit. **Supplementary Figure S6** shows alternative schemes for age inference, including the oldest extant *species* with a gapless alignment to a human TOP (at various threshold % identities), or the oldest *clade* wherein 60% of species have a gapless alignment to the human TOP (at various threshold % identities). We acquired the age of each clade from TimeTree[58].

Determination of TF ages. To infer TF ages, we acquired all vertebrate ortholog annotations and ortholog quality statistics for each TF from Ensembl[100], and ages of each pair of species from TimeTree[58]. The age of a TF was taken as the oldest ortholog annotated as having a 1-1 relationship with human, or having a gene order conservation (GOC, a metric of synteny) score >= 50 and classification as a high-confidence ortholog by Ensembl.

Repeat enrichment. To calculate enrichment for each TF and repeat pair, we identified the intersections of the peak summits from ChIP-seq peaks (or TOPs) and the middle position of GHT-SELEX TOPs with the 2022-10-18 version of the UCSC Genome Browser RepeatMasker track[101]. The enrichment significance between GHT-SELEX and ChIP-seq TOPs and each repeat family was calculated using Fisher's Exact Test implemented in SciPy[102]*.* The contingency table took the form of:

	TF y summit +	**TF y summit -**
Repeat x +	# of repeat x and TF y summit intersections	# of repeat x bases with no TF y summit intersections
Repeat x -	# of TF y summits with no repeat x intersections	# of repeat-annotated bases, excluding # of repeat x bases and TF y summits

DATA AVAILABILITY

The sequencing raw data for the experiments have been deposited into the SRA database under identifiers PRJEB78913 (ChIP-seq), PRJEB76622 (GHT-SELEX), and PRJEB61115 (HT-SELEX). Genomic interval information generated for the ChIP-seq and GHT-SELEX have been deposited into GEO under accessions GSE280248 and GSE278858, respectively. Information on constructs, experiments, analyses, processed data, comparison tracks, with many accessory files and browsable results is available at https://codebook.ccbr.utoronto.ca. Larger collection of motifs generated for these experiments in an accompanying study[36] can be browsed at https://mex.autosome.org and downloaded at https://doi.org/10.5281/ZENODO.8327372.

ACKNOWLEDGEMENTS

We thank the IT Group of the Institute of Computer Science at Halle University for computational resources and Maximilian Biermann for valuable technical support. This work was supported by the following:

- Canadian Institutes of Health Research (CIHR) grants FDN-148403, PJT-186136, PJT-191768, and PJT-191802, and NIH grant R21HG012258 to T.R.H.
- CIHR grant PJT-191802 to T.R.H. and H.S.N.
- Natural Sciences and Engineering Research Council of Canada (NSERC) grant RGPIN-2018-05962 to H.S.N.
- Russian Science Foundation grant 20-74-10075 to I.V.K.
- Russian Science Foundation grant 24-14-20031 to F.A.K.
- Swiss National Science Foundation grant (no. 310030_197082) to B.D.
- Marie Skłodowska-Curie (no. 895426) and EMBO long-term (1139-2019) fellowships to J.F.K.
- NIH grants R01HG013328 and U24HG013078 to M.T.W., T.R.H., and Q.M.
- NIH grants R01AR073228, P30AR070549, and R01AI173314 to M.T.W.
- NIH grant P30CA008748 partially supported Q.M.
- Canada Research Chairs funded by CIHR to T.R.H. and H.S.N.
- Ontario Graduate Scholarships to K.U.L and I.Y.
- A.J. was supported by Vetenskapsrådet (Swedish Research Council) Postdoctoral Fellowship (2016-00158)
- The Billes Chair of Medical Research at the University of Toronto to T.R.H.
- EPFL Center for Imaging
- Institutional funding from EPFL
- Resource allocations from the Digital Research Alliance of Canada

SUPPLEMENTAL TABLES AND DOCUMENTS

Table S1. Overview of the tested TFs. This table lists the TFs that were tested in this study using ChIP-seq.

Table S2. List of all the ChIP-seq experiments. This table lists the ChIP experiments, their approval status, and related produced files.

Table S3. List of 217 TFs. This table lists the TFs with either “approved” ChIP-seq experiments or significant overlap between replicates, together with the list of ChIP-seq samples used in “merged” peaks for each TF.

Table S4. Binding category of the TFs (i.e. Promoter TFs, Enhancer TF, Dark TF, and Others) for 217 TFs with successful ChIP-seq experiments. For the TFs with available GHT-SELEX data (hence TOP sites, the number of optimized ChIP-seq peaks (i.e. Triple peaks), number of TOP ChIP-seq peaks, fraction of direct binding sites (i.e. #TOP peaks divided by #Triple peaks), number of TOPs, number of CTOPs, fraction of conserved TOPs (i.e. #CTOPs / #TOPs), and the median age of the TOP sites are also included. Note that the number of TOP ChIP-seq peaks might be different (less) than TOPs (referring to triple-overlap PWM hits), since each peak might comprise multiple PWM hits.

Table S5. Consolidated functional information for Dark TFs. This table provides the data underlying **Figure 7** including the references in the literature.

Document S1. Heatmaps of conservation/phyloP score across TOPs for 137 TFs. This document provides the same analysis as **Figure 4A,** for all TFs of the study, heatmaps of phyloP scores in PWM hits (middle column) and flanking sequences of tops are displayed. Bars to the right indicate which tests of conservation are satisfied (Likelihood-ratio, Correlation, Wilcoxon), along with overlaps with promoters (P) and specific repeat families if applicable.

REFERENCES

1. Partridge, E.C. *et al.* Occupancy maps of 208 chromatin-associated proteins in one human cell type. *Nature* **583**, 720-728 (2020).
2. Long, H.K., Prescott, S.L. & Wysocka, J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**, 1170-1187 (2016).
3. Lambert, S.A. *et al.* The Human Transcription Factors. *Cell* **175**, 598-599 (2018).
4. Sullivan, P.F. *et al.* Leveraging base-pair mammalian constraint to understand genetic variation and human disease. *Science* **380**, eabn2937 (2023).
5. Lenhard, B. *et al.* Identification of conserved regulatory elements by comparative genome analysis. *J Biol* **2**, 13 (2003).
6. Gumucio, D.L. *et al.* Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. *Mol Cell Biol* **12**, 4919-29 (1992).
7. Dermitzakis, E.T. & Clark, A.G. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol* **19**, 1114-21 (2002).
8. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476-82 (2011).
9. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
10. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304-51 (2001).
11. Nobrega, M.A., Ovcharenko, I., Afzal, V. & Rubin, E.M. Scanning human gene deserts for long-range enhancers. *Science* **302**, 413 (2003).
12. Nobrega, M.A., Zhu, Y., Plajzer-Frick, I., Afzal, V. & Rubin, E.M. Megabase deletions of gene deserts result in viable mice. *Nature* **431**, 988-93 (2004).
13. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-93 (2009).
14. Cosby, R.L. *et al.* Recurrent evolution of vertebrate transcription factors by transposase capture. *Science* **371**(2021).
15. Schmidt, D. *et al.* Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**, 335-48 (2012).
16. Chuong, E.B., Elde, N.C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083-7 (2016).
17. Cohen, C.J., Lock, W.M. & Mager, D.L. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene* **448**, 105-14 (2009).
18. Ernst, J. *et al.* Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat Biotechnol* **34**, 1180-1190 (2016).
19. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42**, 631-4 (2010).
20. Jacobs, F.M. *et al.* An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* **516**, 242-5 (2014).
21. Bannister, A.J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410**, 120-4 (2001).

