Codebook: sequence specificity and genomic binding of poorly-characterized human transcription factors

Arttu Jolma[1]*, Kaitlin U. Laverty[1,2]*, Ali Fathi[1,3]*, Ally W.H. Yang[1]*, Isaac Yellan[1,3]*, Ilya E. Vorontsov[4]*, Sachi Inukai[5,6], Judith F. Kribelbauer-Swietek[5,6], Antoni J. Gralak[5,6], Rozita Razavi[1], Mihai Albu[1], Alexander Brechalov[1], Zain M. Patel[1,3], Vladimir Nozdrin[7], Georgy Meshcheryakov[8], Andrey Buyan[8], Ivan Kozin[4,7,8], Sergey Abramov[4,9], Alexandr Boytsov[4,9], The Codebook Consortium, Matthew T. Weirauch[10], Oriol Fornes[11], Vsevolod J. Makeev[4,#], Jan Grau[12], Ivo Grosse[12], Philipp Bucher[6], Bart Deplancke[5,6]**, Ivan V. Kulakovskiy[4,8, 13]**, and Timothy R. Hughes[1,3]**

[1]Donnelly Centre, University of Toronto, Toronto, ON M5S 3E1, Canada
[2]Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[3]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada
[4]Vavilov Institute of General Genetics, Russian Academy of Sciences, 119991, Moscow, Russia
[5]Laboratory of Systems Biology and Genetics, Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne, 1015, Lausanne, Switzerland
[6]Swiss Institute of Bioinformatics, 1015, Lausanne, Switzerland
[7]Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, 119991, Moscow, Russia
[8]Institute of Protein Research, Russian Academy of Sciences, 142290, Pushchino, Russia
[9]Altius Institute for Biomedical Sciences, Seattle, WA 98121, USA
[10]Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA; Divisions of Allergy and Immunology, Biomedical Informatics, Human Genetics, and Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA; Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA
[11]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
[12]Institute of Computer Science, Martin Luther University Halle-Wittenberg, 06099, Halle, Germany
[13]Institute of Biochemistry and Genetics, Ufa Federal Research Centre of Russian Academy of Sciences, 450054, Ufa, Russia

[#]Present address: Cancer Research UK National Biomarker Centre, University of Manchester, Manchester, M20 4BX, UK

*These authors contributed equally
**To whom correspondence should be addressed: bart.deplancke@epfl.ch, ivan.kulakovskiy@gmail.com, t.hughes@utoronto.ca

The Codebook consortium

Philipp Bucher[6], Bart Deplancke[5,6], Oriol Fornes[11], Jan Grau[12], Ivo Grosse[12], Timothy R. Hughes[1,3], Arttu Jolma[1], Fedor A. Kolpakov[14,15], Ivan V. Kulakovskiy[4,8,12], Vsevolod J. Makeev[4], Mihai Albu[1], Marjan Barazandeh[1], Alexander Brechalov[1], Zhenfeng Deng[1], Ali Fathi[1,3], 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[6], Giovanna Ambrosini[16], Antoni J. Gralak[9,10], Sachi Inukai[5,6,17], Judith F. Kribelbauer-Swietek[5,6], Marie-Luise Plescher[12], Semyon Kolmykov[14], Ivan Yevshin[18], Nikita Gryzunov[8], Ivan Kozin[4,7,8], Mikhail Nikonov[7], Vladimir Nozdrin[7], Arsenii Zinkevich[7], Katerina Faltejskova[19], Pavel Kravchenko[20], Sergey Abramov[4,9], Alexandr Boytsov[4,9], Vasilii Kamenets[4], Dmitry Penzar[4,7], Anton Vlasov[21], Ilya E. Vorontsov[4], Aldo Hernandez-Corchado[22], Hamed S. Najafabadi[22], Quaid Morris[2], Xiaoting Chen[10], Matthew T. Weirauch[10]

[14] Department of Computational Biology, Sirius University of Science and Technology, 354340, Sirius, Krasnodar region, Russia
[15] Bioinformatics Laboratory, Federal Research Center for Information and Computational Technologies, 630090, Novosibirsk, Russia
[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] Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, 160 00 Praha 6, Czech Republic; Computer Science Institute, Faculty of Mathematics and Physics, Charles University, 118 00 Praha 1, Czech Republic
[20] Max Planck Institute of Biochemistry, 82152, Planegg, Germany
[21] TUD Dresden University of Technology, Center for Molecular and Cellular Bioengineering (CMCB), Biotechnologisches Zentrum (BIOTEC), Tatzberg 47/49, 01307 Dresden, Germany
[22] Department of Human Genetics, McGill University, Montréal, QC H3A 0C7, Canada; Victor P. Dahdaleh Institute of Genomic Medicine, Montréal, QC H3A 0G1, Canada

SUMMARY

Gene expression is regulated by transcription factors (TFs), which recognize specific DNA sequence motifs. Several hundred putative human TFs, identified mainly by an apparent DNA-binding domain, lack known binding motifs[1], and even for well-characterized TFs, it remains controversial to what degree motifs accurately reflect binding sites in living cells[2,3]. Here, we describe a systematic effort ("Codebook") to determine the sequence specificity of 332 putative and poorly characterized human TFs. Over 4,000 independent experiments, encompassing multiple *in vitro* and *in vivo* assays, produced motifs for just over half (177, or 53%), of which most are unique to a single protein, thereby extending the vocabulary of sequence recognition encoded by human TFs by ~100 distinct motifs. Moreover, binding motifs identified *in vitro* are strongly enriched within cellular binding sites. Collectively, the data reveal tens of thousands of previously unknown, conserved, and direct TF binding sites across the human genome. These sites are concentrated in promoter regions, and are predictive of gene expression, illustrating that this new data atlas provides an important step forward in decoding the human genome.

KEYWORDS: Transcription factor, TF, ChIP-seq, HT-SELEX, GHT-SELEX, SELEX, SMiLE-seq, Motif, DNA-binding specificity, PWM, PBM, Codebook

Introduction and motivations

A 2018 survey of putative human transcription factors (TFs) concluded that over a quarter of the estimated ~1,600 possible TFs lacked established binding motifs[1]. This is a striking deficit given that most of the conserved DNA in the human genome is noncoding[4], and that a primary hypothesis for the function of noncoding DNA is gene regulation[5]. Moreover, most of the uncharacterized and putative TFs are not close paralogs of any established TF. Thus, despite possessing either literature evidence of DNA-binding or protein domains that would typically bind DNA, their binding motifs cannot be readily inferred, although it also cannot be excluded that these putative TFs lack DNA sequence specificity entirely.

TF DNA-binding motifs are commonly modelled as Position Weight Matrices (PWMs), which describe the relative preference of a TF for each nucleotide base pair in the binding site[6,7], and can be visualized as a sequence logo[8]. Different methods for measuring TF binding, and for deriving PWMs from the resultant data, have different inherent limitations and biases[6]. There is also long-standing controversy regarding the contribution of a TF's inherent sequence specificity to its cellular binding, relative to the influence of chromatin and cofactors[2]. These uncertainties represent fundamental hurdles for the analysis of gene regulation, as well as a myriad of related tasks in genome analysis, including the interpretation of conserved genomic elements and sequence variants.

To address these issues, we analyzed a large majority of the poorly characterized human TFs[1] (defined here as having no confidently known binding motif), as well as several dozen previously studied control TFs[9,10], using a panel of assays that provide different perspectives on DNA sequence specificity. We refer to this international collaborative project as the "Codebook/GRECO-BIT Collaboration": the reagent set and laboratory experiments were initiated as the "Codebook Project", alluding to the fact that TFs decode individual "words" in the genome, and the Gene REgulation COnsortium Benchmarking IniTiative, GRECO-BIT rooted in GRECO[11], was then engaged for much of the data analysis and benchmarking.

In this paper, we present an overview of the Codebook data, major outcomes and findings of the study, and examples of prevalent phenomena and applications. These include:

- Just over half of the 332 putative (i.e., "Codebook") TFs (177) display DNA sequence specificity, most with the same motifs observed *in vitro* and in living cells.
- The resulting 177 motifs are largely different from one another, and from the motifs of previously-studied human TFs.
- Tens of thousands of genomic binding sites for these TFs display evidence of purifying selection at the motif matches, indicating that they are functional.
- These conserved binding sites are abundant in the promoters of coding genes, and promoter binding sites of the Codebook TFs are predictive of gene expression across tissues and cell types.

- Many TFs bind to genomic dark matter, such as transposons, or are derived from transposons, suggesting they may have had roles in adaptive evolution.

We also introduce web resources that can be used to access the primary and processed data, including the PWMs. Accompanying manuscripts and SI provide greater technical detail and depth, including biological findings, new assays, and intriguing TF families, as well as methods for identifying and benchmarking PWMs[12-15].

Overview of the Codebook project

Fig. 1 provides a schematic of the Codebook project. We analyzed 393 proteins (332 putative "Codebook" TFs, and 61 control TFs) (**Supplementary Table 1**) using up to five different assays encompassing both *in vitro* and living cellular contexts: HT-SELEX, SMiLE-seq, Protein Binding Microarrays (PBMs), ChIP-seq (with inducible tagged transgenes in HEK293 cells), and GHT-SELEX, a novel variant of HT-SELEX in which the selections are performed using fragmented genomic DNA[12], instead of random sequences. The GHT-SELEX, SMiLE-seq, and ChIP-seq data yielded significant and extensive biological insight on their own that is detailed in accompanying manuscripts[12,13,15], with highlights and joint conclusions summarized here. Different assays accommodate different protein tags and expression systems, and many TFs were analyzed using multiple constructs (e.g., full-length and DNA-binding domain(DBD)(s)) and multiple expression systems, leading to variable numbers of constructs analyzed and experiments per protein. The study employed 622 and 93 protein-coding inserts cloned into 1267 and 118 protein expression constructs for the Codebook TFs and controls, respectively (**Supplementary Table 2** and **Supplementary Table 3**). In total, we performed 4,804 experiments, 4,377 for Codebook proteins and 427 for control TFs (**Supplementary Table 4**).

Because the Codebook proteins had no known motif, the process of determining which experiments were successful was, by necessity, coupled to the process of deriving motifs from each of the data sets, as described in detail in a separate manuscript[14]. Briefly, if we obtained similar motifs for the same protein from two or more different experimental platforms, and these motifs were predictive of data across platforms, then we considered the experiments producing these motifs to be successful. In total, 1,073 experiments for Codebook proteins and 280 for controls were deemed successful (i.e., produced a motif that was reproducible across experiments) (**Supplementary Table 4**). In total, we obtained motifs for 177/332 Codebook proteins and 58/61 control TFs using the same process, suggesting a low overall false-negative rate. The number of Codebook proteins successful in a particular assay was 130 for ChIP-seq (41% of those tested), 144 for HT-SELEX (44% of those tested), 139 for GHT-SELEX (42% of those tested), 64 for SMiLE-seq (23% of those tested), and 39 for PBMs (27% of those tested) (**Extended Data Fig. 1a**,**b**, **Supplementary Table 1**). For each of the proteins, we selected a single representative PWM that scored highly across data types, for use in subsequent analyses (see **Methods**, **Supplementary Table 5**).

The 177 successful Codebook TFs are dominated by the large and diverse C2H2 zinc finger (C2H2-zf) class, for which 67% (121/180) produced motifs. C2H2-zf proteins can

180 bind RNA, protein, or other ligands[16-18]; this outcome indicates that most of the
181 Codebook C2H2-zf proteins are DNA-binding proteins, although it does not exclude
182 other functions. Among the Codebook TFs with other DBD types, roughly half (49%,
183 50/103) also produced motifs. The Codebook TFs also encompassed 49 proteins that
184 lack a well-established DBD, which were included due to literature evidence for DNA
185 sequence specificity. Among these, only six (12%) yielded motifs. Upon closer
186 inspection, all six do, in fact, contain plausible DNA-binding regions or domains (see
187 **Supplementary Discussion**). For example, DACH1 and DACH2, which recognize the
188 same motif, are homologs of Drosophila *Dachshund*, sharing the SKI/SNO/DAC
189 domain. There is prior evidence of DNA-binding by SKI/SNO/DAC domains[19,20],
190 although it is also known to have other functions[21]. Subsequent to the main Codebook
191 study, we tested all five additional human SKI/SNO/DAC domain-containing proteins
192 with HT-SELEX, and identified a motif for one, SKIDA1, providing further support for
193 SKI/SNO/DAC acting as a DNA-binding domain (**Extended Data Fig. 2a-c** and
194 **Supplementary Discussion**).

195 Nearly half of the Codebook proteins did not yield motifs (PWMs) that passed our
196 success criteria. False negatives could arise from the lack of an obligate binding
197 partner, a requirement for epigenetically modified DNA, lack of requisite post-
198 translational modification in our experiments, or other limitations of the methods. We
199 manually re-assessed each of the failed proteins, and deemed that the majority (83/155)
200 likely represent true negatives. C2H2-zf proteins with a single C2H2-zf domain, or
201 composed only of unusual C2H2-zf domains, typically failed in our assays, as did those
202 that lack the canonical spacing of C2H2-zf domains that supports binding of tandem
203 C2H2-zf arrays to longer sites[22] (see **Supplementary Discussion**, **Extended Data Fig.**
204 **3**, **Supplementary Table 6**, and **Supplementary Document 1** for discussion of the
205 unsuccessful proteins), suggesting that these C2H2-zf domains may have roles other
206 than DNA-binding. Some subtypes of other bona fide DBD classes also lack sequence
207 specificity (e.g., HMG); we found almost two dozen such cases. In addition, the 155
208 putative TFs that did not yield a motif in any assay are enriched for those with no known
209 DBD (of which 43/49 failed); among these, 33 have no evidence for DNA binding
210 beyond a single study, most of which were published over two decades ago.

211 **Motifs identified *in vitro* are highly enriched at cellular binding sites and predict**
212 **differential binding to SNVs**

213 The Codebook project generated technically successful ChIP-seq data for 130 of the
214 177 successful Codebook TFs[13], allowing us to gauge utilization of the motif in a cellular
215 context. There is clear enrichment of motif matches within ChIP-seq peaks, indicating
216 that the intrinsic sequence specificities of the TFs are used in the selection of cellular
217 binding sites (**Fig. 2**). Some TFs display motif enrichment in a wide region around the
218 peak summit due to multiple repeated motif matches; this phenomenon is particularly
219 common at CpG island promoters (discussed below). An accompanying manuscript[12]
220 demonstrates that genomic binding *in vitro* and *in vivo* are more similar than generally
221 believed. This conclusion could only be reached by a combination of GHT-SELEX and
222 ChIP-seq data coupled with optimization of PWM derivation and scanning methods.

223 The Codebook project was conducted over a period of nearly six years, and during this
224 time, several large-scale studies aimed at systematic ChIP-seq analysis of human TFs
225 (e.g., ENCODE) were published, mainly employing cell lines other than HEK293[23-25].
226 Peaks from these external datasets overlap only partially with the Codebook ChIP-seq
227 peaks for the same TF, which is expected because chromatin and cofactors can impact
228 site accessibility and selection (**Extended Data Fig. 4a-d**, **Supplementary Table 7**).
229 Nonetheless, Codebook PWMs were generally able to discriminate between binding
230 sites in these other ChIP-seq experiments and random sequences (median AUROC of
231 0.71 on HEK293, and slightly lower on other cell types), illustrating that the motifs are
232 predictive across studies and cell types (**Extended Data Fig. 4e**).

233 Sequence variants can disrupt TF binding sites, impacting gene regulation[26]. We
234 examined allelic read imbalance[27] in ChIP-seq and GHT-SELEX peak sets (**Extended**
235 **Data Fig. 5**, **Supplementary Table 8**), detecting a total of 12,060 apparent allele-
236 specific binding (ASB) events at 10,009 common single-nucleotide polymorphisms
237 (SNPs) for 190 individual TFs at 5% FDR. Importantly, for highly motif-disrupting
238 variants (with at least two-fold difference in P-values between PWM predictions for the
239 reference and the alternative alleles), the allelic preference of ASBs was concordant
240 with PWM predictions for 1,682 out of 2,260 (74%) of ASBs. Thus, while the specific
241 ASBs detected experimentally are limited to SNPs present in HEK293 cells, the
242 Codebook motifs can be applied to predict differential TF binding to any other sequence
243 variants. Considering aggregated ASBs across all tested TFs and experiments (see
244 Methods), the corresponding SNPs were enriched for previously characterized ASBs[28]
245 (2.4 odds-ratio, P-value 3.90e-162), disease-associated SNPs[29] (1.35 odds-ratio, P-
246 value 3.79e-18,), and GTEx eQTLs[30] (1.34 odds-ratio, P-value 1.77e-41) (two-sided
247 Fisher's exact test) (**Supplementary Table 9**).

