Extensive binding of nebulous 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[4,5,6], Philipp Bucher[7], Quaid Morris[2], Hamed S. Najafabadi[3,8], and Timothy R. Hughes[1]**

[1]Donnelly Centre and Department of Molecular Genetics, 160 College Street, Toronto, ON M5S 3E1, Canada
[2]Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 25 10065, USA
[3]Victor P. Dahdaleh Institute of Genomic Medicine, Montréal, QC 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]Institute of Biochemistry and Genetics, Ufa Federal Research Centre of Russian Academy of Sciences, 450054, Ufa, Russia
[7]Swiss Institute of Bioinformatics, 1015, Lausanne, Switzerland
[8]Department of Human Genetics, McGill University, Montréal, QC H3A 0C7, Canada

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

The Codebook consortium

Philipp Bucher[7], Bart Deplancke[7,9], Oriol Fornes[10], Jan Grau[11], Ivo Grosse[11], Timothy R. Hughes[1], Arttu Jolma[1], Fedor A. Kolpakov[12,13], Ivan V. Kulakovskiy[4,5,6], Vsevolod J. Makeev[4,14], Mihai Albu[1], Marjan Barazandeh[1], Alexander Brechalov[1], Zhenfeng Deng[1], Ali Fathi[1], Chun Hu[1], Samuel A. Lambert[1], Kaitlin U. Laverty[1,2], Zain M. Patel[1], Sara E. Pour[1], Rozita Razavi[1], Mikhail Salnikov[1], Ally W.H. Yang[1], Isaac Yellan[1], Hong Zheng[1], Georgy Meshcheryakov[5], Giovanna Ambrosini[15,16], Antoni J. Gralak[7,9], Sachi Inukai[17], Judith F. Kribelbauer-Swietek[7,9], Marie-Luise Plescher[11], Semyon Kolmykov[12], Ivan Yevshin[18], Nikita Gryzunov[5], Ivan Kozin[4,19], Mikhail Nikonov[19], Vladimir Nozdrin[19], Arsenii Zinkevich[19], Katerina Faltejskova[20,21], Pavel Kravchenko[22], Sergey Abramov[4,23], Alexandr Boytsov[4,23], Vasilii Kamenets[4], Dmitry Penzar[4,19], Anton Vlasov[24], Ilya E. Vorontsov[4], Aldo Hernandez-Corchado[3], Hamed S. Najafabadi[3,8], Quaid Morris[2], Xiaoting Chen[25], Matthew T. Weirauch[25]

[9] Laboratory of Systems Biology and Genetics, Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne, 1015, Lausanne, Switzerland
[10] Department of Medical Genetics, Centre for Molecular Medicine and Therapeutics, BC Children's Hospital Research Institute, University of British Columbia, Vancouver, BC V5Z 4H4, Canada
[11] Institute of Computer Science, Martin Luther University Halle-Wittenberg, 06099, Halle, Germany
[12] Department of Computational Biology, Sirius University of Science and Technology, 354340, Sirius, Krasnodar region, Russia
[13] Bioinformatics Laboratory, Federal Research Center for Information and Computational Technologies, 630090, Novosibirsk, Russia
[14] Current address: Cancer Research UK National Biomarker Centre, University of Manchester, M20 4BX, Manchester, UK
[15] Bioinformatics Competence Center, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
[16] Bioinformatics Competence Center, Université de Lausanne, 1015 Lausanne, Switzerland
[17] Chugai Pharmaceutical Co., Ltd, Tokyo, 103-8324, Japan
[18] Biosoft.Ru LLC, 630058, Novosibirsk, Russia
[19] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, 119991, Moscow, Russia
[20] Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, 160 00 Praha 6, Czech Republic
[21] Computer Science Institute, Faculty of Mathematics and Physics, Charles University, 118 00 Praha 1, Czech Republic
[22] Max Planck Institute of Biochemistry, 82152, Planegg, Germany
[23] Altius Institute for Biomedical Sciences, 98121, Seattle, WA, USA
[24] TUD Dresden University of Technology, Center for Molecular and Cellular Bioengineering (CMCB), Biotechnologisches Zentrum (BIOTEC), Tatzberg 47/49, 01307 Dresden, Germany
[25] Cincinnati Children's Hospital, Cincinnati, OH 45229, USA

SUMMARY

The functional impact of a large portion of the human genome known as "dark matter DNA", which is composed mainly of repeat sequences, remains enigmatic. The genome also encodes hundreds of putative and poorly characterized transcription factors (TFs). Here, we determined genomic binding locations of 166 poorly characterized human TFs in living cells. Nearly half of them associate strongly with known regulatory regions such as promoters and enhancers, frequently co-localizing with each other at conserved motif matches. The other half often associate with genomic dark matter, however, at largely non-overlapping (i.e., unique) sites, via intrinsic sequence recognition. Fifty-four of the latter half, which we term "Dark TFs", mainly bind within regions of closed chromatin, with each recognizing a unique set of repeat sequences. The Dark TFs include many KZNFs, which are known to bind and silence TEs, and other TFs with apparent repressive functions. By contrast, some may be pioneers: we find that induction of TPRX1, a known regulator of zygotic preimplantation, leads to chromatin opening at many of its binding sites in the dark matter genome. Altogether, our results shed light on a large fraction of poorly characterized human TFs and simultaneously illuminate the diversity of function within the dark matter genome.

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

INTRODUCTION

For decades, the term “dark matter”, borrowed from astronomy and cosmology, has been applied to describe the excess of seemingly non-functional regions of many genomes[1]. These regions are dominated by repetitive sequences such as transposable elements (TEs) and endogenous retroelements (EREs), which largely account for the extreme differences in genome size across eukaryotes. There are many potential roles for genomic dark matter[1,2], but it is also conceivable that it is mainly inert or simply provides spacing. An initial proposal for the function of the dark matter was gene regulation[3], and indeed, while some so-called “gene deserts” appear to be dispensable, others contain transcriptional enhancers[4-6]. Function of genomic dark matter in gene regulation is also supported by the existence of transcription factors (TFs) that are capable of binding TE and ERE sequences[7].

It is well-established that TEs can rewire regulatory circuitry by introducing transcription factor binding sites, thus spawning novel cis-regulatory elements[8-10]. In some cases, these elements derive from the promoter of the TE (e.g., endogenous retrovirus long terminal repeats (ERV LTRs))[11,12]. Other cases may represent inadvertent matches to a TF motif within a TE[13]; over evolutionary time, regulatory sites tend to turn over rapidly, arising seemingly at random, and TEs/EREs are a known contributor to this process[14]. In addition, mammalian genomes encode a large family of TFs, the KRAB-containing C2H2 zinc-finger (KZNF) class (~350 members in human), which has evolved rapidly in parallel to ERE classes bound by its members[15]. Different KZNFs typically bind to different specific subclasses of ERE in ChIP-seq[7,16]. The KRAB domain is best known for recruiting KAP1/TRIM28, which associates physically with both readers (e.g., HP1/CBX proteins) and writers (SETDB1)[17,18] of the H3K9me3 mark that defines constitutive heterochromatin[19,22]. Silencing EREs, presumably to limit their spread, appears to be the major driver of KZNF evolution, but the fact that so many KZNFs have been retained, often tens of millions of years after the EREs have been reduced to molecular fossils, suggest that the KZNFs often contribute to adaptive evolution[20,21]. At least some have clearly taken on regulatory roles that benefit the genome in which they reside[21,22].

There are also many poorly characterized human TFs that, like the dark matter genome, remain nebulous in regard to DNA-binding, which is the defining property of TFs: 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 have no known DNA binding motif[23]. TFs overall are best known for their association with promoters and/or enhancers[24,25], although the largest class in human is the KZNFs, which appear primarily to act as repressors. TFs also vary greatly in their ability to bind nucleosomal DNA[26], and only a subset of TFs, termed “pioneers”, can access binding sequences within closed chromatin and recruit or allow access to other TFs[27]. There are many possible mechanisms of cooperation between TFs and TF binding sites, of which pioneering is just one example[28]. Presumably as a consequence, decoding gene regulation is a long-standing challenge in molecular biology and computational genomics, which is compounded by the fact that the sequence preferences of many TFs remain unknown.

145 In addition, even for TFs with known motifs, it is uncertain whether the motif is an
146 accurate representation of the TF's sequence preferences. The KZNFs, for example,
147 are known for the potential to have very long binding sites, enabled by their long C2H2-
148 zf domain arrays[29], and the DNA binding preferences of these proteins have proven
149 difficult to characterize precisely due to generally poor functionality in biochemical
150 assays. ChIP-seq reveals what families of EREs are bound by C2H2-zf proteins, but it is
151 difficult to accurately determine their precise sequence specificity because the repeat
152 elements are related by common descent, which confounds motif discovery and
153 testing[30]. In addition, *in vivo* measurements include the effects of indirect recruitment,
154 nonspecific binding, and chromatin[31,32]. For KZNFs, it has not been proven that the
155 sequence specificity of the C2H2-zf array on its own is sufficient to identify specific ERE
156 classes within the genome: the predictions from scans with existing motifs yield an
157 excess of matches outside EREs[33].

158 As part of an international initiative termed the "Codebook consortium", aimed at
159 obtaining binding motifs for all human TFs[34], we analyzed 315 human TFs (i.e., with
160 unknown DNA sequence specificity) by ChIP-seq in HEK293 cells, together with 58
161 controls. Roughly 60% of these TFs (n=217) yielded reliable results in our study (i.e.,
162 reproducible and/or enriched for TF motif matches). We evaluated the data in
163 conjunction with other data from the Codebook project, which includes binding to the
164 human genome *in vitro*[35], and use of multiple *in vitro* data types to define binding
165 motifs[34,36], thus allowing base-level identification of direct binding sites in ChIP-seq.
166 Previous ChIP-seq analyses have focused mainly on preferential association of TFs
167 with promoters vs. enhancers[24], and consistent with this expectation, roughly half of the
168 TFs with successful data in our study recognize such sites. Surprisingly, however, the
169 other half of the Codebook TFs, including most KZNFs as well as other TF families,
170 bound to apparently unique sites dispersed across the genome that lack active
171 chromatin marks prior to introduction of the TF. Fifty-four of these TFs bind mainly to
172 sequences located in closed chromatin. We refer to this subset as "Dark TFs" and show
173 that most bind preferentially to specific subfamilies of TEs and EREs. Our further
174 investigation into the binding sites for three Dark TFs revealed significant
175 reconfiguration of chromatin architecture at these genomic locations. Multiple lines of
176 evidence suggest diverse biochemical, cellular, and physiological functions of the Dark
177 TFs, and by extension, shed light on potential functions of the dark matter genome.

178 **RESULTS**

179 **Generation of ChIP-seq data for hundreds of putative TFs**

180 We surveyed the genomic binding sites of 315 poorly characterized, putative human
181 TFs (the "Codebook" set, derived from Lambert 2018[37], and described in detail
182 elsewhere[34]), and 58 previously characterized TFs as controls (selected from Isakova
183 2017[38] and Schmitges 2016[33]), using ChIP-seq in HEK293 cells (**Fig. 1a**,
184 **Supplementary Table 1**). We used an established inducible eGFP-tagged transgene
185 system (**Fig. 1b**) previously employed in ChIP-seq and protein-protein interaction
186 studies[16,33,39]. Of particular relevance here, this system was effective at revealing
187 specific binding of KZNFs to classes of retroelements, confirming that the system can

188 readily assess binding to inactive or repressed regions of the genome[16,33]. The present
189 study includes biological ChIP replicates for each TF performed by different
190 experimentalists, such that the resulting dataset includes 678 ChIP-seq experiments for
191 Codebook TFs, and 112 experiments for control TFs (**Supplementary Table 1**). A full
192 list of experiments is given in **Supplementary Table 2**. Representative motifs obtained
193 from control TFs of various families illustrate that the assay recovers known sequence-
194 binding preferences, as expected (**Fig. 1c** and **Supplementary Fig. 1**).

195 We used two criteria to determine which experiments were successful. First, the data
196 were analyzed as part of a larger Codebook benchmarking effort, which is described in
197 more detail in accompanying manuscripts[34,40]. The Codebook benchmarking included
198 expert curation that relied mainly on obtaining similar motifs for the same TF from
199 different *in vivo* and *in vitro* data types (ChIP-seq, Protein Binding Microarrays[41], SMiLE-
200 seq[38], and several variants of HT-SELEX[42]) as evidence of direct, sequence-specific
201 DNA binding. This Codebook motif benchmarking identified 130 Codebook TFs and 49
202 controls with successful, motif-enriched ChIP-seq data, meaning that sequence-specific
203 DNA binding observed in ChIP-seq (*in vivo*) is supported in almost all cases by *in vitro*
204 experimental data (**Supplementary Table 1**).

205 Second, we identified experiments in which the peak overlaps of biological replicates
206 exceeded what is expected at random (i.e., with TF identities permuted). ChIP-seq
207 experiments that were classified as successful based on the motif similarity analysis
208 described above displayed a higher overlap between TF replicates relative to
209 mismatched TFs (median Kulczynski II coefficient of 0.57 vs. 0.034; **Supplementary**
210 **Fig. 2a**; this statistic is a modified Jaccard value that compensates for class imbalance).
211 A Kulczynski II coefficient threshold of 0.4 captures 79% of approved experiments with
212 replicates, and 90% of controls, but eliminates 94.5% of mismatched experiments; thus,
213 we used this value as a threshold for success of replicate experiments, independent of
214 motifs. By this measure, an additional 36 putative TFs (and two controls) were
215 considered technically successful, despite not displaying motif enrichment
216 (**Supplementary Tables 1** and **2**). The DNA binding in these cases may be indirect,
217 i.e., the proteins may be DNA-associated chromatin factors, rather than TFs.
218 Alternatively, they may recognize properties of the DNA sequence that are not captured
219 by common motif models. Nevertheless, we included these 38 TFs in subsequent
220 analyses. Cautioning that these proteins may be chromatin factors, for simplicity, we
221 refer to the entire set of 217 successful proteins (130 putative + 49 control TFs
222 successful in the first round, and 36 putative + 2 controls with matching replicates;
223 **Supplementary Fig. 3a** and **Supplementary Table 3**) as "TFs" in the text.

224 We merged the peaks from TF replicates among the 427 ChIP-seq experiments
225 deemed successful to produce a dataset for downstream analysis (**Supplementary**
226 **Table 3**). This dataset encompasses 217 proteins (166 Codebook putative TFs, and 51
227 controls), with a median of 12,200 peaks per protein (range 7-145,470) (using a MACS
228 threshold of $P<10^{-10}$; see **Methods** for explanation of threshold choice)[43].

229 **ChIP-seq data illustrate that half of the Codebook TFs bind genomic dark matter**

230 To begin characterizing the ChIP-seq data, we surveyed for preferential association of
231 the putative TFs with promoters and/or enhancers. **Fig. 2a** shows the fraction of peaks
232 from each protein that overlaps with protein-coding promoters (defined by RefSeq[44])
233 and enhancers (defined by HEK293 chromatin state[45], and mainly carrying H3K4me1
234 signal; see **Methods** and **Supplementary Fig. 4a**). Indeed, many proteins are highly
235 associated with promoters, and a smaller number with enhancers. On average, the
236 HEK293-derived enhancer set yielded higher overlaps than the larger, more universal
237 "GeneHancer" set[46] (**Supplementary Fig. 4b**), indicating that this lower number of
238 enhancer-favoring TFs (relative to promoter-favoring) is not due to incomplete enhancer
239 annotations. Moreover, for control TFs with publicly available endogenous data in
240 multiple cell types[47], the fraction of peaks overlapping GeneHancer enhancers, HEK293
241 enhancers, or promoters shows a consistent trend with Codebook data, suggesting that
242 the low number of enhancer-enriched TFs is not particularly due to HEK293-specific
243 chromatinization (**Supplementary Fig. 4c-e**). Surprisingly, a large fraction of proteins
244 did not associate with either promoters or enhancers (**Fig. 2a)**.

245 To gain further insight into the properties of the ChIP-seq binding sites, we compared
246 the peak sets for each of the proteins to all other proteins in the dataset, along with an
247 extensive panel of genome annotations, protein type, and expression levels (**Fig. 2b**
248 and **Supplementary Fig. 5)**. **Fig. 2b** (bottom) shows a symmetric heatmap of Jaccard
249 values (intersection/union) between the peak sets of the 217 successful TFs, providing
250 an overview of similarities in genomic localization. The heatmap at the top of **Fig. 2b**
251 shows the fraction of each corresponding TF peak set that overlaps with each type of
252 genome annotation. The chromatin states were derived mainly from public-domain data
253 for unperturbed HEK293 cells; we therefore expect them to reflect the state of the
254 chromosomes prior to induction of the tagged TF transgene. This state could be
255 involved in recruiting the TF, but it could also result from endogenous expression of the
256 native, untagged protein, as most of the studied TFs are already expressed in HEK293
257 cells[33] (**Supplementary Fig. 5**).

258 **Fig. 2b** reflects and expands upon trends observed in **Fig. 2a**. The large bright square
259 in the lower right quadrant of the bottom heatmap corresponds to TFs that associate
260 primarily with open chromatin (ATAC-seq) and H3K4me3. These TFs also often
261 associate with many promoters (median 8,813 coding gene promoters; see below),
262 leading to high overlap between the peak sets. The observation that many TFs co-bind
263 promoters and/or enhancers is prevalent in the literature (e.g.[24]); we note, however, that
264 the Codebook proteins were not previously known to bind these sequences, and
265 therefore it appears that even well-known regulatory elements often contain previously
266 unidentified TF binding sites.

267 A second main feature of **Fig. 2b** is the diagonal line in the upper left quadrant of the
268 heatmap. These are cases in which each peak set has very little overlap with any other
269 peak set. In addition, the individual peaks bound by each TF are often unique, i.e., they
270 do not overlap with any peak from any other protein in the Codebook ChIP-seq dataset
271 (**Supplementary Fig. 2b**). The unique binding profiles are not explained by
272 experimental error or random events since there is strong peak overlap between

273 replicates for the same protein (**Supplementary Fig. 2c**), and these proteins often bind
274 to the same unique sequences both *in vivo* and *in vitro* (see below).

275 The ChIP-seq peaks for these proteins also tend to be outside open chromatin and
276 outside of either promoters or apparent enhancers in HEK293 cells (**Fig. 2b,** *top*). In
277 fact, roughly half of these TFs' peak sets are either enriched for marks that characterize
278 heterochromatin or lack any of the diagnostic marks of promoter or enhancer activity
279 (i.e., "empty" ChromHMM regions, **Supplementary Fig. 4a**). Peaks for a subset of
280 these TFs (n=87) are mainly associated with the Hi-C "B" compartment, and many
281 associate with specific classes of repeats, particularly TEs (**Fig. 2b** and see below).
282 This is an expected property of KZNFs, which constitute more than half (48/87) of these
283 TFs[16] (bar in the middle of **Fig. 2b**).

284 **Classification of TFs based on genomic annotations of binding sites**

285 We sought to gain a better understanding of the properties and characteristics of TFs
286 (and their binding sites) that represent the major patterns shown in **Fig. 2**. As a starting
287 point, we defined four mutually exclusive groups (see **Supplementary Table 3** for TF
288 labels) based on the dominant annotation types that overlap with their genomic binding
289 sites (i.e., ChIP-seq peaks). One group we named "Promoter TFs" (56 proteins); for
290 these proteins, more than 45% of peaks exclusively overlap with promoters (this
291 threshold captures the most prominent features in **Fig. 2a**). Another group was
292 designated "Enhancer TFs" (12 proteins); for these proteins, >37% of peaks overlap
293 with enhancers (this threshold corresponds to the visual separation of data points on the
294 vertical axis in **Fig. 2a**). A third group we named "Dark TFs", after the genomic "dark
295 matter" (54 proteins); for these, most peaks lie within either the "empty", "constitutive
296 heterochromatin", or "facultative heterochromatin" states (i.e. in HEK293 cells, they are
297 outside of the states that represent promoters, enhancers, insulators, or gene bodies),
298 and fewer than half of the peaks overlap with ATAC-seq peaks in unperturbed HEK293
299 cells. The remaining ~40% of TFs we labeled as "Other" (i.e., not Promoter, Enhancer,
300 or Dark TFs); they include 30 control TFs and 65 Codebook proteins, which display
301 diverse combinations of binding sites within the defined categories that do not reach our
302 classification thresholds. The "Other TFs" thus present a rich landscape for additional
303 exploration, but we did not attempt to further subclassify them here.

304 The choice of features and thresholds to define the categories of TFs was to readily
305 draw distinctions between TFs via interpretable binding behaviors. Unbiased clustering
306 of the TFs using all the tracks presented on the top in **Fig. 2b** results in the nearly same
307 groups, but with less stringency (i.e. some "Other" TFs being labeled as Promoter TFs,
308 Enhancer TFs, or Dark TFs, **Supplementary Fig. 6**). Our thresholds do exclude some
309 TFs that may associate significantly with specific regions but also bind many other
310 locations; for example, the control TF YY1 bound 58% of all promoters in human, but it
311 also had 40,764 additional binding sites (72% of all its binding sites) outside promoters.
312 Similarly, some KZNFs were not classified as Dark TFs because they had many binding
313 sites within open chromatin.

314 Since the Promoter TFs and Dark TFs bind distinct and prominent genomic features
315 and contain most of the Codebook TFs, we focused mainly on contrasting these two
316 groups for the remainder of our analyses. It is important to note that the Promoter and
317 Dark TF classifications could not be accounted for by technical artifacts despite the
318 known bias of ChIP-seq towards open chromatin[32]: the fraction of reads in peaks (FRiP)
319 for Dark TFs is comparable to those of Promoter TFs (**Supplementary Fig. 7a**),
320 indicating that there is no systematic difference between the quality of the peaks called
321 within promoter regions or elsewhere. Moreover, the absence of peaks within empty or
322 heterochromatin regions for Promoter TFs or other non-Dark TFs is not due to the lack
323 of read coverage in those regions: for all Promoter TFs (except CXXC4), >85% of ChIP-
324 seq reads are outside promoters (**Supplementary Fig. 7b**). In addition, the
325 reproducibility of the Dark TF peaks is comparable to that of the Promoter TF peaks
326 (median optimized Jaccard between the replicates is 0.59 for Dark TFs, and 0.55 for
327 Promoter TFs) (**Supplementary Table 3**, **Supplementary Fig. 2a**). **Supplementary**
328 **Fig. 7c** illustrates that, in many cases, the imperfect overlap between peak sets is at
329 least partially accounted for by peaks near the threshold, which pass the cutoff in one
330 experiment but not the other. Altogether, both Promoter and Dark TF peak sets are
331 technically robust by conventional measures.

332 **Recognition of repeat families by Dark TFs is direct and sequence-specific**

333 Dark TF ChIP-seq peaks in general tend to overlap genomic repeats, and 36 of the 54
334 Dark TFs are KZNFs, which are expected to bind to specific ERE classes (**Fig. 2b**). We
335 further interrogated Dark TFs' peaks to ask whether each TF displays preferences for
336 specific repeat types. Indeed, in addition to a large fraction of peaks overlapping
337 repeats, there is a clear tendency for the ChIP-seq peak sets of each Dark TF to display
338 strong statistical enrichment with at least one specific and unique repeat family, mainly
339 corresponding to TEs and EREs (**Fig. 3a**). In contrast, many fewer Promoter TF peaks
340 overlap repeats, and the most enriched repeat is typically "G-rich", which itself is
341 enriched at promoters (2,425 in promoters out of 13,493 in the genome).