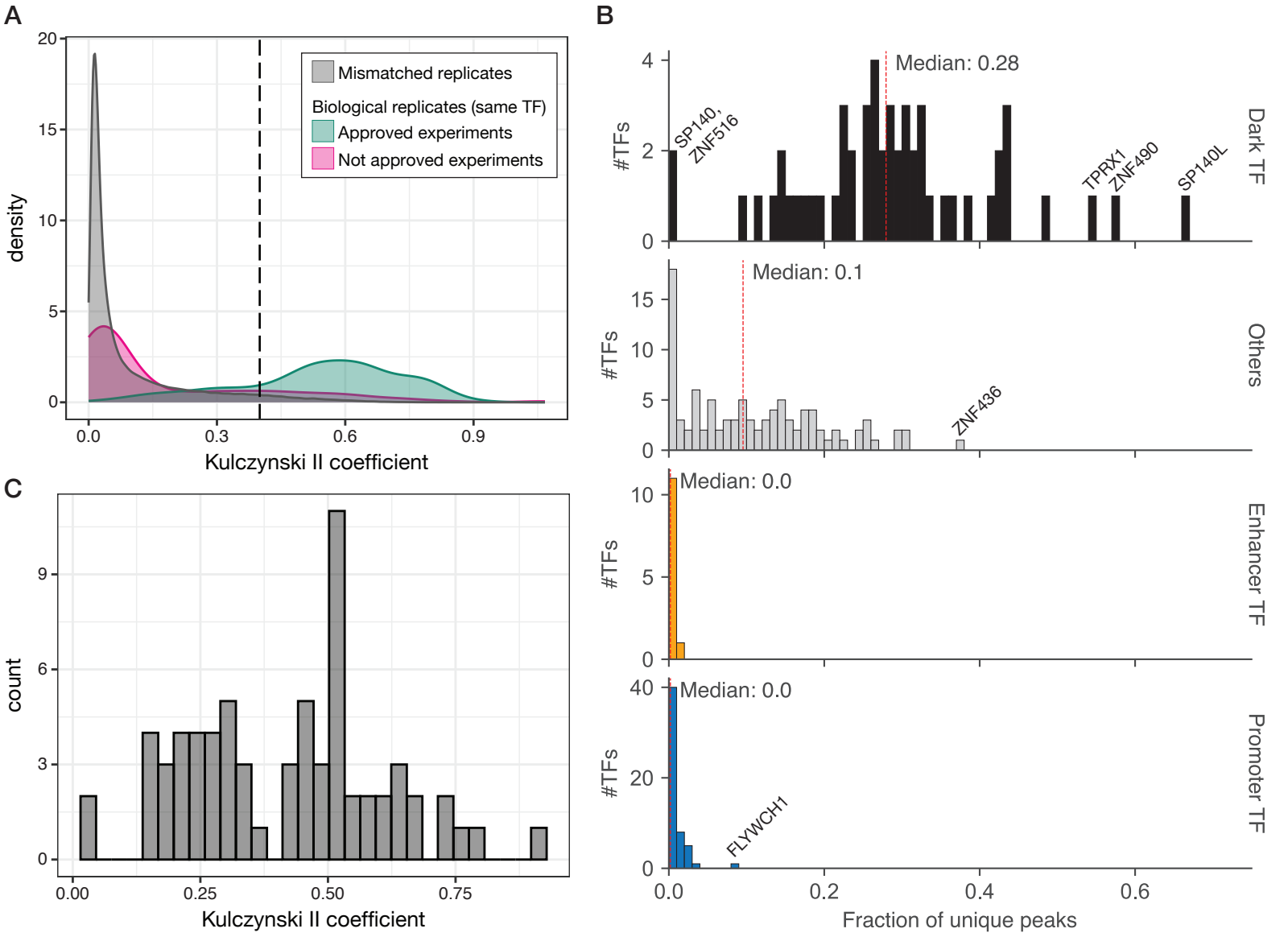
22. Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**, 116-20 (2001).
23. Ayyanathan, K. *et al.* Regulated recruitment of HP1 to a euchromatic gene induces mitotically heritable, epigenetic gene silencing: a mammalian cell culture model of gene variegation. *Genes Dev* **17**, 1855-69 (2003).
24. Stubbs, L., Sun, Y. & Caetano-Anolles, D. Function and Evolution of C2H2 Zinc Finger Arrays. *Subcell Biochem* **52**, 75-94 (2011).
25. Jolma, A. *et al.* DNA-Binding Specificities of Human Transcription Factors. *Cell* **152**, 327-39 (2013).
26. Imbeault, M., Helleboid, P.Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550-554 (2017).
27. Schmitges, F.W. *et al.* Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Res* **26**, 1742-1752 (2016).
28. Najafabadi, H.S., Albu, M. & Hughes, T.R. Identification of C2H2-ZF binding preferences from ChIP-seq data using RCADE. *Bioinformatics* **31**, 2879-81 (2015).
29. Worsley Hunt, R. & Wasserman, W.W. Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. *Genome Biol* **15**, 412 (2014).
30. Auerbach, R.K. *et al.* Mapping accessible chromatin regions using Sono-Seq. *Proc Natl Acad Sci U S A* **106**, 14926-31 (2009).
31. Jolma, A. *et al.* Perspectives on Codebook: sequence specificity of uncharacterized human transcription factors. *bioRxiv*, 2024.11.11.622097 (2024).
32. Lambert, S.A. *et al.* The Human Transcription Factors. *Cell* **172**, 650-665 (2018).
33. Isakova, A. *et al.* SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nat Methods* **14**, 316-322 (2017).
34. Najafabadi, H.S. *et al.* C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat Biotechnol* (2015).
35. Kean, M.J. *et al.* Structure-function analysis of core STRIPAK Proteins: a signaling complex implicated in Golgi polarization. *J Biol Chem* **286**, 25065-75 (2011).
36. Vorontsov, I.E. *et al.* Cross-platform DNA motif discovery and benchmarking to explore binding specificities of poorly studied human transcription factors. *bioRxiv*, 2024.11.11.619379 (2024).
37. Berger, M.F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* **24**, 1429-35 (2006).
38. Jolma, A. *et al.* Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res* **20**, 861-73 (2010).
39. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
40. O'Leary, N.A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-45 (2016).
41. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215-6 (2012).
42. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* **2017**(2017).