248 **Most of the Codebook TF motifs are unique, and many require degeneracy for**
249 **highest motif performance**

250 The Codebook protein sequences are typically not closely related to each other or to
251 those of known TFs, suggesting that their DNA sequence preferences may also differ[1].
252 **Fig. 3** shows an overview of similarity[31] among the representative Codebook PWMs,
253 illustrating that most of the Codebook TF motifs are indeed unique. The majority are
254 also dissimilar to any other known PWM[14] (**Extended Data Fig. 6**). This result is partly
255 explained by the large number of C2H2-zf proteins, which differ in their DNA-contacting
256 "specificity residues"[32]. Small clusters along the diagonal in **Fig. 3** mostly correspond to
257 the handful of paralogs analyzed (e.g., TIGD4 and 5, SP140 and SP140L, DACH1 and
258 2, CAMTA1 and 2, and ZXDA, B, and C). In addition, the middle of **Fig. 3** contains a set
259 of eight TFs for which the motif is dominated by a single CG sequence. Five of them
260 (CXXC4, FBXL1, TET3, MBD1, and KDM2A) contain CXXC-zf domains, which are
261 expected to bind clusters of unmethylated CpGs[33]. Similarly, in the lower right is a group
262 of five AT-hook proteins that prefer A/T containing sequences. Overall, however, the
263 177 Codebook TFs encompass 129 distinct motifs (**Fig. 3**, **Methods**). When we
264 performed motif clustering jointly with previously-catalogued, representative motifs for
265 1,211 human TFs, we obtained only 4.75-fold more distinct motif clusters (613), of

266 which 92 (15%) contained only Codebook proteins (**Extended Data Fig. 6**,
267 **Supplementary Table 10**). These numbers are robust to methodology: utilization of a
268 different public motif database, a different similarity metric and a different clustering
269 procedure led to a similar outcome, with a total of 653 motif clusters, 135 (21%) of
270 which contained only Codebook motifs (**Supplementary Table 11, Methods**). Thus, the
271 Codebook data expands the lexicon of human TFs by at least 15%, providing ~100
272 previously unknown motifs.

273 For dozens of TFs, the representative PWM, selected for highest accuracy across data
274 sets, had a degenerate quality, i.e. there were few or no positions at which a specific
275 base is absolutely required (**Extended Data Fig. 7a**). Artificially increasing the
276 information content (IC) (i.e., "unflattening" the sequence logo, and thus increasing the
277 apparent "specificity" of the low-IC representative PWMs) almost universally reduced
278 motif performance (**Extended Data Fig. 7b,c**), indicating that the degeneracy is
279 required for accuracy across data types. We also found that, overall, IC is not predictive
280 of motif performance[14]. It is counterintuitive that degeneracy (i.e., lower apparent
281 specificity) would lead to better predictive capacity, but similar findings by others
282 support the validity of the result[34-36]. The degeneracy may be a consequence of forcing
283 a single PWM to represent the specificity of TFs that, in reality, recognize multiple
284 related motifs[14]. It is also conceivable that TFs having the ability to tolerate mutations in
285 binding sites is beneficial for evolution[37].

286 **Codebook TF binding sites suggest functions for tens of thousands of conserved**
287 **elements, which often co-occur in promoters**

288 Together, the Codebook assays and PWMs pinpoint genomic loci that are bound
289 directly by each TF *in vivo* (i.e., in ChIP-seq), by identifying sites that are also bound *in*
290 *vitro* (i.e., GHT-SELEX), and that contain a PWM match, thus allowing base-level
291 resolution. We refer to these as "triple overlap" (TOP) sites, which are taken as the
292 overlap of the three sets (ChIP-seq, GHT-SELEX, and PWM matches) after applying
293 optimized score thresholds for each (see **Methods** for details). To further gauge
294 functionality of the TOP sites, we examined whether the pattern of per-nucleotide
295 conservation[38] at each site is consistent with the TF's sequence preference driving local
296 sequence constraint. To do this, we tested whether the motif matches at TOPs are more
297 highly conserved than neighboring sequence, and whether the degree of conservation
298 at each position within the motif match correlates with the contribution of that base to
299 predicted relative affinity (see **Methods** for details). **Fig. 4a** shows several examples
300 illustrating apparent conservation of PWM matches, for both control and Codebook TFs.
301 Overall, 85 of the 101 Codebook TFs (as well as 31/36 controls) with successful GHT-
302 SELEX and ChIP-seq data displayed conservation of at least one TOP site (FDR < 0.1).
303 In total we identified 113,577 such conserved TOP sites ("CTOP" sites) (82,760 for
304 Codebook TFs and 30,817 for controls), encompassing 1,394,530 bases within the
305 human genome. These results provide strong support for the functional importance of
306 Codebook TF binding sites in the genome.

307 Many of the CTOP sites were either overlapping or adjacent to other CTOP sites for the
308 same or different TFs, suggesting that these sites may be part of the same regulatory
309 element. We therefore grouped the CTOPs that were within 100bp from each other,
310 yielding 41,529 CTOP clusters (of which 24,733 are singletons and 16,796 contain
311 multiple CTOPs). The majority (35,583, or 85.7%) overlapped with the UCSC
312 PhastCons track. Thus, binding of these TFs presents a likely biochemical function for
313 these previously-identified conserved elements. The remaining 5,946 CTOPs appear to
314 represent new functional annotations that were not readily detected by comparative
315 genomics alone. It is known that TF binding motifs can provide additional functional
316 information (e.g., degenerate positions within motif matches) that can explain complex
317 conservation patterns[39].

318 Codebook TFs with the largest number of CTOP sites were typically located within
319 promoters (36.9% of all CTOP clusters) and/or CpG islands (47.3% of CTOP clusters)
320 (**Fig. 4b**). The majority of CpG islands that are within protein-coding genes (60.4%,
321 8,111/13,427) contained CTOP sites, with an average of 4.55 CTOP sites per CpG
322 island. Moreover, 56/101 of Codebook TFs analyzed had at least one CTOP site within
323 a CpG island. An example CTOP that overlaps a CpG island is shown in **Fig. 4c**. The
324 extent of specific, conserved, and intrinsic occupancy of CpG islands by many TFs of
325 diverse classes is surprising: CpG islands are known to be enriched for binding sites for
326 specific TFs (SP1, NRF1, E2F, and ETS classes), which we do observe (**Fig. 4c-e**), but
327 CpG islands typically have little long-range conservation[40]. The CXXC proteins are
328 known to specifically recognize unmethylated CG dinucleotides and to modulate
329 chromatin at promoters[41], and we do observe that the CXXC proteins KDM2A, CXXC4,
330 FBXL19, and TET3 predominantly bind CpG islands. The abundance of CG
331 dinucleotides in CpG islands has been attributed primarily to their lack of methylation in
332 the germline, however, rather than primary sequence constraint or binding to TFs[42]. The
333 Codebook data thus provide a new lens for study of CpG islands. Intriguingly, many of
334 the Codebook TFs with CTOP sites in CpG islands recognize elaborate C/G rich motifs,
335 often with degenerate positions and long gaps between invariant bases (**Fig. 4c**).

336 CTOP clusters were also found in non-CpG island protein-coding promoters (**Fig. 4b**)
337 (678/6,606 such promoters, defined as -1000 to +500 relative to TSS). These clusters
338 are not dominated by any specific TFs**,** although some TFs are more prevalent than
339 others (e.g. controls ELF3 and CTCF, and Codebook TF ZNF648) (**Fig. 4d**). **Fig. 4e**
340 shows an example of a CTOP cluster in a non-CpG promoter, occurring early in the first
341 intron of the *TSPAN31* gene, which exhibits apparent conserved spacing and
342 orientation of multiple Codebook TF binding sites. In contrast, CTOP clusters outside of
343 promoters and CpG islands often contain just one or two CTOP sites (**Fig. 4b)**. One
344 example is a very strongly conserved ZNF689 binding site found in a lncRNA intron and
345 overlapping an L1M4 transposon (**Fig. 4f**).

346 CTOP clusters also overlapped with known and putative enhancers: 1,825 overlap with
347 HEK293 enhancers (defined by ChromHMM[13]), and 4,057 are found in the extensive
348 GeneHancer annotation set[43]. These lower numbers, relative to promoters, could be
349 attributed to the relatively rapid evolution of enhancers[44], lack of complete knowledge of
350 enhancer identities, depletion of enhancer-related functions among the Codebook TFs,

351 or to other, non-enhancer functions of these conserved sites, such as control of
352 heterochromatin or other aspects of chromosome architecture.

353 **Codebook TF motifs predict gene expression levels across tissues and cell types**

354 Conservation of TF binding sites is a strong indicator that they are functional elements,
355 but it does not inherently reveal their biological purpose, i.e. in what conditions they are
356 active, and what processes they regulate. To link Codebook motifs to biological
357 functions, we employed Motif Activity Response Analysis (MARA)[45,46], a framework that
358 identifies relationships between the promoter motif occurrences and gene expression in
359 different biological samples. For MARA, we chose a representative motif from each of
360 the 653 motif clusters (**Supplementary Table 11**), scanned the FANTOM5 promoters[47]
361 with the representative motifs, and used the resulting scores to predict the promoter
362 activity (normalized CAGE tag counts) in 583 biological samples (tissues, primary cells,
363 and cell lines) using a linear model (**See Methods, Extended Data Fig. 8a,**
364 **Supplementary Table 12**). Next, for each motif cluster, we calculated the Pearson
365 correlation between the predicted motif activity and the total expression of the
366 corresponding TFs across all samples. The distribution of resulting correlation
367 coefficients for 135 motif clusters containing only Codebook motifs is similar to that of
368 motif clusters for previously known TFs (**Extended Data Fig. 8b**), supporting the notion
369 that the Codebook proteins are bona fide regulatory factors.

370 To explore the cellular functions the Codebook TFs might regulate, we examined the 25
371 Codebook-only motif clusters with the highest variance in activity across the FANTOM5
372 samples. Bi-clustering activities yielded four major TF groups (**Fig. 5**): (**1**) TFs targeting
373 genes that are preferentially expressed in immune cells, including transcriptional
374 repressors ZBTB8A[48] and ZFTA, whose targets are specifically repressed in the brain,
375 and are often activated in ependymomas caused by ZFTA fusion with RELA[49]; (**2**) TFs
376 active in both immune cells and cancer, including proteins that bind unmethylated CG-
377 containing motifs (CGGBP1, CXXC4, and SP140), the transcriptional activator
378 ZNF131[50], and MBD1; (**3**) TFs with target genes specifically expressed in nervous
379 tissues, including neuron maturation regulators CAMTA1[51] and ZNF292[52],
380 neurogenesis-associated repressors MBD3[53] and DNTTIP1[54], as well as MYRFL, the
381 paralog of myelin regulatory factor (MYRF), critical for central nervous system
382 myelination[55], exclusively active in the adult nervous tissues; (**4**) TFs mostly restricted to
383 mesodermal primary cells, including rRNA transcription regulator TTF1[56] and zinc finger
384 and BTB domain-containing proteins ZBTB41 and ZBTB5, which were found to
385 stimulate cell proliferation[57,58].

386 We obtained similar outcomes when we analyzed the global patterns of activity among
387 all 135 Codebook-only motif clusters and FANTOM5 samples using PCA: the first
388 component mainly separates nervous system from all other samples, while the second
389 one separates the developing nervous systems from adult (**Extended Data Fig. 8c**).
390 Taken together, the motif activities of Codebook TFs distinguish FANTOM5 tissues and
391 cell types, and provide many candidates for new transcriptional regulators of genes
392 involved in the inflammatory response, cell cycle control, and nervous system
393 development.

394 **Codebook TFs and the dark matter genome**

395 TF binding sites in promoters and enhancers are easily rationalized as likely being
396 involved in the regulation of adjacent genes. A principal result of the Codebook project
397 is that dozens of Codebook TFs instead bind preferentially to other regions of the
398 genome[13]. An accompanying manuscript[13] focuses on the finding that roughly half of the
399 examined TFs bind predominantly outside of promoters and enhancers in ChIP-seq,
400 with most TFs having a very distinct set of binding sites that are supported by both
401 GHT-SELEX and motif matches. Among these TFs are KRAB-containing C2H2 zinc
402 fingers (KZNFs), which are best known for binding to specific classes of endogenous
403 retroelements and nucleate regions of H3K9 trimethylation[59]. Until now, it has not been
404 proven that the precise binding of KZNFs to specific retroelement subtypes is defined
405 entirely by the sequence specificity of the KZNFs alone[60], but the Codebook GHT-
406 SELEX data abundantly demonstrate that this is indeed the case[12].

407 The Codebook data also underscore the fact that transposons contribute to the TF
408 repertoire itself[61]. Sixteen of the Codebook TFs (and two controls) that successfully
409 yielded motifs possess a DBD that has itself been derived from a DNA transposase.
410 These include CGGBP1[62], six proteins containing BED-zf domains[63], six with the related
411 CENBP or Brinker domains[64], three with transposon-derived Myb/SANT/MADF
412 domains[65], and FLYWCH1[66]. The PWMs obtained for CENPB/Brinker TFs are often
413 long and information-rich, consistent with their presumed ancestral role in binding
414 specifically to the transposons within a large genome (**Extended Data Fig. 9a,**
415 **Supplementary Discussion**). A striking example is JRK, a TF that is found broadly in
416 mammals[67] and derived from an ancient domesticated Tigger DNA transposon[68]. All
417 DNA transposons, including Tigger, have been transpositionally inactive in the human
418 lineage for over 40 million years, and are presumably "fossils"[69]. Genomic binding of
419 JRK is enriched for the same family of Tigger elements from which it appears to be
420 derived (**Supplementary Discussion**). The consensus sequence for this Tigger
421 element family has PWM-predicted binding sites for JRK in the terminal repeats
422 (**Extended Data Fig. 9b**), consistent with its presumed ancestral role in transposition.
423 To our knowledge, it is only the second human protein known to retain this quality (the
424 other is SETMAR[70]). These cases apparently represent the simultaneous introduction of
425 a multitude of *cis*-regulatory elements, and the TF that binds them, by the same
426 transposon.

427 **Completing the human TF codebook**

428 The primary purpose of Codebook was to assess 332 uncharacterized and putative
429 human TFs for sequence-specific DNA binding activity, and produce trustworthy motifs
430 for them, in order to assemble a complete human TF motif collection. This goal was
431 mostly achieved. Codebook alone succeeded for slightly more than half of the tested
432 TFs, and there is reason to believe that many of the unsuccessful putative TFs are likely
433 not DNA-binding proteins or bind DNA with little or no sequence specificity (see

434 **Supplementary Discussion** and **Supplementary Document 1**). As noted above, a
435 subset of the Codebook TFs, as well as other poorly characterized TFs, have also been
436 analyzed by others since our study began. To evaluate the current scope of known
437 human TF specificities, we surveyed other databases for PWMs for putative TFs that
438 were not included in this study or were not found among the 177 Codebook successes.
439 A conservative manual curation identifies 33 additional human TFs (i.e., beyond the 177
440 that were successful in Codebook) that have at least one plausible motif available in
441 datasets that have been released since our 2018 TF census[1] (**Supplementary Table**
442 **13**), leading to a new total of 1,421 human TFs with characterized sequence
443 specificities (**Fig. 6** and **Supplementary Table 14**). **Supplementary Data 1** contains
444 representative PWMs for all 1,421 TFs (see **Supplementary Table 10** for motif details).

445 This enumeration leaves 175 proteins, mostly with conventional DBDs, that lack known
446 sequence specificity. Not all proteins with such domains are necessarily TFs, as noted
447 above, and our re-curation of TFs that were unsuccessful in our assays suggests that
448 83 of these – nearly half – appear unlikely to be bona fide TFs (**Supplementary**
449 **Document 1**). At the same time, new DBD classes continue to be identified (e.g.
450 WH[71,72]). In addition, some TFs may bind only to methylated DNA, a possibility we
451 explored in a separate manuscript in this collection[15]. Ongoing advances in the
452 prediction of protein and protein-DNA structures[73] have the potential to identify
453 additional candidates for sequence-specific DNA binding. Thus, while completion of the
454 objective to obtain a motif for every human TF now appears much closer, the list of
455 likely human TFs continues to evolve.

456 An important technical demonstration of the Codebook project is that the simultaneous
457 application of multiple experimental strategies and multiple motif-derivation and motif-
458 scoring strategies was highly beneficial, with no method either dominating or
459 dispensable[14], consistent with previous benchmarking efforts[36,74]. As a result, many of
460 the Codebook TFs are now among the best characterized human DNA-binding proteins
461 in terms of their sequence specificity. In particular, obtaining *in vivo* and *in vitro* genomic
462 binding profiles, together with assessing binding to randomly-generated artificial
463 sequences, is advantageous in several ways: it facilitates disentanglement of direct and
464 indirect binding, the contribution of the cellular environment, and the highly non-random
465 nature of the genome sequence. Only for a small handful of the 1,000+ previously
466 characterized TFs is there such a rich combination of data types. A much better
467 perspective on human gene regulation and genome function, and evolution could
468 presumably be obtained from the generation of such data for all human TFs.