342 ChIP-seq can detect both direct binding (i.e., the TF intrinsically recognizes the bound
343 DNA sequences) and indirect binding (e.g., recruitment by another factor)[31]. ChIP-seq
344 also readily detects non-specific DNA-binding (e.g., by histones), and is biased towards
345 open chromatin since the sonication step preferentially releases these regions[32]. To
346 determine whether the ChIP-seq peaks are due to direct sequence-specific binding of
347 the TFs to the genome, and to accurately identify these direct binding sites, we
348 employed data from a novel assay, GHT-SELEX (Genomic HT-SELEX; described in
349 detail in [48]), which was available from the Codebook project. GHT-SELEX surveys the
350 binding of synthetic TFs to fragmented, purified, and unmodified genomic DNA *in vitro*,
351 yielding peaks that resemble those from ChIP-seq. The GHT-SELEX data differs from
352 ChIP-seq in several ways, however: 1) it has greater resolution, due to the smaller DNA
353 fragment lengths (~64 bases), and 2) GHT-SELEX represents only the intrinsic binding
354 of the TF to naked genomic DNA, devoid of chromatin and other cellular cofactors,
355 which can modulate binding *in vivo*. As a result, only a subset of the ChIP-seq and
356 GHT-SELEX peaks typically overlap (see below). Nonetheless, the general properties of
357 binding sites across the genome can be readily compared.

358 **Fig. 3b** provides an overview of the statistical enrichment of specific TE families for all
359 Codebook TFs (i.e., excluding control TFs) for which >30% of the ChIP-seq peaks
360 overlap any kind of repeat (**Supplementary Document 1** shows an expanded version
361 with all rows labelled), and for which GHT-SELEX data was obtained[35]. Strikingly, the
362 enriched TEs in the GHT-SELEX data closely correspond to those in the ChIP-seq data
363 (shown adjacent to each other in **Fig. 3b**), even though one assay is *in vivo* and one is
364 *in vitro*. Thus, specific binding to TEs is an intrinsic and independent property of
365 individual TFs, including many KZNFs. Most of these associations are also supported
366 by the presence of binding motif matches (see below). Previous motif models derived
367 from other KZNFs' ChIP-seq data were unable to specify precise direct genomic binding
368 to their target elements[33]. To our knowledge, this is the first experimental demonstration
369 that KZNFs possess autonomous and sufficient sequence specificity to discriminate
370 ERE subfamilies from the rest of the genome.

371 **Fig. 3b** also illustrates that many TFs in the "Other" class bind to specific repeat
372 families, of which most (9/17) are KZNFs. Thus, the high association of KZNFs with TEs
373 is not limited to Dark TFs: among the 28 KZNFs binding to specific TE classes, 19 are
374 Dark TFs, and nine are Other TFs. In addition, four of the Promoter TFs in the
375 Codebook set are KZNFs. Their peaks overlap with fewer repeats, but they are
376 nonetheless enriched for specific classes: ZKSCAN4 tends to bind MER53; ZNF676
377 significantly recognizes LTR12C; and ZNF746 and ZNF500 are biased towards G-rich
378 repeats (**Fig. 3a**). Thus, while the Dark TFs are enriched for KZNFs, and both Dark TFs
379 and KZNFs tend to bind to TEs, these associations are far from absolute.

380 **Direct and sequence-specific binding of Promoter TFs to promoter regions**

381 We next asked whether the GHT-SELEX data displayed enrichment for the same broad
382 genome annotations observed in ChIP-seq. The fraction of intrinsic (i.e., GHT-SELEX)
383 and cellular (i.e., ChIP-seq) sites for each TF that overlap with protein-coding gene
384 promoters, repeat sequences (of any kind), and the combination of the "empty" and
385 "heterochromatin" states are shown in **Figs. 4a**, **b**, and **c,** respectively. In each case,
386 there is preferential binding *in vitro* that corresponds to what is observed in cells, with
387 Promoter TFs having much higher intrinsic preference for promoter DNA, and Dark TFs
388 having higher preference for repeats and empty/heterochromatin (as expected from
389 their binding to specific repeat classes). We note that many Promoter and Enhancer
390 TFs have a greater tendency to bind the "empty" and "heterochromatin" state DNA *in*
391 *vitro* than in cells (**Fig. 4c**). This could be due to (a) a *bona fide* preference for open
392 chromatin (and/or inability to access "empty" and "heterochromatin" regions), (b)
393 functional binding at these loci in other cell types (but not HEK293), and/or (c)
394 preferential extraction of these proteins at open chromatin in ChIP-seq experiments.

395 In addition, dozens of Promoter TFs specifically displayed intrinsic preference for the
396 regions that overlap or are just upstream of transcription start sites (TSS), by multiple
397 measures (**Fig. 4d,** *left*), similar to that described for TFs in a variety of genomes[49-51].
398 This observation is consistent with functional roles for these TFs in promoter definition,
399 delineation of TSS location, and/or gene regulation.

400 **Distinct conservation patterns of Dark TF vs Promoter TF binding sites**

401 To examine the potential functionality of direct genomic binding sites, we examined the
402 primary sequence conservation of sites that are bound directly. This analysis first
403 requires the identification of high-resolution, high confidence binding sites, enabling
404 conservation patterns to be attributed to individual bases within the binding site while
405 excluding inconclusive relations arising from indirect TF recruitment. To make a
406 conservative assessment of direct binding, we considered a ChIP-seq peak to be bound
407 directly by a TF if the peak overlapped with a GHT-SELEX peak for the same TF as well
408 as a high-scoring motif match. We obtained genomic scores for TF DNA binding motifs,
409 modeled as Position Weight Matrices (PWMs), from Codebook consortium data; PWM
410 derivation and benchmarking are described in more detail in accompanying
411 manuscripts[34,40]. To further increase confidence in direct binding, we adjusted the
412 significance thresholds for all three (ChIP-seq and GHT-SELEX peaks, and PWM hits
413 treated as peaks) to maximize the Jaccard value between all three peak sets; we refer
414 to these as "triple overlap" or TOP sites (see **Methods** for details). This approach
415 maximizes the probability that ChIP-seq peaks are due to direct, sequence-specific
416 binding; however, false negatives will arise due to experimental error of any type or
417 inaccuracy of motif models. Thus, the numbers of direct binding sites obtained are likely
418 to be underestimates. In addition, this analysis could be conducted only on 137 of the
419 217 TFs (101 Codebook TFs and 36 controls; see **Supplementary Table 3**) because
420 37% of the Codebook proteins lacked GHT-SELEX data, and/or did not yield motifs
421 (**Supplementary Fig. 3a, b**).

422 For these 137 TFs, the fraction of ChIP-seq peaks that could be conservatively
423 attributed to direct binding ranged from 0% for SP140 (indicating that its ChIP-seq sites
424 are dominated by indirect recruitment) to 62.55% for ZNF728 (which has a unique 21-
425 base motif), with a median of 8.5% (at the optimized threshold). The fraction is similar
426 for the Promoter TFs (11.5%) and Dark TFs (11.1%) (**Fig. 4e**), and the absolute number
427 of direct binding sites is similarly high for a subset of both groups (**Fig. 4f**). We conclude
428 that there is no systematic difference between Promoter TFs and Dark TFs in direct
429 binding characteristics, and that many of the observed TF binding sites are direct. The
430 Promoter TF TOPs tend to be strongly aligned with TSS (**Fig. 4d**, *middle*).

431 We then examined the conservation of the TOP sites, producing an estimate of whether
432 each site is under purifying selection. In essence, for many TFs, the TOP sites in
433 aggregate display conservation patterns that mimic the selectivity of each base position
434 in the TF's PWM. **Fig. 5a** shows a graphic demonstration: when TOP sites are aligned
435 to the PWM hit, and displayed as heatmaps that show base-level conservation scores
436 (here, phyloP[52]), there are often vertical blue lines. These lines represent positions in
437 the PWM hits that are preferentially conserved across many TOP sites. Similar to
438 previous observations made with well-characterized TFs[52], the positions with the
439 highest conservation often correspond to tall letters in the sequence logo (i.e., high
440 information content), indicating selection on the binding site to match the sequence
441 preferences of the TF.

442 We developed three heuristics to discriminate conserved vs. unconserved TOP sites
443 (see **Methods**). Two of them test for a relationship between the information content at
444 each base position of the PWM and the conservation score, while the third tests for
445 higher overall conservation at the PWM hit than in the immediate flanking sequence. As
446 shown in **Fig. 5a**, and in similar diagrams for all 137 TFs for which these tests could be
447 run (**Supplementary Document 2**), these tests together detected sites that appear
448 plausible by visual inspection (i.e., apparent conservation signal relative to flanks). We
449 considered a site to be conserved if any of the three criteria were met, and at least one
450 nucleotide in the PWM hit had an FDR-corrected phyloP score $\geq$1.

451 By these criteria, conservation of TOPs is observed for both Promoter TFs and Dark
452 TFs (**Fig. 5a** and **Supplementary Fig. 8**), but the fraction and absolute number of
453 conserved TOP sites for Promoter TFs are much higher (**Fig. 5b, c** and
454 **Supplementary Table 3**). This outcome suggests that many Promoter TF binding sites
455 are functional, and that the corresponding TFs have conserved functions at promoters.
456 An individual conserved TOP site (hereafter, referred to as "CTOP" sites,
457 **Supplementary Fig. 3c**), for the Promoter TF ZNF407, is shown in **Fig. 5d**; like many
458 Promoter TF TOPs and CTOPs (**Fig. 4d***, right*), it overlaps with a transcription start site.
459 Promoter CTOPs are also often found adjacent to other CTOPs; an example of multiple
460 sites for ZNF131 and YY1 is shown in **Fig. 5e**.

461 Despite their lower numbers, there are still thousands of CTOPs for Dark TFs: in
462 aggregate, the criteria used here yielded 2,546 (**Supplementary Fig. 8** and
463 **Supplementary Table 3**). They tend to be distant from promoters, or each other (e.g.,
464 1,556 are > 1000 bp away from any other CTOP), and they tend to have lower phyloP
465 scores than CTOPs for Promoter TFs. It is conceivable that there is lower statistical
466 power to detect Dark TF binding sites because they typically overlap with TEs and are
467 thus only present in species that contain the ancestral insertion. Nonetheless, striking
468 examples emerged. The Dark TF ZBTB40 recognized 846 CTOP sites, the vast majority
469 of which correspond to remnants of *hAT/Charlie* DNA transposons (**Fig. 5a, f**). Its most
470 strongly conserved CTOP falls outside of a transposon, however, and instead is within
471 the PRKACA 3' UTR (**Fig. 5g**), which may be relevant to its apparent function in
472 spermatogenesis (see below and **Supplementary Document 3**). The Dark TF ZNF689
473 has 35 CTOP sites (out of 353 TOPs) and is enriched for binding L1M5 elements (**Fig.
474 5a**; example CTOP shown in **Fig. 5h**). Overall, these analyses indicate that Dark TFs
475 occupy a unique and expansive fraction of the genome, and thousands of their direct
476 binding sites show indications of conserved function.

477 **Functions of Dark TFs**
478
479 Finally, we considered potential functions of the Dark TFs. As a starting point, for three
480 of the Dark TFs (SOX2, TPRX1, ZNF836), we examined their direct functional impact
481 on chromatin by ChIP sequencing of H3K9me3 and ATAC-seq following ectopic
482 induction in HEK293 cells (using the same transgene cell lines described earlier). SOX2
483 was included as a positive control. It is a well-known pioneer, but only 5% of its TOP
484 sites overlap with ATAC-seq peaks in unperturbed HEK293 cells, while 66% of its TOP
485 sites are in "empty" chromatin, which we reasoned might be impacted upon induction.

486 TPRX1 and ZNF836 were chosen because of their relatively low endogenous
487 expression in HEK293, yet like SOX2, most of their ChIP-seq binding sites are in
488 "empty" or heterochromatin. In addition, TPRX1 has recently been described as a
489 master regulator in zygotic genome activation[53], consistent with potential pioneer
490 activity, while ZNF836 is a KZNF that binds the MaLR LTR both *in vitro* and *in vivo* (see
491 **Fig. 3b**).

492 These experiments revealed widespread impact of these TFs on chromatin at their
493 genomic binding sites (TOPs). The ATAC signal at SOX2 TOPs increased by 2.5-fold
494 on average, with most TOP sites clearly impacted (**Fig. 6a**). TPRX1 increased ATAC
495 signal at its TOPs by 1.4-fold, overall, although the impact is concentrated on ~10% of
496 its 3,882 TOPs (**Fig. 6b**), suggesting that additional factors are required for TPRX1 to
497 initiate chromatin opening. ZNF836, in contrast, strongly increased the H3K9me3 signal
498 over its binding sites, with less impact at sites that already reside in heterochromatin
499 prior to the induction (**Fig. 6c**). We conclude that at least a subset of Dark TFs can
500 readily influence the chromatin state at their binding loci.

501 We also examined existing literature and databases to survey known and potential
502 functions for the Dark TFs in relation to our new finding **Fig. 7** provides a graphical
503 overview of various properties of these proteins such as their interactions with
504 chromatin remodelers and cofactors, properties of their binding sites, and previously-
505 described biological functions (see **Supplementary Document 3** for a detailed
506 description, and **Supplementary Table 4** for a table of known protein-protein
507 interactions and other information). In summary, most Dark TFs have apparent roles in
508 repression of transcription, via at least six different mechanisms that are different from
509 recruitment of KAP1[33,54]. Intriguingly, 22 of the 47 C2H2-zf proteins, including both
510 KRAB and non-KRAB C2H2-zf proteins (as well as transposon-derived ZBED9, SOX2,
511 and SP100), interact with other C2H2-zf proteins, often extensively[55], suggesting a
512 potentially widespread role in the organization of chromosome topology. Not all Dark
513 TFs are apparent repressors, however. In addition to TPRX1, SALL3 is a known master
514 regulator, controlling the differentiation of hiPSCs into cardiomyocytes vs neural cells[56].
515 Two other Dark TFs have been described as impacting DNA metabolism; ZNF384 binds
516 Ku and recruits non-homologous end joining factors to double-strand breaks[57], while
517 ZNF146 depletion slows the replication fork[58].

518 The distinct binding sites and diversity of apparent effector mechanisms and cellular
519 roles of the Dark TFs suggest that they may each regulate specific biological functions,
520 and that they may also be multifunctional. Indeed, physiological consequences that
521 have been reported for perturbation of the Dark TFs vary widely (**Fig. 7**, right column;
522 **Supplementary Table 4** provides the values and sources), ranging from basic cellular
523 processes to development. For example, the KZNF ZNF689, shown in **Fig. 5a** to bind
524 35 conserved sites enriched for L1M5 (**Fig. 5g**), also binds promoters of various L1
525 subtypes, preventing genomic instability conferred by L1 retrotransposition[59]. ZBTB40,
526 which we observe almost exclusively at DNA *hAT/Charlie* transposons (**Fig. 5e**), and
527 which is one of the oldest Dark TFs (**Fig. 7**), was recently shown to bind telomeric
528 dsDNA breaks and maintain telomeric length[60]. In mouse, Zbtb40 deficiency impacts
529 spermatogenesis through disrupted telomeric lengthening and maintenance in

530 spermatocytes[61]. *hAT/Charlie* transposons are enriched in telomeric regions of human
531 DNA[62], suggesting that this function may be conserved.

532 **Complex patterns of coevolution among Dark TFs and their binding sites**

533 Dark TF binding sites typically feature specific classes of TEs, which are the youngest
534 insertions in the genome. The Dark TFs themselves are dominated by KZNFs, which
535 are the youngest class of TFs in human. TPRX1, a homeodomain protein, is also
536 relatively recent (it appears to be a eutherian-specific duplicate of the adjacent CRX
537 gene[63]). KZNFs both bind and co-evolve with TEs they suppress but are also assumed
538 to adopt functions beyond TE silencing. Given these clear observations and expected
539 tendencies, we asked whether we could systematically derive evolutionary principles
540 that would characterize the Dark TFs and their binding sites, relative to other TF
541 classes. To do this, we compared the estimated ages of the TFs to the ages of their
542 binding sites, which we defined by seeking ungapped alignment in ancestral genomes,
543 regardless of sequence conservation, or whether the sequence is a TE
544 (**Supplementary Fig. 9a**). Simultaneously, we considered whether the binding sites of
545 TFs within TEs suggest either a TF-specific evolutionary history or reflect more general
546 trends (e.g., TFs binding TEs of the same age, or specific classes of TEs being
547 predominant).

548 This analysis is described in detail in **Supplementary Document 4.** No obvious
549 systematic patterns emerged that could be described as rules, but a wealth of
550 tendencies and specific observations were revealed. First, regardless of TE binding, or
551 TF class, a clear trend is that TFs tend to be older than the sites they bind: Promoter
552 TFs have a median age of 429 MYA, while Dark TFs have a median age of 99.19 MYA
553 (**Supplementary Fig. 9b**), which is still older than a typical binding site even for
554 Promoter TFs, and older than most human TEs (**Supplementary Fig. 9c**). These
555 results suggest that the established phenomenon of TF binding site turnover[14] is
556 prevalent for the Dark TFs, just as it is for other TF classes. Second, TEs of all major
557 classes (LINE, SINE, LTR/ERV, and DNA transposons) are recognized by specific TFs.
558 Moreover, for all four major classes of TEs, there are cases in which greater than 10%
559 of a TF's TOP sites overlap one TE type and are conserved (**Supplementary Fig. 10a**).
560 These are all apparent examples in which the corresponding TF/TE combination has
561 systematically contributed to the regulatory function of the host genome.

562 Third, the encompassed TEs span a very wide age range, from human-specific AluY
563 elements to L2, L3, and MIR, which pre-date eutherian mammals. These associations
564 can provide insight into the evolution and molecular function(s) of specific TFs. For
565 example, ZNF836 and ZNF841 (which are both Dark TFs and KZNFs) are paralogs that
566 arose from a pre-simian duplication event[64] and bind to distinct subtypes of the closely
567 related, simian-specific MaLR LTR elements. They bind distinct motifs that specify the
568 differing base identities at homologous positions within the diverged LTR, suggesting
569 neofunctionalization and retention to maintain silencing of both LTR subtypes
570 (**Supplementary Fig. 10b**).

571 There are also many cases in which the TFs and TEs they bind are grossly mismatched
572 in age, however, suggesting that TE binding is fortuitous. For example, ZNF286B is a
573 human-specific duplicate of ZNF286A which has lost its KRAB domain[65,66], but its
574 binding sites are enriched for LINE-3 (L3), an ancient element found across all
575 mammals, suggestive of coincidental adaptation (**Fig. 3b**). In contrast, ZNF362 and
576 ZNF384 (both non-KZNFs) are products of a duplication ~429 MYA (the duplication is
577 found across bony vertebrates), but the binding sites for both proteins are enriched for
578 the much younger, primate-specific Alu elements, as well as poly-A repeats, consistent
579 with their DNA binding motifs (**Supplementary Fig. 10c**). These proteins have the
580 largest number of TOP sites within the entire Codebook dataset, and it is possible that
581 the recently-expanded target range of these proteins is a coincidental liability, as
582 rearrangements of both ZNF362 and ZNF384 genes (most commonly as fusion proteins
583 with activating TFs and cofactors) are found frequently in leukemia[67,68].

584 **DISCUSSION**

585 The Codebook ChIP-seq data provide cellular binding sites for 130 putative TFs,
586 defined as previously lacking PWMs or other models of sequence specificity. It
587 represents a valuable resource for studying TF function and evolution in the context of
588 regulatory genomics. For a large majority of the proteins assayed, we have also now
589 identified a binding motif which is supported by independent assays[40]. Thirty-six of the
590 proteins did not produce a motif and may not be *bona fide* TFs. Their ChIP-seq profiles
591 are nonetheless informative: enrichment of ChIP-seq peaks at different types of
592 genomic features (e.g. promoters, repeats) or chromatin states, as well as co-
593 occurrence with peaks for other proteins (e.g. TFs), can yield clues as to potential
594 function.

595 Previous large-scale ChIP-seq analyses have mainly focused on the established roles
596 of TFs in binding to promoters and enhancers (e.g. Partridge *et al.*[24]), although studies
597 of KZNFs typically focus on binding to TEs, and specifically EREs[7,16,33]. The analysis
598 scheme described here considers these models of TF function as hypotheses with
599 equal weight. The expected binding categories are clearly present, including
600 preferences for promoters and enhancers, as well as the strong tendency for KZNFs to
601 bind specific classes of EREs, and within constitutive heterochromatin. Binding to EREs
602 is not the only function of KZNFs, however, and other TF classes can clearly bind
603 genomic dark matter. Just over half (36/63; 57%) of the KZNFs we analyzed are Dark
604 TFs - as are roughly 10% of non-KZNF control TFs (4/35). If these proportions were to
605 hold for the 350 human KZNFs and the 781 non-KZNF TFs with a known DNA-binding
606 motif, then there would be a total of ~277 human Dark TFs among the ~1,600 human
607 TFs (i.e., 17% of all human TFs), among which roughly a third would not be KZNFs.

608 Overall, TF behavior with respect to chromatin states and genomic landmarks appears
609 more varied than a simple categorization scheme would imply. We did not
610 systematically explore the “Other” category, used here as a catch-all. However, like the
611 Dark TFs, Other TFs appear to encompass proteins that bind both *in vitro* and *in vivo* to
612 many regions that are labeled as inactive and/or closed chromatin in HEK293 cells prior
613 to induction of the TF, and to satisfy some expectations of “pioneer” activity. One

614 example is the non-KRAB protein ZSCAN2, for which we cataloged 90 CTOPs and
615 found they are enriched for mammal-wide LINE-3 (L3) elements. ZSCAN2 is shown to
616 be expressed only in early stages of embryogenesis as well as post-meiotic round
617 spermatids[69], and to affect fertility in mice (and perhaps human)[70].

618 Establishing physiological functions, if any, for individual TF binding sites is a long-
619 standing and difficult problem in regulatory genomics. It is becoming clear that even
620 KZNFs exert specific physiological outcomes, beyond ERE repression[22], but few are
621 currently known. Biochemical functions of TFs may be more straightforward to identify:
622 our observation of ATAC signal enrichment within TPRX1 TOPs upon ectopic
623 expression of the TF suggests a straightforward approach. Biochemical function does
624 not equate to physiological purpose, however. A particular challenge with TFs is that the
625 level of binding site turnover observed on evolutionary timescales requires that binding
626 sites arise at random, many of which are presumably irrelevant for gene regulation or
627 reproductive fitness (at least initially), despite being biochemically functional. By this
628 reasoning, we expect that many biochemically verified, direct TF binding sites should be
629 physiologically non-functional, and indeed, we find that, overall, most TOP sites are not
630 conserved, even for Promoter TFs. Lack of conservation cannot be taken as lack of
631 biological purpose; nonetheless, conserved TOP sites would seem most likely to yield
632 interpretable results in targeted laboratory studies. More generally, the Codebook TOP
633 and CTOP catalogue will provide a rich resource for future efforts in examining genome
634 function, e.g., employing techniques similar to GTEx[71] to relate variants at the TOP sites
635 to chromatin and gene expression.

636 The lower overall primary sequence conservation of Dark TF binding sites, relative to
637 those of Promoter TFs, could have several explanations. One possibility is that very few
638 of the binding sites are functional; in theory, only a single binding site that confers a
639 modest selective advantage would be sufficient to drive retention of both the site and
640 the TF, with all other sites arising at random (and under neutral evolutionary pressure,
641 provided they are not detrimental). Another possibility is that the exact positioning of the
642 sites is not critical to their function, unlike Promoter TFs, which by definition must be
643 close to TSSs. Dark TF functions could simply require that binding sites are distributed
644 widely across non-functional DNA and thus could be redundant over large sequence
645 windows. Such functions would not preclude a small subset of sites being co-opted for
646 regulation of host genes, which would become constrained (i.e., conserved).
647 Regardless of what biochemical, cellular, and physiological functions are revealed, the
648 Dark TFs represent a new contribution to the decades-old odyssey into the function and
649 significance of the dark matter genome.