43. Jolma, A. *et al.* GHT-SELEX demonstrates unexpectedly high intrinsic sequence specificity and complex DNA binding of many human transcription factors. *bioRxiv*, 2024.11.11.618478 (2024).
44. Badis, G. *et al.* A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol Cell* **32**, 878-87 (2008).
45. Core, L.J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet* **46**, 1311-20 (2014).
46. FitzGerald, P.C., Shlyakhtenko, A., Mir, A.A. & Vinson, C. Clustering of DNA sequences in human promoters. *Genome Res* **14**, 1562-74 (2004).
47. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110-21 (2010).
48. Barazandeh, M., Lambert, S.A., Albu, M. & Hughes, T.R. Comparison of ChIP-Seq Data and a Reference Motif Set for Human KRAB C2H2 Zinc Finger Proteins. *G3 (Bethesda)* **8**, 219-229 (2018).
49. Thomas, J.H. & Schneider, S. Coevolution of retroelements and tandem zinc finger genes. *Genome Res* **21**, 1800-12 (2011).
50. Nowick, K. *et al.* Gain, loss and divergence in primate zinc-finger genes: a rich resource for evolution of gene regulatory differences between species. *PLoS One* **6**, e21553 (2011).
51. Nowick, K., Hamilton, A.T., Zhang, H. & Stubbs, L. Rapid sequence and expression divergence suggest selection for novel function in primate-specific KRAB-ZNF genes. *Mol Biol Evol* **27**, 2606-17 (2010).
52. Alexander, T.B. *et al.* The genetic basis and cell of origin of mixed phenotype acute leukaemia. *Nature* **562**, 373-379 (2018).
53. Arber, D.A. *et al.* International Consensus Classification of Myeloid Neoplasms and Acute Leukemias: integrating morphologic, clinical, and genomic data. *Blood* **140**, 1200-1228 (2022).
54. Iouranova, A. *et al.* KRAB zinc finger protein ZNF676 controls the transcriptional influence of LTR12-related endogenous retrovirus sequences. *Mob DNA* **13**, 4 (2022).
55. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554-66 (2015).
56. Christmas, M.J. *et al.* Evolutionary constraint and innovation across hundreds of placental mammals. *Science* **380**, eabn3943 (2023).
57. Harrison, P.W. *et al.* Ensembl 2024. *Nucleic Acids Res* **52**, D891-D899 (2024).
58. Kumar, S. *et al.* TimeTree 5: An Expanded Resource for Species Divergence Times. *Mol Biol Evol* **39**(2022).
59. Helleboid, P.Y. *et al.* The interactome of KRAB zinc finger proteins reveals the evolutionary history of their functional diversification. *EMBO J* **38**, e101220 (2019).
60. Huttlin, E.L. *et al.* Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell* **184**, 3022-3040 e28 (2021).
61. Silva, F.P., Hamamoto, R., Furukawa, Y. & Nakamura, Y. TIPUH1 encodes a novel KRAB zinc-finger protein highly expressed in human hepatocellular carcinomas. *Oncogene* **25**, 5063-70 (2006).

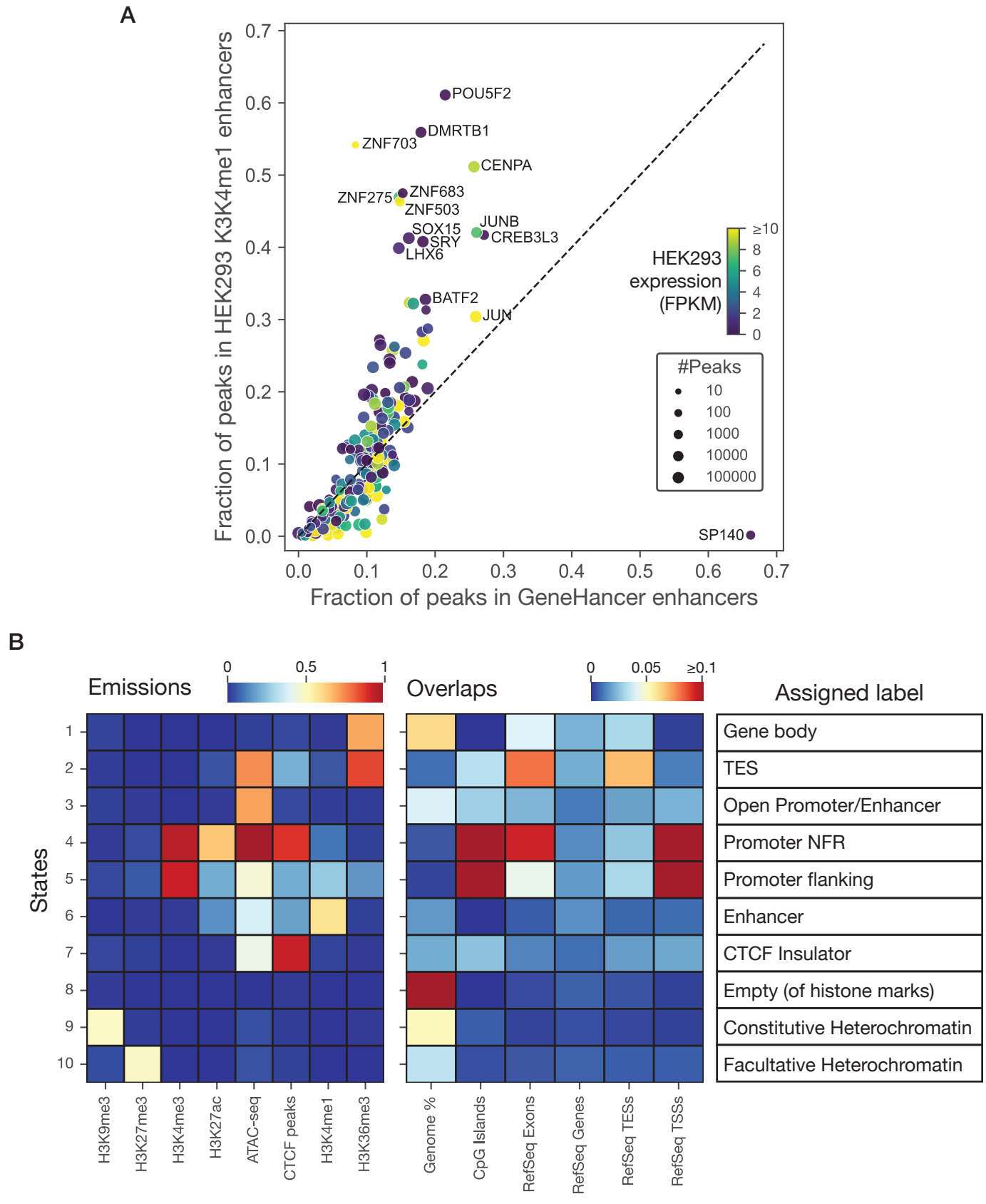
62. Kim, J.J. *et al.* Systematic bromodomain protein screens identify homologous recombination and R-loop suppression pathways involved in genome integrity. *Genes Dev* **33**, 1751-1774 (2019).
63. Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci* **30**, 187-200 (2021).
64. Agricola, E., Randall, R.A., Gaarenstroom, T., Dupont, S. & Hill, C.S. Recruitment of TIF1gamma to chromatin via its PHD finger-bromodomain activates its ubiquitin ligase and transcriptional repressor activities. *Mol Cell* **43**, 85-96 (2011).
65. Maier, V.K. *et al.* Functional Proteomic Analysis of Repressive Histone Methyltransferase Complexes Reveals ZNF518B as a G9A Regulator. *Mol Cell Proteomics* **14**, 1435-46 (2015).
66. Li, L. *et al.* ZNF516 suppresses EGFR by targeting the CtBP/LSD1/CoREST complex to chromatin. *Nat Commun* **8**, 691 (2017).
67. Huggenvik, J.I. *et al.* Characterization of a nuclear deformed epidermal autoregulatory factor-1 (DEAF-1)-related (NUDR) transcriptional regulator protein. *Mol Endocrinol* **12**, 1619-39 (1998).
68. Fraschilla, I. & Jeffrey, K.L. The Speckled Protein (SP) Family: Immunity's Chromatin Readers. *Trends Immunol* **41**, 572-585 (2020).
69. Tumber, A. *et al.* Potent and Selective KDM5 Inhibitor Stops Cellular Demethylation of H3K4me3 at Transcription Start Sites and Proliferation of MM1S Myeloma Cells. *Cell Chem Biol* **24**, 371-380 (2017).
70. Huttlin, E.L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425-440 (2015).
71. Soufi, A., Donahue, G. & Zaret, K.S. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. *Cell* **151**, 994-1004 (2012).
72. Zou, Z. *et al.* Translatome and transcriptome co-profiling reveals a role of TPRXs in human zygotic genome activation. *Science* **378**, abo7923 (2022).
73. Kuroda, T. *et al.* SALL3 expression balance underlies lineage biases in human induced pluripotent stem cell differentiation. *Nat Commun* **10**, 2175 (2019).
74. Singh, J.K. *et al.* Zinc finger protein ZNF384 is an adaptor of Ku to DNA during classical non-homologous end-joining. *Nat Commun* **12**, 6560 (2021).
75. Feu, S. *et al.* OZF is a Claspin-interacting protein essential to maintain the replication fork progression rate under replication stress. *FASEB J* **34**, 6907-6919 (2020).
76. Ge, L.P. *et al.* ZNF689 deficiency promotes intratumor heterogeneity and immunotherapy resistance in triple-negative breast cancer. *Cell Res* **34**, 58-75 (2024).
77. Zhou, M. *et al.* ZBTB40 is a telomere-associated protein and protects telomeres in human ALT cells. *J Biol Chem* **299**, 105053 (2023).
78. Cui, Y., Zhou, M., He, Q. & He, Z. Zbtb40 Deficiency Leads to Morphological and Phenotypic Abnormalities of Spermatocytes and Spermatozoa and Causes Male Infertility. *Cells* **12**(2023).
79. Campos-Sanchez, R., Kapusta, A., Feschotte, C., Chiaromonte, F. & Makova, K.D. Genomic landscape of human, bat, and ex vivo DNA transposon integrations. *Mol Biol Evol* **31**, 1816-32 (2014).