METHODS

Plasmids and inserts. Sequences and accompanying information are given in **Supplementary Tables 2**, **3**, and **4**. Briefly, we selected Codebook TFs (and their DNA-binding domains) from information accompanying Lambert 2018[1] and posted at https://humantfs.ccbr.utoronto.ca. Inserts named with an "-FL" suffix correspond to the full-length ORF of a representative isoform of the protein. Those with a "-DBD" suffix contain all of the predicted DBDs in the protein flanked by either 50 amino acids, or up to the N or C-terminus of the protein. Those with a "-DBD1", "-DBD2" or "-DBD3" suffix contain a subset of the DBDs present in the proteins; these were designed manually, mainly for large C2H2-zf arrays. Inserts were obtained as recoded synthetic ORFs (BioBasic, US) flanked by AscI and SbfI sites, and subcloned into up to three plasmids: (i) pTH13195, a tetracycline-inducible, N-terminal eGFP-tagged expression vector with FLiP-in recombinase sites[10]; (ii) pTH6838, a T7-promoter driven, N-terminal GST-tagged bacterial expression vector[75], and (iii) pTH16500 (pF3A-ResEnz-egfp), an SP6-promoter driven, N-terminal eGFP-tagged bacterial expression vector, modified from pF3A–eGFP[9] to contain the two restriction sites after the eGFP.

Protein production. Each experiment used a protein expressed from one of the following systems: (a) FLiP-in HEK293 cells (catalog number: R78007), induced with Doxycycline for 24 hours, used for inserts in pTH13195; (b) PURExpress T7 recombinant IVT system (NEB Cat.#E6800L), for inserts in pTH6838; or (c) SP6-driven wheat germ extract-based IVT (Promega Cat#L3260), for inserts in pTH16500.

DNA binding assays. We followed previously-described protocols for ChIP-seq[10], PBMs[36], and SMiLE-seq[9]. Detailed descriptions of GHT-SELEX, HT-SELEX, ChIP-seq, and SMiLE-seq data collection and initial analysis are provided in the accompanying papers[14,15,76,77]. For PBMs, we analyzed proteins on two different PBM arrays (HK and ME), with differing probe sequences[78].

Data processing and motif derivation. The accompanying paper[14] describes motif derivation and evaluation in detail. Briefly, after initial preprocessing, we obtained a set of 'true positive' (likely bound) sequences for each individual experiment. (721 / 4,873) experiments were removed at this step, due to a low number of peaks, or other technical issues, as documented in **Supplementary Table 4**. We then applied a suite of tools (listed in **Supplementary Table 15**) to a training subset of the data from each experiment, and tested the resulting motifs on a test subset of the data from the same experiment, and also on the independent data for the same TF (i.e. the test sets from all other experiments done for the same TF). We employed a binary classification regime for all experiments and all motifs, and scored the motifs by a variety of criteria such as the areas under the receiver operating characteristic (AUROC) or the precision-recall curve (AUPRC).

Systematic filtering of artefactual motifs. While curating the datasets and assembling a reliable motif collection, we accounted for enrichment of similar artifact motifs. These recurrent DNA patterns were detected due to systematic experimental noise or peculiarities of particular motif discovery tools. For example, in HT-SELEX experiments

511 the "ACGACG" motif is often enriched. The sequence matches the constant flanking
512 region of the ligand and likely abundant due to selective pressure for partially single-
513 stranded DNA. In experiments with cell lysates, motifs of abundant native proteins in
514 HEK293 cells were sometimes enriched (e.g., NFI, YY1, ETS-family TFs). To minimize
515 the influence of these artifacts, we (**1**) manually compiled a list of known artifact motifs
516 (**Supplementary Table 16**), (**2**) scanned the entire motif collection with MACRO-APE[79],
517 and removed motifs that were highly similar to those in our catalog of artifacts.
518 Additionally, we removed motifs that matched the constant, non-variable regions of the
519 DNA used in HT-SELEX, GHT-SELEX, and SMiLE-Seq. We did not filter out ETS-related
520 motifs for ETS-family positive controls, such as ELF3, FLI1, and GABPA. Subsequent to
521 the expert curation, we confirmed that enriched k-mers in HT-SELEX experiments did not
522 correspond to potential artifacts associated with individual expression systems[76].

523 **Motif discovery success rate evaluation.** To estimate the success rate of different motif
524 discovery tools across different platforms, we started with the successful experiments and
525 TFs. For each combination of a platform X (e.g. ChIP-seq) and a motif discovery tool Y
526 (e.g. Autoseed), we computed the number of experiments (**Extended Data Fig. 1c**) or
527 TFs (**Extended Data Fig. 1d**) that yielded a highly similar motif to the reference motif (i.e.
528 manually curated) in that combination. For that, for a specific experiment or TF, we took
529 the whole set of candidate motifs generated by that combination of platform X and tool Y,
530 and checked whether any of those motifs pass the similarity threshold using a one-pass
531 scan of MACRO-APE[79] ScanCollection command with the following parameters: -c 0.05
532 --rough-discretization 10 --precise 100500. Note that prior to comparison, the motifs were
533 converted to log-odds PWMs (as in [14]).

534 **Experiment evaluation by expert curation.** To gauge the success of individual
535 experiments, we employed an “expert curation” workflow with an initial voting scheme in
536 which a committee of annotators gauged whether individual experiments should be
537 “successful”, i.e. included in subsequent analyses. All experiments were examined by at
538 least three annotators. A subcommittee (AJ, IVK, and TRH) jointly resolved all cases of
539 disagreement among initial annotators (~300 experiments) and then reviewed all
540 successful experiments. Annotators had available an early version of the MEX portal
541 (https://mex.autosome.org, described in [14]) containing results of all PWMs scored
542 against all experiments, and were tasked with gauging whether the experiments yielded
543 PWMs that were similar across experiments, or scored highly across experiments.
544 Annotators also considered whether the motif was consistent with those for other
545 members of their protein family (e.g. BHLHA9 yielded an E-box-like motif, CAnCTG),
546 and/or similar between closely related paralogs (e.g. ZXDA, ZXDB, and ZXDC all
547 yielded similar motifs). We also considered whether (and how many) “peaks” were
548 obtained from ChIP-seq or GHT-SELEX, and whether these peaks were common to
549 independent experiments (e.g. both ChIP-seq and GHT-SELEX). Annotators were
550 further given a measure of similarity between Codebook PWMs and any PWMs in the
551 public domain, as well as enrichment of known or suspected common contaminant
552 motifs in any experiment.

553 **Post-evaluation peak processing.** After identification of successful experiments, we
554 re-derived peak sets for ChIP-seq and GHT-SELEX experiments in order to obtain a

555 single peak set for each TF, as described in the accompanying papers[12,13]. Briefly, for
556 ChIP-seq we repeated the peak calling using MACS2 and experiment-specific
557 background sets, using a procedure previously described[10], then merged the peak sets
558 for replicates of the same TF with BEDTools merge[80] (see accompanying manuscript[13]:
559 "ChIP peak replicate analysis and merging"). We derived GHT-SELEX peaks using a
560 novel method, MAGIX, that calculates enrichment of reads in each cycle, and treats
561 different experiments as independent statistical samples in order to obtain a single
562 enrichment coefficient per peak[12].

563 **Expert motif curation.** For this study, to identify a single representative PWM for each
564 TF, we first compiled a set of the highest-scoring candidate PWMs for each TF (as
565 summarized above and elsewhere[14]), then ran additional tests with them, utilizing the
566 reprocessed peak data, and manually evaluated the outputs. We started with the union
567 of three sets of 20 PWMs for each TF: the 20 PWMs with the highest AUROC (as
568 calculated elsewhere[14]) on (i) any successful ChIP-seq experiment for the given TF, (ii)
569 any successful GHT-SELEX experiment for the given TF, and (iii) any successful HT-
570 SELEX experiment for the given TF. These PWMs were selected regardless of the data
571 set from which they were derived. We then reassessed these PWMs against ChIP-seq
572 and GHT-SELEX data with two parallel methodologies. First, we recalculated AUROC
573 for each of the candidate top PWMs on the merged, thresholded sets of ChIP-seq
574 peaks ($P < 10^{-10}$)[13] using AffiMX[31] to score each peak. We generated negative sets
575 using BEDTools shuffle[80] with the *-noOverlapping* option to create sets of random
576 genomic regions with the same number of peaks, and the same peak width distribution
577 as the corresponding ChIP peak sets. We used the same technique to calculate
578 AUROC values for GHT-SELEX, with thresholded peak sets (using a "kneedle"[81]
579 specificity value of 30 in the sorted enrichment values[13]). In parallel, we calculated the
580 Jaccard index to measure the overlap between PWM matches (identified by MOODS[82]
581 with -p 0.001) vs. ChIP-seq peaks and GHT-SELEX peaks, as two separate
582 measures. The overlap in each case was maximized by applying different thresholds on
583 the peak sets and choosing the cutoff at which the Jaccard index was the highest[12]. We
584 then applied expert curation (by a committee consisting of AJ, TRH, KUL, AF, RR, MA,
585 and IY) to choose a single representative PWM with high performance on all compiled
586 scores that, all else equal, also reflected reasonable expectation from the DBD class
587 (including recognition-code predicted motifs, see accompanying manuscript[12]) and did
588 not have low information content.

589 **Motif similarity analysis.** We took two different approaches to determine the number
590 of distinct motifs represented by the Codebook TFs, and the number of novel motifs
591 added to the human TF repertoire. In the first approach, we identified the similarity
592 between 1,582 PWMs representing the Codebook TFs (177 PWMs, 177 TFs) and TFs
593 with previously known specificities (1,405 PWMs, 1,211 TFs). The Codebook TF PWM
594 set is the set of representative PWMs for the 177 Codebook proteins with identified
595 motifs. For TFs with previously known specificities, the set of PWMs identified as the
596 "best" in Lambert et al. 2018[1] was used, except for two TFs (MTF2 and PHF1) that did
597 not have an assigned PWM. For these two TFs, the PWMs were retrieved from CisBP[83].
598 We used the correlation between pairwise affinities to 150,000 random sequences of
599 length 100, as calculated by MoSBAT[31], as the metric for PWM similarity. We then

clustered the 177 Codebook TFs based on these PWM similarities with Pearson correlation and average linkage and identified the ideal number of clusters (129) by determining the optimal silhouette value[84], which corresponded to splitting PWMs into clusters at a distance of 0.76. We then clustered the full set of 1,582 motifs and used the same distance to split the PWMs into clusters, resulting in 613 clusters, 92 of which contained only Codebook TFs. The set of motifs used in the clustering and their cluster membership are provided in **Supplementary Table 10**.

Independently, and to obtain a non-redundant set of motifs for MARA analysis, we followed the procedure as in[85] with the addition of Codebook motifs. First, we merged HOCOMOCO[38] v12 motifs with those evaluated in Codebook, preferably selecting those built with ChIPMunk and ranking high in benchmarking[14] to maintain consistency with HOCOMOCO v12, which was fully built with ChIPMunk. Having the joint motif collection, we estimated the motif similarities with MACRO-APE[79] at the motif P-value cutoff of 0.0005 and a default matrix discretization parameter (-d) of 1 (increased to 10 for improved precision only for motif pairs with the Jaccard similarity over 0.01 at -d 1). Next, using the pairwise motif similarity matrix, we performed agglomerative clustering ('average' linkage) with sklearn. The number of clusters was taken to maximize the silhouette score. Clusters of low-quality motifs (HOCOMOCO 'D') were discarded. Finally, for each cluster, a single representative motif was taken according to the highest average similarity to all other motifs in the cluster (**Supplementary Table 11**). The non-redundant set of representative motifs is available for download from Zenodo (https://zenodo.org/records/17307845) [doi:10.5281/zenodo.17307845], and the motif clusters annotation is available in **Supplementary Table 11.**

Motif degeneracy analysis. To explore if low information content (IC) is an intrinsic feature of some binding motifs, we adjusted the IC of PWMs and tested prediction accuracy of ChIP-seq and GHT-SELEX binding sites. PWM IC was adjusted on a per-base-pair basis, by iteratively scaling probabilities for each base at each position until the PWM reached an average IC of 1 bit per base pair. The script, “logo_rescale.pl”, is available at https://gitlab.sib.swiss/EPD/pwmscan.

Comparison to external peak sets and PWMs. For comparison, we downloaded ChIP-seq peak sets from GTRD[86] and ENCODE (4.12.2023)[87], for all Codebook TFs. We then divided these data into four categories corresponding to the cell type: HEK293/HEK293T, HepG2, K562, and other cells. We preferentially selected the peak sets from GTRD, because it contained more diverse peak sets: since GTRD has processed the majority of ENCODE consortium experiments, together with many non-ENCODE experiments. When multiple experiments were available for a TF in a cell type category, we selected the experiment with higher peak counts. If multiple computational methods had been used to derive peak sets for the selected experiment, we chose the peak set using a preferential order MACS, GEM, SISSRS, PICS and PEAKZILLA. See **Supplementary Table 7** for identifiers and metadata of the reference datasets.

For PWM scoring, the external peak sets were used as downloaded, with the exception of peak sets that were generated with the GEM peak caller, which have a peak width of 1 (summit only), and were therefore expanded 250 bases in both directions. For

643 Codebook data, we used the merged and thresholded Codebook ChIP peak sets as in
644 "Expert motif curation". We generated negative peak sets for each ChIP-seq peak set
645 using BEDTools shuffle[80] with the *-noOverlapping* option to create sets of random
646 genomic regions with the same number of peaks and the same peak width distribution
647 as the corresponding ChIP peak sets. We downloaded PWMs for all Codebook TFs
648 from JASPAR[88] (2024 version), HOCOMOCO[38] (v12) and Factorbook[89] (downloaded
649 15.12.2023) (**Supplementary Table 13**). We scanned Codebook and external peak
650 sets (and corresponding negative sets) with the representative (i.e., expert-curated)
651 Codebook motifs (PWMs) using AffiMX[31], and calculated AUROC values. Additionally,
652 for the 19 Codebook TFs with a successful Codebook ChIP-seq experiment, a
653 Codebook PWM, an external ChIP-seq experiment, and an external PWM, we
654 compared the performance of PWMs across the different peak sets as follows. We first
655 selected a single external PWM for each of the 20 TFs by scanning each PWM for
656 a given TF on each external peak set for the same TF and identifying the PWM that
657 produced the highest AUROC. We then used these highest scoring PWMs to scan the
658 corresponding Codebook data and calculate AUROC values.

659 **Curation of external motifs for previously uncharacterized TFs.** For this analysis we
660 considered all proteins that were annotated as putative TFs in Lambert et al. 2018[1] but
661 did not then have a credible motif. We downloaded all motifs for these TFs from
662 JASPAR[88] (2024 version), HOCOMOCO[38] (v12) and Factorbook[89] (downloaded
663 15.12.2023), resulting in a set of 484 PWMs, that were then assessed manually by AJ
664 considering: 1) Whether similar motifs were obtained for the protein from multiple
665 independent datasets; 2) If the motif was consistent with the structural class of the TF
666 and; 3) If the motif was likely to describe inherent specificity of the protein and not, for
667 example, reflect a target site of another TF or a likely artifact. The curated motifs are
668 listed in **Supplementary Table 13** with more details in **Supplementary Table 10**. The
669 PWMs themselves are available in **Supplementary Data 1**, and examples of artifactual
670 and correct external motifs are shown in **Extended Data Fig. 10**.

671 **Motif activity response analysis (MARA).** The promoter locations and activity data
672 were taken from FANTOM5[47]. After filtering (see **Supplementary Methods**), the set
673 comprised 209,374 individual promoters and 1,020 samples (including replicates),
674 belonging to 583 unique samples (142 tissues, 187 primary cells, and 254 cell lines, see
675 **Supplementary Table 12** and Zenodo [https://zenodo.org/records/17307845]). The
676 sum-occupancy scores[90] were computed in [250 bp upstream; 10 bp downstream]
677 regions relative to the FANTOM5 TSS locations by scanning with a subset of 632 non-
678 redundant motifs representing TFs with expression detectable in the FANTOM5 data.
679 We employed MARADONER (MARA-done-right) v 0.13, a command-line Python
680 reimplementation of MARA available in PyPI (pip install maradoner). Conceptually,
681 MARADONER extends the original logic of MARA[45,46] to model the promoter activity in
682 each sample as a linear function of the motifs' scores and sample-specific motif
683 activities, where each motif represents a set of TFs with shared binding specificity.
684 Compared to the classical MARA, MARADONER explicitly models sample and motif
685 means (see **Supplementary Methods** for more details**)**.