650

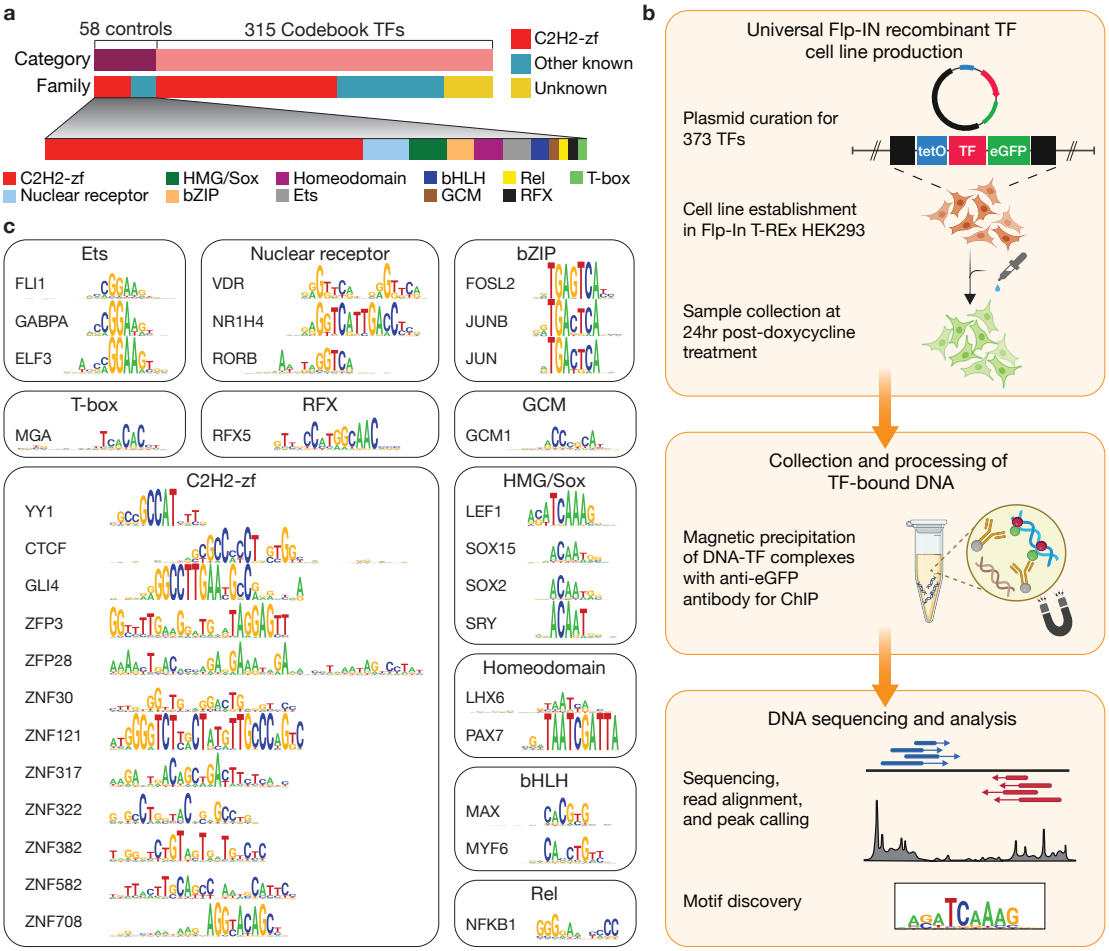
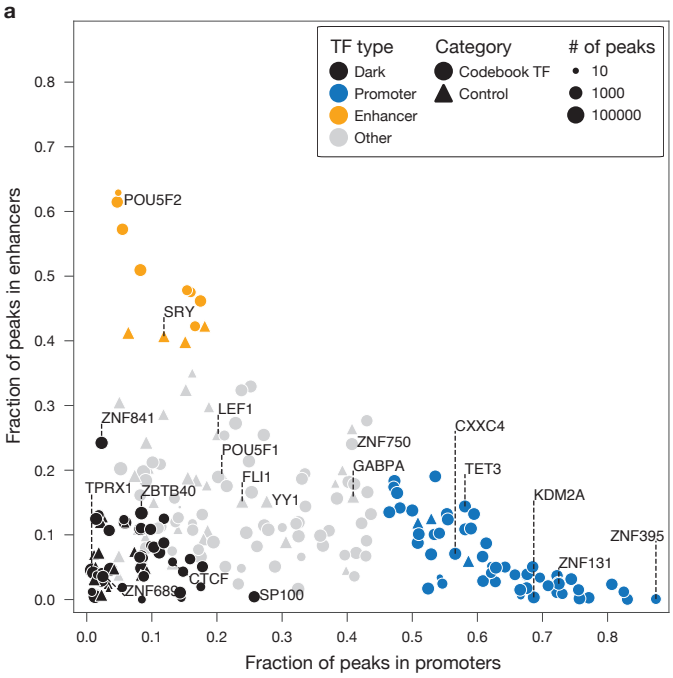

a
58 controls
315 Codebook TFs
Category
Family
C2H2-zf
Other known
Unknown
C2H2-zf
HMG/Sox
Homeodomain
bHLH
Rel
T-box
Nuclear receptor
bZIP
Ets
GCM
RFX
b
Universal Flp-IN recombinant TF cell line production
Plasmid curation for 373 TFs
tetO
TF
eGFP
Cell line establishment in Flp-In T-REx HEK293
Sample collection at 24hr post-doxycycline treatment
Collection and processing of TF-bound DNA
Magnetic precipitation of DNA-TF complexes with anti-eGFP antibody for ChIP
DNA sequencing and analysis
Sequencing, read alignment, and peak calling
Motif discovery
c
Ets
FLI1
GABPA
ELF3
Nuclear receptor
VDR
NR1H4
RORB
bZIP
FOSL2
JUNB
JUN
T-box
MGA
RFX
RFX5
GCM
GCM1
C2H2-zf
YY1
CTCF
GLI4
ZFP3
ZFP28
ZNF30
ZNF121
ZNF317
ZNF322
ZNF382
ZNF582
ZNF708
HMG/Sox
LEF1
SOX15
SOX2
SRY
Homeodomain
LHX6
PAX7
bHLH
MAX
MYF6
Rel
NFKB1

Figure 1. Overview of ChIP-seq strategy to identify genomic binding sites for 315 nebulous human TFs. (**a**) Summary of human TF categories (38 control and 315 Codebook) and protein families (with known and unknown classifications) that were selected for ChIP-seq. The “other known” category includes non-C2H2-zf protein families ten of which are included in the control TFs. (**b**) A schematic of the experimental pipeline for production of 373 inducible EGFP-labelled TF cell lines, universal Chromatin immunoprecipitation with eGFP antibody, and derivation of TF binding sites for each line. (**c**) Samples of representative motifs obtained for each protein family of the control TFs show expected matches.


a
TF type
Dark
Promoter
Enhancer
Other
Category
Codebook TF
Control
# of peaks
10
1000
100000
POU5F2
SRY
ZNF841
LEF1
POU5F1
FLI1
YY1
TPRX1
ZBTB40
CTCF
ZNF689
SP100
ZNF750
GABPA
CXXC4
TET3
KDM2A
ZNF395
ZNF131
Fraction of peaks in enhancers
Fraction of peaks in promoters
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

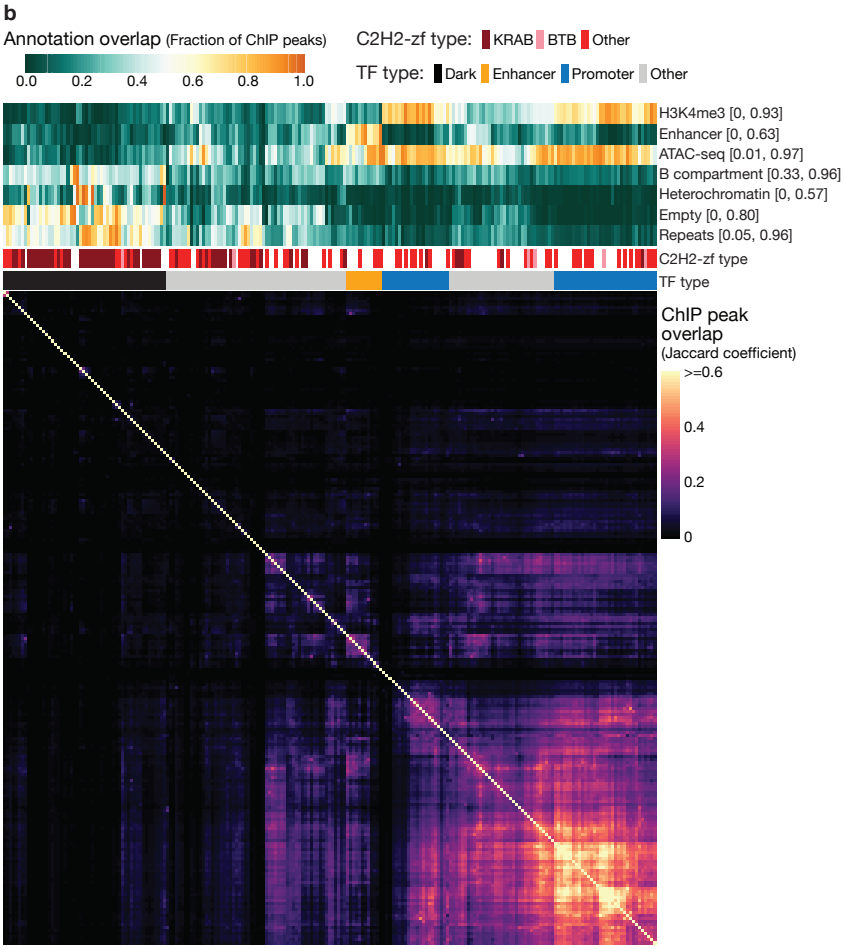
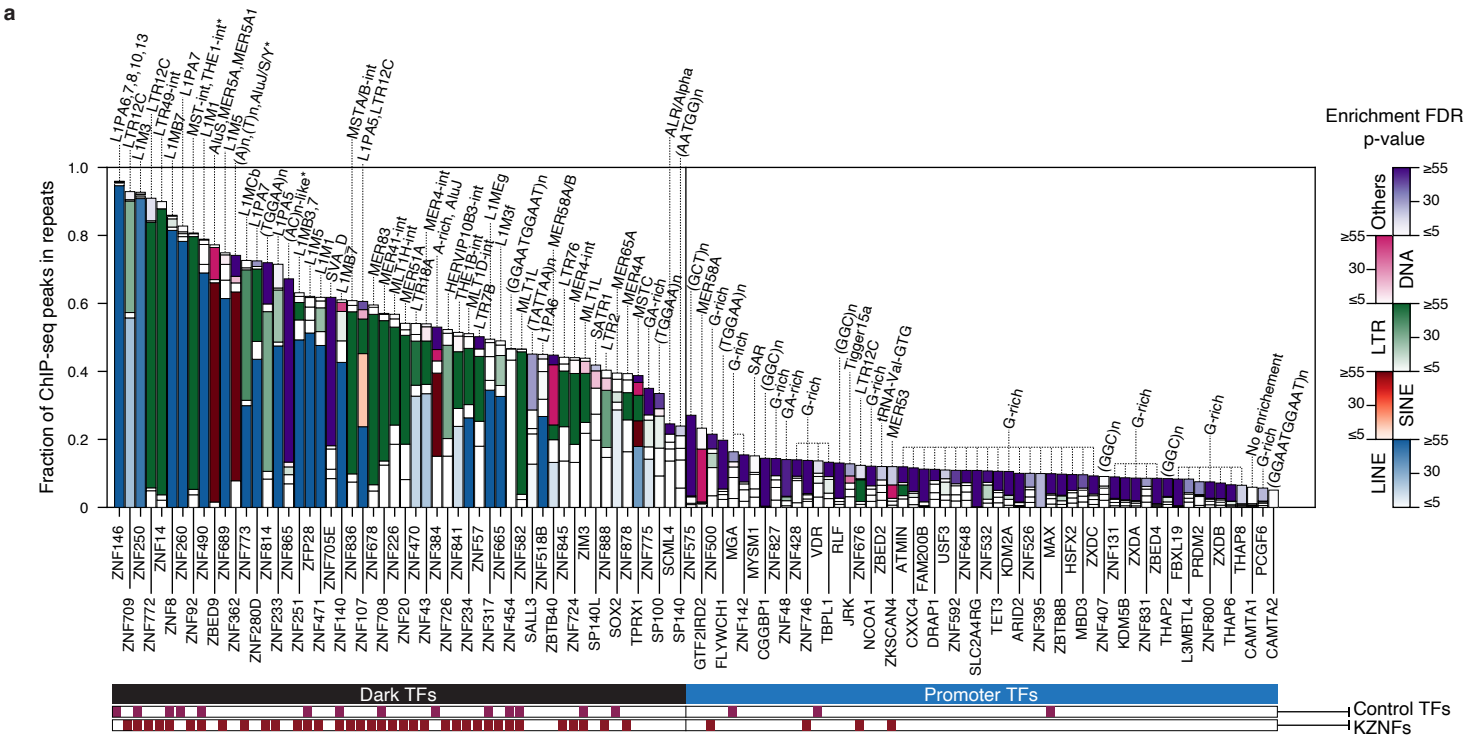

b
Annotation overlap (Fraction of ChIP peaks)
0.0
0.2
0.4
0.6
0.8
1.0
C2H2-zf type: KRAB BTB Other
TF type: Dark Enhancer Promoter Other
H3K4me3 [0, 0.93]
Enhancer [0, 0.63]
ATAC-seq [0.01, 0.97]
B compartment [0.33, 0.96]
Heterochromatin [0, 0.57]
Empty [0, 0.80]
Repeats [0.05, 0.96]
C2H2-zf type
TF type
ChIP peak overlap
(Jaccard coefficient)
>=0.6
0.4
0.2
0

Figure 2. Distinct genomic binding patterns identifies four different TF types with varying preferences for accessible or closed chromatin. (**a**) distribution of ChIP-seq peak fractions for 217 TFs in protein-coding promoters (x-axis) and HEK293 enhancers (y-axis) used for discrimination of Promoter and Enhancer TFs. Dark TFs and “Other” categories are also labelled here. Point sizes are proportional to the number of peaks for each TF (log scale). (**b**) *Bottom:* (square) heatmap of Jaccard similarity coefficient between ChIP-seq peaks of all TF pairs showcasing the colocalization patterns of binding sites. The bright corner aligns with Promoter TFs. *Top:* heatmap of ChIP-seq peak fractions that fall within various annotated genomic tracks of chromatin accessibility as indicated, as well as protein 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 for '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.


a
Fraction of ChIP-seq peaks in repeats
Enrichment FDR p-value
Others
DNA
LTR
SINE
LINE
Dark TFs
Promoter TFs
Control TFs
KZNFs

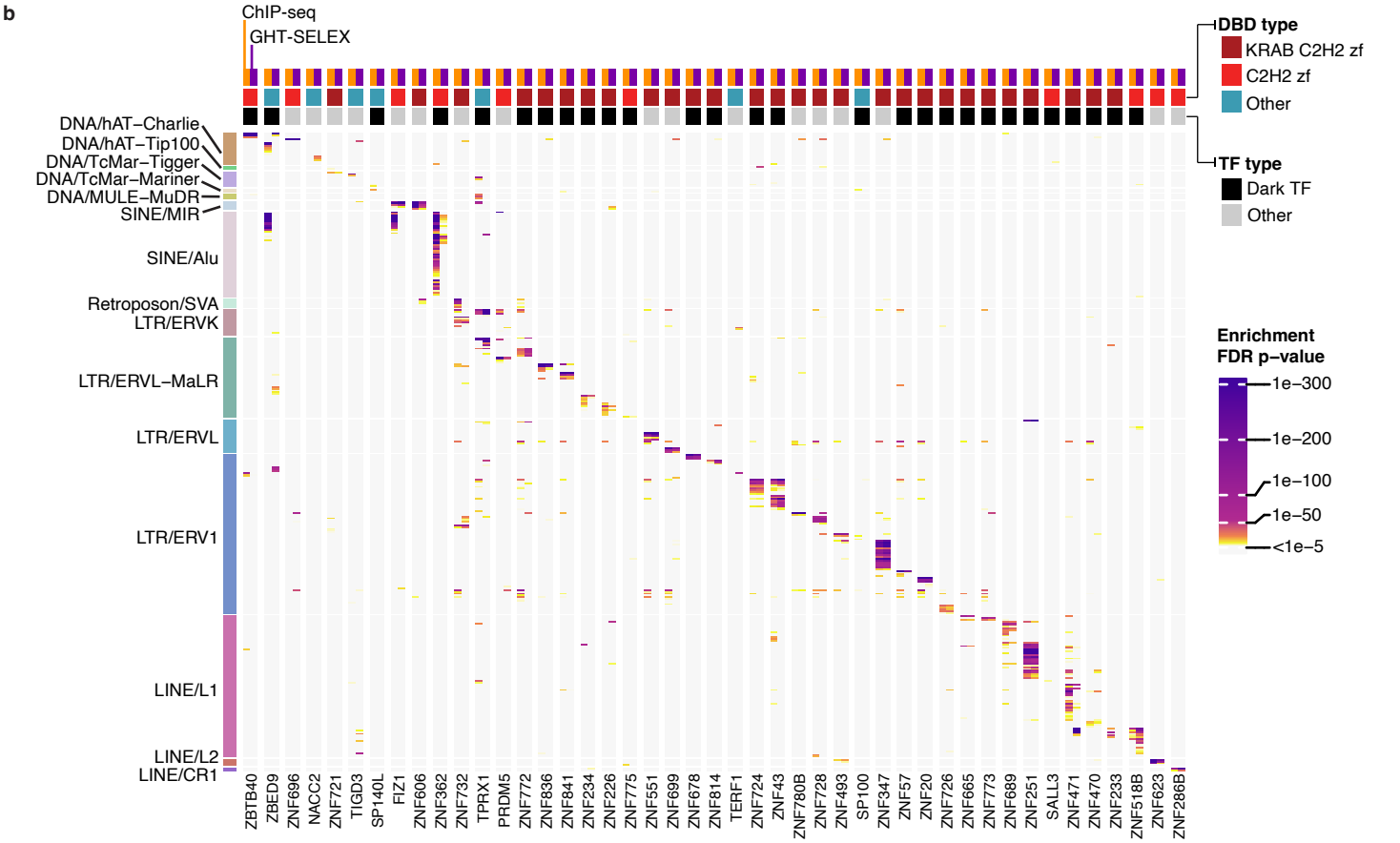

b
ChIP-seq
GHT-SELEX
DBD type
KRAB C2H2 zf
C2H2 zf
Other
TF type
Dark TF
Other
DNA/hAT−Charlie
DNA/hAT−Tip100
DNA/TcMar−Tigger
DNA/TcMar−Mariner
DNA/MULE−MuDR
SINE/MIR
SINE/Alu
Retroposon/SVA
LTR/ERVK
LTR/ERVL−MaLR
LTR/ERVL
LTR/ERV1
LINE/L1
LINE/L2
LINE/CR1
Enrichment FDR p−value
1e−300
1e−200
1e−100
1e−50
<1e−5

Figure 3. Dark TFs recognize specific families of transposable elements. (**a**) Stacked bar chart showing the fraction of ChIP-seq peaks (at MACS2.0 $P<10^{-10}$ threshold) overlapping with repeats, partitioned into 5 superclasses: LINEs (blue), SINEs (red), LTRs (green), DNA transposons (orange), and other types (purple), for all Dark TFs and Promoter TFs. Color intensities represent the highest enrichment (–log10 FDR-corrected p-values) between the peak sets of a TF and the repeat families within that corresponding superclass. The most enriched repeat family is labeled above each bar. (**b**) Heatmap showing the enrichment (–log10 FDR-corrected p-values) of ChIP-seq (yellow) and GHT-SELEX (purple) peaks for binding to specific TE families (y-axis). Only Codebook TFs (i.e. controls excluded) having both datasets and with >30% of their ChIP-seq peaks overlapping any repeat are shown, using universal thresholds (MACS2.0 $P<10^{-10}$ for ChIP-seq and knee specificity ≥30 for GHT-SELEX; see **Methods**). Labels on the left indicate TE superfamilies and families.

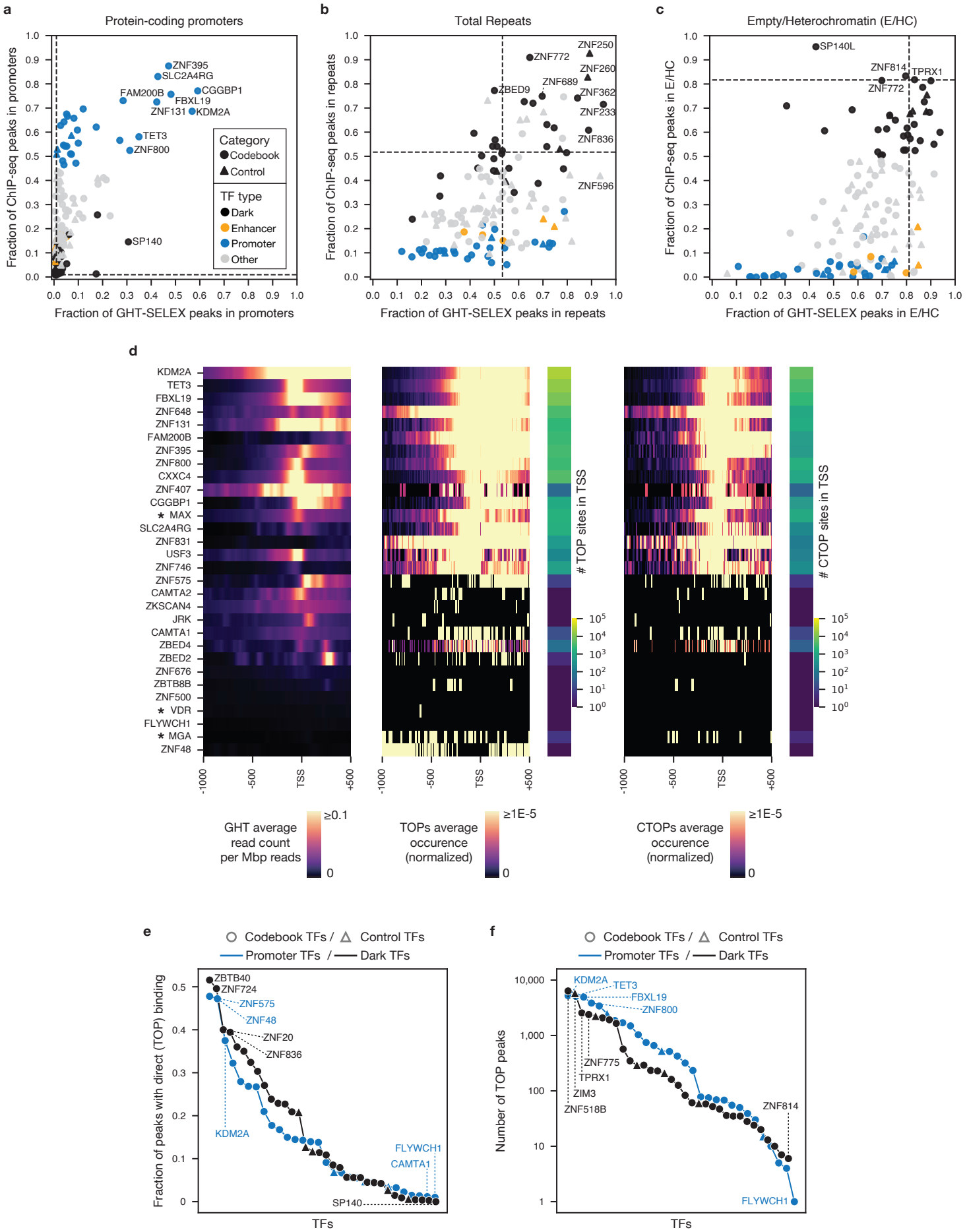

a
Protein-coding promoters
Fraction of ChIP-seq peaks in promoters
Fraction of GHT-SELEX peaks in promoters
ZNF395
SLC2A4RG
FAM200B
CGGBP1
FBXL19
ZNF131
KDM2A
TET3
ZNF800
SP140
Category
Codebook
Control
TF type
Dark
Enhancer
Promoter
Other
b
Total Repeats
Fraction of ChIP-seq peaks in repeats
Fraction of GHT-SELEX peaks in repeats
ZNF250
ZNF772
ZNF260
ZNF689
ZBED9
ZNF362
ZNF233
ZNF836
ZNF596
c
Empty/Heterochromatin (E/HC)
Fraction of ChIP-seq peaks in E/HC
Fraction of GHT-SELEX peaks in E/HC
SP140L
ZNF814
TPRX1
ZNF772
d
KDM2A
TET3
FBXL19
ZNF648
ZNF131
FAM200B
ZNF395
ZNF800
CXXC4
ZNF407
CGGBP1
* MAX
SLC2A4RG
ZNF831
USF3
ZNF746
ZNF575
CAMTA2
ZKSCAN4
JRK
CAMTA1
ZBED4
ZBED2
ZNF676
ZBTB8B
ZNF500
* VDR
FLYWCH1
* MGA
ZNF48
-1000
-500
TSS
+500
TOP sites in TSS
CTOP sites in TSS
GHT average read count per Mbp reads
TOPs average occurence (normalized)
CTOPs average occurence (normalized)
≥0.1
≥1E-5
0
e
Codebook TFs / Control TFs
Promoter TFs / Dark TFs
Fraction of peaks with direct (TOP) binding
ZBTB40
ZNF724
ZNF575
ZNF48
ZNF20
ZNF836
KDM2A
FLYWCH1
CAMTA1
SP140
TFs
f
Codebook TFs / Control TFs
Promoter TFs / Dark TFs
Number of TOP peaks
KDM2A
TET3
FBXL19
ZNF800
ZNF775
TPRX1
ZIM3
ZNF518B
ZNF814
FLYWCH1
TFs

Figure 4. Evaluation of the genomic binding sites that are recognized by Promoter TFs, Enhancer TFs, and Dark TFs. (**a, b, c**) Fraction of GHT-SELEX (x-axis) and ChIP-seq (y-axis) peaks falling in the specified genomic regions; protein-coding promoters (**a**), repeats (**b**), and empty or heterochromatin (**c**), using the peaks at the universal threshold (MACS2.0 $P<10^{10}$ for ChIP-seq and knee specificity of 30 for GHT-SELEX, see **Methods**), highlighting the extent of overlaps accounted by sequence specificity. Dashed lines show the expected fraction if peaks were distributed at random. (**d**) 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 by sidebars at the right of each heatmap. Control TFs are marked by asterisks. (**e**) and (**f**) compare the fraction and absolute number of peaks with direct binding (i.e. TOP sites) for Promoter TFs and Dark TFs respectively. In (**e**), for each TF, the number of TOP (i.e. triple optimized overlap) ChIP-seq peaks is divided by the total number of ChIP-seq peaks at the same optimized threshold (i.e. no overlapping). SP140 with no TOPs has been removed from (**f**).