80. Burton, K.A. *et al.* Haploinsufficiency at the protein kinase A RI alpha gene locus leads to fertility defects in male mice and men. *Mol Endocrinol* **20**, 2504-13 (2006).
81. Stielow, B. *et al.* The SAM domain-containing protein 1 (SAMD1) acts as a repressive chromatin regulator at unmethylated CpG islands. *Sci Adv* **7**(2021).
82. Lord, T. *et al.* A novel high throughput screen to identify candidate molecular networks that regulate spermatogenic stem cell functionsdagger. *Biol Reprod* **106**, 1175-1190 (2022).
83. Schmidt, D. *et al.* ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. *Methods* **48**, 240-8 (2009).
84. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-9 (2012).
85. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
86. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**(2021).
87. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X.S. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* **7**, 1728-40 (2012).
88. Liu, T. Use model-based Analysis of ChIP-Seq (MACS) to analyze short reads generated by sequencing protein-DNA interactions in embryonic stem cells. *Methods Mol Biol* **1150**, 81-95 (2014).
89. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-2 (2010).
90. Grandi, F.C., Modi, H., Kampman, L. & Corces, M.R. Chromatin accessibility profiling by ATAC-seq. *Nat Protoc* **17**, 1518-1552 (2022).
91. Consortium, E.P. *et al.* Perspectives on ENCODE. *Nature* **583**, 693-698 (2020).
92. Frankish, A. *et al.* GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res* **51**, D942-D949 (2023).
93. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-80 (2012).
94. Satopaa, V., Albrecht, J., Irwin, D. & Raghavan, B. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. in *2011 31st international conference on distributed computing systems workshops* 166-171 (IEEE, 2011).
95. Korhonen, J., Martinmaki, P., Pizzi, C., Rastas, P. & Ukkonen, E. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics* **25**, 3181-2 (2009).
96. Armstrong, J. *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246-251 (2020).
97. Ramirez, F., Dundar, F., Diehl, S., Gruning, B.A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42**, W187-91 (2014).
98. Amemiya, H.M., Kundaje, A. & Boyle, A.P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**, 9354 (2019).
99. Hickey, G., Paten, B., Earl, D., Zerbino, D. & Haussler, D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341-2 (2013).
100. Martin, F.J. *et al.* Ensembl 2023. *Nucleic Acids Res* **51**, D933-D941 (2023).
101. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* **16**, 418-20 (2000).