686 **TOP (Triple Overlap) and CTOP (Conserved Triple Overlap) peak set analyses.** To
687 obtain the TOP sites, we first identified thresholds for ChIP-seq peaks, GHT-SELEX
688 peaks, and PWM-derived "peaks" (see below) that maximize the three-way Jaccard
689 metric (overlap/union) of the three sets, with the thresholds calculated for each TF
690 independently. We converted PWM matches (derived from MOODS[82] using a P-value
691 cut-off of 0.001) into peaks by merging neighboring matches with a distance less than
692 200 bp and re-scoring them using the sum-occupancy for clusters. We then identified
693 TOPs as peaks exceeding these thresholds in all three sets and overlap in all three. To
694 obtain the CTOP sites, we then extracted PhyloP scores for each base at each TOP site
695 (and 100 flanking bases) from the Zoonomia consortium[91], removed sites overlapping
696 the ENCODE Blacklist[92] or protein coding sequences (due to the skew in phyloP scores
697 caused by codons), and applied three different statistical tests for significance of phyloP
698 scores over the PWM match: two that tested correlation between the IC and the phyloP
699 value at each base position of the PWM (using either Pearson correlation or linear
700 regression), and one that tested for higher phyloP scores over the PWM match
701 (Wilcoxon test). Greater detail on these specific operations is given in the
702 accompanying manuscripts[12,13].

703 **Intersection of TOPs/CTOPs and genomic features.** We first clustered all CTOPs
704 using BEDTools merge[80], with a max distance of 100 bp, then intersected with the
705 following genomic feature sets: basic canonical protein coding promoters from
706 GENCODE v44[93], defined as 1000 bp upstream and 500 bp downstream of the
707 canonical TSS; the "Unmasked CpG Island" track, PhastCons Conserved Elements
708 from the Multiz 470 Mammalian alignment, and RepeatMasker track from UCSC[94];
709 ChromHMM HEK293 enhancers[13]. We classified promoters as CpG island or non-CpG
710 island based on the GENCODE basic TSS being within +/- 50 bp of a CpG island from
711 the unmasked track. We classified the CTOP clusters as associated with a single type
712 of genomic feature in the following order of priority: CpG island associated with a protein
713 coding promoter; other CpG islands; a non-CpG island-associated protein-coding
714 promoter; an enhancer; clusters containing a CTCF binding site but not overlapping a
715 CpG island, promoter or enhancer; overlapping a transposable element and none of the
716 previous categories; overlapping a non-TE repeat and none of the prior categories; and
717 "Other" for CTOP clusters not intersecting any examined features.

718 **Analysis of allele-specific binding.** The ASB analysis was performed as in Buyan et
719 al.[27]. Briefly, we performed SNP calling with bcftools[95], corrected mapping bias with
720 *WASP*[96], generated the maps of background allelic dosage with *BABACHI*[97] (Abstract
721 O3), called ASBs with *MIXALIME*[27], and performed PWM scoring with *PERFECTOS-*
722 *APE*[98], see **Supplementary Methods** for details. To obtain the non-redundant set of
723 ASB-SNPs across Codebook data for intersecting with existing regulatory SNP
724 databases (GTEx v.8[30], ADASTRA v.6.1[28], and EBI GWAS Catalog v1 e115_r2025-12-
725 03_full)[29], we performed a joint aggregation of the allelic imbalance P-values over all
726 processed datasets (ChIP-Seq and GHT-SELEX combined).

DATA AVAILABILITY

The sequencing raw data for the ChIP-seq, GHT-SELEX, and HT-SELEX 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 ChIP-seq and GHT-SELEX have been deposited into GEO under accessions GSE280248 (ChIP-seq) and GSE278858 (GHT-SELEX). PWMs can be browsed at https://mex.autosome.org and downloaded at from Zenodo (https://zenodo.org/records/15667805) [doi:10.5281/zenodo.15667805]. An updated list of human TFs is available at https://humantfs.ccbr.utoronto.ca. The final motifs generated in this study are available in build 3.0 of the CisBP database[99]. Information on constructs, experiments, analyses, processed data, comparison tracks, and browsable pages with information and results for each TF is available at https://codebook.ccbr.utoronto.ca. SNP calls and ASBs are available at Zenodo (https://doi.org/10.5281/zenodo.18224872) [doi:10.5281/zenodo.18224871]. The data underlying MARA are available at Zenodo https://zenodo.org/records/17307845.

CODE AVAILABILITY

The script for adjusting PWM information content, logo_rescale.pl, is available at https://gitlab.sib.swiss/EPD/pwmscan, the code for the identification of TOPs and cTOPs is available at https://github.com/imyellan/Codebook_CTOP_scripts. The code used for the SNP-centric analysis is available at GitHub: https://github.com/autosome-ru/perspectives-on-codebook-allele-specificity, the MIXALIME is available on GitHub: https://github.com/autosome-ru/mixalime. MARADONER is available on GitHub: https://github.com/autosome-ru/MARADONER.

ACKNOWLEDGEMENTS

We thank the IT Group of the Institute of Computer Science at Halle University for computational resources, Maximilian Biermann for valuable technical support, Gherman Novakovsky for providing feedback, Berat Dogan for testing earlier versions of RCADEEM, and Debashish Ray for assistance with database depositions.

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.
- A Swiss National Science Foundation grant (no. 310030_197082) to B.D.
- Assignment 125091010189-3 to I.V.K.
- MSHERF grant number no. 075-15-2025-014 (previously no. 075-15-2024-666)
- 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.D.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's Center for Imaging
- Resource allocations from Digital Research Alliance of Canada

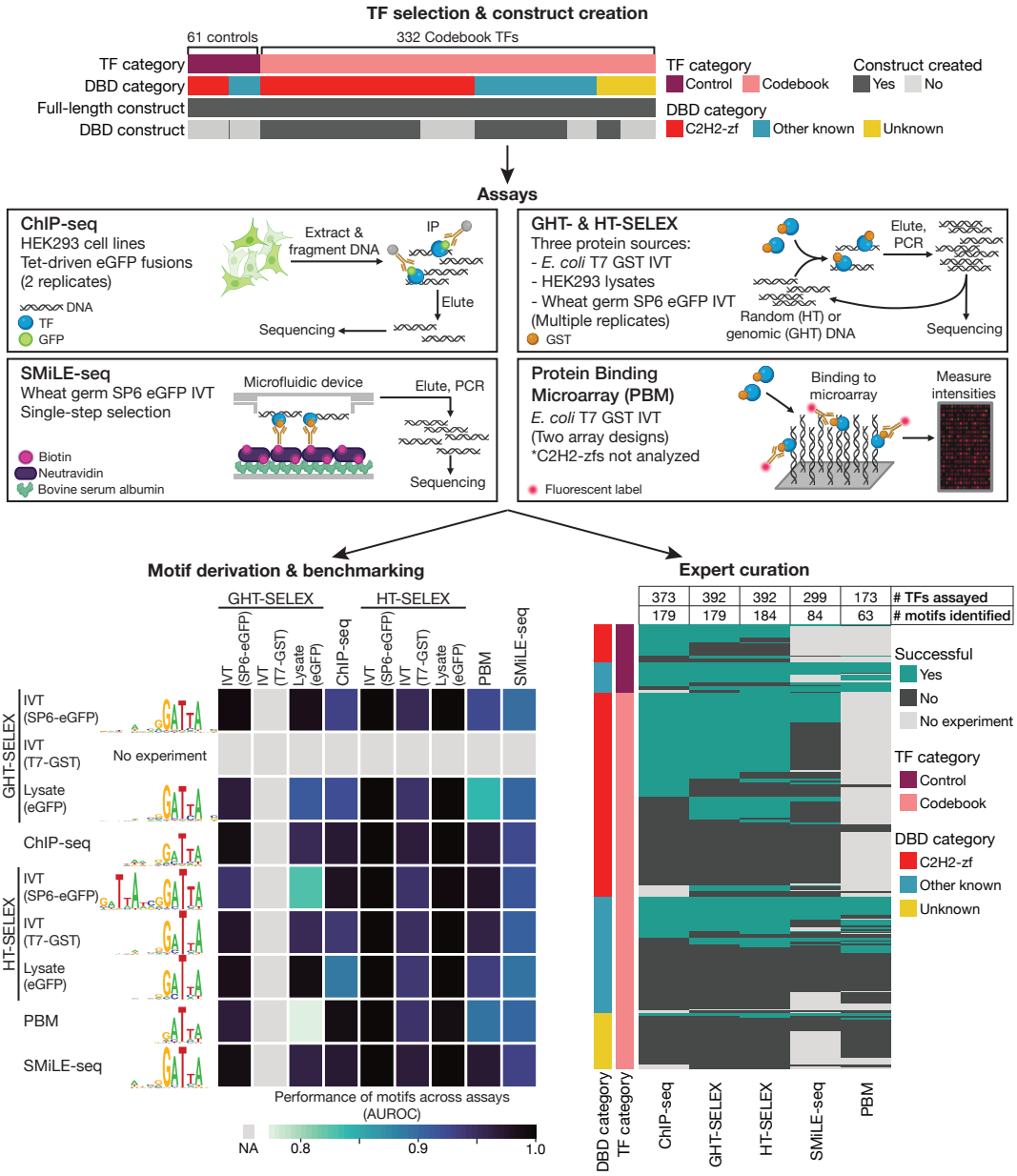
AUTHOR CONTRIBUTIONS

T.R.H. conceived of the Codebook Consortium. T.R.H. and B.D. directed the laboratory work. I.V.K. led the GRECO-BIT data analysis team. A.J., O.F., V.J.M., J.G., I.G., P.B., B.D., I.V.K., and T.R.H. oversaw the data analysis. Experiments were performed by: A.J. and A.W.H.Y (HT- and GHT-SELEX); A.W.H.Y (PBM); R.R., A.Bo., A.Br., H.Z., and M.B (ChIP-seq), and A.J.G, J.F.K.S., S.I, A.W.H.Y., and B.D. (SMiLE-seq). J.G., P.B., O.F., V.J.M., I.G., I.K., I.V.K., K.U.L., A.F., S.A., A.Br., A.Bo., J.F.K.S., A.J.G., M.A., M.T.W., I.V., and the Codebook Consortium members contributed to PWM derivation and motif analysis. All authors contributed to the initial assessment of experiments. I.V.K., A.J., and T.R.H. performed the secondary assessment of experiments. K.U.L, A.F., I.Y, V.N., A.B., G.M., and A.J. performed statistical analysis and created illustrations. The paper was drafted and finalized by T.R.H., A.J., K.U.L., A.F., I.Y., V.N., A.B., V.J.M., P.B., and I.V.K. All authors contributed to data analysis and reviewed the manuscript.

DECLARATION OF COMPETING INTERESTS

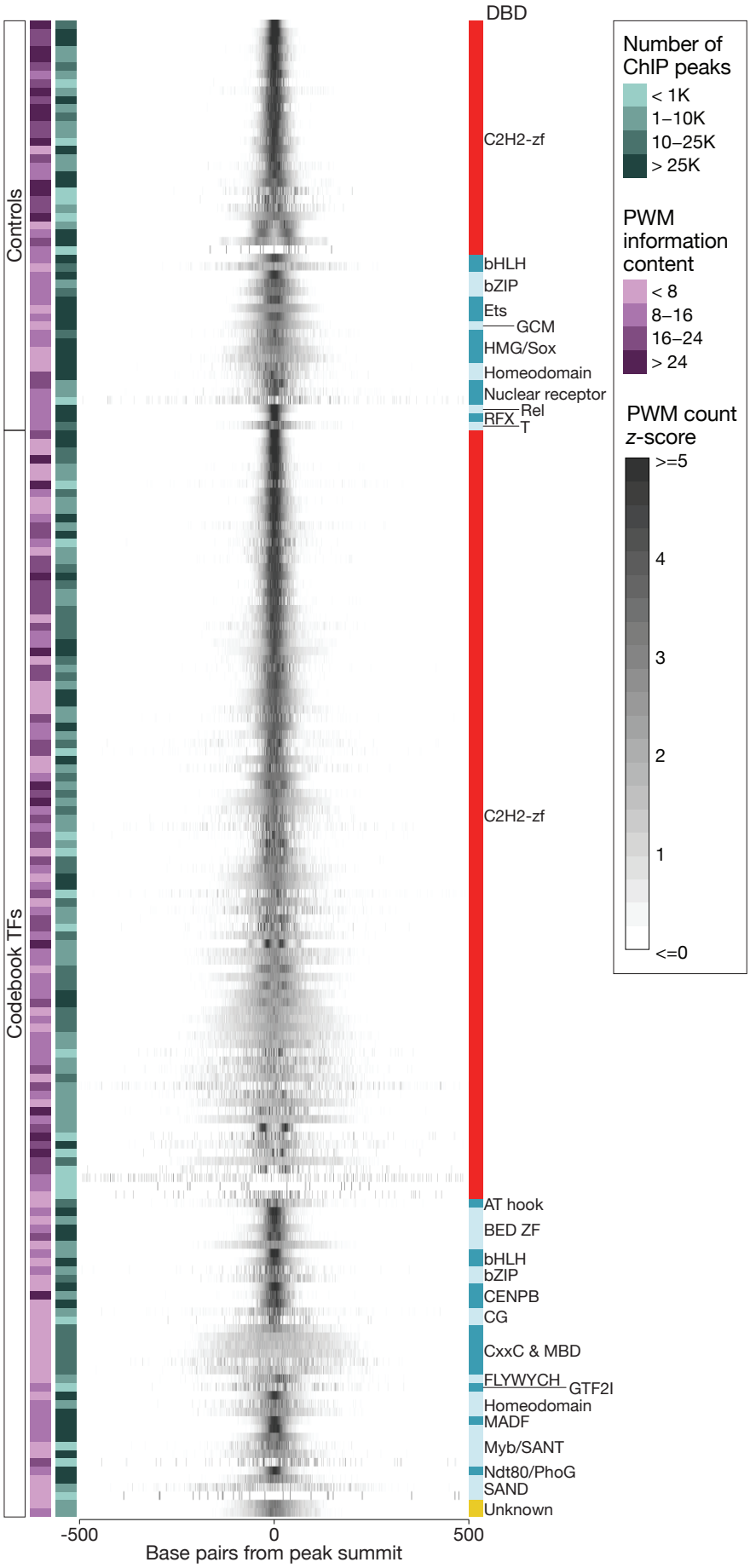
O.F. is employed by Roche.