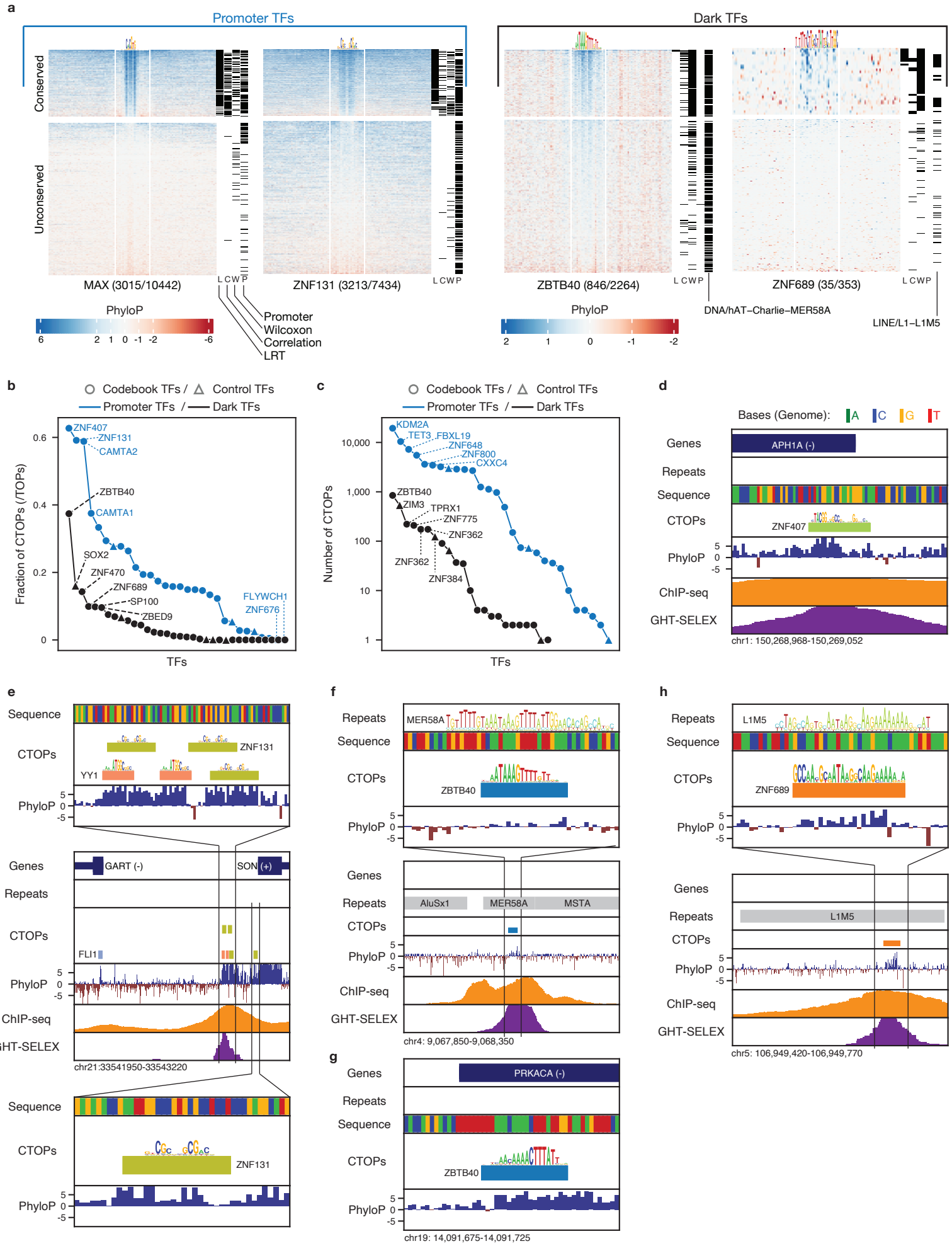

a
Promoter TFs
Dark TFs
Conserved
Unconserved
MAX (3015/10442)
ZNF131 (3213/7434)
ZBTB40 (846/2264)
ZNF689 (35/353)
L C W P
PhyloP
Promoter
Wilcoxon
Correlation
LRT
DNA/hAT−Charlie−MER58A
LINE/L1−L1M5
b
Codebook TFs / Control TFs
Promoter TFs / Dark TFs
Fraction of CTOPs (/TOPs)
TFs
ZNF407
ZNF131
CAMTA2
ZBTB40
CAMTA1
SOX2
ZNF470
ZNF689
SP100
ZBED9
FLYWCH1
ZNF676
c
Number of CTOPs
KDM2A
TET3
FBXL19
ZNF648
ZNF800
CXXC4
ZIM3
TPRX1
ZNF775
ZNF362
ZNF384
d
Bases (Genome): A C G T
Genes
APH1A (-)
Repeats
Sequence
CTOPs
ZNF407
PhyloP
ChIP-seq
GHT-SELEX
chr1: 150,268,968-150,269,052
e
Sequence
CTOPs
ZNF131
YY1
PhyloP
Genes
GART (-)
SON (+)
Repeats
FLI1
ChIP-seq
GHT-SELEX
chr21:33541950-33543220
f
Repeats
MER58A
ZBTB40
AluSx1
MER58A
MSTA
chr4: 9,067,850-9,068,350
g
PRKACA (-)
chr19: 14,091,675-14,091,725
h
L1M5
ZNF689
chr5: 106,949,420-106,949,770

Figure 5. Conservation patterns of binding sites of Promoter TFs versus Dark TFs. (**a**) heatmaps of FDR-corrected phyloP scores across the TOP sites (rows) of two Promoter (MAX, ZNF131) and two Dark TFs (ZBTB40, ZNF689), split into top and bottom segments that contain conserved and unconserved sites. 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. 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. Comparison of fraction (**b**) and absolute number (**c**) of conserved TOPs (CTOPs), for Promoter TFs and Dark TFs. TFs with no CTOPs are not shown in (**c**). (**d, e, f, g, h**) Genome track alignments (genes, sequence, PhyloP, ChIP-seq and GHT-SELEX signals) of CTOPs for ZND407 (**d**), ZNF131 and YY1 (**e**), ZBTB40 (**f, g**) and ZNF689 (**h**). (**d**) is an example of a Promoter TF CTOP (ZNF407, which has the highest fraction of CTOPs) residing within the promoter of APH1A gene. (**e**) shows the four CTOPs of Promoter TF, ZNF131, colocalizing with YY1 in the prompter region of SON. (**f**) shows a CTOP of the Dark TF, ZBTB40, located within a *hAT/Charlie* (MER58A) element. ZBTB40 most-conserved TOP resides at the PRKACA promoter (**g**). (**h**) shows for another DARK TF, ZNF689, its CTOP is located within an L1M5 element. The repeat sequence logos shown in MER58A (**f**) and L1M1 (**h**) were produced in Dfam[91].

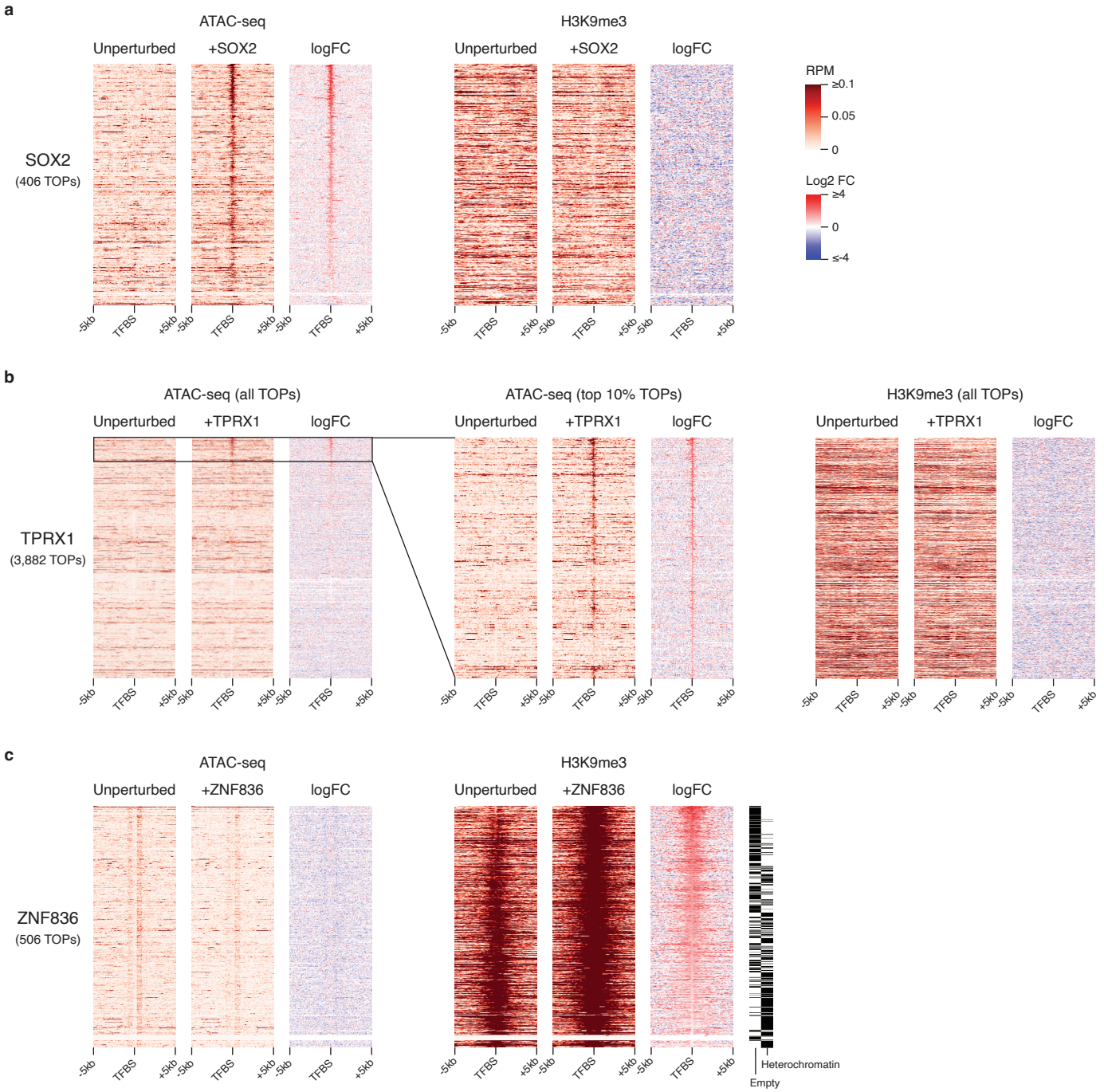

a
ATAC-seq
H3K9me3
Unperturbed
+SOX2
logFC
SOX2
(406 TOPs)
RPM
≥0.1
0.05
0
Log2 FC
≥4
0
≤-4
-5kb
TFBS
+5kb
b
ATAC-seq (all TOPs)
ATAC-seq (top 10% TOPs)
H3K9me3 (all TOPs)
+TPRX1
TPRX1
(3,882 TOPs)
c
+ZNF836
ZNF836
(506 TOPs)
Heterochromatin
Empty

Figure 6. Chromatin reorganization by Dark TFs and their pioneer activity at TOP sites. Comparison of ATAC-seq and H3K9me3 ChIP-seq signals in unperturbed HEK293 cells with three over-expressed Dark TFs SOX2 (**a**), TPRX1 (**b**), and ZNF836 (**c**), measured in a window of 10 kb around their TOPs. For ATAC-seq (left panels), heatmaps show the pileup of Tn5 tag insertion sites (reads per million; RPM) extended to 50 bp on each side. Right panels show the heatmaps for the pileup of H3K9me3 ChIP-seq fragments (RPM). The logFC heatmap illustrates $\log_2$(TF overexpressed/unperturbed HEK293). For TPRX1 ATAC-seq, the heatmap of its dominant 10% TOPs is enlarged in the middle panel for clarification. For better visualization, SOX2 and TPRX1 TOPs in all panels are sorted by their ATAC-seq $\log_2$FC at the TFBS center, while ZNF836 TOPs are sorted by H3K9me3 signal.

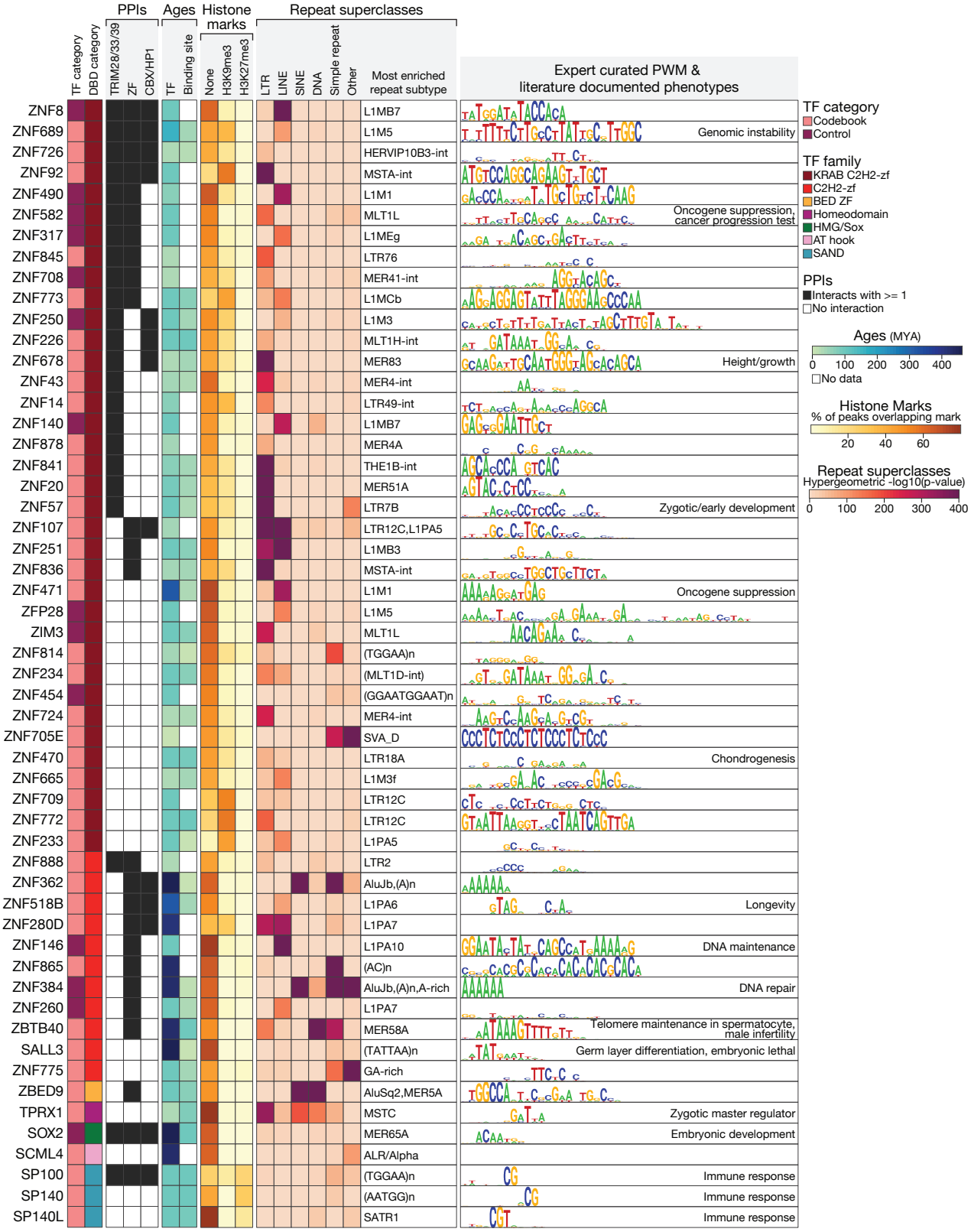

PPIs
Ages
Histone marks
Repeat superclasses
TF category
DBD category
TRIM28/33/39
ZF
CBX/HP1
TF
Binding site
None
H3K9me3
H3K27me3
LTR
LINE
SINE
DNA
Simple repeat
Other
Most enriched repeat subtype
Expert curated PWM & literature documented phenotypes
ZNF8 L1MB7
ZNF689 L1M5 Genomic instability
ZNF726 HERVIP10B3-int
ZNF92 MSTA-int
ZNF490 L1M1
ZNF582 MLT1L Oncogene suppression, cancer progression test
ZNF317 L1MEg
ZNF845 LTR76
ZNF708 MER41-int
ZNF773 L1MCb
ZNF250 L1M3
ZNF226 MLT1H-int
ZNF678 MER83 Height/growth
ZNF43 MER4-int
ZNF14 LTR49-int
ZNF140 L1MB7
ZNF878 MER4A
ZNF841 THE1B-int
ZNF20 MER51A
ZNF57 LTR7B Zygotic/early development
ZNF107 LTR12C,L1PA5
ZNF251 L1MB3
ZNF836 MSTA-int
ZNF471 L1M1 Oncogene suppression
ZFP28 L1M5
ZIM3 MLT1L
ZNF814 (TGGAA)n
ZNF234 (MLT1D-int)
ZNF454 (GGAATGGAAT)n
ZNF724 MER4-int
ZNF705E SVA_D
ZNF470 LTR18A Chondrogenesis
ZNF665 L1M3f
ZNF709 LTR12C
ZNF772 LTR12C
ZNF233 L1PA5
ZNF888 LTR2
ZNF362 AluJb,(A)n
ZNF518B L1PA6 Longevity
ZNF280D L1PA7
ZNF146 L1PA10 DNA maintenance
ZNF865 (AC)n
ZNF384 AluJb,(A)n,A-rich DNA repair
ZNF260 L1PA7
ZBTB40 MER58A Telomere maintenance in spermatocyte, male infertility
SALL3 (TATTAA)n Germ layer differentiation, embryonic lethal
ZNF775 GA-rich
ZBED9 AluSq2,MER5A
TPRX1 MSTC Zygotic master regulator
SOX2 MER65A Embryonic development
SCML4 ALR/Alpha
SP100 (TGGAA)n Immune response
SP140 (AATGG)n Immune response
SP140L SATR1 Immune response
TF category
Codebook
Control
TF family
KRAB C2H2-zf
C2H2-zf
BED ZF
Homeodomain
HMG/Sox
AT hook
SAND
PPIs
Interacts with >= 1
No interaction
Ages (MYA)
0 100 200 300 400
No data
Histone Marks
% of peaks overlapping mark
20 40 60
Repeat superclasses
Hypergeometric -log10(p-value)
0 100 200 300 400

Figure 7. Consolidated information for functions and other properties of Dark TFs. Compiled protein-protein interactions(PPIs)[92] information obtained from available studies and grouped into three categories with potential regulatory functions: 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" (or None) state are shown in the middle. For the repeats, 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 PWMs), along with the corresponding phenotype for any TF with known biological function through literature review (in the same block).

651 **METHODS**

652 **Plasmids.** The Codebook project design is described elsewhere[34]. Putative TFs were
653 those from Lambert et al.[23], with TFs we had already attempted as part of ENCODE
654 removed. We attempted ChIP-seq analysis of all Codebook TFs. We designed full-
655 length ORFs for synthesis (BioBasic.ca) and used conventional restriction cloning to
656 insert them into Flp-In destination plasmid pTH13195 (a modified pDEST
657 pcDNA5/FRT/TO-eGFP vector), which places ORFs under the control of a tet-on, CMV-
658 driven promoter, with an N-terminally EGFP tag. We obtained the 58 controls
659 independently[33,38] and cloned them into pTH13195 by the same process. See
660 accompanying manuscript[34] for the sequences and other information about the inserts.

661 **Cell line production.** We used a previously established protocol for creating individual
662 cell lines for each TF[16]. In brief, we cultured and maintained HEK293 Flp-In T-REx cells
663 (Invitrogen) in Dulbecco's modified Eagle's medium with 10% fetal bovine serum and
664 antibiotics. We created individual cell lines for each TF (between one to five cell lines
665 per construct as biological replicates, see **Supplementary Table 2**) by flp
666 recombination, in which individual destination plasmids were co-transfected into Flp-In
667 T-REx 293 cells together with the pOG44 Flp recombinase expression plasmid using
668 FuGene (Roche, 11814443001). We then selected cells for FRT site-specific
669 recombination into the genome using selection media containing hygromycin (200μg/ml)
670 for 1 to 4 weeks. For each TF cell line, we confirmed expression of EGFP by fluorescent
671 microscopy after 24 hours of Doxycycline treatment (1ug/ml), at which point 10M-20M
672 cells were used for downstream experiments. Individual cell lines were
673 immunoprecipitated between one to five times, in different batches. For each batch of
674 ChIP-seq experiments, we also collected "input DNA" (i.e., non-IP sonicated chromatin),
675 chosen at random for ~10% of all the experiments (see **Supplementary Table 2**).

676 **Chromatin immunoprecipitation.** We fixed ~20M cells on 15cm plates using 1%
677 paraformaldehyde for 10 min on ice, followed by 10 min quenching with 0.125 M
678 glycine. We washed fixed cells twice with cold PBS, scrape-collected, pelleted, flash-
679 froze, and stored the cells at -80 °C. Upon completion of cell collection for a panel of
680 TFs, we thawed cell pellets on ice, lysed them as previously described[72], and sonicated
681 them using a BioRuptor to shear chromatin. We then trapped the protein-DNA
682 complexes on Dynabeads using an anti-EGFP or anti-H3k9me3 (Ab290 and Ab8898,
683 respectively, Abcam). Following wash, crosslink reversal, and elution, we assessed the
684 size and concentration of DNA fragments with an Agilent bioanalyzer and Qubit, prior to
685 sequencing.