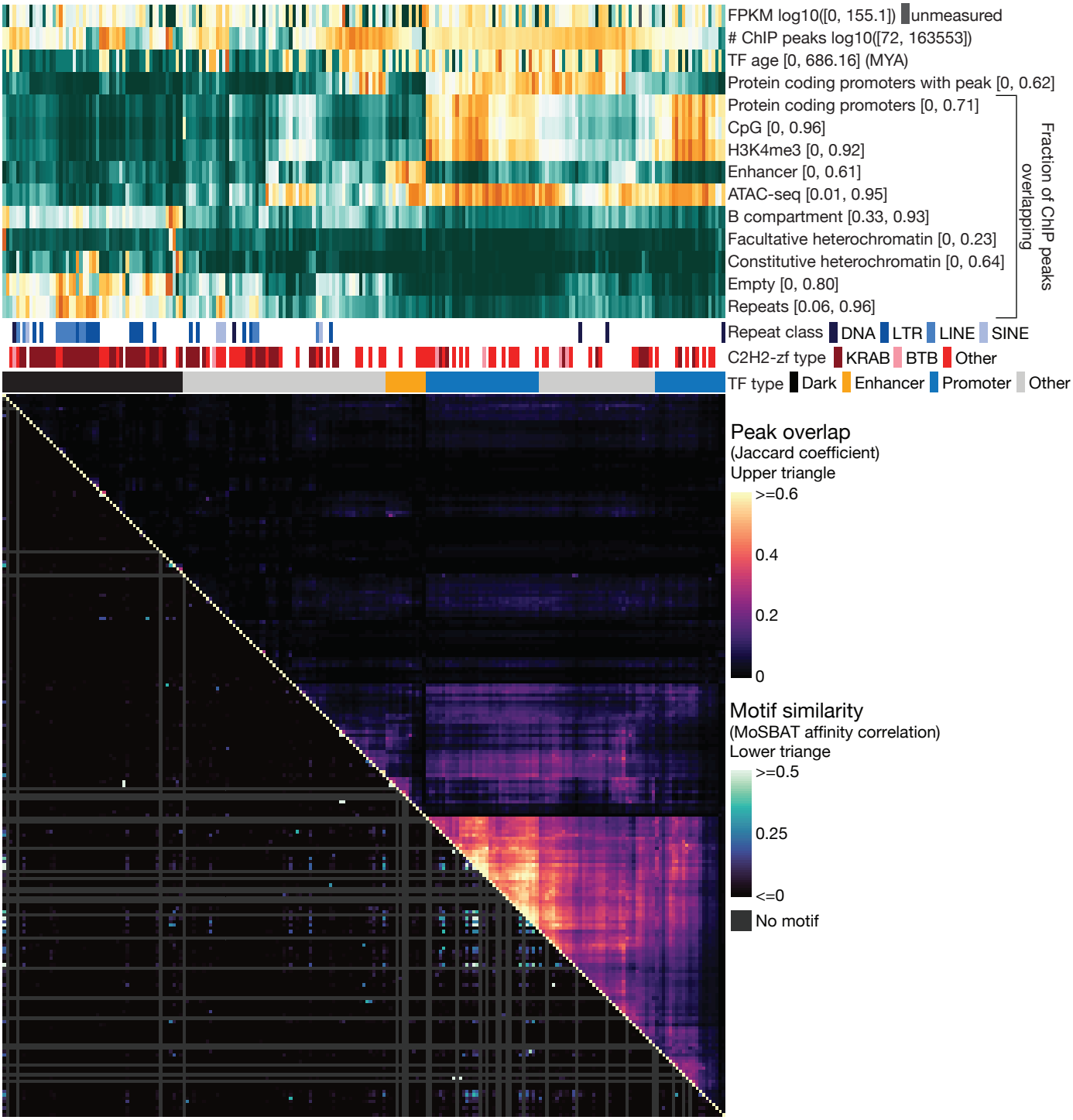
102. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**, 261-272 (2020).
103. Storer, J., Hubley, R., Rosen, J., Wheeler, T.J. & Smit, A.F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA* **12**, 2 (2021).
104. Lambert, S.A., Albu, M., Hughes, T.R. & Najafabadi, H.S. Motif comparison based on similarity of binding affinity profiles. *Bioinformatics* **32**, 3504-3506 (2016).



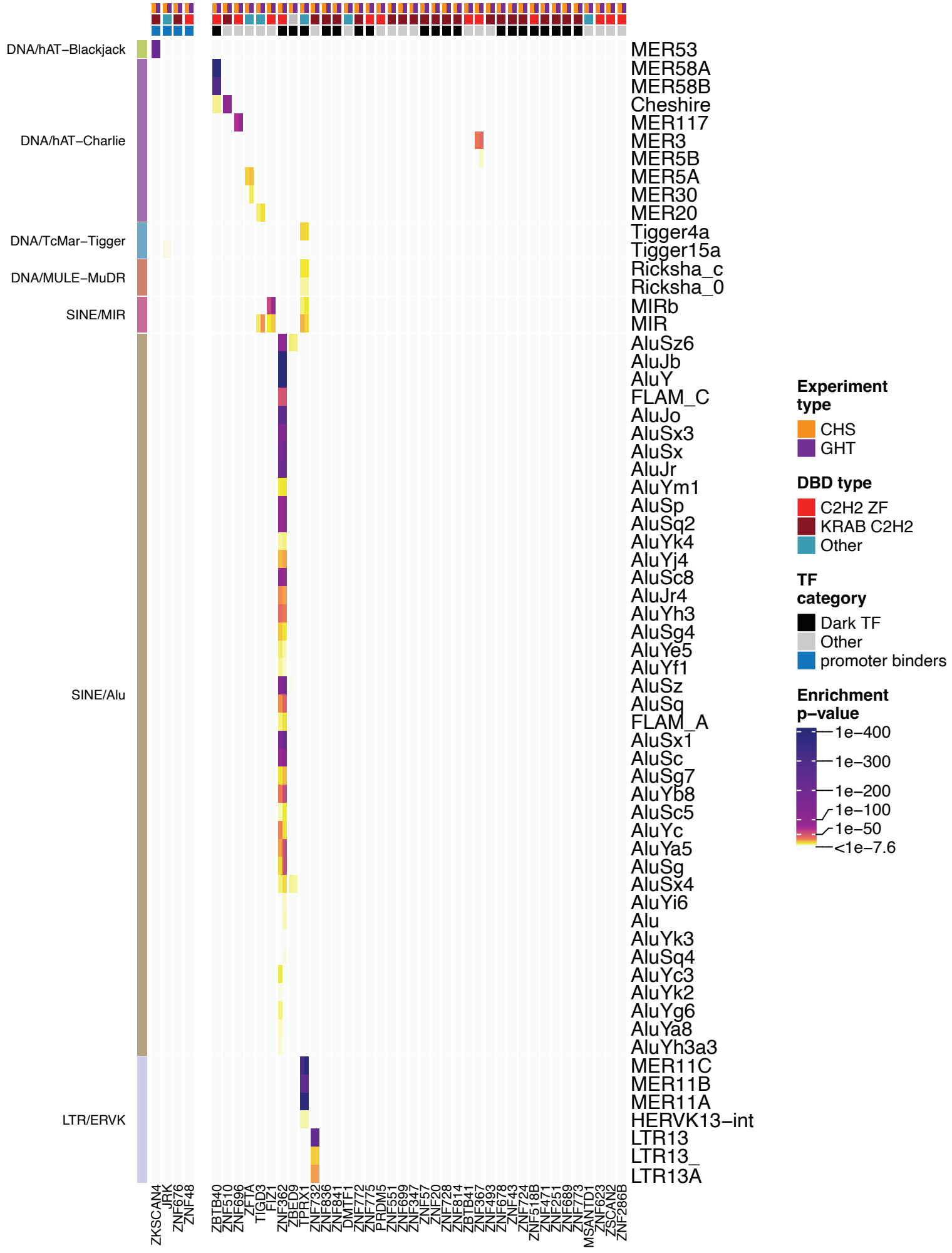
Supplementary Figure S1. Evaluation of ChIP-seq success by peak overlap assessment between experimental replicates. (**A**) Distribution of peak overlap between ChIP-seq replicates, for approved experiments (i.e., produced a motif), not approved experiments (i.e., did not produce a motif), and mismatch replicates (i.e., TF identities permuted), calculated by Kulczynski II similarity metric (i.e. average of overlaps). The dotted line indicates the threshold at which pairs of not approved experiments were considered successful and thus could be included in downstream analyses. (**B**) Distribution of the uniqueness of peaks for different categories of TFs, measured as the fraction of ChIP-seq peaks (at the universal threshold) not overlapping with *any* peak from any other TF in this study. (**C**) Distribution of Kulczynski II similarity metric between ChIP-seq replicates (as in (**A**)), restricted to the TFs that have a low peak overlap with other TFs (specifically, the 94 TFs in the upper left darker region of the square matrix in **Figure 2B**).