793 **FIGURES & LEGENDS**


TF selection & construct creation
61 controls
332 Codebook TFs
TF category
DBD category
Full-length construct
DBD construct
TF category
Control
Codebook
Construct created
Yes
No
DBD category
C2H2-zf
Other known
Unknown
Assays
ChIP-seq
HEK293 cell lines
Tet-driven eGFP fusions
(2 replicates)
DNA
TF
GFP
Extract & fragment DNA
IP
Elute
Sequencing
GHT- & HT-SELEX
Three protein sources:
- E. coli T7 GST IVT
- HEK293 lysates
- Wheat germ SP6 eGFP IVT
(Multiple replicates)
GST
Elute, PCR
Random (HT) or genomic (GHT) DNA
Sequencing
SMiLE-seq
Wheat germ SP6 eGFP IVT
Single-step selection
Biotin
Neutravidin
Bovine serum albumin
Microfluidic device
Elute, PCR
Sequencing
Protein Binding Microarray (PBM)
E. coli T7 GST IVT
(Two array designs)
*C2H2-zfs not analyzed
Fluorescent label
Binding to microarray
Measure intensities
Motif derivation & benchmarking
GHT-SELEX
HT-SELEX
IVT (SP6-eGFP)
IVT (T7-GST)
Lysate (eGFP)
ChIP-seq
PBM
SMiLE-seq
No experiment
Performance of motifs across assays (AUROC)
NA
0.8
0.9
1.0
Expert curation
373
392
392
299
173
TFs assayed
179
179
184
84
63
motifs identified
Successful
Yes
No
No experiment
TF category
Control
Codebook
DBD category
C2H2-zf
Other known
Unknown
DBD category
TF category
ChIP-seq
GHT-SELEX
HT-SELEX
SMiLE-seq
PBM

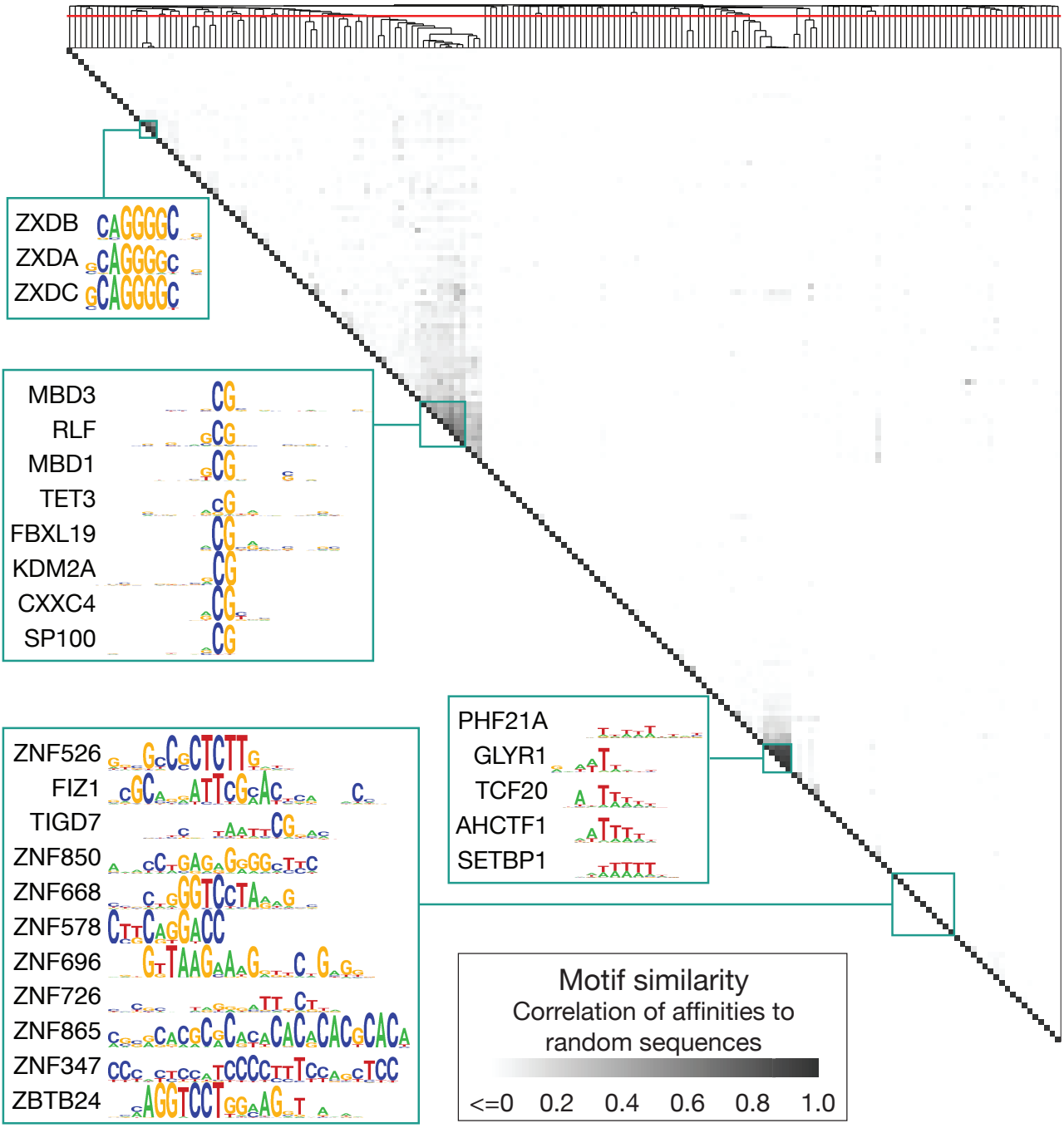
794 **Figure 1. Codebook project overview.** ***Top***, Categories of 393 TFs assayed and their
795 associated constructs. ***Middle***, Graphical summary of assays employed. ***Bottom left***,
796 Example of performance (as AUROC) of the best performing PWM for TPRX1, for each
797 combination of experiment type – one for PWM derivation (rows), and one for PWM
798 testing (columns). ***Bottom right***, Heatmap shows successful experiments for all 393
799 TFs across all experiment types. AUROC: area under precision-recall curve, PWM:
800 position weight matrix, TF: transcription factor, C2H2-zfs: C2H2 zinc finger proteins, IP:
801 immunoprecipitation, DBD: DNA-binding domain, IVT: in vitro transcription/translation

Figure 2

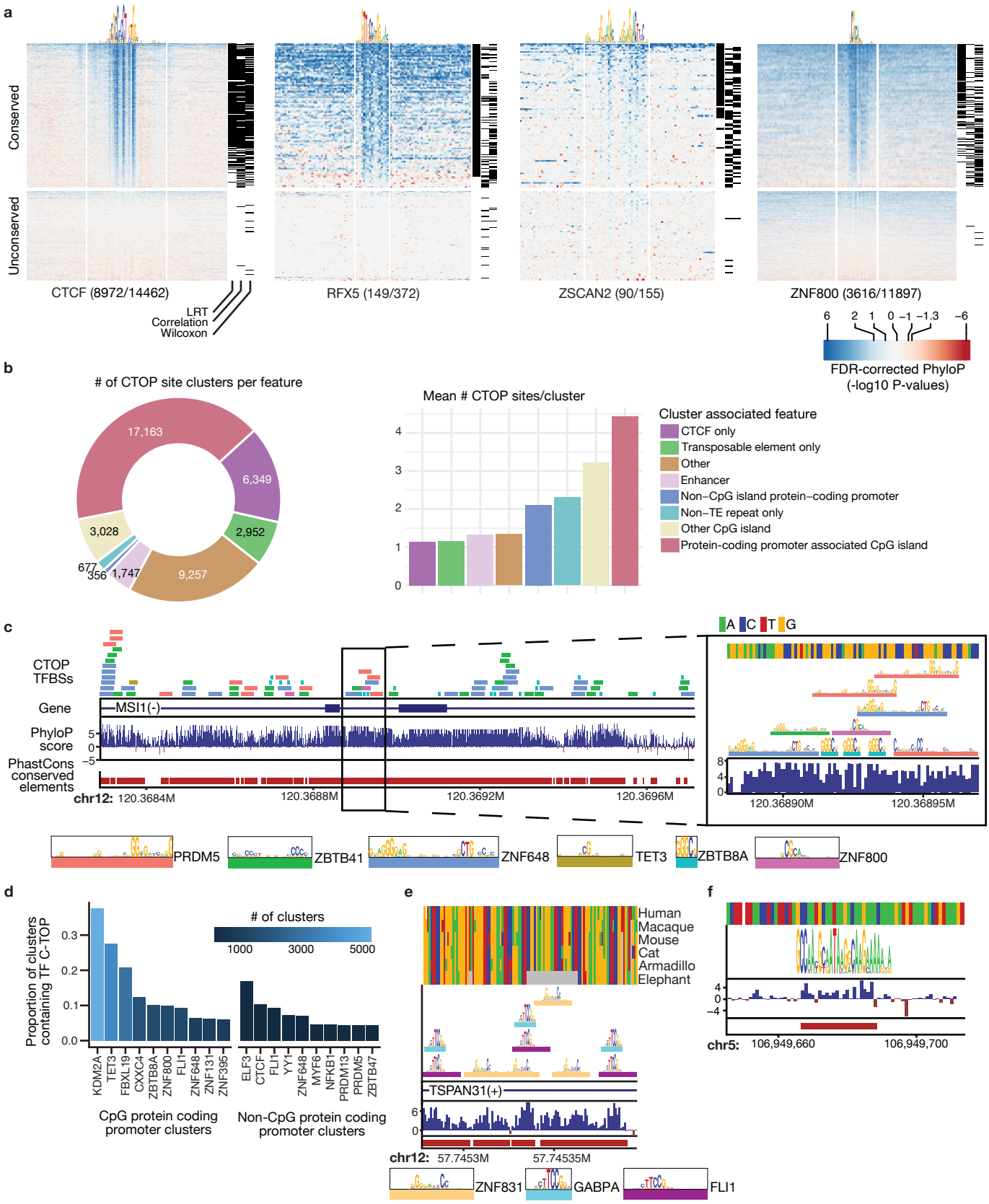

DBD
Controls
Codebook TFs
C2H2-zf
bHLH
bZIP
Ets
GCM
HMG/Sox
Homeodomain
Nuclear receptor
Rel
RFX
T
C2H2-zf
AT hook
BED ZF
bHLH
bZIP
CENPB
CG
CxxC & MBD
FLYWYCH
GTF2I
Homeodomain
MADF
Myb/SANT
Ndt80/PhoG
SAND
Unknown
-500
0
500
Base pairs from peak summit
Number of ChIP peaks
< 1K
1–10K
10–25K
> 25K
PWM information content
< 8
8–16
16–24
> 24
PWM count z-score
>=5
4
3
2
1
<=0

802 **Figure 2. Motif enrichment at ChIP-seq binding sites.** Heatmap displays the
803 standardized count of representative PWM matches at each base +/-500 bp from the
804 TF's ChIP-seq peak summits. PWM matches were identified with MOODS (P-value <
805 0.001) and aggregated for each TF at each base position. For visualization, counts were
806 standardized (z-score across all base positions) for each TF. PWM information content
807 and the number of ChIP-seq peaks for each TF are indicated on the left. Control and
808 Codebook TFs are separated and grouped by DBD type. PWM: position weight matrix,
809 DBD: DNA-binding domain, ChIP: chromatin immunoprecipitation, TF: transcription
810 factor

Figure 3


ZXDB
ZXDA
ZXDC
MBD3
RLF
MBD1
TET3
FBXL19
KDM2A
CXXC4
SP100
ZNF526
FIZ1
TIGD7
ZNF850
ZNF668
ZNF578
ZNF696
ZNF726
ZNF865
ZNF347
ZBTB24
PHF21A
GLYR1
TCF20
AHCTF1
SETBP1
Motif similarity
Correlation of affinities to random sequences
<=0
0.2
0.4
0.6
0.8
1.0

811 **Figure 3. Similarity of Codebook TF motifs.** Symmetric heatmap displays the
812 similarity between representative PWMs for Codebook TFs, clustered by Pearson
813 correlation with average linkage. PWM similarity is the correlation between pairwise
814 affinities to 150,000 random sequences of length 100, as calculated by MoSBAT[31]. The
815 red line across the dendrogram indicates the clustering threshold determined by the
816 optimal silhouette value[84]. Pullouts and labels illustrate specific points in the main text.
817 TF: transcription factor, PWM: position weight matrix


a
Conserved
Unconserved
CTCF (8972/14462)
RFX5 (149/372)
ZSCAN2 (90/155)
ZNF800 (3616/11897)
LRT
Correlation
Wilcoxon
6 2 1 0 −1 −1.3 −6
FDR-corrected PhyloP
(-log10 P-values)
b
# of CTOP site clusters per feature
17,163
6,349
2,952
9,257
1,747
356
677
3,028
Mean # CTOP sites/cluster
Cluster associated feature
CTCF only
Transposable element only
Other
Enhancer
Non−CpG island protein−coding promoter
Non−TE repeat only
Other CpG island
Protein-coding promoter associated CpG island
c
CTOP TFBSs
Gene
MSI1(-)
PhyloP score
PhastCons conserved elements
chr12: 120.3684M
120.3688M
120.3692M
120.3696M
A C T G
120.36890M
120.36895M
PRDM5
ZBTB41
ZNF648
TET3
ZBTB8A
ZNF800
d
Proportion of clusters containing TF C-TOP
# of clusters
1000 3000 5000
KDM2A TET3 FBXL19 CXXC4 ZBTB8A ZNF800 FLI1 ZNF648 ZNF131 ZNF395
ELF3 CTCF FLI1 YY1 ZNF648 MYF6 NFKB1 PRDM13 PRDM5 ZBTB47
CpG protein coding promoter clusters
Non-CpG protein coding promoter clusters
e
Human
Macaque
Mouse
Cat
Armadillo
Elephant
TSPAN31(+)
chr12: 57.7453M
57.74535M
ZNF831
GABPA
FLI1
f
chr5: 106,949,660
106,949,700

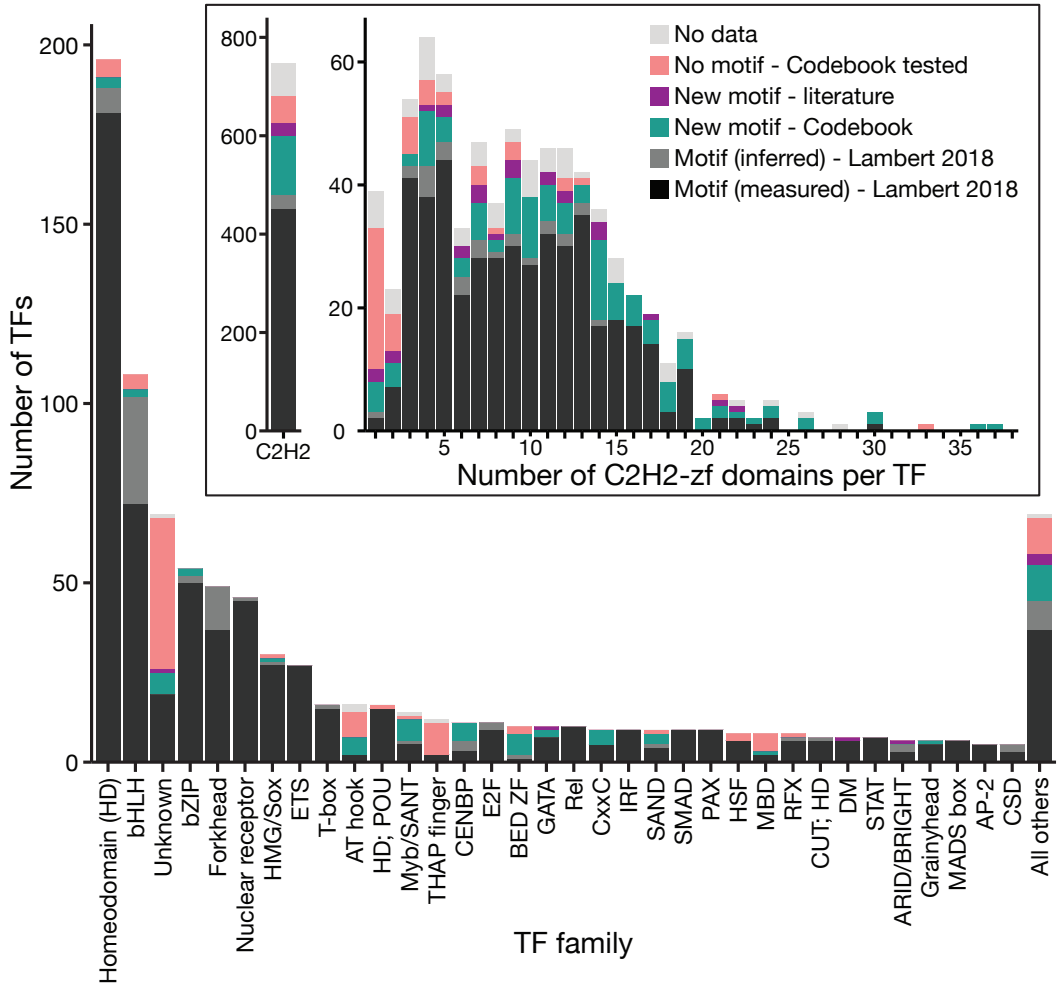
818 **Figure 4. Conservation of Codebook TF binding sites and association with**
819 **genomic features. a,** Heatmaps of FDR-corrected phyloP scores over the PWM match
820 and 50 bp flanking for TOP sites for four TFs (two controls and two Codebook TFs).
821 Statistical test results (see main text and **Methods**) are indicated at right. **b,** *Left,* Donut
822 plot displays the proportion and number of clusters of conserved TOP (CTOP) sites that
823 overlap the genomic features indicated. *Middle,* Bar plot displays the mean # of
824 individual CTOPs contained within clusters that overlap the examined genomic regions.
825 **c,** A 1,420-base, CpG-island-overlapping CTOP cluster (chr12:120368293-120369713).
826 Zoonomia 241-mammal phyloP scores and Multiz 471 Mammal alignment PhastCons
827 Conserved Elements are shown. **d,** Bar plot of the frequency of TFs with CTOPs that
828 occur most frequently in CTOP clusters that overlap CpG and non-CpG protein coding
829 promoters, respectively. **e,** CTOP cluster overlapping the non-CpG promoter at
830 chr12:57,745,278-57,745,396. **f,** CTOP site for the KRAB-C2H2-zf protein ZNF689,
831 overlapping an L1M4-type L1 element located at chr5:106949666-106949692. TF:
832 transcription factor, TOP: triple-overlap, CTOP: conserved triple-overlap