686 **Library preparation and sequencing.** DNA library preparation and sequencing were
687 performed at three different facilities (Memorial Sloan Kettering Cancer Center, SickKids
688 Hospital in Toronto, and the Donnelly Centre at the University of Toronto) over a period
689 of four years. The facilities prepared DNA libraries using the NEBNext® Ultra™ II DNA
690 Library Prep Kit for Illumina. Samples were paired-end sequenced (50-150bp) with a
691 target depth of 20M reads (for EGFP) and 50M reads (for H3K9me3) per sample. 183
692 samples have two sequencing replicates (specified as D and M replicates in
693 **Supplementary Table 2** under the Technical Seq. Rep # column).

694 **TF ChIP-seq data preprocessing.** We mapped raw ChIP-seq reads to the human
695 genome build hg38 with *bowtie2*[73] (options: *--very-sensitive*, and *--no-unal*). We used
696 Samtools[74,75] (options: *-q 30*, and *-F 1548*) to remove reads that were unmapped, failed
697 platform/vendor quality checks, were PCR duplicates, or had a mapping quality <30.
698 This stringent threshold on mapping quality ensures that reads mapped to multiple
699 locations are removed; hence, the peak locations (see below) are unambiguous, and
700 our analysis of repeat elements is not skewed. This elimination may, however, result in
701 missing potential peaks in young repeat families with identical or near-identical copies
702 (e.g., L1HS). This could lead to an underestimation of the proportion of binding sites in
703 the dark genome, as repeat elements are often found in those regions, but it would only
704 result in false negatives.

705 **TF ChIP-seq background generation and peak calling.** We generated sample-
706 specific background models using a previously established approach[33], with minor
707 modifications. For each ChIP-seq pull-down experiment, our goal was to construct a
708 background control that closely matched its nonspecific signal while maintaining high
709 coverage by pooling reads from multiple "input" experiments in proportions that best
710 reproduced the background characteristics of the given pull-down.

711 First, we identified genomic regions showing high background signal in at least one
712 input DNA experiment. To do so, we performed peak calling on each input sample using
713 MACS2 (options: p-value < 0.001, and *--nomodel*)[43,76,77] and collected the resulting
714 peaks as "background hotspots". We merged hotspots whose summits were within 50
715 bp across experiments to produce a unified set of background regions.

716 Next, we quantified read counts overlapping each hotspot. For each input DNA
717 experiment, we counted the number of reads overlapping every region in the unified
718 hotspot set, creating a read count matrix with rows being the hotspots and columns
719 being input experiments. For each pull-down experiment, we also calculated read
720 counts over the same hotspot regions; these served as the response variable in the
721 following modeling step. We modeled the pull-down read counts using a non-negative
722 Poisson regression, with the input DNA read count matrix as predictors. This regression
723 estimated a set of non-negative coefficients (one coefficient per input experiment)
724 representing the weighted combination of controls that best reconstructs the
725 background signal observed in the pull-down.

726 Finally, we used these coefficients to generate a synthetic background dataset.
727 Specifically, we sampled reads from each input BAM file in proportion to its
728 corresponding regression weight and pooled them to form a single, pulldown-specific
729 control file. This customized control was then used as the background input for peak
730 calling on the pull-down sample with MACS2[43,76,77].

731 **ChIP peak replicate analysis and merging.** For each TF with one or more replicates,
732 we calculated the Kulczynski II similarity metric for each pair of replicates
733 (**Supplementary Fig. 2a**). We used the Kulczynski II metric in place of Jaccard as it is
734 less affected by the uneven size of the peak sets. We additionally calculated the
735 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 (**Supplementary Fig. 2a**). 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[78] merge to generate new, wider peaks, with the sum of component peak -log(p-values) assigned as the new peak score, and the 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.

H3K9me3 ChIP-seq data processing. Adapter sequences were first trimmed using *cutadapt*. The resulting reads were then mapped to the human genome (hg38) using *bowtie2*[73], followed by the creation of BAM files using *samtools*[74,75] *view*, and sorting with *samtools sort*. Unintended or duplicated reads were removed using *sambamba view -h -f bam -F "[XS] == null and not unmapped and not duplicate and proper_pair and mapping_quality >= 30"* and *picard MarkDuplicates*. Pileups (bdg) for pulled down fragments were obtained using *macs3*[43,76,77] *callpeak* with the options *--format BAMPE -g hs --scale-to small –bdg* using all the replicates of that experiment for option *-t*, followed by *bedGraphToBigWig*. Finally, to get the signal over TOP sites, DeepTools[79] *computeMatrix* was used, and plotted by custom Python scripts.

ATAC-seq experiment. We performed ATAC-seq in HEK293 cells in two separate sets. The first set included only unperturbed HEK293 cells with four replicates in order to establish the baseline status for chromatin state (see Chromatin state analysis section below). The second set included three HEK293 cell lines (SOX2, TPRX1, and ZNF836) with two biological replicates for each TF. We performed the well-established OMNI-ATAC protocol as previously described[80]. Briefly, 50,000 viable cells were pelleted (500 RCF at 4°C for 5 min) collected and palleted in 1ml cold PBS After removing the supernatant, the cells were lysed in 50 µl of cold ATAC–Lysis buffer 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. 1 ml of cold ATAC-resuspension buffer containing 0.1% Tween 20 was then added to the lysate followed by 10 min centrifugation in 4°C at 500with to pallet the nuclei Transposition was performed using either Tagment DNA TDE1 Enzyme and Buffer Kit (Illumina) or loaded Tn5 enzyme and 2x tagmentation buffer (Diagenode, C01070012 and C01019043) in 50µl volume of transposition mixture that was prepared according to the manufacturer’s protocol. After mixing well by pipetting up and down, the mixture was incubated for 30 min at 37 °C. 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 either Nextera DNA Library Prep kit (Illumina, for unperturbed HEK293 replicates) or Diagenode’s tagmentase compatible primers “24 UDI for tagmented libraries sets (I, II, III)” (C01011034, C01011036, and C01011037, for the three TFs as well as two unperturbed HEK293 replicates) with an average number of 10 cycles per sample.

780 Subsequent sequencing was performed at the Donnelly Center sequencing facility using
781 100bp paired end sequencing at 60M reads per sample (for unperturbed HEK293
782 replicates in the first set) or at the McMaster Genomic facility (McMaster University)
783 using 50bp paired end sequencing at 60M reads per sample (for SOX2, TPRX1,
784 ZNF836, and unperturbed control HEK293 in the second set).

785 **ATAC-seq data analysis**. Adapter sequences were first trimmed using *cutadapt* to
786 remove the Tn5 adapter CTGTCTCTTATACACATCT. The resulting reads were then
787 mapped to the human genome (hg38) using *bowtie2*[73], followed by the creation of BAM
788 files using *samtools*[74,75] *view*, and sorting with *samtools sort*. Unintended or duplicated
789 reads were removed using *sambamba view -h -f bam -F "[XS] == null and not*
790 *unmapped and not duplicate and proper_pair and mapping_quality >= 30"* and *picard*
791 *MarkDuplicates*. The first set of ATAC-seq experiments went through peak calling on
792 the sorted BAM reads by running *macs2*[43,76,77] *callpeak* with the options *-f BAMPE, -g*
793 *hs, -B*, and *-q 0.01*. Then, to generate a single peak file for HEK293 open chromatin, all
794 the peak files were merged using *bedtools*[78] *merge*. For the second set of experiments
795 (SOX2, TPRX1, and ZNF836), the BAM files were first adjusted by *alignmentSieve –*
796 *ATACshift* to eliminate the Tn5 bias by +5bp (+ strand) and -4bp (- strand). In order to
797 make pileups (bdg) for the Tn5 tag locations, *macs3*[43,76,77] *callpeak* was used with the
798 options *--format BAM -g hs --nomodel --shift -50 --extsize 100 --keep-dup all --scale-to*
799 *small –bdg* using all the replicates of that experiment for option *-t*, followed by
800 *bedGraphToBigWig*. Finally, to get the signal over TOP sites, we used DeepTools[79]
801 *computeMatrix*, followed by a custom Python script for plotting the fold change.

802 **Chromatin state analysis.** We obtained chromatin states by training a ChromHMM[45]
803 with ten states (see **Supplementary Fig. 4a**) on marks H3K4me1, H3H4me3,
804 H3K36me3, H3K27ac, H3K9me3, and H3K27me3, collected in HEK293 cells by
805 ENCODE[81], plus ATAC-seq and CTCF peaks from HEK293 cells generated as part of
806 this project. We also employed promoter regions derived from the GENCODE
807 annotation (release 44)[82] for our analyses (-1000 to +500). For Hi-C B compartment
808 annotations, we labeled genomic regions with a Hi-C first eigenvector value less than
809 0.4 in ENCODE data for HAP1 cells[83], comprising 65% of the genome.

810 **Overlap of ChIP-seq peaks between all pairs of TFs.** Jaccard similarity is taken as O
811 / (N1+N2-O), where O is the number of intersecting peaks and N1 and N2 are the sizes
812 of each set. We utilized BEDTools[78] to calculate overlaps. To prevent miscounting of the
813 cases in which one peak in one set overlaps with multiple peaks in another set, we used
814 the average of overlapping peaks (O = (O1+O2)/2 where O1 is the number of peaks in
815 set 1 overlapping with any peak in set 2 and vice versa) to calculate the intersection in
816 Jaccard. The same methodology was used to calculate the overlap of ChIP-seq peaks
817 with the chromatin tracks.

818 **Selection of universal ChIP-seq and GHT-SELEX thresholds.** We calculated the
819 Jaccard similarity from all 137 pairs of TFs with ChIP-seq and GHT-SELEX data, using
820 the merged ChIP-seq peaks. We performed a grid search for all TFs simultaneously,
821 sampling ChIP-seq P-value and GHT-SELEX cutoffs (determined by selecting different

822 "knee" values[84] in the graph of sorted enrichment coefficients[48]), to identify a pair of
823 thresholds that maximizes median Jaccard. Two ChIP-seq cutoffs (10^{-10} and 10^{-20})
824 yielded an almost identical maximum; we chose 10^{-10} as it includes a larger number of
825 peaks. A corresponding knee specificity threshold of 30 emerged for the GHT-SELEX
826 knee-based cutoff.

827 **Filtering peaks in genome unmappable regions.** Before comparing ChIP-seq and
828 GHT-SELEX peak sets to define TOPs (see **Derivation of TOP sites** below), we first
829 filtered out the peaks overlapping with genomic regions that are problematic for any of
830 the assays. The following sources were used to make a "genomic whitelist" for our
831 study:

832 1. hg38 36-bp mappable regions on human genome: Bismap[85], downloaded
833 from UCSC: https://genome-euro.ucsc.edu/cgi-
834 bin/hgTables?db=hg38&hgta_group=map&hgta_track=umap&hgta_table=um
835 ap36&hgta_doSchema=describe+table+schema. 36-bp was chosen to match
836 GHT-SELEX fragment length.
837 2. Codebook GHT-SELEX-covered regions: all the genomic regions with at least
838 one read coverage in any GHT-SELEX experiment across the entire study.
839 3. Codebook ChIP-seq-covered regions: all the genomic regions with at least
840 one read coverage in any ChIP-seq experiment across the entire study.
841 4. ENCODE blacklist v2.
842
843 The Codebook whitelist was then made by intersecting the first three tracks and the
844 complement of the fourth. Any peak (ChIP-seq, GHT-SELEX, or PWM-predictions) with
845 less than 25% of its length overlapping with the whitelist was removed from the further
846 steps.

847 **Derivation of TOP sites.** To define binding sites supported by ChIP-seq, GHT-SELEX,
848 and PWM hits, we optimized the cut-offs of all three to maximize the overlap between all
849 three data types. We first sorted the peaks based on their statistical scores, i.e., merged
850 p-values for ChIP-seq peaks, cycle enrichment coefficient for GHT-SELEX peaks (see
851 accompanying manuscript[48]), and sum-of-affinities for clusters of PWM hits with a p-
852 value < 0.001 (from MOODS[86]), merged with neighboring hits in the case of having a
853 distance less than 200 bp. Then, for different values of *N*, we took the top *N* peaks and
854 calculated the overlap (measured as the Jaccard index; intersection of all three divided
855 by the union of all) using the top *N* ChIP-seq peaks, top *N* GHT-SELEX peaks, and top
856 *N* merged PWM hits. The *N* that maximizes the Jaccard overlap (for *N* between
857 max(100, 0.01 x n_max) and n_max) was taken as the optimized threshold, and the
858 overlap of all three sets at this threshold (*N*) is referred to as triple overlap or "TOP"
859 sites.

860 **Analysis of purifying selection and classification as conserved and unconserved**
861 **binding sites.** We extracted phyloP scores[52] for each PWM hit, and for flanking regions
862 of equal length (for a total of 100 bp including the PWM hit and its flanks) from the 241
863 eutherian mammal Zoonomia alignment[87] using DeepTools[79]. We excluded PWM hits
864 overlapping with ENCODE Blacklist sites[88] or protein coding sequences, due to the

skew in phyloP scores caused by codons. All phyloP scores 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 the 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 one-sided Wilcoxon tests, one comparing values in the PWM hits to those in the 25bp downstream flank, and the same for the 25bp upstream flank. The alternative hypothesis was that the PWM hit phyloP values were greater than the compared 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).

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[89]. The enrichment significance between GHT-SELEX and ChIP-seq TOPs and each repeat family was calculated using a one-sided Fisher's Exact Test (alternative hypothesis being greater overlap), implemented in SciPy[90]. 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 has 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. A larger collection of motifs generated for these experiments in an accompanying study[40] can be browsed at https://mex.autosome.org and downloaded at https://doi.org/10.5281/ZENODO.8327372. Source Data are provided with this paper.

CODE AVAILABILITY

The custom scripts for extracting phyloP scores for TOPs and classification as CTOPs by performing the three statistical tests are available at GitHub: https://github.com/imyellan/Codebook_CTOP_scripts.

ACKNOWLEDGEMENTS

We thank the IT Group of the Institute of Computer Science at Halle University for computational resources, Maximilian Biermann for valuable technical support, and Hadi Tabarraei for aiding with ATAC-seq experiments. 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.
- MSHERF grant 075-15-2025-014 (prev. 075-15-2024-666)
- Assignment 125091010189-3 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

AUTHOR CONTRIBUTIONS

T.R.H. conceived of the study. R.R. designed the study. R.R., A.B., H.Z., M.B., and C.H. prepared cell lines and performed the ChIP-seq experiments. A.J. and A.W.H.Y. performed and analyzed the SELEX experiments. A.F., I.Y., A.B., A.H.C. and K.U.L. performed statistical and sequence analyses. I.E.V. and Z.M.P. contributed to motif discovery and assessment, and initial evaluation of experiments. K.U.L, I.Y., A.F., and R.R. prepared the illustrations. M.A. assisted with data deposition. I.V.K., P.B., Q.M., H.S.N., and H.S.N. organized the study and oversaw the data collection and data processing. T.R.H., R.R., A.F., and I.Y. wrote the paper. All authors contributed to data analysis and reviewed the manuscript.

SUPPLEMENTAL TABLES AND DOCUMENTS

Supplementary Table 1. Overview of the tested TFs. This table lists the TFs that were tested in this study using ChIP-seq.

Supplementary Table 2. List of all the ChIP-seq experiments. This table lists the ChIP experiments, their approval status, and related produced files.

Supplementary Table 3. Summary of 217 TFs with successful ChIP-seq data. This table lists the TFs with successful ChIP-seq experiments either "Motif-enriched" or significant overlap between replicates, together with the list of ChIP-seq samples used in "merged" peaks for each TF and their binding category (i.e. Promoter TF, Dark TF, Enhancer TF, or Others)**.** 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.

Supplementary Table 4. Consolidated functional information for Dark TFs. This table provides the data underlying **Fig. 7** including the references in the literature.

Supplementary Document 1. Enrichment of transposable elements in TOPs, with expanded TE family classification. Heatmap is from **Fig. 3b**, with expanded labels for specific elements enriched in TOPs of each TF

Supplementary Document 2. Heatmaps of conservation/phyloP score across TOPs for 137 TFs. This document provides the same analysis as **Fig. 5a,** 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.

Supplementary Document 3. Overview of Dark TFs' function. This document provides a summary of functional impact of Dark TFs. It is based on a survey (**Fig. 7**) of Dark TF interaction with chromatin remodelers, other zinc-finger proteins, the specific categories and age of enriched TEs within each TFs TOPs, their motifs, as well as documented functions based on literature review.

Supplementary Document 4. Analysis of binding site ages for TFs. This document provides a summary of evolutionary analysis performed to estimate the age of each TOP site for all the TFs in this study, as well as the age of the TFs.

1267 **REFERENCES**

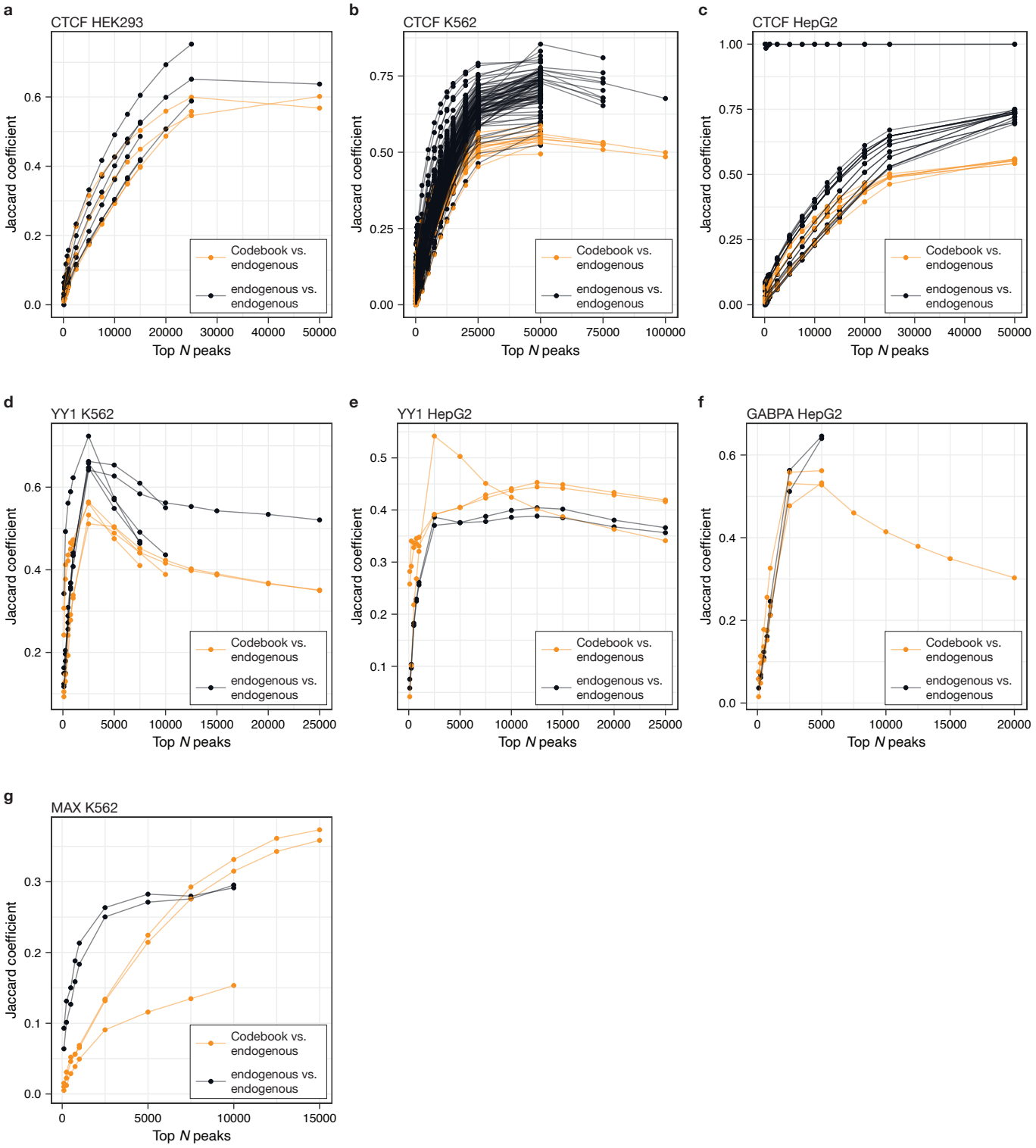
1268 1. Gemmell, N.J. Repetitive DNA: genomic dark matter matters. *Nat Rev Genet* **22**, 342
1269 (2021).
1270 2. Alexander, R.P., Fang, G., Rozowsky, J., Snyder, M. & Gerstein, M.B. Annotating non-
1271 coding regions of the genome. *Nat Rev Genet* **11**, 559–71 (2010).
1272 3. Britten, R.J. & Davidson, E.H. Gene regulation for higher cells: a theory. *Science* **165**,
1273 349–57 (1969).
1274 4. Nobrega, M.A., Ovcharenko, I., Afzal, V. & Rubin, E.M. Scanning human gene deserts for
1275 long-range enhancers. *Science* **302**, 413 (2003).
1276 5. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–51 (2001).
1277 6. Nobrega, M.A., Zhu, Y., Plajzer-Frick, I., Afzal, V. & Rubin, E.M. Megabase deletions of
1278 gene deserts result in viable mice. *Nature* **431**, 988–93 (2004).
1279 7. Imbeault, M., Helleboid, P.Y. & Trono, D. KRAB zinc-finger proteins contribute to the
1280 evolution of gene regulatory networks. *Nature* **543**, 550–554 (2017).
1281 8. Cosby, R.L. *et al.* Recurrent evolution of vertebrate transcription factors by transposase
1282 capture. *Science* **371**(2021).
1283 9. Schmidt, D. *et al.* Waves of retrotransposon expansion remodel genome organization
1284 and CTCF binding in multiple mammalian lineages. *Cell* **148**, 335–48 (2012).
1285 10. Chuong, E.B., Elde, N.C. & Feschotte, C. Regulatory evolution of innate immunity
1286 through co-option of endogenous retroviruses. *Science* **351**, 1083–7 (2016).
1287 11. Cohen, C.J., Lock, W.M. & Mager, D.L. Endogenous retroviral LTRs as promoters for
1288 human genes: a critical assessment. *Gene* **448**, 105–14 (2009).
1289 12. Ernst, J. *et al.* Genome-scale high-resolution mapping of activating and repressive
1290 nucleotides in regulatory regions. *Nat Biotechnol* **34**, 1180–1190 (2016).
1291 13. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of
1292 human embryonic stem cells. *Nat Genet* **42**, 631–4 (2010).
1293 14. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–66
1294 (2015).
1295 15. Jacobs, F.M. *et al.* An evolutionary arms race between KRAB zinc-finger genes ZNF91/93
1296 and SVA/L1 retrotransposons. *Nature* **516**, 242–5 (2014).
1297 16. Najafabadi, H.S. *et al.* C2H2 zinc finger proteins greatly expand the human regulatory
1298 lexicon. *Nat Biotechnol* (2015).
1299 17. Bannister, A.J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the
1300 HP1 chromo domain. *Nature* **410**, 120–4 (2001).
1301 18. Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3
1302 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**, 116–20 (2001).
1303 19. Ayyanathan, K. *et al.* Regulated recruitment of HP1 to a euchromatic gene induces
1304 mitotically heritable, epigenetic gene silencing: a mammalian cell culture model of gene
1305 variegation. *Genes Dev* **17**, 1855–69 (2003).
1306 20. Emerson, R.O. & Thomas, J.H. Adaptive evolution in zinc finger transcription factors.
1307 *PLoS Genet* **5**, e1000325 (2009).
1308 21. Senft, A.D. & Macfarlan, T.S. Transposable elements shape the evolution of mammalian
1309 development. *Nat Rev Genet* **22**, 691–711 (2021).