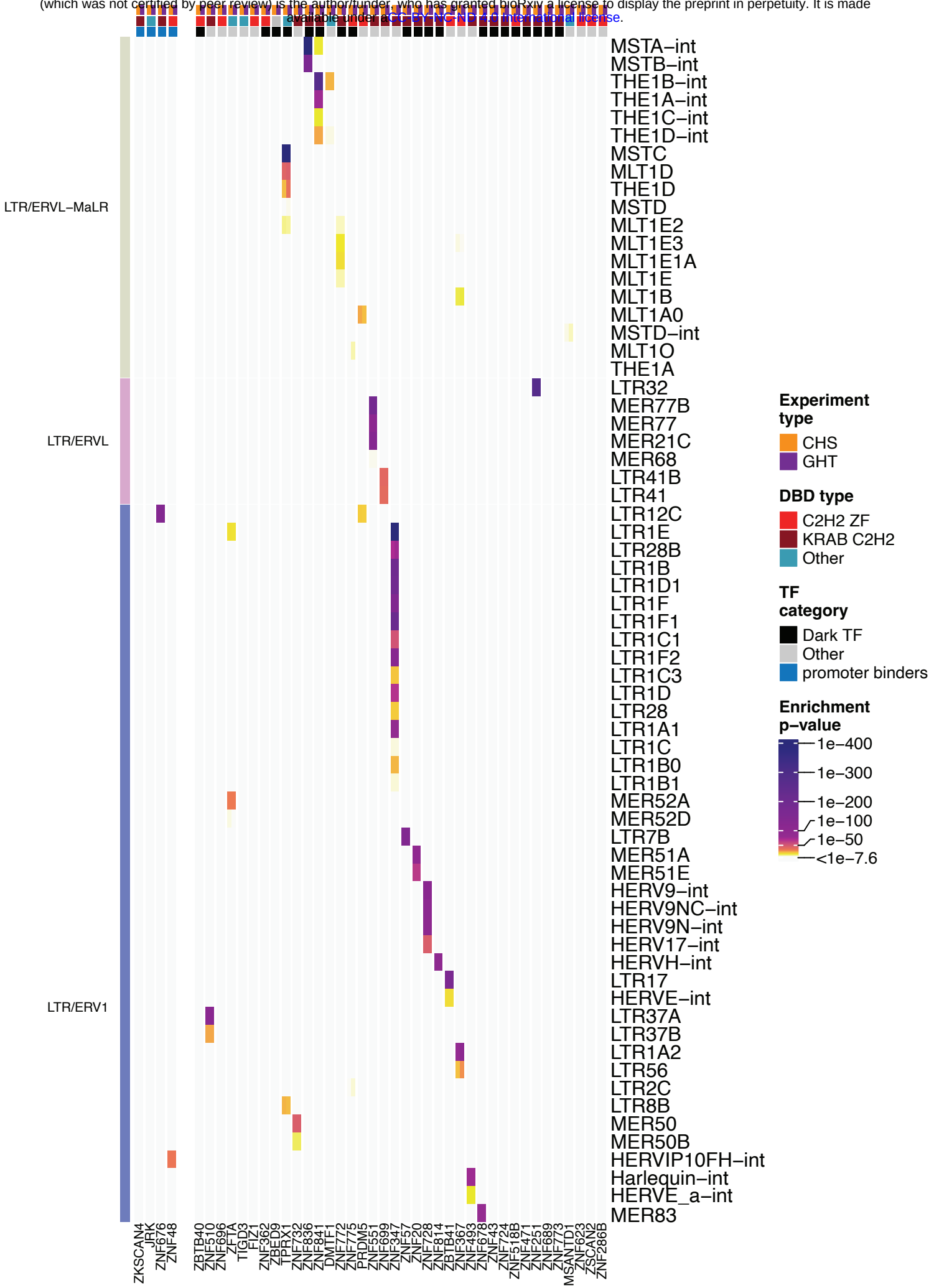


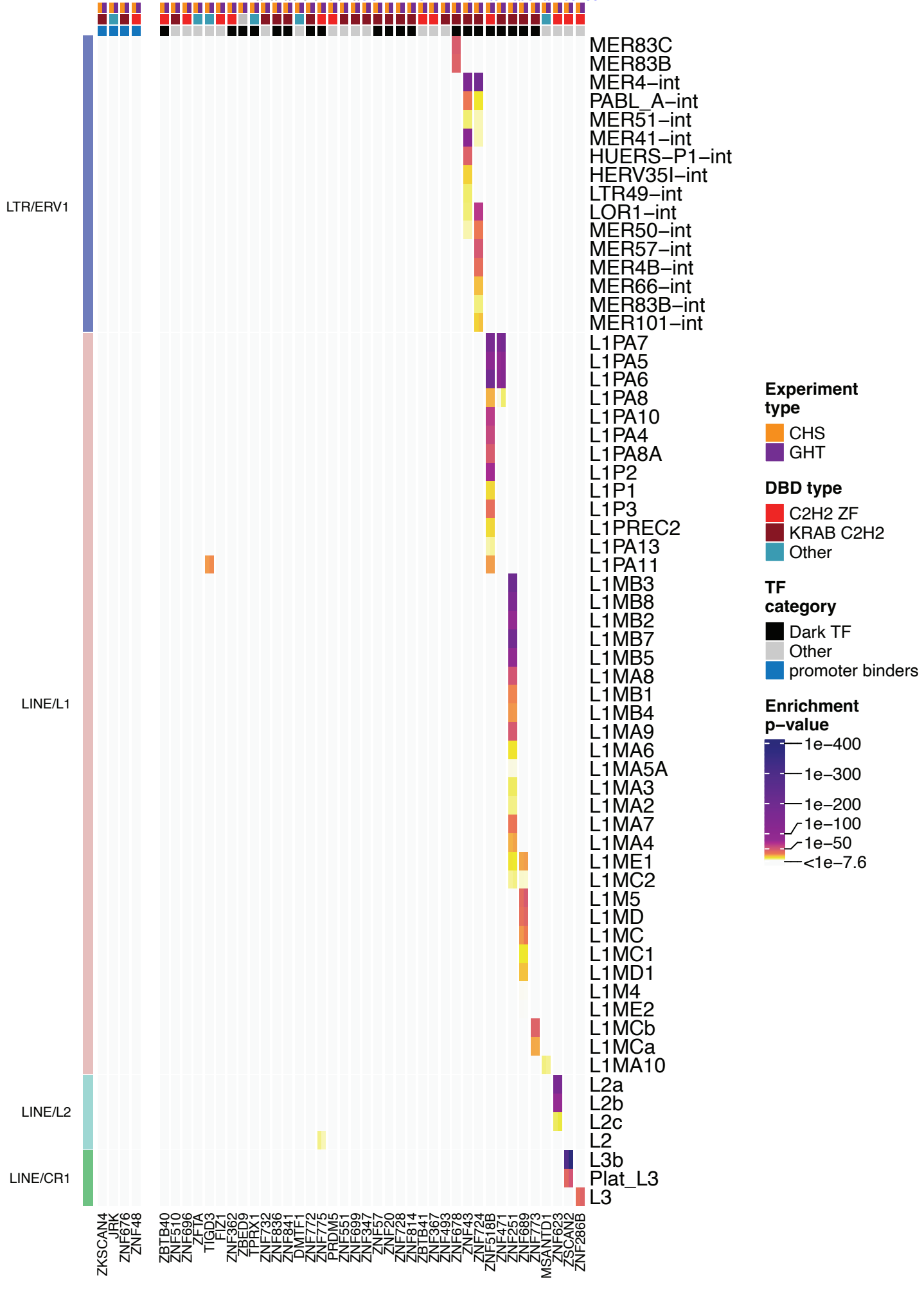
Supplementary Figure S2. Overlap of ChIP-seq peaks with different enhancer sets and ChromHMM tracks. (**A**) Fraction of ChIP-seq peaks overlapping with GeneHancer annotated enhancers (x-axis) and HEK293 enhancers (defined by H3K4me1-positive regions from ChromHMM; y-axis). Points (TFs) are scaled based on their number of peaks. Colors also display the expression of TFs in HEK293 cells[27]. (**B**) Characterization of the states of a ChromHMM model with 10 states trained on various HEK293 chromatin data (i.e., H3K9me3, H3K27me3, H3K4me1, H3K4me3, H3K36me3, and H3K27ac from ENCODE, and ATAC-seq and CTCF peaks from this study). Based on the correspondence between emissions and the chromatin marks and genome annotations, the states were assigned to Gene body, TES, Open Promoter/Enhancer, Promoter NFR (nucleosome-free regions), Promoter flanking, Enhancer, CTCF Insulator, Empty (of histone marks), Constitutive Heterochromatin, and Facultative Heterochromatin.