Figure 5

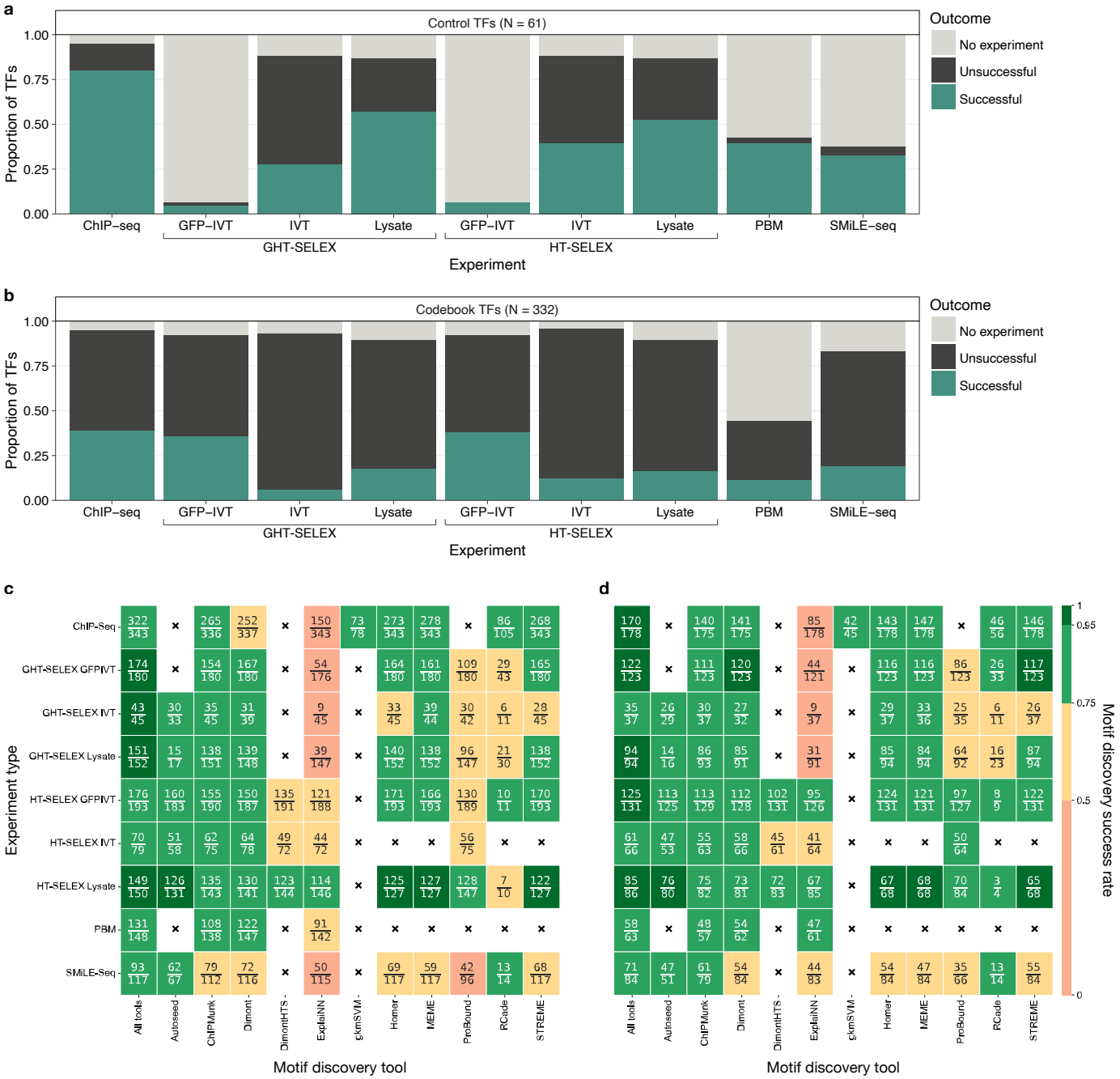
Motif activity
-20 -10 0 10 20
FANTOM5 ontology
Disease
Carcinoma
Immune system cancer
Other/mixed
Cell
Epithelial cell
Mesenchymal cell
Other/mixed
Organ
Brain
Skeletal element
Other/mixed
Tissue
Neurectoderm
Mesoderm
Endoderm
Other/mixed
PCC with expression
0.6
-0.4
Disease
Cell
Organ
Tissue
Domain
BED ZF
C2H2 ZF
BED ZF
CxxC
C2H2 ZF
C2H2 ZF
CxxC
SAND
C2H2 ZF
unknown
Ndt80/PhoG
CG-1
MBD
C2H2 ZF
C2H2 ZF
C2H2 ZF
C2H2 ZF
C2H2 ZF
C2H2 ZF
BED ZF
C2H2 ZF
C2H2 ZF
Myb/SANT
SAND
C2H2 ZF
ZBED5
ZBTB8A
ZFTA
(1)
MBD1
ZNF721
RLF
CXXC4
SP140
ZNF131
CGGBP1
(2)
MYRFL
CAMTA1
MBD3
ZNF800
DNTTIP1
ZNF292
ZNF215
ZNF48
ZFP91
(3)
FAM200B
FIZ1
ZBTB41
TTF1
SP100
ZBTB5
(4)
Lymphoid lineage cells, lymphoma & myeloid leukemia cell lines
Myeloid lineage cells
Carcinoma & mesothelioma cell lines
Epithelial cells
MSCs, fibroblasts & smooth muscle cells
Carcinoma, blastoma & sarcoma cell lines
Tissues (except nervous)
Fetal & newborn nervous system
Adult nervous system

833 **Figure 5. Motif activity response analysis highlights the Codebook motifs'**
834 **functionality**. Heatmap of bi-clustered MARA-predicted motif activities. X-axis: 583
835 FANTOM5 samples, clusters are labeled manually according to the predominant cell
836 and tissue types. Y-axis: TFs representing 25 Codebook motif clusters with the most
837 significant differential activity across FANTOM5 samples. Additional columns, left-to-
838 right: DNA-binding domain of the representative motif and the correlation between TF
839 expression and motif activity. Extra rows, top-to-bottom: FANTOM5 ontology-based
840 sample classification using the selected groups from disease state (DOID:4), cell type
841 (CL:0000003), organ (UBERON:0000062), and tissue (UBERON:0000479) terms,
842 clusters are colored by the predominant term if assigned to at least 40% of the samples.
843 PCC: Pearson correlation coefficient, TF: transcription factor

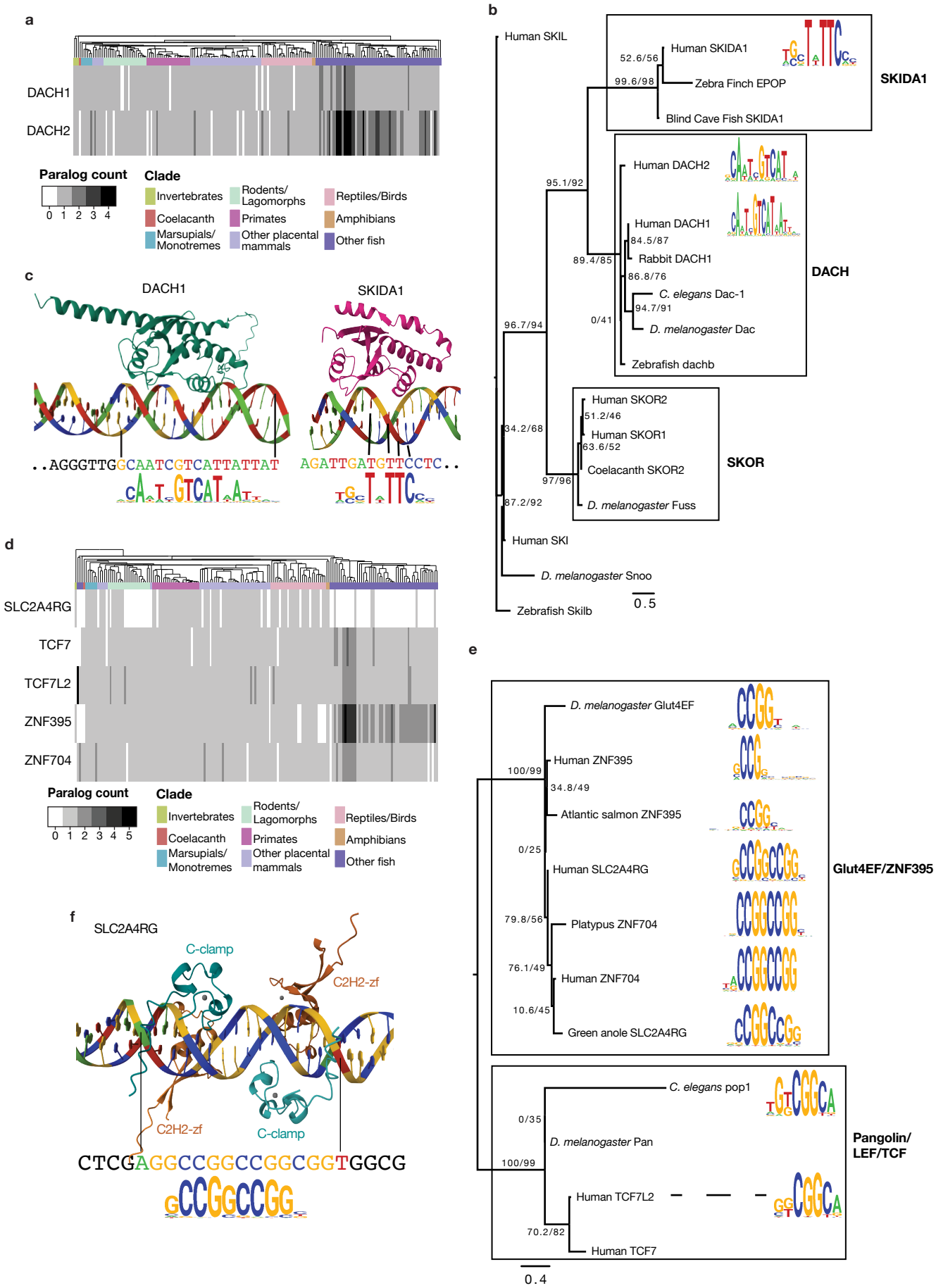
Figure 6


No data
No motif - Codebook tested
New motif - literature
New motif - Codebook
Motif (inferred) - Lambert 2018
Motif (measured) - Lambert 2018
Number of TFs
C2H2
Number of C2H2-zf domains per TF
TF family
Homeodomain (HD)
bHLH
Unknown
bZIP
Forkhead
Nuclear receptor
HMG/Sox
ETS
T-box
AT hook
HD; POU
Myb/SANT
THAP finger
CENBP
E2F
BED ZF
GATA
Rel
CxxC
IRF
SAND
SMAD
PAX
HSF
MBD
RFX
CUT; HD
DM
STAT
ARID/BRIGHT
Grainyhead
MADS box
AP-2
CSD
All others

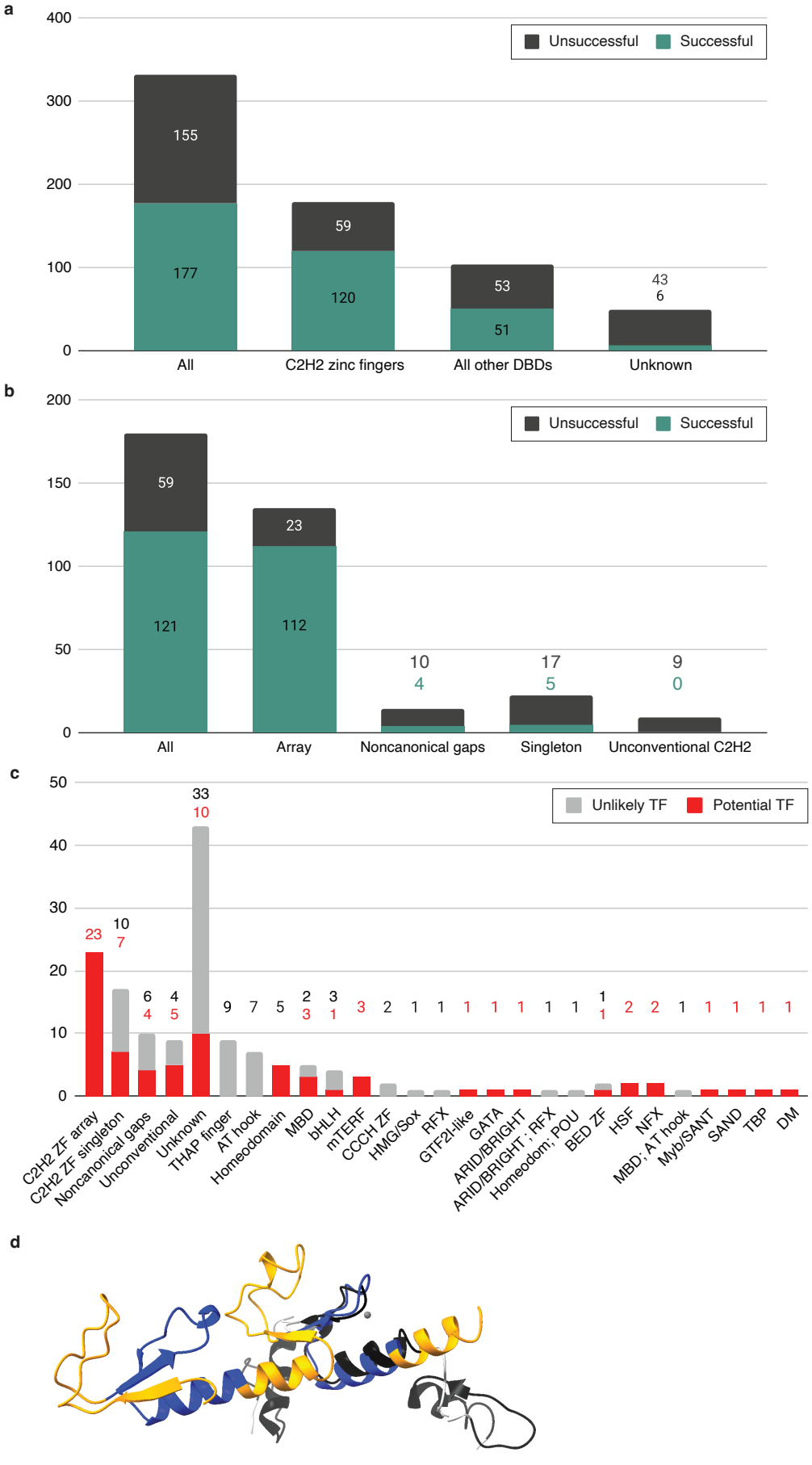
844 **Figure 6. Motif coverage of human TFs by DBD family.** TFs are categorized into
845 structural classes from Lambert et al.[1]. See **Supplementary Table 14** for underlying
846 information. TF: transcription factor, DBD: DNA-binding domain


a
Control TFs (N = 61)
Proportion of TFs
1.00
0.75
0.50
0.25
0.00
ChIP–seq
GFP–IVT
IVT
Lysate
GFP–IVT
IVT
Lysate
PBM
SMiLE–seq
GHT-SELEX
HT-SELEX
Experiment
Outcome
No experiment
Unsuccessful
Successful
b
Codebook TFs (N = 332)
Proportion of TFs
1.00
0.75
0.50
0.25
0.00
ChIP–seq
GFP–IVT
IVT
Lysate
GFP–IVT
IVT
Lysate
PBM
SMiLE–seq
GHT-SELEX
HT-SELEX
Experiment
Outcome
No experiment
Unsuccessful
Successful
c
Experiment type
ChIP-Seq
GHT-SELEX GFPIVT
GHT-SELEX IVT
GHT-SELEX Lysate
HT-SELEX GFPIVT
HT-SELEX IVT
HT-SELEX Lysate
PBM
SMiLE-Seq
All tools
Autoseed
ChIPMunk
Dimont
DimontHTS
ExplaiNN
gkmSVM
Homer
MEME
ProBound
RCade
STREME
Motif discovery tool
d
All tools
Autoseed
ChIPMunk
Dimont
DimontHTS
ExplaiNN
gkmSVM
Homer
MEME
ProBound
RCade
STREME
Motif discovery tool
Motif discovery success rate
1
0.95
0.75
0.5
0