1310 22. Milovanović, D. *et al.* Tissue-specific restriction of TE-derived regulatory elements
1311 safeguards cell-type identity. *bioRxiv*, 2025.05.13.653700 (2025).
1312 23. Lambert, S.A. *et al.* The Human Transcription Factors. *Cell* **175**, 598–599 (2018).
1313 24. Partridge, E.C. *et al.* Occupancy maps of 208 chromatin-associated proteins in one
1314 human cell type. *Nature* **583**, 720–728 (2020).
1315 25. Long, H.K., Prescott, S.L. & Wysocka, J. Ever-Changing Landscapes: Transcriptional
1316 Enhancers in Development and Evolution. *Cell* **167**, 1170–1187 (2016).
1317 26. Zhu, F. *et al.* The interaction landscape between transcription factors and the
1318 nucleosome. *Nature* **562**, 76–81 (2018).
1319 27. Zaret, K.S. & Carroll, J.S. Pioneer transcription factors: establishing competence for gene
1320 expression. *Genes Dev* **25**, 2227–41 (2011).
1321 28. Courey, A.J. Cooperativity in transcriptional control. *Curr Biol* **11**, R250–2 (2001).
1322 29. Stubbs, L., Sun, Y. & Caetano-Anolles, D. Function and Evolution of C2H2 Zinc Finger
1323 Arrays. *Subcell Biochem* **52**, 75–94 (2011).
1324 30. Najafabadi, H.S., Albu, M. & Hughes, T.R. Identification of C2H2-ZF binding preferences
1325 from ChIP-seq data using RCADE. *Bioinformatics* **31**, 2879–81 (2015).
1326 31. Worsley Hunt, R. & Wasserman, W.W. Non-targeted transcription factors motifs are a
1327 systemic component of ChIP-seq datasets. *Genome Biol* **15**, 412 (2014).
1328 32. Auerbach, R.K. *et al.* Mapping accessible chromatin regions using Sono-Seq. *Proc Natl*
1329 *Acad Sci U S A* **106**, 14926–31 (2009).
1330 33. Schmitges, F.W. *et al.* Multiparameter functional diversity of human C2H2 zinc finger
1331 proteins. *Genome Res* **26**, 1742–1752 (2016).
1332 34. Jolma, A. *et al.* Perspectives on Codebook: sequence specificity of uncharacterized
1333 human transcription factors. *bioRxiv*, 2024.11.11.622097 (2024).
1334 35. Jolma, A. *et al.* GHT-SELEX demonstrates unexpectedly high intrinsic sequence specificity
1335 and complex DNA binding of many human transcription factors. *bioRxiv*,
1336 2024.11.11.618478 (2024).
1337 36. Vorontsov, I.E. *et al.* Cross-platform DNA motif discovery and benchmarking to explore
1338 binding specificities of poorly studied human transcription factors. *bioRxiv*,
1339 2024.11.11.619379 (2024).
1340 37. Lambert, S.A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).
1341 38. Isakova, A. *et al.* SMiLE-seq identifies binding motifs of single and dimeric transcription
1342 factors. *Nat Methods* **14**, 316–322 (2017).
1343 39. Kean, M.J. *et al.* Structure-function analysis of core STRIPAK Proteins: a signaling
1344 complex implicated in Golgi polarization. *J Biol Chem* **286**, 25065–75 (2011).
1345 40. Vorontsov, I.E. *et al.* Cross-platform DNA motif discovery and benchmarking to explore
1346 binding specificities of poorly studied human transcription factors. *bioRxiv*,
1347 2024.11.11.619379 (2024).
1348 41. Berger, M.F. *et al.* Compact, universal DNA microarrays to comprehensively determine
1349 transcription-factor binding site specificities. *Nat Biotechnol* **24**, 1429–35 (2006).
1350 42. Jolma, A. *et al.* Multiplexed massively parallel SELEX for characterization of human
1351 transcription factor binding specificities. *Genome Res* **20**, 861–73 (2010).
1352 43. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).

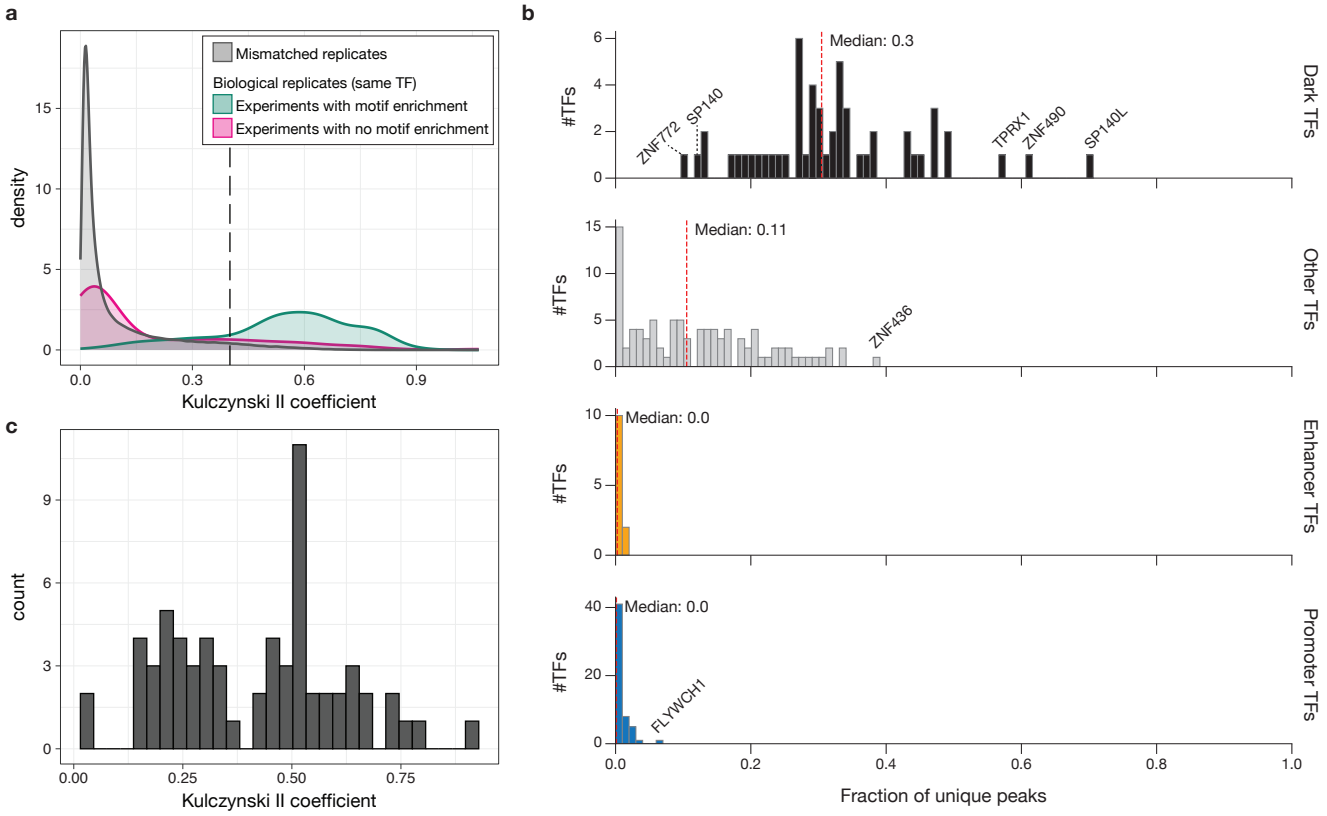
1353 44. O'Leary, N.A. *et al.* Reference sequence (RefSeq) database at NCBI: current status,
1354 taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733–45 (2016).
1355 45. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and
1356 characterization. *Nat Methods* **9**, 215–6 (2012).
1357 46. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target
1358 genes in GeneCards. *Database (Oxford)* **2017**(2017).
1359 47. Kolmykov, S. *et al.* GTRD: an integrated view of transcription regulation. *Nucleic Acids*
1360 *Res* **49**, D104–D111 (2021).
1361 48. Jolma, A. *et al.* GHT-SELEX demonstrates unexpectedly high intrinsic sequence specificity
1362 and complex DNA binding of many human transcription factors. *bioRxiv*,
1363 2024.11.11.618478 (2024).
1364 49. Badis, G. *et al.* A library of yeast transcription factor motifs reveals a widespread
1365 function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol Cell* **32**, 878–87
1366 (2008).
1367 50. Core, L.J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation
1368 regions at mammalian promoters and enhancers. *Nat Genet* **46**, 1311–20 (2014).
1369 51. FitzGerald, P.C., Shlyakhtenko, A., Mir, A.A. & Vinson, C. Clustering of DNA sequences in
1370 human promoters. *Genome Res* **14**, 1562–74 (2004).
1371 52. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral
1372 substitution rates on mammalian phylogenies. *Genome Res* **20**, 110–21 (2010).
1373 53. Zou, Z. *et al.* Translatome and transcriptome co-profiling reveals a role of TPRXs in
1374 human zygotic genome activation. *Science* **378**, abo7923 (2022).
1375 54. Helleboid, P.Y. *et al.* The interactome of KRAB zinc finger proteins reveals the
1376 evolutionary history of their functional diversification. *EMBO J* **38**, e101220 (2019).
1377 55. Huttlin, E.L. *et al.* Dual proteome-scale networks reveal cell-specific remodeling of the
1378 human interactome. *Cell* **184**, 3022–3040 e28 (2021).
1379 56. Kuroda, T. *et al.* SALL3 expression balance underlies lineage biases in human induced
1380 pluripotent stem cell differentiation. *Nat Commun* **10**, 2175 (2019).
1381 57. Singh, J.K. *et al.* Zinc finger protein ZNF384 is an adaptor of Ku to DNA during classical
1382 non-homologous end-joining. *Nat Commun* **12**, 6560 (2021).
1383 58. Feu, S. *et al.* OZF is a Claspin-interacting protein essential to maintain the replication
1384 fork progression rate under replication stress. *FASEB J* **34**, 6907–6919 (2020).
1385 59. Ge, L.P. *et al.* ZNF689 deficiency promotes intratumor heterogeneity and
1386 immunotherapy resistance in triple-negative breast cancer. *Cell Res* **34**, 58–75 (2024).
1387 60. Zhou, M. *et al.* ZBTB40 is a telomere-associated protein and protects telomeres in
1388 human ALT cells. *J Biol Chem* **299**, 105053 (2023).
1389 61. Cui, Y., Zhou, M., He, Q. & He, Z. Zbtb40 Deficiency Leads to Morphological and
1390 Phenotypic Abnormalities of Spermatocytes and Spermatozoa and Causes Male
1391 Infertility. *Cells* **12**(2023).
1392 62. Campos-Sanchez, R., Kapusta, A., Feschotte, C., Chiaromonte, F. & Makova, K.D.
1393 Genomic landscape of human, bat, and ex vivo DNA transposon integrations. *Mol Biol*
1394 *Evol* **31**, 1816–32 (2014).

1395 63. Lewin, T.D., Royall, A.H. & Holland, P.W.H. Dynamic Molecular Evolution of Mammalian
1396 Homeobox Genes: Duplication, Loss, Divergence and Gene Conversion Sculpt PRD Class
1397 Repertoires. *J Mol Evol* **89**, 396–414 (2021).
1398 64. Thomas, J.H. & Schneider, S. Coevolution of retroelements and tandem zinc finger
1399 genes. *Genome Res* **21**, 1800–12 (2011).
1400 65. Nowick, K. *et al.* Gain, loss and divergence in primate zinc-finger genes: a rich resource
1401 for evolution of gene regulatory differences between species. *PLoS One* **6**, e21553
1402 (2011).
1403 66. Nowick, K., Hamilton, A.T., Zhang, H. & Stubbs, L. Rapid sequence and expression
1404 divergence suggest selection for novel function in primate-specific KRAB-ZNF genes. *Mol*
1405 *Biol Evol* **27**, 2606–17 (2010).
1406 67. Alexander, T.B. *et al.* The genetic basis and cell of origin of mixed phenotype acute
1407 leukaemia. *Nature* **562**, 373–379 (2018).
1408 68. Arber, D.A. *et al.* International Consensus Classification of Myeloid Neoplasms and Acute
1409 Leukemias: integrating morphologic, clinical, and genomic data. *Blood* **140**, 1200–1228
1410 (2022).
1411 69. Denny, P. & Ashworth, A. A zinc finger protein-encoding gene expressed in the post-
1412 meiotic phase of spermatogenesis. *Gene* **106**, 221–7 (1991).
1413 70. Lord, T. *et al.* A novel high throughput screen to identify candidate molecular networks
1414 that regulate spermatogenic stem cell functionsdagger. *Biol Reprod* **106**, 1175–1190
1415 (2022).
1416 71. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–5
1417 (2013).
1418 72. Schmidt, D. *et al.* ChIP-seq: using high-throughput sequencing to discover protein-DNA
1419 interactions. *Methods* **48**, 240–8 (2009).
1420 73. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods*
1421 **9**, 357–9 (2012).
1422 74. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
1423 2078–9 (2009).
1424 75. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**(2021).
1425 76. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X.S. Identifying ChIP-seq enrichment using
1426 MACS. *Nat Protoc* **7**, 1728–40 (2012).
1427 77. Liu, T. Use model-based Analysis of ChIP-Seq (MACS) to analyze short reads generated
1428 by sequencing protein-DNA interactions in embryonic stem cells. *Methods Mol Biol*
1429 **1150**, 81–95 (2014).
1430 78. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic
1431 features. *Bioinformatics* **26**, 841–2 (2010).
1432 79. Ramirez, F., Dundar, F., Diehl, S., Gruning, B.A. & Manke, T. deepTools: a flexible
1433 platform for exploring deep-sequencing data. *Nucleic Acids Res* **42**, W187–91 (2014).
1434 80. Grandi, F.C., Modi, H., Kampman, L. & Corces, M.R. Chromatin accessibility profiling by
1435 ATAC-seq. *Nat Protoc* **17**, 1518–1552 (2022).
1436 81. Consortium, E.P. *et al.* Perspectives on ENCODE. *Nature* **583**, 693–698 (2020).
1437 82. Frankish, A. *et al.* GENCODE: reference annotation for the human and mouse genomes
1438 in 2023. *Nucleic Acids Res* **51**, D942–D949 (2023).

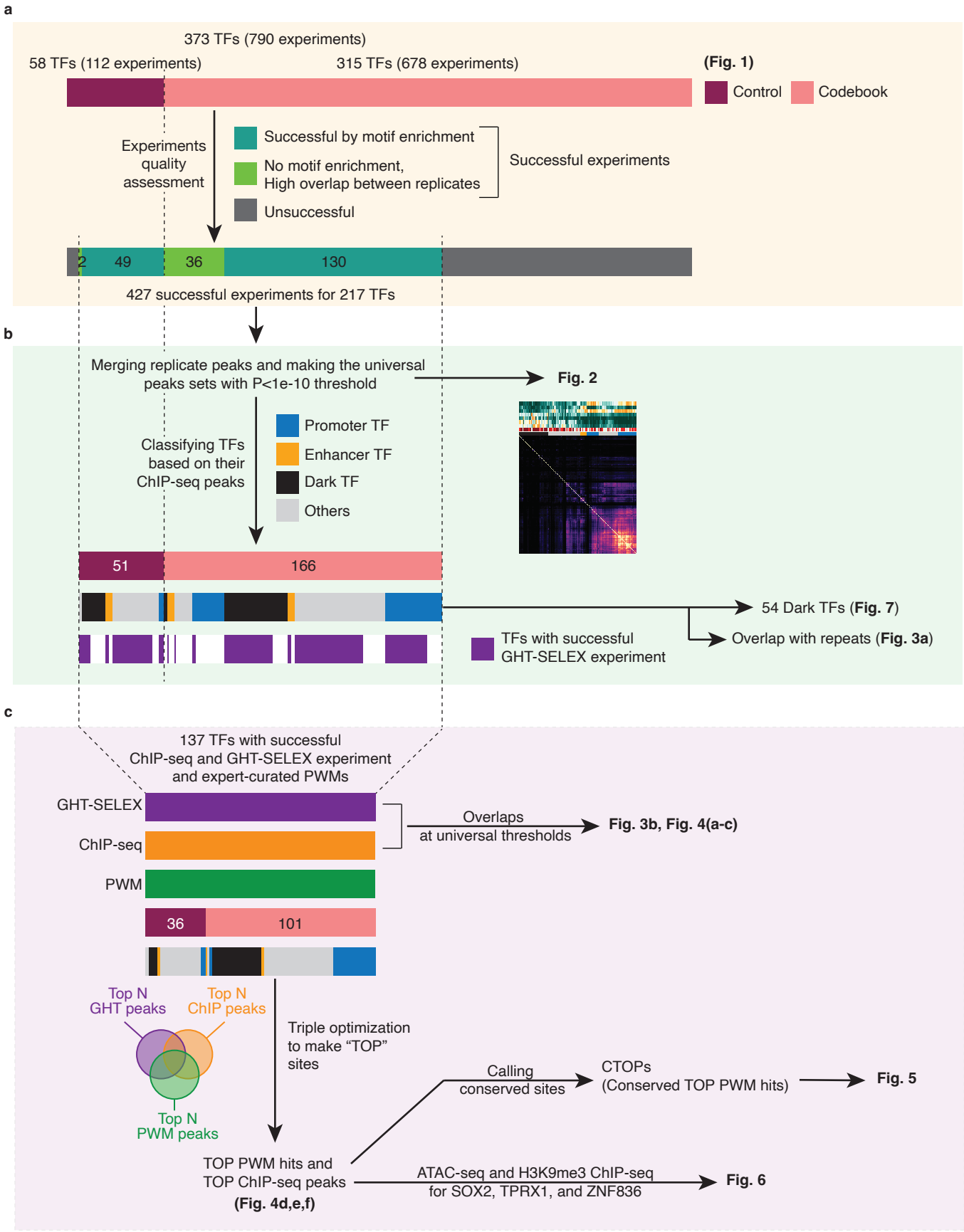
1439 83. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of
1440 chromatin interactions. *Nature* **485**, 376–80 (2012).
1441 84. Satopaa, V., Albrecht, J., Irwin, D. & Raghavan, B. Finding a" kneedle" in a haystack:
1442 Detecting knee points in system behavior. in *2011 31st international conference on*
1443 *distributed computing systems workshops* 166–171 (IEEE, 2011).
1444 85. Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M.M. Umap and Bismap: quantifying
1445 genome and methylome mappability. *Nucleic Acids Res* **46**, e120 (2018).
1446 86. Korhonen, J., Martinmaki, P., Pizzi, C., Rastas, P. & Ukkonen, E. MOODS: fast search for
1447 position weight matrix matches in DNA sequences. *Bioinformatics* **25**, 3181–2 (2009).
1448 87. Armstrong, J. *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-
1449 genome era. *Nature* **587**, 246–251 (2020).
1450 88. Amemiya, H.M., Kundaje, A. & Boyle, A.P. The ENCODE Blacklist: Identification of
1451 Problematic Regions of the Genome. *Sci Rep* **9**, 9354 (2019).
1452 89. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements.
1453 *Trends Genet* **16**, 418–20 (2000).
1454 90. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python.
1455 *Nat Methods* **17**, 261–272 (2020).
1456 91. Storer, J., Hubley, R., Rosen, J., Wheeler, T.J. & Smit, A.F. The Dfam community resource
1457 of transposable element families, sequence models, and genome annotations. *Mob DNA*
1458 **12**, 2 (2021).
1459 92. Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of
1460 curated protein, genetic, and chemical interactions. *Protein Sci* **30**, 187–200 (2021).
1461 93. Lambert, S.A., Albu, M., Hughes, T.R. & Najafabadi, H.S. Motif comparison based on
1462 similarity of binding affinity profiles. *Bioinformatics* **32**, 3504–3506 (2016).
1463


a
CTCF HEK293
b
CTCF K562
c
CTCF HepG2
d
YY1 K562
e
YY1 HepG2
f
GABPA HepG2
g
MAX K562
Jaccard coefficient
Top N peaks
Codebook vs. endogenous
endogenous vs. endogenous

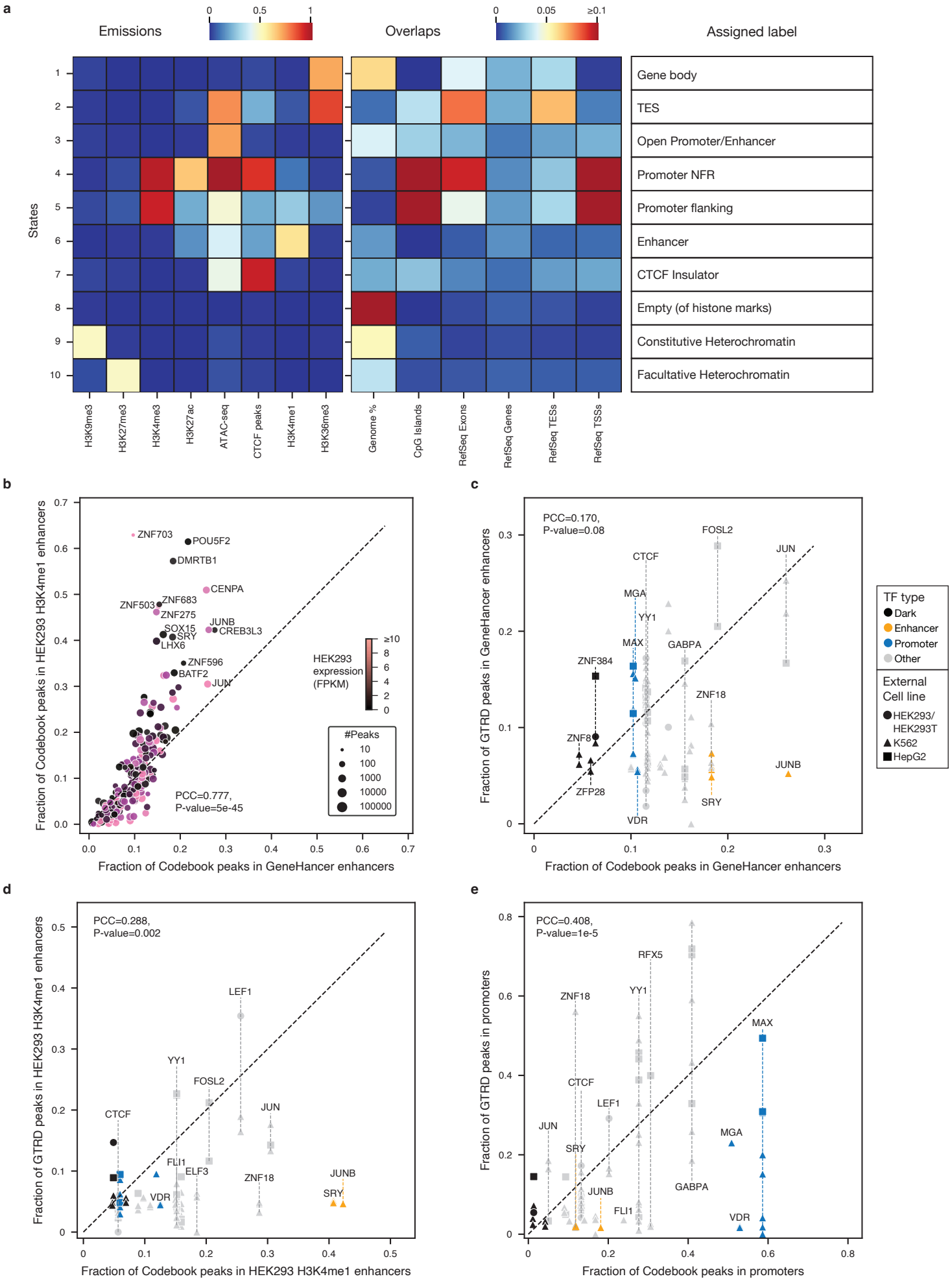
Supplementary Fig. 1. Comparison of ChIP-seq peaks obtained from ectopically expressed TF in Codebook project with ChIP-seq peaks from previous studies using endogenous TFs. Individual plots show the overlap (Jaccard coefficient) between peak sets for pairs of ChIP experiments. ChIP peaks are ordered by their scores and the Jaccard coefficient is calculated using the top *N* peaks from each experiment, with different values of *N*. Comparisons between pairs of ChIP experiments that IP the endogenous TF (endogenous vs. endogenous) are distinguished from comparisons between a Codebook ChIP experiment and a ChIP experiment that IPs the endogenous TF (Codebook vs endogenous). All Codebook ChIP experiments are performed in HEK293 cells, each panel shows comparisons for a specific TF with endogenous IP experiments performed in a specific cell line. (**a**), CTCF, HEK293 cells. (**b**) CTCF, K562 cells. (**c)** CTCF, HepG2 cells. (**d**) YY1, K562 cells. (**e**) YY1, HepG2 cells. (**f**) GABPA, HepG2 cells. (**g**) MAX, K572 cells.