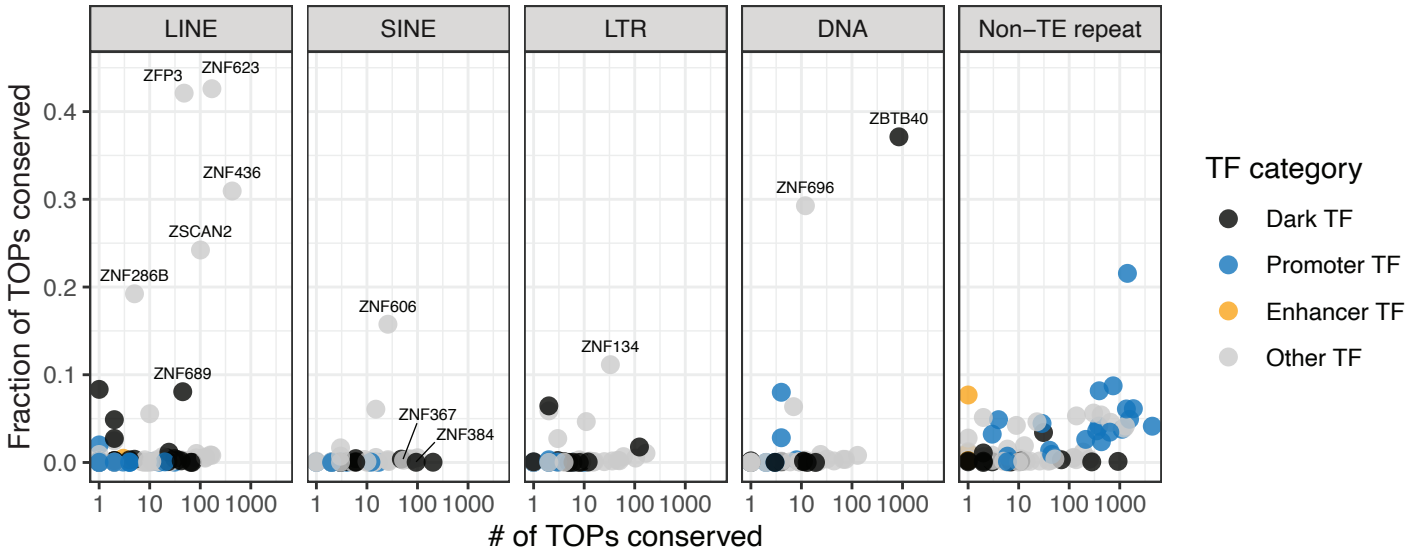
Supplementary Figure S3. A detailed version of **Figure 2** including additional tracks, such as gene expression in HEK293 cells (FPKM)[27], number of total ChIP-seq peaks (at the universal threshold of MACS2 P-value≤10^{-10}), TF age, fraction of human protein-coding promoters (out of 20,052) covered by TF peaks, fraction of ChIP-seq peaks falling within: CpG islands, H3K4me3-positive regions, facultative heterochromatin, and constitutive heterochromatin, with the main repeat class bound by the TFs included. The upper triangle in the bottom square is the same as **Figure 2**, however, the lower triangle here is the similarity between PWMs for each pair of TFs, calculated by MoSBAT[104]. Gray stripes correspond to the TFs without a selected PWM in the Codebook set.