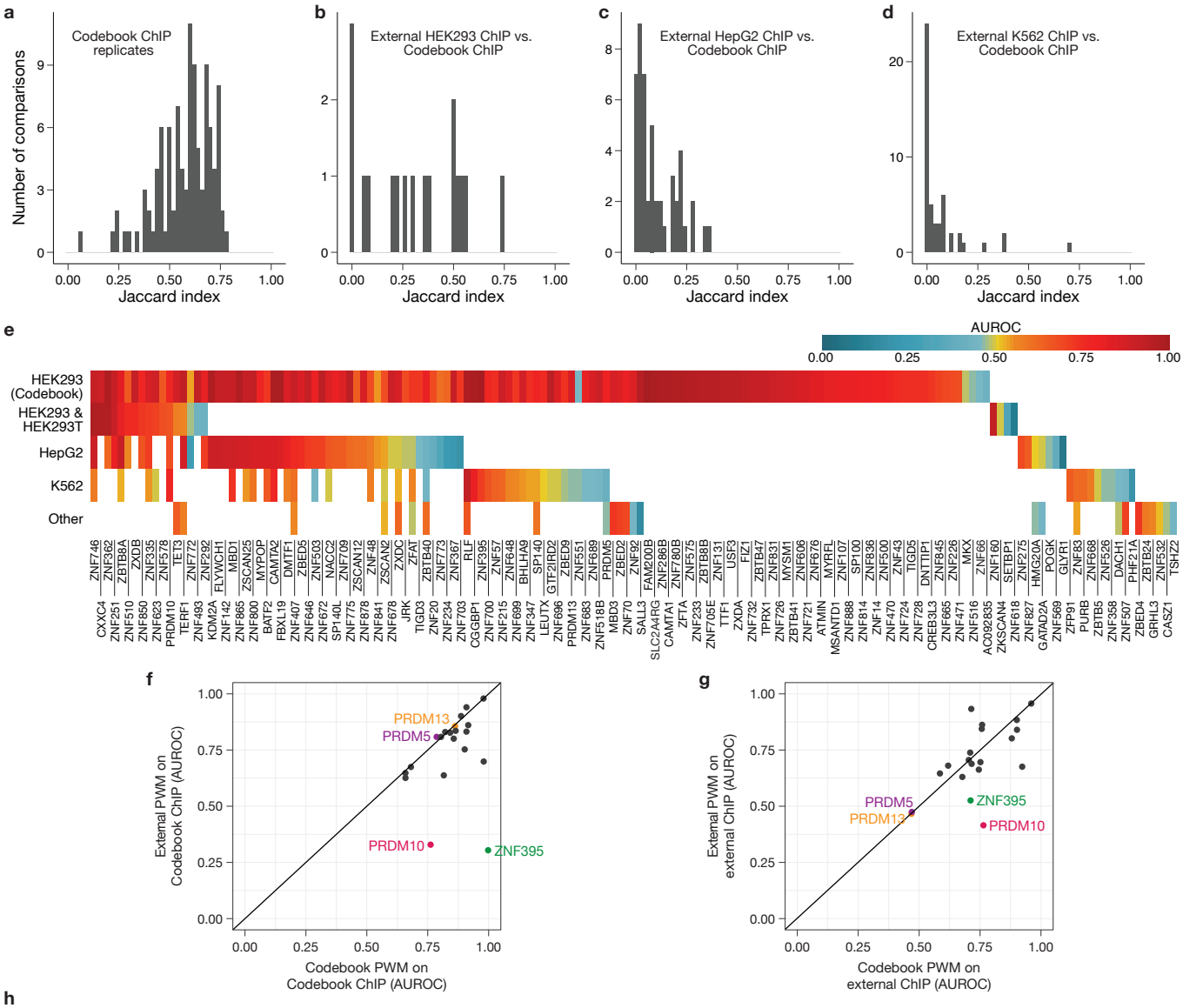
847 **Extended Data Figure 1. Motif discovery success rates. a**, **b**, Proportion of control
848 (**a**) and Codebook (**b**) TFs for which a motif was successfully identified across
849 experiment types. GHT-SELEX and HT-SELEX experiments are separated according to
850 protein expression system. The proportion of untested TFs is also indicated. **c, d**, Rate
851 of successful motif discovery across experiment types and motif discovery tools.
852 Numerator in each cell represents the number of successes (cases of correctly
853 identified motifs), and denominator represents the number of experiments (**c**) or TFs (**d**)
854 for a particular combination of a tool and an assay. The first column in **c** and **d** displays
855 the joint estimates across all tools. IVT: in vitro transcription, TF: transcription factor


a
DACH1
DACH2
Paralog count
0 1 2 3 4
Clade
Invertebrates
Rodents/ Lagomorphs
Reptiles/Birds
Coelacanth
Primates
Amphibians
Marsupials/ Monotremes
Other placental mammals
Other fish
b
Human SKIL
52.6/56
Human SKIDA1
99.6/98
Zebra Finch EPOP
Blind Cave Fish SKIDA1
SKIDA1
95.1/92
Human DACH2
Human DACH1
84.5/87
Rabbit DACH1
89.4/85
86.8/76
C. elegans Dac-1
94.7/91
0/41
D. melanogaster Dac
Zebrafish dachb
DACH
96.7/94
Human SKOR2
51.2/46
Human SKOR1
63.6/52
Coelacanth SKOR2
97/96
D. melanogaster Fuss
SKOR
34.2/68
87.2/92
Human SKI
D. melanogaster Snoo
Zebrafish Skilb
0.5
c
DACH1
SKIDA1
..AGGGTTGGCAATCGTCATTATTAT
AGATTGATGTTCCTC..
d
SLC2A4RG
TCF7
TCF7L2
ZNF395
ZNF704
Paralog count
0 1 2 3 4 5
Clade
Invertebrates
Rodents/ Lagomorphs
Reptiles/Birds
Coelacanth
Primates
Amphibians
Marsupials/ Monotremes
Other placental mammals
Other fish
e
D. melanogaster Glut4EF
100/99
Human ZNF395
34.8/49
Atlantic salmon ZNF395
0/25
Human SLC2A4RG
79.8/56
Platypus ZNF704
76.1/49
Human ZNF704
10.6/45
Green anole SLC2A4RG
Glut4EF/ZNF395
C. elegans pop1
0/35
D. melanogaster Pan
100/99
Human TCF7L2
70.2/82
Human TCF7
Pangolin/ LEF/TCF
0.4
f
SLC2A4RG
C-clamp
C2H2-zf
C2H2-zf
C-clamp
CTCGAGGCCGGCCGGCGGTGGCG

856 **Extended Data Figure 2. Undercharacterized DNA-binding domains.** Overview of
857 new motifs for undercharacterized TF families. **a,** Number of DACH1 and DACH2
858 orthologs (union of one-to-one and one-to-many) across Ensembl v111 vertebrates and
859 selected invertebrates. Species order reflects the Ensembl species tree. **b,** Sequence
860 logos (from the main Codebook project, and follow-up study described in the
861 **Supplementary Discussion**) and sequence relationships of human and representative
862 animal SKI/SNO/DAC domains. All motifs from successful HT-SELEX assays are
863 shown, and a representative set of the DBDs tested is included in the phylogenetic tree.
864 Tree is an unrooted maximum-likelihood phylogram from IQTree[100] using parameters --
865 alrt 1000 -B 1000 (1000 replicates to compute SH-aLRT (a branch support value from 0-
866 100%), 1000 replicates for ultrafast bootstrap), and automatic best-fit model selection
867 with ModelFinder. Branch values are SH-aLRT/ultrafast bootstrap. **c,** AlphaFold3-
868 predicted structure of the DACH1 SKI/SNO/DAC region (residues 130 – 390) bound to
869 an HT-SELEX ligand sequence with a high-scoring PWM match. **d,** Number of orthologs
870 (same counting process as **a**) for human C-clamp proteins. **e**, Sequence logos (from the
871 main Codebook project and follow-up study) and sequence relationships of human and
872 representative animal C-Clamp domains (*ZNF704 motif from [101]). The tree was
873 produced identically to the tree in **b**. **f,** AlphaFold3-predicted structure of two full-length
874 SLC2A4RG proteins bound to a CTOP sequence with flanking sequences
875 (chr17:48,048,369-48,048,401), and four Zn^{2+} ions (grey). The remainder of the proteins
876 (beyond the C-clamp and C2H2-zf domains) are hidden, for visual simplicity. DBD:
877 DNA-binding domain


a
Unsuccessful
Successful
400
300
200
100
0
155
177
59
120
53
51
43
6
All
C2H2 zinc fingers
All other DBDs
Unknown
b
Unsuccessful
Successful
200
150
100
50
0
59
121
23
112
10
4
17
5
9
0
All
Array
Noncanonical gaps
Singleton
Unconventional C2H2
c
Unlikely TF
Potential TF
50
40
30
20
10
0
23
10
7
6
4
4
5
33
10
9
7
5
2
3
3
1
3
2
1
1
1
1
1
1
1
1
1
2
2
1
1
1
1
1
C2H2 ZF array
C2H2 ZF singleton
Noncanonical gaps
Unconventional
Unknown
THAP finger
AT hook
Homeodomain
MBD
bHLH
mTERF
CCCH ZF
HMG/Sox
RFX
GTF2I-like
GATA
ARID/BRIGHT
ARID/BRIGHT ; RFX
Homeodom; POU
BED ZF
HSF
NFX
MBD ; AT hook
Myb/SANT
SAND
TBP
DM
d

878 **Extended Data Figure 3. Description and re-assessment of unsuccessful proteins.**
879 **a**, Number of successful and unsuccessful Codebook proteins by category. **b,** Number
880 of successful and unsuccessful C2H2-zf proteins categorized based on: whether a
881 protein has an array of at least two DBDs separated by 3-10 AA; has multiple C2H2-zf
882 domains but with other kinds of spacings; has only a single domain; or if it has a
883 potential C2H2-zf domain based on literature but none are predicted by our standard
884 HMM based prediction. **c,** Summary of re-assessment (see **Supplementary Table 6**
885 and **Supplementary Document 1** for details) of unsuccessful proteins by structural
886 class or category. Colors indicate if the protein is still assessed as a potential TF (red)
887 and is therefore a potential false negative, or if it is now assessed as an unlikely TF
888 (gray). **d**, shows a comparison of the CASZ1 "para-ZNF" domain[102] zinc fingers 1 and 2
889 against an example of canonical C2H2-zf domain of (GLI2, PDB:2GLI), with the
890 predicted C2H2-zf region CASZ1 DBD2 structurally aligned to DBD4 of GLI2. TF:
891 transcription factor, DBD: DNA-binding domain, HMM: hidden Markov model

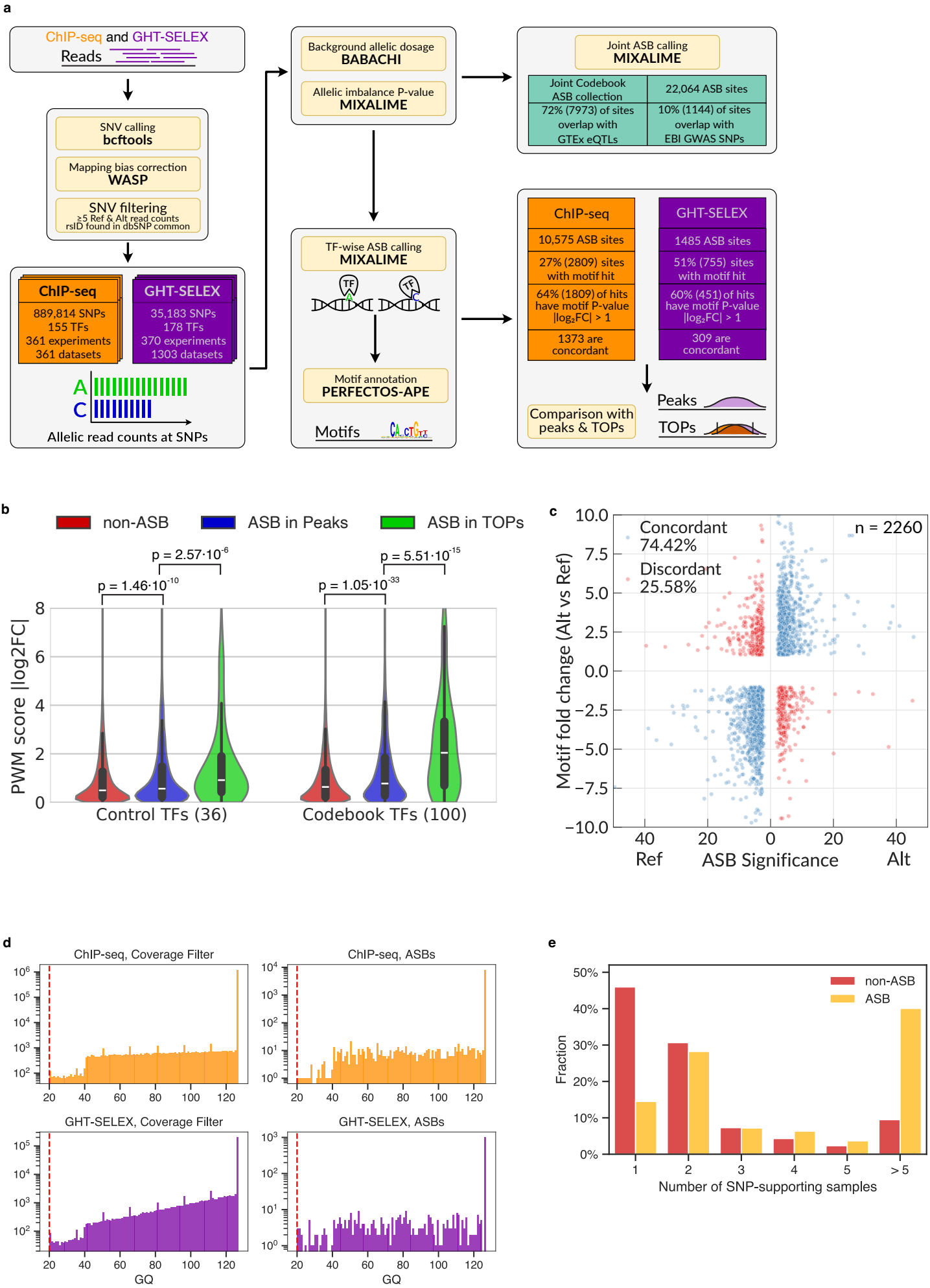


h

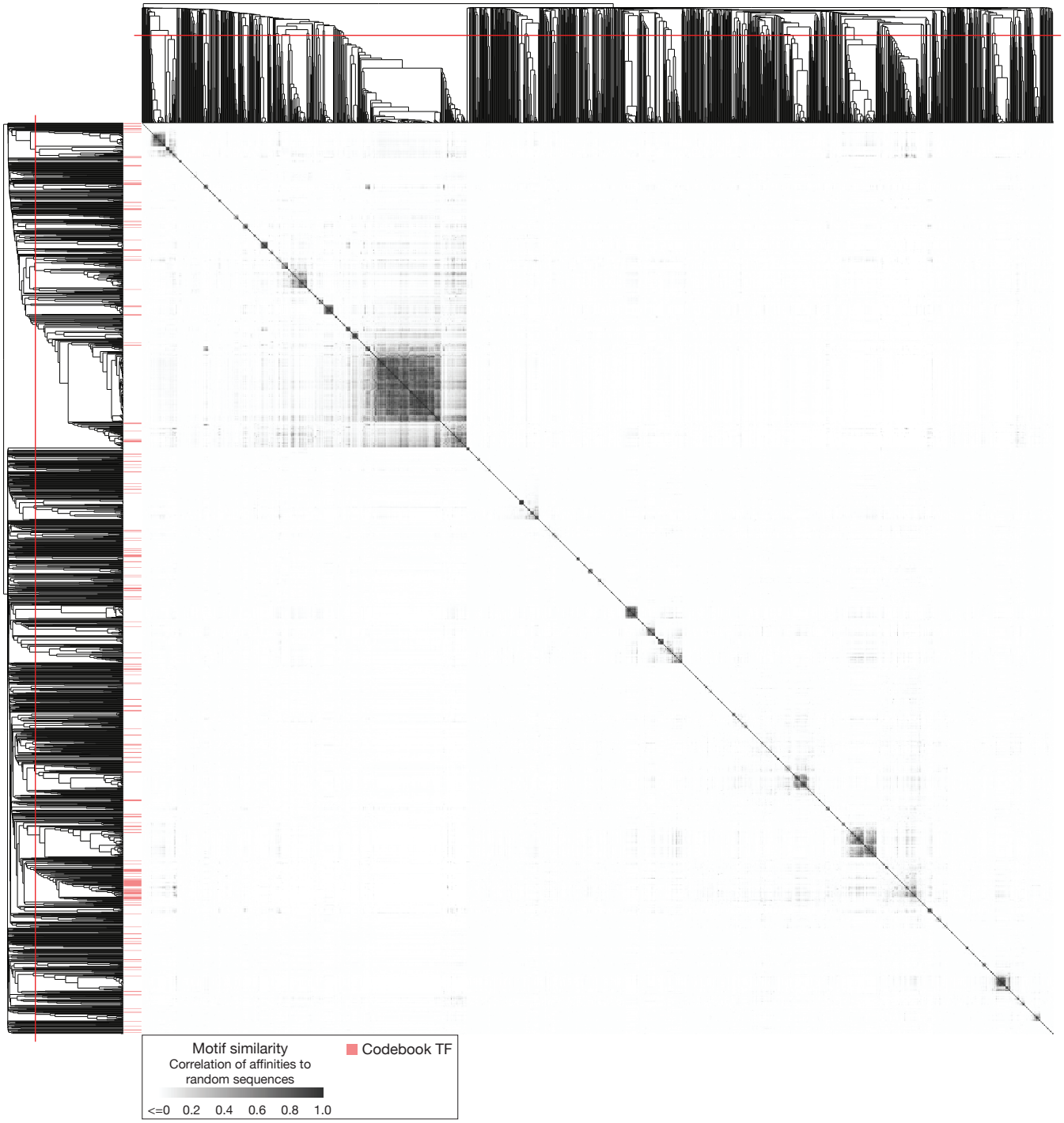
TF	Codebook PWM Codebook ChIP	External PWM Codebook ChIP	Codebook PWM external ChIP	External PWM external ChIP	Codebook PWM	External PWM
PRDM5	0.787	0.808	0.470	0.474		
PRDM13	0.862	0.856	0.468	0.466		
ZNF395	0.998	0.304	0.711	0.526		
PRDM10	0.762	0.329	0.764	0.415		