a
Mismatched replicates
Biological replicates (same TF)
Experiments with motif enrichment
Experiments with no motif enrichment
density
Kulczynski II coefficient
b
Median: 0.3
ZNF772
SP140
TPRX1
ZNF490
SP140L
Dark TFs
Median: 0.11
ZNF436
Other TFs
Median: 0.0
Enhancer TFs
Median: 0.0
FLYWCH1
Promoter TFs
#TFs
Fraction of unique peaks
c
count
Kulczynski II coefficient

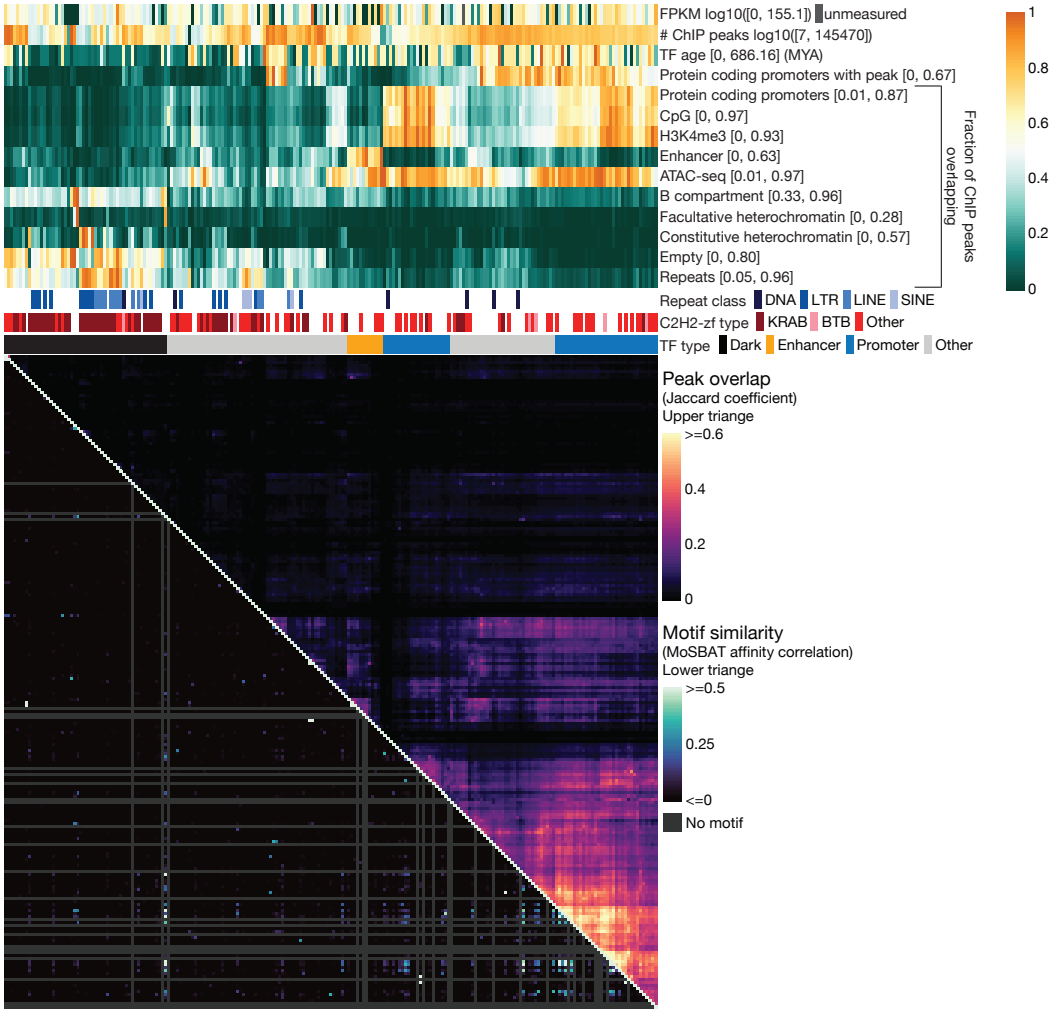
Supplementary Fig. 2. Peak overlap assessment of ChIP-seq as a measure of success. (**a**) Distribution of peak overlap between ChIP-seq replicates, for the experiments yielded a motif (green), experiments that did not yield a motif (red), and the mismatched (i.e., TF identities permuted) replicates (grey). Values are calculated Kulczynski II similarity metric (i.e. average of overlaps). Dotted line indicates the threshold at which the peak overlap of experimental replicates without motif enrichment considered successful and were included in downstream analyses. (**b**) Distribution peak uniqueness for defined TF categories (Promoter, Enhancer, Dark, and Other), 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 successful ChIP-seq replicates (as in (**a**)), restricted to the TFs that have a low peak overlap with other TFs (specifically, the 87 TFs in the upper left darker region of the square matrix in **Fig. 2b**).


a
373 TFs (790 experiments)
58 TFs (112 experiments)
315 TFs (678 experiments)
(Fig. 1)
Control
Codebook
Experiments quality assessment
Successful by motif enrichment
No motif enrichment, High overlap between replicates
Unsuccessful
Successful experiments
2
49
36
130
427 successful experiments for 217 TFs
b
Merging replicate peaks and making the universal peaks sets with P<1e-10 threshold
Fig. 2
Classifying TFs based on their ChIP-seq peaks
Promoter TF
Enhancer TF
Dark TF
Others
51
166
54 Dark TFs (Fig. 7)
Overlap with repeats (Fig. 3a)
TFs with successful GHT-SELEX experiment
c
137 TFs with successful ChIP-seq and GHT-SELEX experiment and expert-curated PWMs
GHT-SELEX
ChIP-seq
PWM
Overlaps at universal thresholds
Fig. 3b, Fig. 4(a-c)
36
101
Top N GHT peaks
Top N ChIP peaks
Top N PWM peaks
Triple optimization to make "TOP" sites
Calling conserved sites
CTOPs (Conserved TOP PWM hits)
Fig. 5
TOP PWM hits and TOP ChIP-seq peaks (Fig. 4d,e,f)
ATAC-seq and H3K9me3 ChIP-seq for SOX2, TPRX1, and ZNF836
Fig. 6

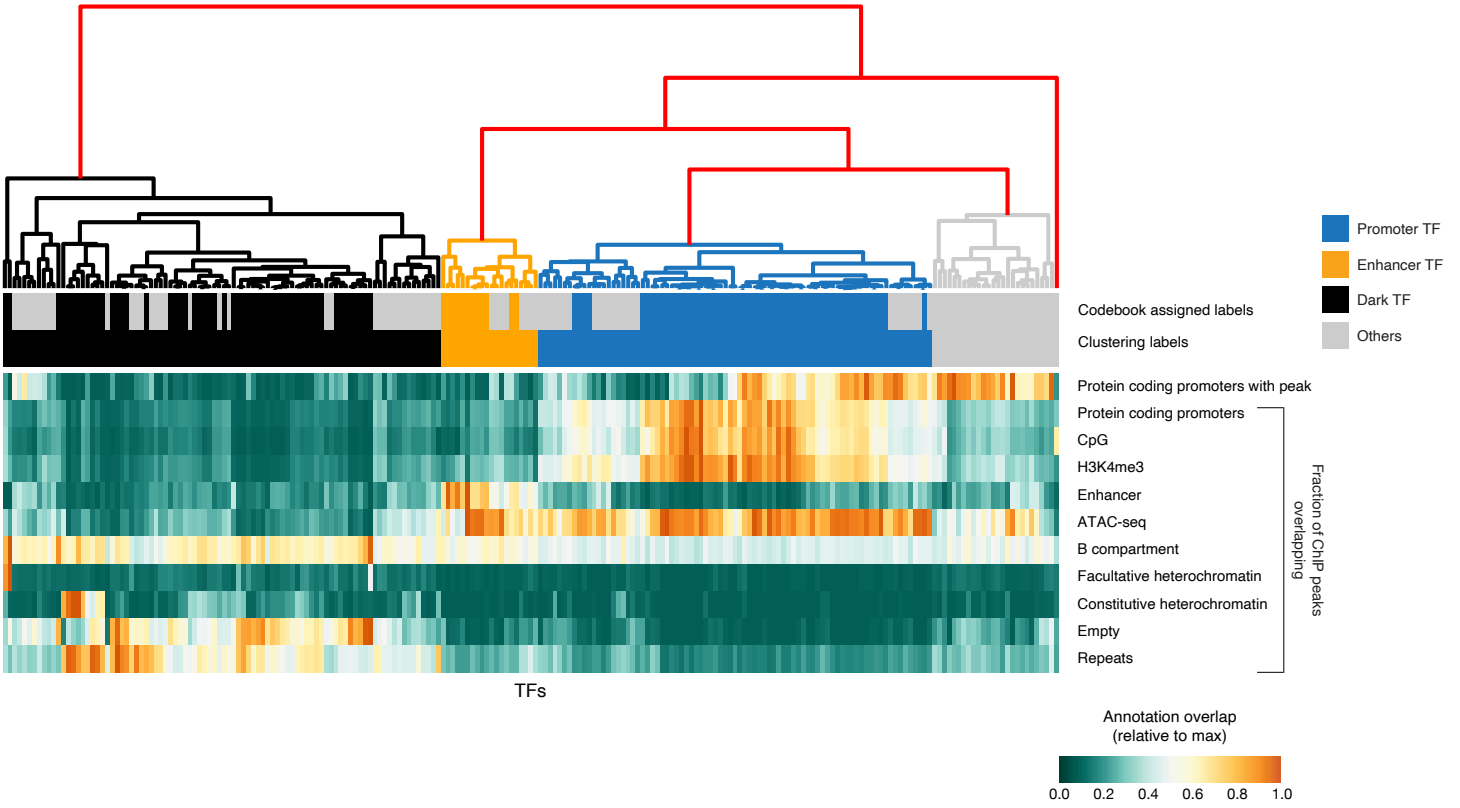
Supplementary Fig. 3. Overview of ChP-seq data processing. (**a**) selection process for successful ChIP-experiments based on replicate overlap and motif enrichment. Data for 790 ChIP experiments (including replicates for 373 TFs) were assessed and 427 (217 TFs) were considered successful. (**b**) peaks for the successful replicates were merged to create one set of peaks for each TF. The overlap of peak sets with each other and various genomic tracks were evaluated and four binding categories of TFs emerged. 137 of these TFs also had successful GHT-SELEX experiments as labelled in purple. (**c**) defining triple optimized overlap peaks (TOPs) and conserved TOPs (CTOPs). TOPs were defined based on overlap of ChIP-seq and GHT-SELEX peaks as well as PWM predicted binding sites at optimized thresholds (see **Methods**). CTOPs were defied based on phyloP conservation pattern of the PWM hit.


a
Emissions
0
0.5
1
Overlaps
0
0.05
≥0.1
Assigned label
States
1
2
3
4
5
6
7
8
9
10
Gene body
TES
Open Promoter/Enhancer
Promoter NFR
Promoter flanking
Enhancer
CTCF Insulator
Empty (of histone marks)
Constitutive Heterochromatin
Facultative Heterochromatin
H3K9me3
H3K27me3
H3K4me3
H3K27ac
ATAC-seq
CTCF peaks
H3K4me1
H3K36me3
Genome %
CpG Islands
RefSeq Exons
RefSeq Genes
RefSeq TESs
RefSeq TSSs
b
Fraction of Codebook peaks in HEK293 H3K4me1 enhancers
Fraction of Codebook peaks in GeneHancer enhancers
ZNF703
POU5F2
DMRTB1
CENPA
ZNF683
ZNF503
ZNF275
JUNB
CREB3L3
SOX15
SRY
LHX6
ZNF596
BATF2
JUN
HEK293 expression (FPKM)
≥10
8
6
4
2
0
#Peaks
10
100
1000
10000
100000
PCC=0.777,
P-value=5e-45
c
PCC=0.170,
P-value=0.08
Fraction of GTRD peaks in GeneHancer enhancers
Fraction of Codebook peaks in GeneHancer enhancers
FOSL2
JUN
CTCF
MGA
YY1
MAX
GABPA
ZNF384
ZNF18
ZNF8
JUNB
ZFP28
SRY
VDR
TF type
Dark
Enhancer
Promoter
Other
External Cell line
HEK293/ HEK293T
K562
HepG2
d
PCC=0.288,
P-value=0.002
Fraction of GTRD peaks in HEK293 H3K4me1 enhancers
Fraction of Codebook peaks in HEK293 H3K4me1 enhancers
LEF1
YY1
FOSL2
JUN
CTCF
FLI1
ELF3
ZNF18
JUNB
SRY
VDR
e
PCC=0.408,
P-value=1e-5
Fraction of GTRD peaks in promoters
Fraction of Codebook peaks in promoters
RFX5
YY1
ZNF18
MAX
CTCF
LEF1
JUN
MGA
SRY
JUNB
GABPA
FLI1
VDR

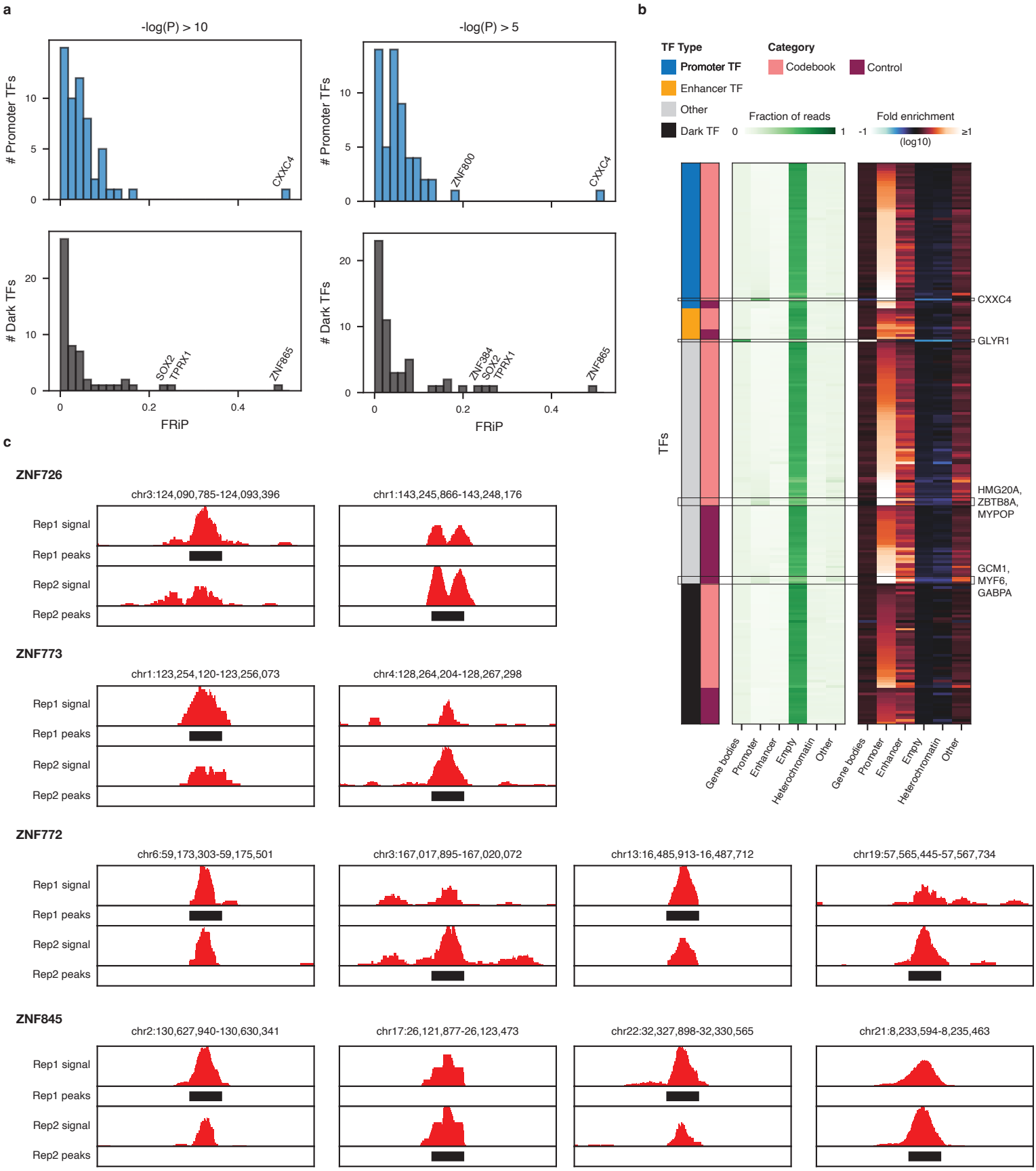
Supplementary Fig. 4. Generating HEK293-specific chromatin states. (**a**) Characterization of output tracks for the ChromHMM with 10 states trained on various HEK293-specific ChIP-seq and ATAC-seq data (i.e., ChIP-seq data for H3K9me3, H3K27me3, H3K4me1, H3K4me3, H3K36me3, and H3K27ac from ENCODE project, and ATAC-seq and CTCF peaks generated here by Codebook project). Based on the correspondence between emissions and the chromatin marks and genome annotations, states were assigned to: Gene body, Transcription end site (TES), Open Promoter/Enhancer, Promoter NFR (nucleosome-free regions), Promoter flanking region, Enhancer, CTCF Insulator, Empty (of assigned histone marks), Constitutive Heterochromatin, and Facultative Heterochromatin. (**b**) the panel shows the fraction of TF ChIP-seq peaks overlap with GeneHancer annotated enhancers (x-axis) and HEK293 enhancers (defined by H3K4me1-positive regions from ChromHMM; y-axis). Circle sizes (TFs) are scaled based on peak numbers and color gradient corresponds to endogenous expression level of TFs in HEK293 cells[33]. (**c, d, e**) Comparison of Codebook ChIP-seq peaks and publicly available ChIP-seq or ChIP-exo data on GTRD, for 21 control TFs, based on the fraction of peaks overlapping with GeneHancer enhancers (**c**), HEK293 defined enhancers (**d**), and protein-coding promoters (**e**). For the Codebook data (x-axis), the merged peak set with the universal threshold has been used.


FPKM log10([0, 155.1]) unmeasured
ChIP peaks log10([7, 145470])
TF age [0, 686.16] (MYA)
Protein coding promoters with peak [0, 0.67]
Protein coding promoters [0.01, 0.87]
CpG [0, 0.97]
H3K4me3 [0, 0.93]
Enhancer [0, 0.63]
ATAC-seq [0.01, 0.97]
B compartment [0.33, 0.96]
Facultative heterochromatin [0, 0.28]
Constitutive heterochromatin [0, 0.57]
Empty [0, 0.80]
Repeats [0.05, 0.96]
Fraction of ChIP peaks overlapping
1
0.8
0.6
0.4
0.2
0
Repeat class DNA LTR LINE SINE
C2H2-zf type KRAB BTB Other
TF type Dark Enhancer Promoter Other
Peak overlap
(Jaccard coefficient)
Upper triange
>=0.6
0.4
0.2
0
Motif similarity
(MoSBAT affinity correlation)
Lower triange
>=0.5
0.25
<=0
No motif

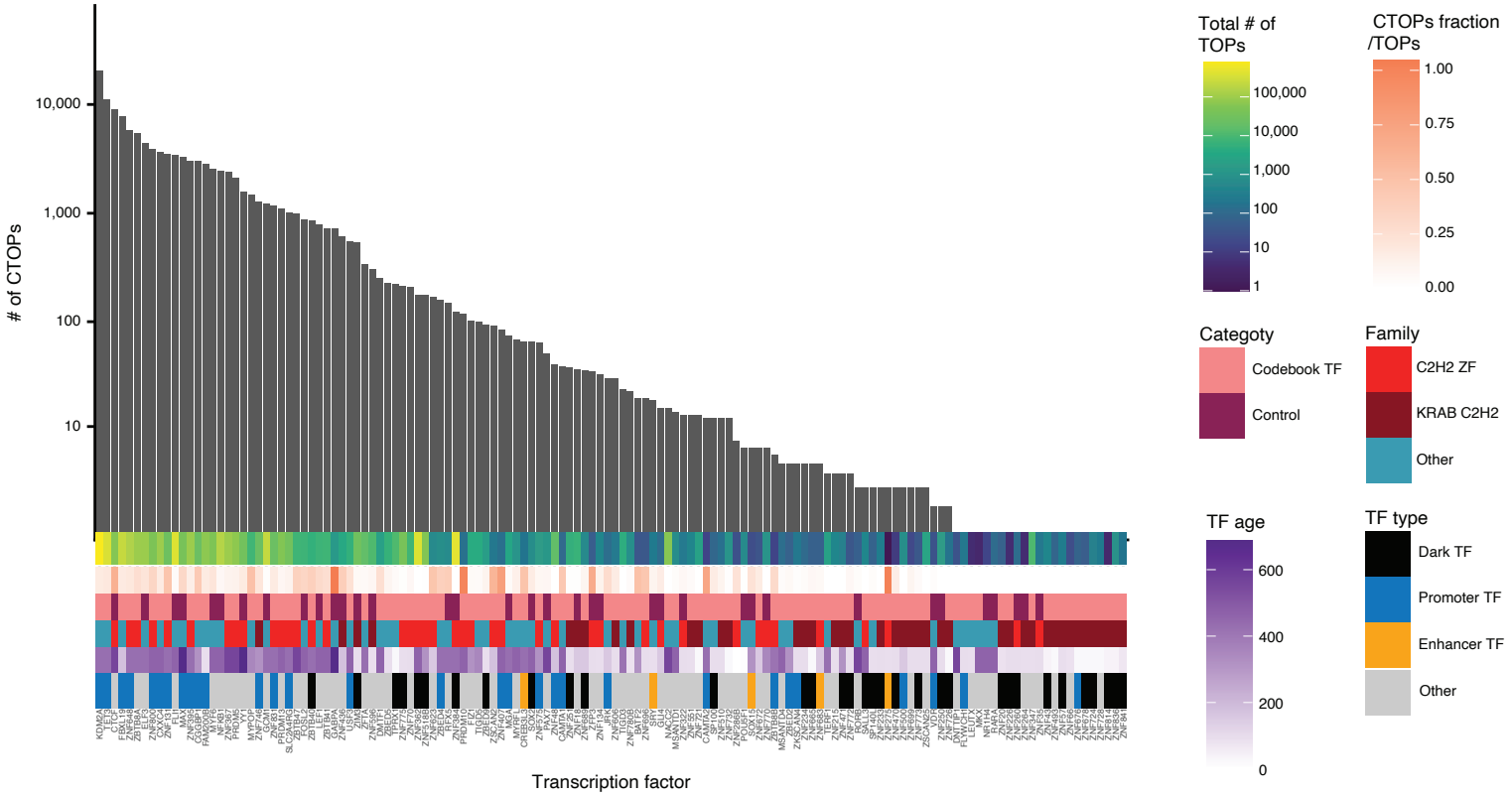
Supplementary Fig. 5. Expanded information for genomic tracks and protein expression data for 217 TF binding sites. This is a more detailed version of **Fig. 2**. Additional tracks, such as gene expression in HEK293 cells (FPKM)[33], number of total ChIP-seq peaks (at the universal threshold of MACS2 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 are included. The upper triangle in the bottom square is the same as **Fig. 2**, however, the lower triangle here shows the similarity between PWMs for each pair of TFs, calculated by MoSBAT[93]. Gray stripes correspond to the TFs without a selected PWM in the Codebook set.