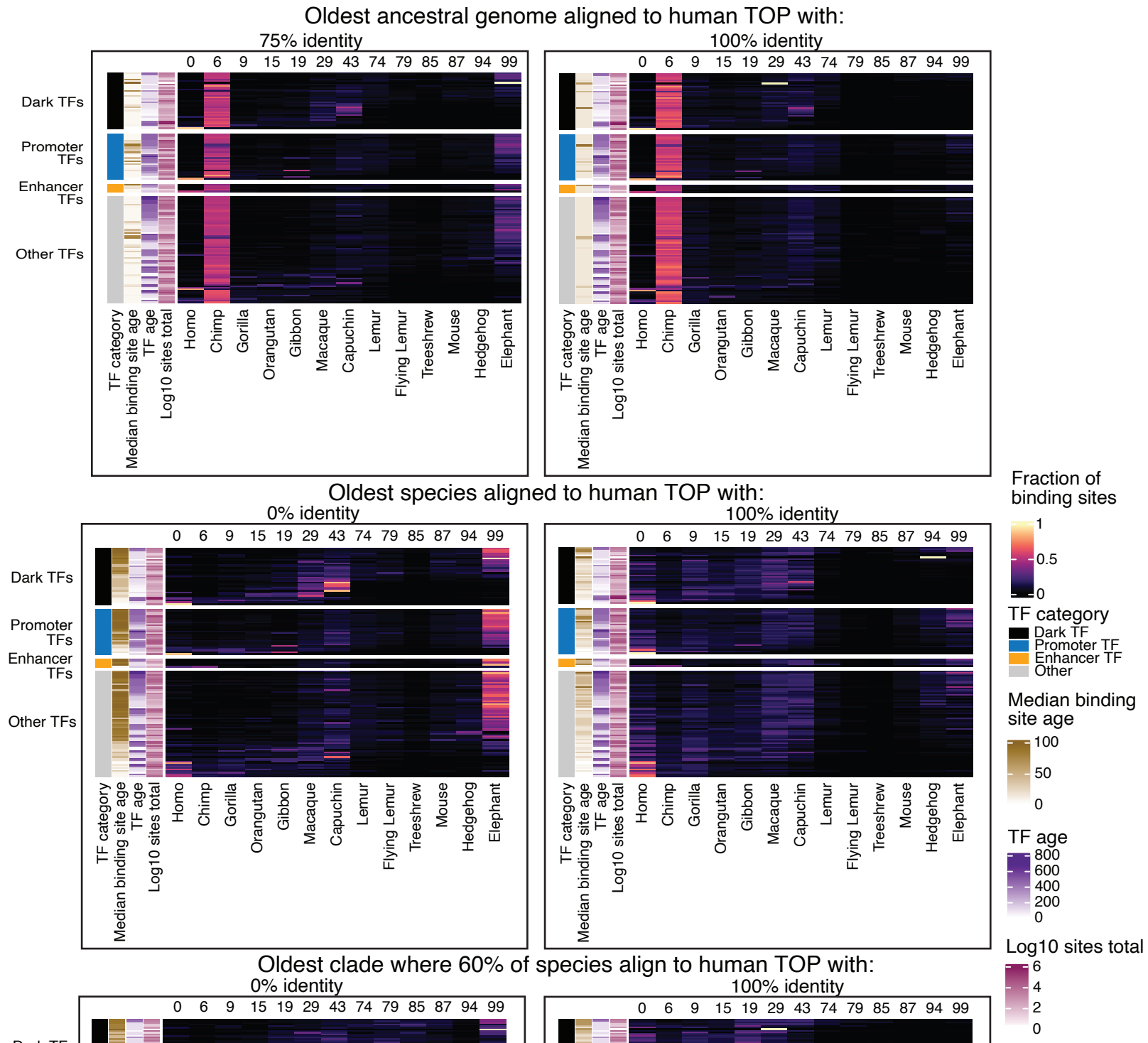




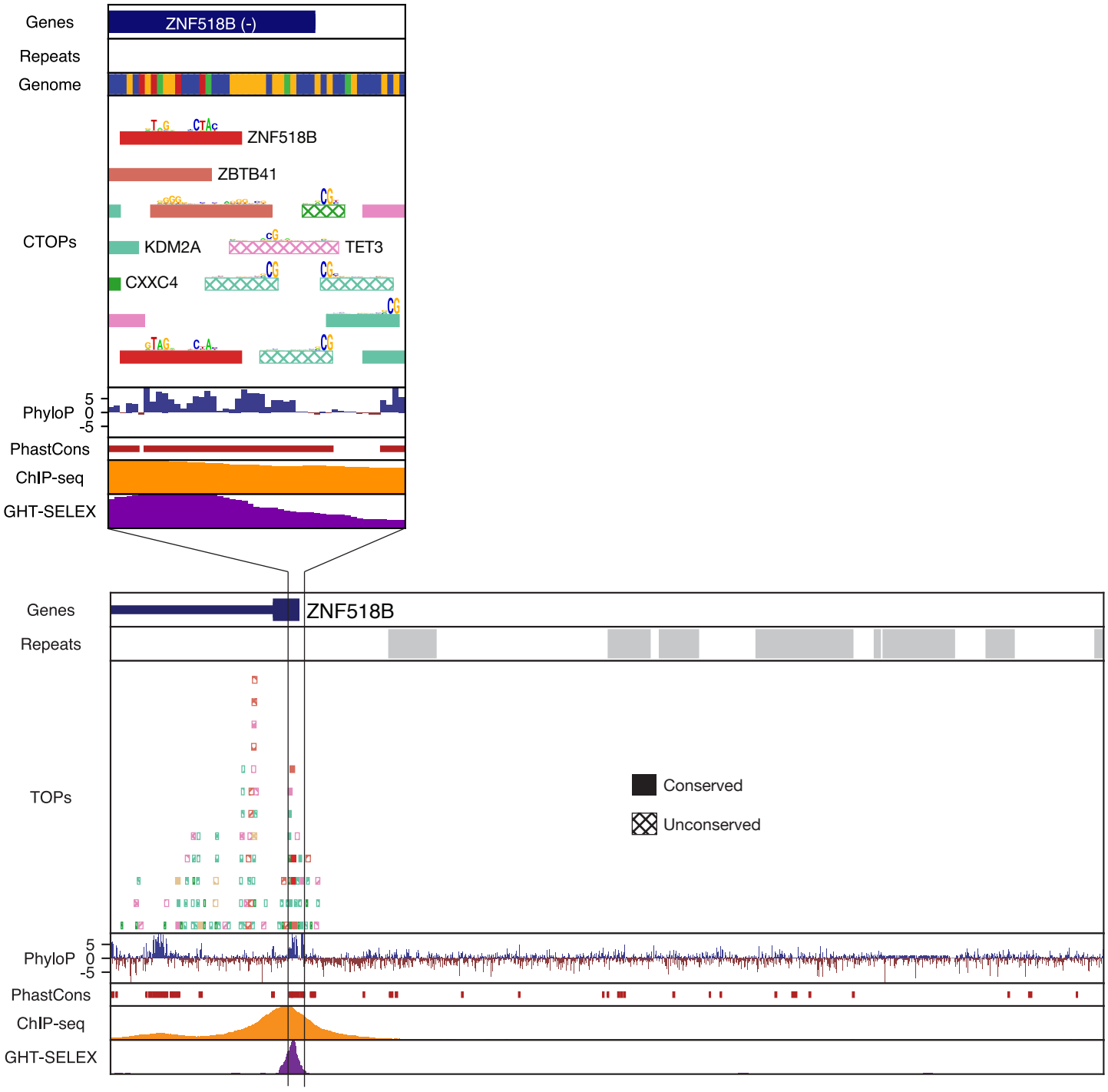
Supplementary Figure S4. Enrichment of transposable elements in TOPs, with expanded TE family classification. Heatmap is from **Figure 6**, with expanded labels for specific elements enriched in TOPs of each TF.



Supplementary Figure S5. Conservation of the binding sites for repeat-binding TFs. Plots showing the fraction of each TF's TOPs that are conserved (i.e. 'CTOPs') and overlap a major class of transposable elements or non-TE repeats. The proportion of TOPs that are conserved and overlap a repeat class is shown on the y-axis, and the log10 count of these sites is shown on the x-axis. Each TF is coloured according to its classification as a Dark TF, Promoter TF, Enhancer and Other TFs. Only proteins with a fraction greater than 0.1 of conserved TOPs that fall in a repeat class are labeled. TFs discussed in the main text are also labeled.



Supplementary Figure S6. Estimation of binding site age using three different test and two different thresholds. Heatmaps show the proportion of each TF's TOPs (rows) inferred to be a certain age, as in Figure 5, but with each panel utilizing a different scheme. *Top row*: Age of each TOP site inferred as that of oldest ancestral genome with a gapless alignment to the human TOP site and minimum 75% identity (left) or 100% identity (right). (**Figure 5** shows this same analysis with a 0% identity threshold). *Middle row*: Age of each TOP site inferred as that of oldest *species* with a gapless alignment to the human TOP site and minimum 0% identity (left) or 100% identity (right). *Bottom row*: Age of each TOP site inferred as that of the oldest *clade* where 60% of the species have a gapless alignment to the human TOP with a minimum 0% identity (left) or 100% identity (right).



Supplementary Figure S7. A conserved binding site of ZNF518B as a potential self-regulator. Conserved binding sites for ZNF518B (red) located in the promoter of ZNF518B itself, and in a predicted enhancer-region ~4kb upstream of its promoter. Binding sites for ZBTB41, KDM2A, TET3, and CXXC4 are also present in this region.