Promoter TF
Enhancer TF
Dark TF
Others
Codebook assigned labels
Clustering labels
Protein coding promoters with peak
Protein coding promoters
CpG
H3K4me3
Enhancer
ATAC-seq
B compartment
Facultative heterochromatin
Constitutive heterochromatin
Empty
Repeats
Fraction of ChIP peaks overlapping
TFs
Annotation overlap (relative to max)
0.0
0.2
0.4
0.6
0.8
1.0

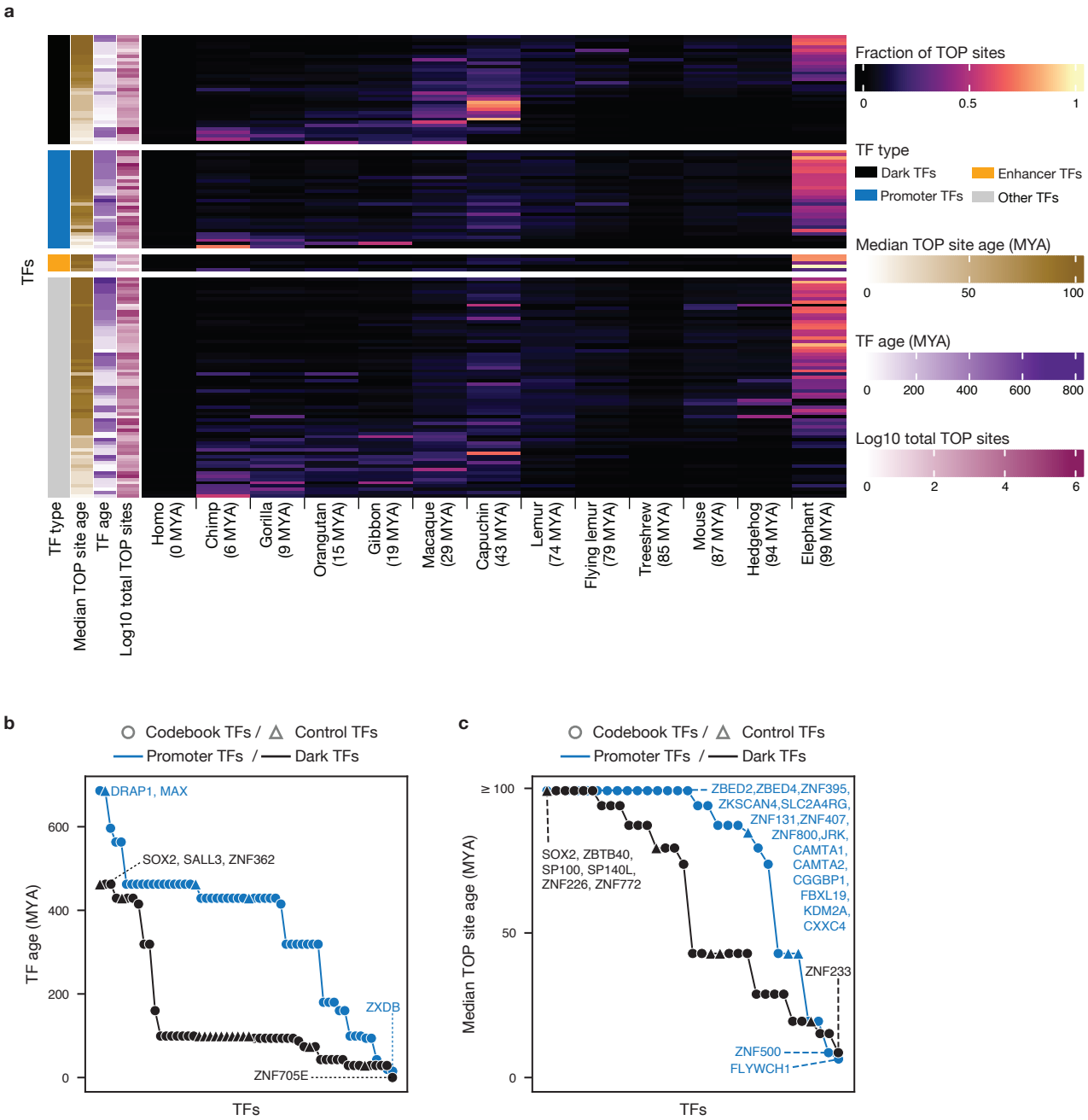
Supplementary Fig. 6. Automated clustering of the TFs and comparison with expert labeling. The 217 TFs with successful ChIP-seq data were clustered using the features presented at the top of **Fig. 2b**, applying hierarchical clustering with a correlation similarity metric and average linkage. The top five major branches (red) were then assigned to Promoter TFs, Dark TFs, and Enhancer TFs, with two branches (one containing only a single member) left as Others (see “Clustering labels”). The “Codebook assigned labels” are shown on top for comparison.


a
-log(P) > 10
-log(P) > 5
Promoter TFs
Dark TFs
FRiP
CXXC4
ZNF800
SOX2
TPRX1
ZNF865
ZNF384
b
TF Type
Promoter TF
Enhancer TF
Other
Dark TF
Category
Codebook
Control
Fraction of reads
Fold enrichment
(log10)
TFs
CXXC4
GLYR1
HMG20A,
ZBTB8A,
MYPOP
GCM1,
MYF6,
GABPA
Gene bodies
Promoter
Enhancer
Empty
Heterochromatin
Other
c
ZNF726
chr3:124,090,785-124,093,396
chr1:143,245,866-143,248,176
ZNF773
chr1:123,254,120-123,256,073
chr4:128,264,204-128,267,298
ZNF772
chr6:59,173,303-59,175,501
chr3:167,017,895-167,020,072
chr13:16,485,913-16,487,712
chr19:57,565,445-57,567,734
ZNF845
chr2:130,627,940-130,630,341
chr17:26,121,877-26,123,473
chr22:32,327,898-32,330,565
chr21:8,233,594-8,235,463
Rep1 signal
Rep1 peaks
Rep2 signal
Rep2 peaks

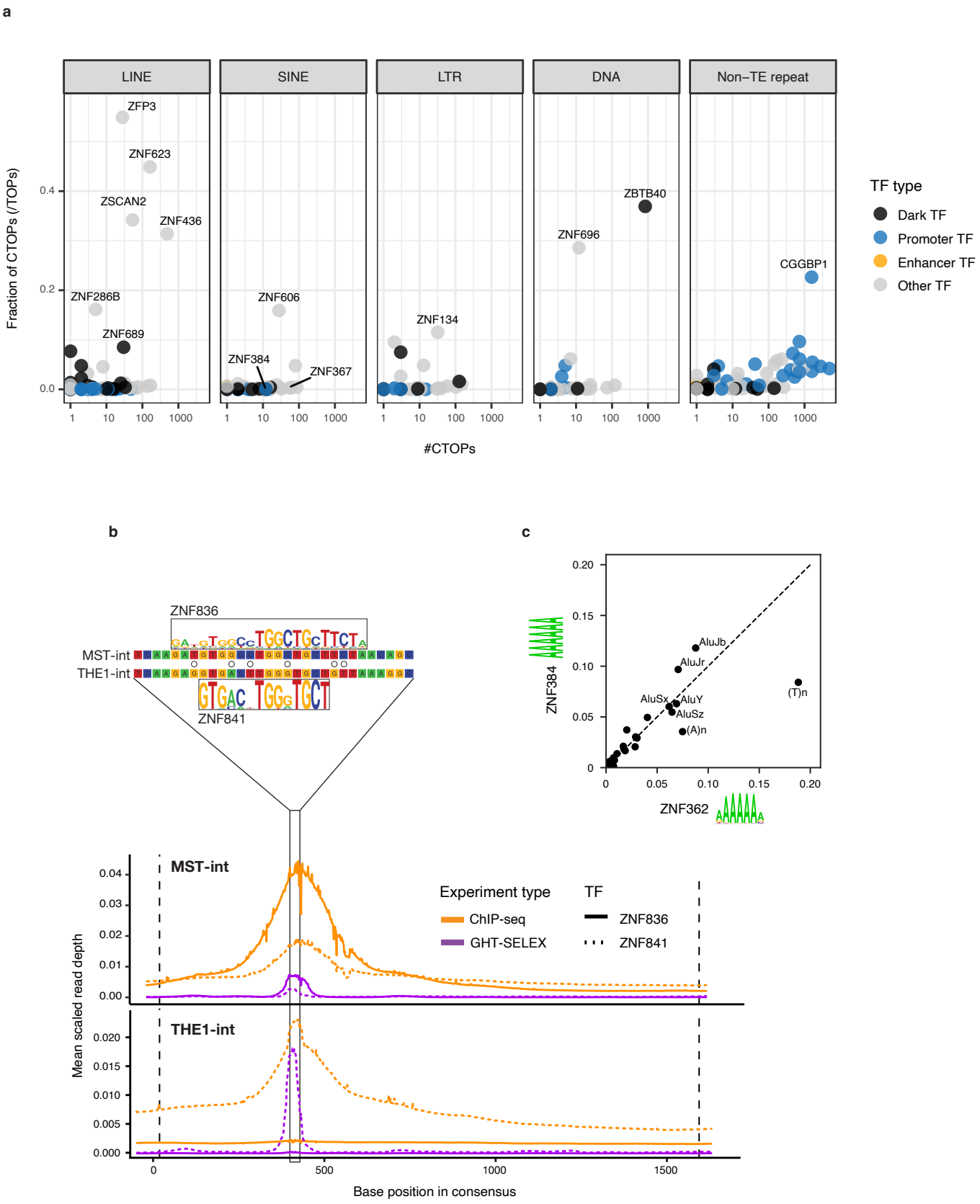
Supplementary Fig. 7. Comparative quality control for ChIP-seq peaks. (**a**) Fraction of Reads in Peaks (FRiP) for Promoter TFs and Dark TFs, at the universal threshold (MACS2 *P*-value < 1e−10) and a less stringent threshold (1e−5). (**b**) Middle heatmap: distribution of raw reads (i.e., before peak calling) in representative chromatin states (from ChromHMM) for 217 TFs with successful data. Right heatmap: $\log_{10}$ fold enrichment of the read counts (in the middle heatmap) relative to the size of each chromatin state. Labels and categories of TFs are indicated on the left. (**c**) Genome browser examples for four Dark TF, showing read pileups (red) for two replicates, at locations where both replicates show a high read coverage, but only one replicate passed the peak calling threshold (black square).


10,000
1,000
100
10
# of CTOPs
Transcription factor
Total # of TOPs
100,000
10,000
1,000
100
10
1
CTOPs fraction /TOPs
1.00
0.75
0.50
0.25
0.00
Categoty
Codebook TF
Control
Family
C2H2 ZF
KRAB C2H2
Other
TF age
600
400
200
0
TF type
Dark TF
Promoter TF
Enhancer TF
Other

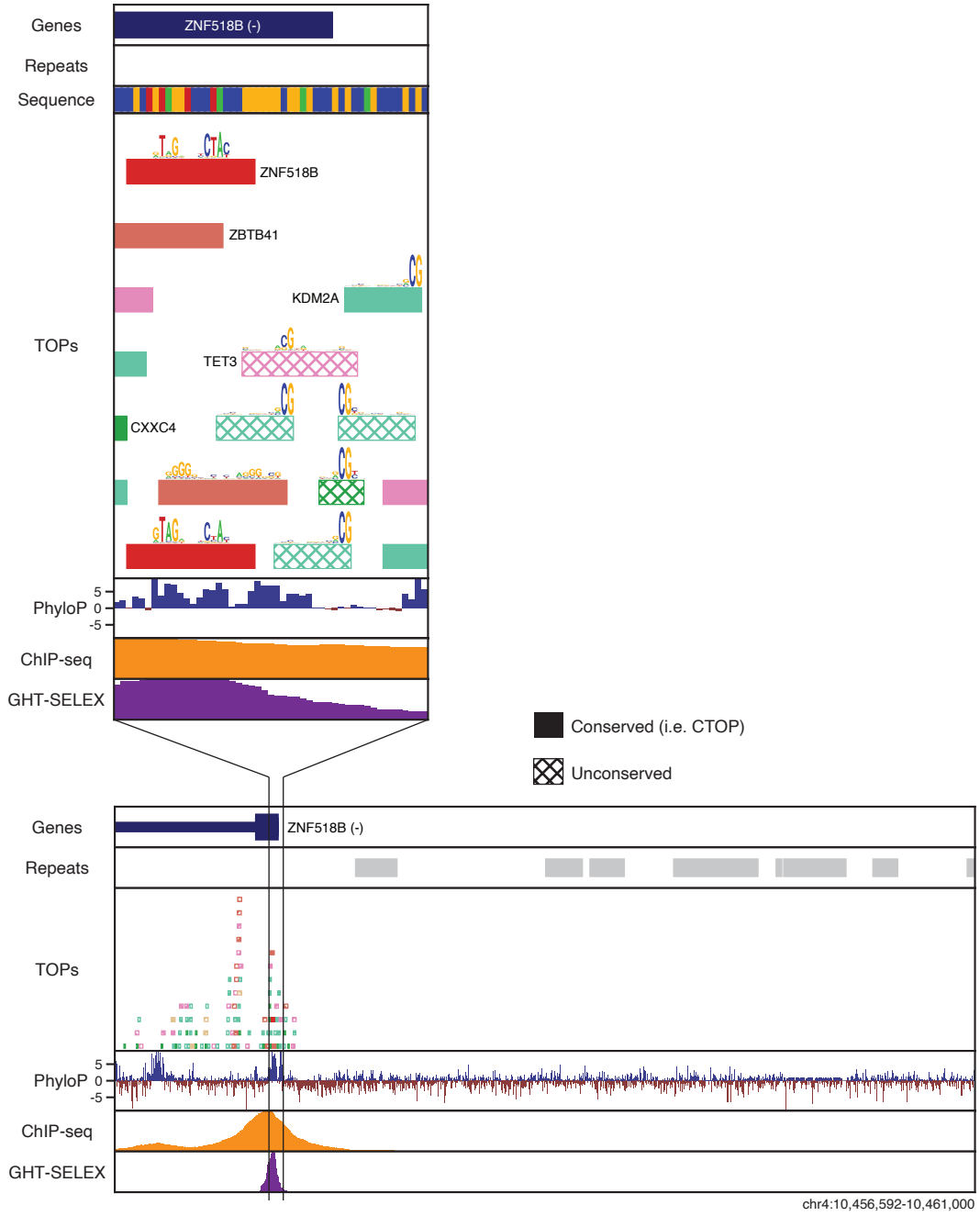
Supplementary Fig. 8. Distribution of CTOP counts for 137 Codebook TFs. Bar graph displays the number of individual CTOP sites obtained for each TF. Other properties for each TF (TOP counts, CTOPs fraction/TOPs, category, DBD family, TF type, and TF age) are annotated below the bar graph. Heatmap scales are shown on the top right.


a
Fraction of TOP sites
0
0.5
1
TF type
Dark TFs
Enhancer TFs
Promoter TFs
Other TFs
Median TOP site age (MYA)
0
50
100
TF age (MYA)
0
200
400
600
800
Log10 total TOP sites
0
2
4
6
TFs
TF type
Median TOP site age
TF age
Log10 total TOP sites
Homo (0 MYA)
Chimp (6 MYA)
Gorilla (9 MYA)
Orangutan (15 MYA)
Gibbon (19 MYA)
Macaque (29 MYA)
Capuchin (43 MYA)
Lemur (74 MYA)
Flying lemur (79 MYA)
Treeshrew (85 MYA)
Mouse (87 MYA)
Hedgehog (94 MYA)
Elephant (99 MYA)
b
Codebook TFs / Control TFs
Promoter TFs / Dark TFs
DRAP1, MAX
SOX2, SALL3, ZNF362
ZXDB
ZNF705E
TF age (MYA)
0
200
400
600
TFs
c
Codebook TFs / Control TFs
Promoter TFs / Dark TFs
≥ 100
50
0
Median TOP site age (MYA)
ZBED2,ZBED4,ZNF395,
ZKSCAN4,SLC2A4RG,
ZNF131,ZNF407,
ZNF800,JRK,
CAMTA1,
CAMTA2,
CGGBP1,
FBXL19,
KDM2A,
CXXC4
SOX2, ZBTB40,
SP100, SP140L,
ZNF226, ZNF772
ZNF233
ZNF500
FLYWCH1
TFs

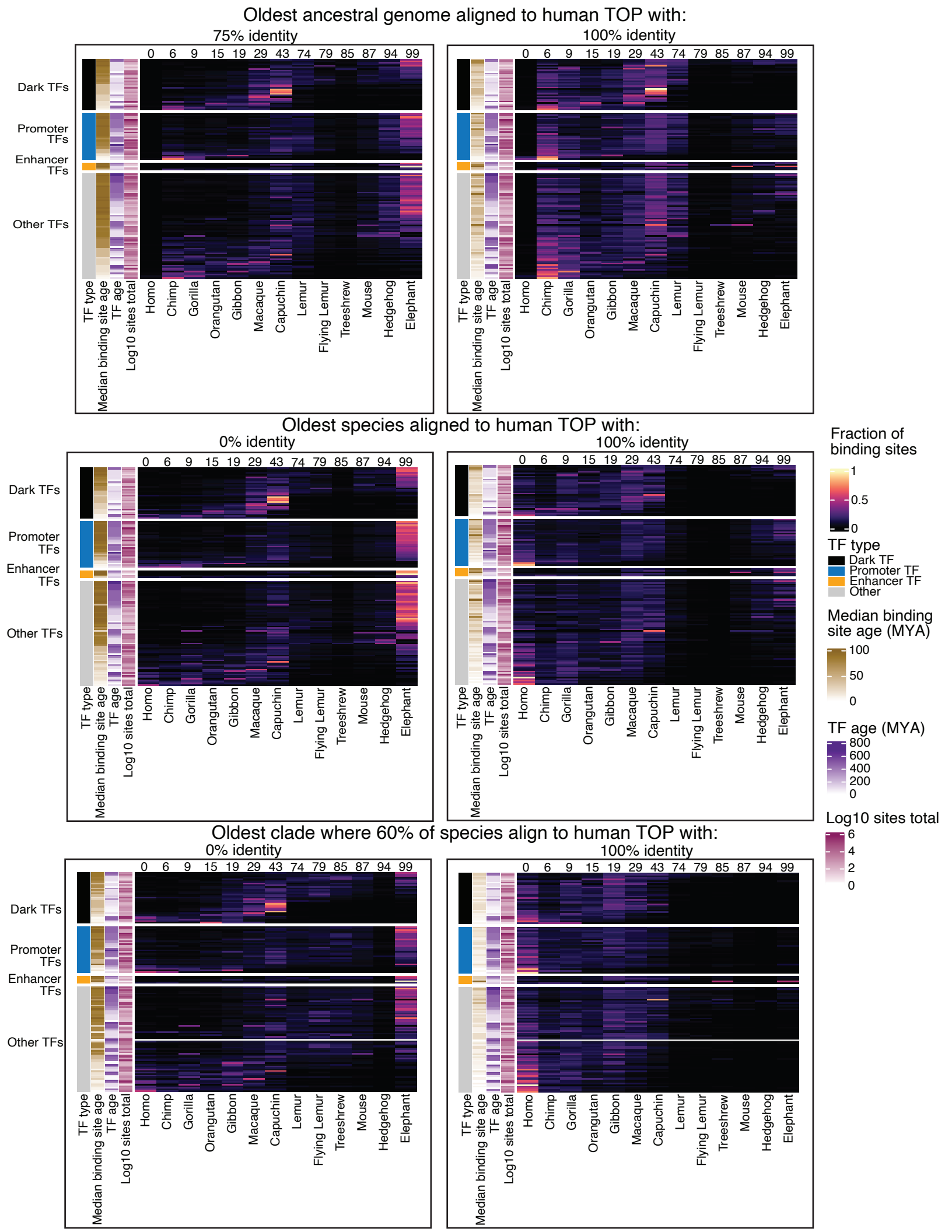
Supplementary Fig. 9. Age distribution relationship between TFs and their binding sites. (**a**) Heatmap showing the fraction of TOP sites for each TF, with an exact match within ancestral genomes available in Zoonomia (13 clades in human lineage up to 99 million years ago; MYA) ordered by age. In the left side for each TOP, the related TF types, median age of TOP sites, TF ages (MYA), and log 10 of total TOP sites are indicated. (**b**) The age comparison of Dark TFs and Promoter TFs. (**c**) Comparing the median age of the TOP sites between Promoter and Dark TFs in both control and Codebook categories.


a
LINE
SINE
LTR
DNA
Non−TE repeat
Fraction of CTOPs (/TOPs)
#CTOPs
ZFP3
ZNF623
ZSCAN2
ZNF436
ZNF286B
ZNF689
ZNF606
ZNF384
ZNF367
ZNF134
ZBTB40
ZNF696
CGGBP1
TF type
Dark TF
Promoter TF
Enhancer TF
Other TF
b
ZNF836
MST-int
THE1-int
ZNF841
MST-int
THE1-int
Mean scaled read depth
Experiment type
ChIP-seq
GHT-SELEX
TF
ZNF836
ZNF841
Base position in consensus
c
ZNF384
ZNF362
AluJb
AluJr
AluSx
AluY
AluSz
(A)n
(T)n

Supplementary Fig. 10. Codebook TFs exhibit diverse co-evolutionary interactions with repeats. (**a**) 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 fraction 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 colored according to its classification as a Dark TF, Promoter TF, Enhancer TF, 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. (**b**) Top plot shows binding motifs of two paralogous Dark TFs, ZNF836 and ZNF841, aligned to the consensus sequences of their specific TE targets, MSTA-int and THE1-int, respectively. They bind to a homologous region in the two different but related LTR families. Circles indicate nucleotide differences between the two TEs. The bottom plot shows the average ChIP-seq and GHT-SELEX signal across all the instances of MST-int and THE1-int. (**c**) compares the fractions of TOP sites for two Dark TFs, ZNF362 and ZNF384, that overlap with various repeat elements. Both bind poly-A sequences; motifs are shown adjacent to the axis labels. Alu elements are the dominant target TEs for both TFs.


Genes
ZNF518B (-)
Repeats
Sequence
ZNF518B
ZBTB41
KDM2A
TOPs
TET3
CXXC4
PhyloP
5
0
-5
ChIP-seq
GHT-SELEX
Conserved (i.e. CTOP)
Unconserved
Genes
ZNF518B (-)
Repeats
TOPs
PhyloP
5
0
-5
ChIP-seq
GHT-SELEX
chr4:10,456,592-10,461,000

Supplementary Fig. 11. Age estimation of TOPs. Three different tests and two different thresholds were applied. Heatmaps show the proportion of each TF's TOPs (rows) inferred to be a certain age, as in **Fig. 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). (**Fig. 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).


Oldest ancestral genome aligned to human TOP with:
75% identity
100% identity
Oldest species aligned to human TOP with:
0% identity
100% identity
Oldest clade where 60% of species align to human TOP with:
0% identity
100% identity
0 6 9 15 19 29 43 74 79 85 87 94 99
Dark TFs
Promoter TFs
Enhancer TFs
Other TFs
TF type
Median binding site age
TF age
Log10 sites total
Homo
Chimp
Gorilla
Orangutan
Gibbon
Macaque
Capuchin
Lemur
Flying Lemur
Treeshrew
Mouse
Hedgehog
Elephant
Fraction of binding sites
1
0.5
0
TF type
Dark TF
Promoter TF
Enhancer TF
Other
Median binding site age (MYA)
100
50
0
TF age (MYA)
800
600
400
200
0
Log10 sites total
6
4
2
0

Supplementary Fig. 12. Dark TF, ZNF518B, is a potential self-regulator. Two palindromic conserved binding sites of ZNF518B (red) are in the promoter of ZNF518B itself. Binding sites of four other TFs (i.e. TOPs and CTOPs for ZBTB41, KDM2A, TET3, and CXXC4) are also present in this region.