892 **Extended Data Figure 4. Comparison of Codebook ChIP-seq datasets and PWMs**
893 **to external ChIP-seq datasets and PWMs. a-d,** Histograms of optimized Jaccard
894 indices measuring the overlap between ChIP-seq peak sets for the same TF: **a,**
895 Codebook ChIP-seq replicates; **b, c, d**: Codebook ChIP-seq vs. external ChIP-seq
896 performed in HEK293 cells (**b**), HepG2 cells (**c**), or K562 cells (**d**). Control TFs are not
897 included. **e,** AUROC values for expert-curated Codebook PWMs (columns),
898 discriminating ChIP-seq peaks vs. random genomic background loci. Rows show
899 different cell types. **f, g,** comparison of Codebook and external PWMs at the task of
900 discriminating ChIP-seq peak sets from random sequences (as in **e**), for the 20 TFs that
901 have Codebook ChIP-seq data, a Codebook PWM, external ChIP-seq data, and an
902 external PWM. **f**, Performance of Codebook PWM vs. the best performing external
903 PWM on Codebook ChIP-seq data. **g**, Performance of Codebook PWM vs. the best
904 performing external PWM on external ChIP-seq data. The four TFs with an AUROC of <
905 0.5 on either axis of either plot are highlighted. **h,** Sequence logos for the four TFs
906 highlighted in **f** and **g**. All expert-curated Codebook PWMs displayed are supported by
907 ChIP-seq, GHT-SELEX, HT-SELEX, and SMiLE-seq data. PWM: position weight matrix,
908 TF: transcription factor, AUROC: area under the receiver operating characteristic

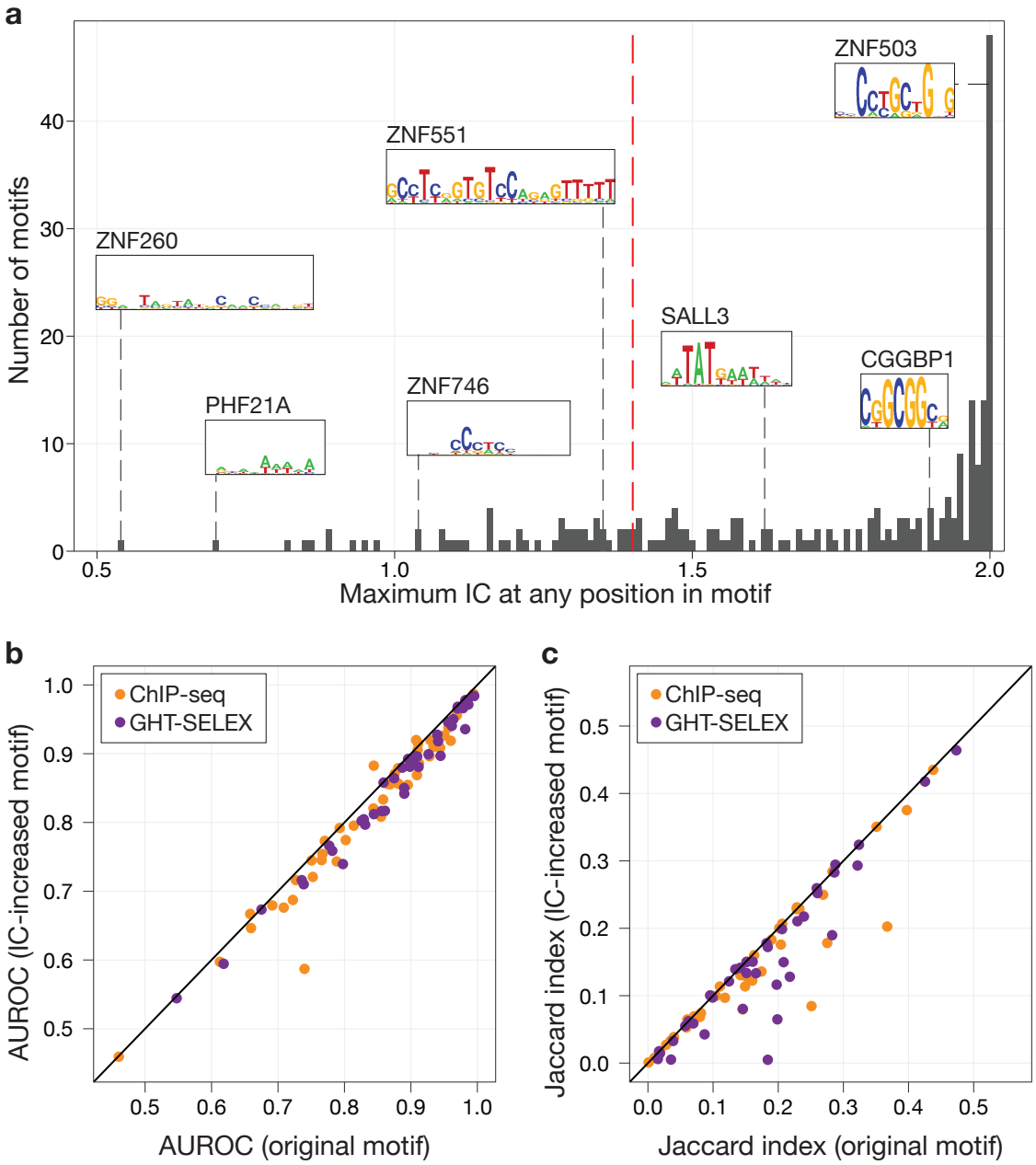
Extended Data Figure 5



Extended Data Figure 5. Identifying allele-specific TF binding sites in Codebook ChIP-seq and GHT-SELEX data. a, Principal scheme of the allele-specific analysis. **b,** Distribution of PWM score P-value ratio (fold change), $|\log_2 FC|$, between alleles for non-ASBs (FDR ≥ 5%), ASBs (FDR < 5%) in peaks, and ASBs in TOPs (FDR < 5%). *Left*, 36 positive control TFs; *Right*, 100 Codebook TFs. P-values: Mann-Whitney U test. **c,** Motif concordance of Codebook ASBs. X-axis: ASB allelic preference and significance ($-\log_{10}$ FDR), *left*: preference for the reference allele (Ref), *right*: preference for the alternative allele (Alt). Y-axis: PWM score P-value ratio ($\log_2$ FC) between Alt and Ref. The plot shows only ASBs with $|\log_2 FC| > 1$. **d,** Genotype quality (GQ) of SNPs called from ChIP-Seq (top) and GHT-SELEX (bottom) data. X-axis: GQ, Y-axis: number of SNPs or ASBs. Red dashed line at 20 is the default GQ threshold used to filter SNP calls. The majority of candidate SNPs and ASBs have GQ much higher than the default threshold**. e,** The number of SNP-supporting ChIP-Seq or GHT-SELEX datasets for ASB and non-ASB SNPs. The majority of ASB SNPs are supported by two or more datasets. ASB: allele-specific binding site, FC: fold change, SNP: single-nucleotide polymorphism, TOP: triple-overlap

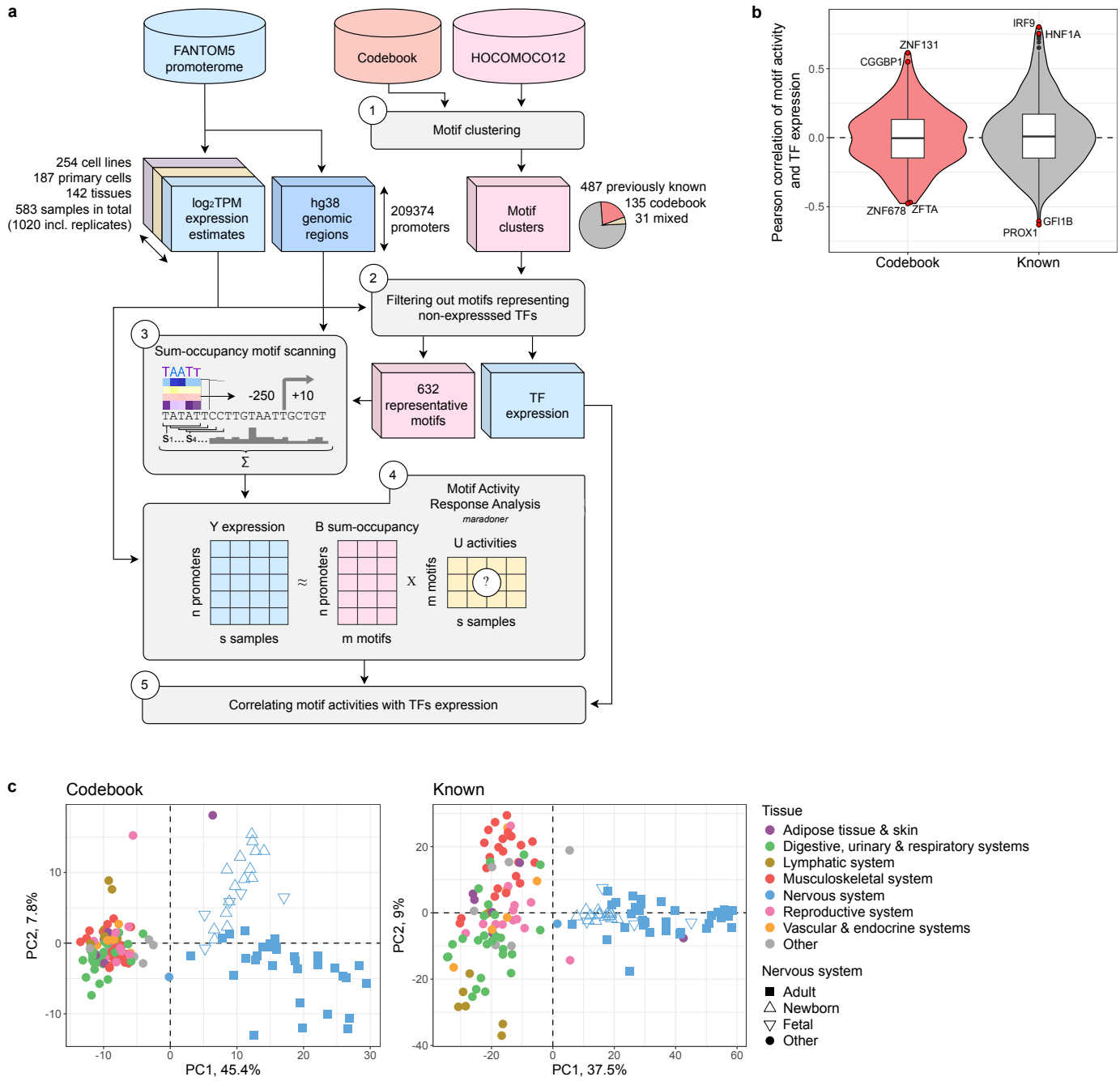

Motif similarity
Correlation of affinities to random sequences
<=0 0.2 0.4 0.6 0.8 1.0
Codebook TF

925 **Extended Data Figure 6. Similarity of TF motifs.** Symmetric heatmap displays the
926 similarity between 1,582 PWMs clustered by Pearson correlation with average linkage.
927 The PWM set is composed of the 177 representative PWMs for Codebook TFs, and
928 1,405 PWMs for 1,211 TFs with previously identified binding specificities. The set of
929 PWMs for non-Codebook TFs was retrieved from Lambert et al. 2018[1]. PWM similarity
930 is the correlation between pairwise affinities to 150,000 random sequences of length
931 100 bp, as calculated by MoSBAT[31]. The red line across the dendrogram indicates the
932 clustering threshold (see **Fig. 3**, **Methods**) determined by the optimal silhouette value[84].
933 PWMs and cluster membership are in **Supplementary Table 10**. PWM: position weight
934 matrix, TF: transcription factor

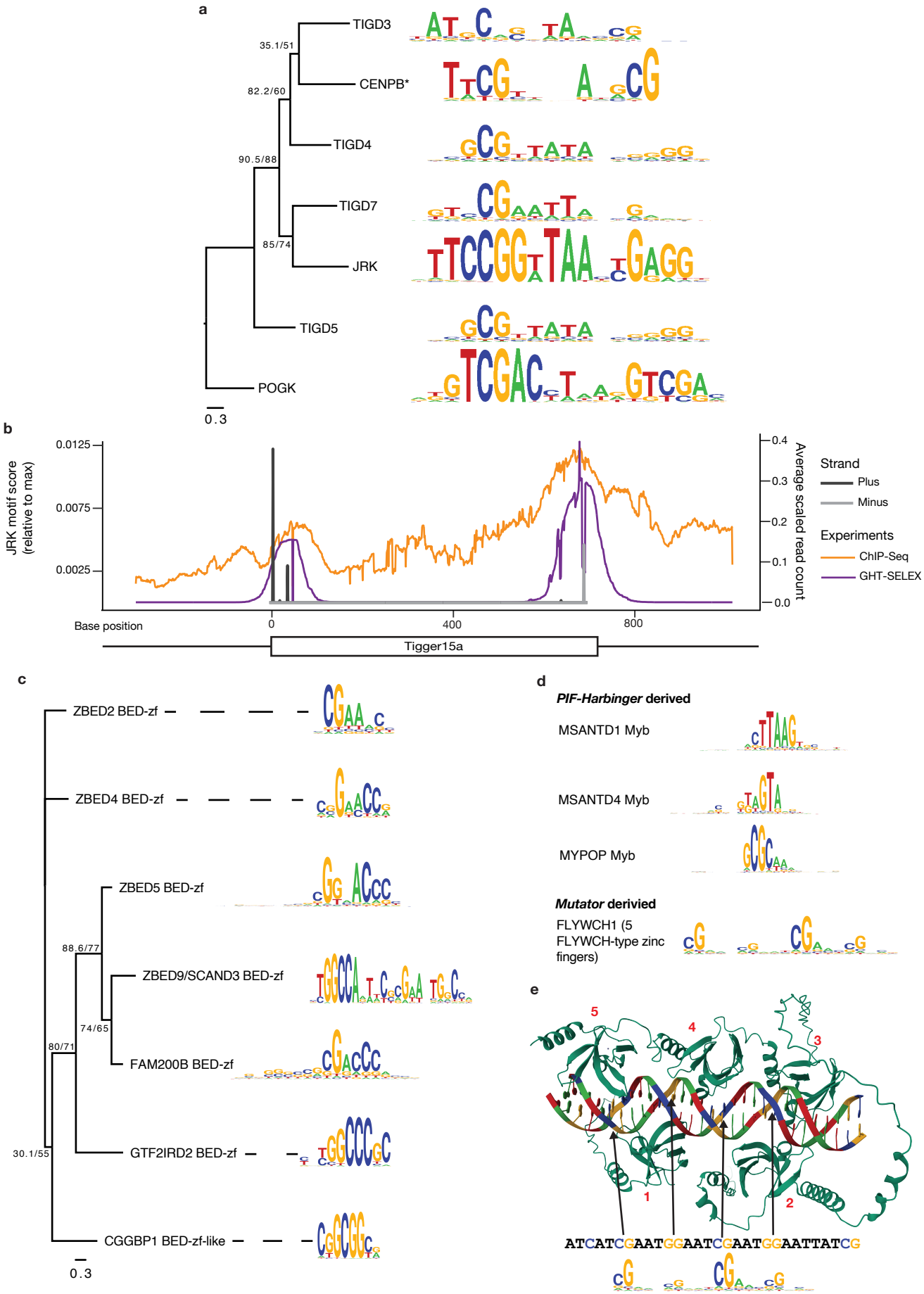

a
Number of motifs
Maximum IC at any position in motif
ZNF503
ZNF551
ZNF260
SALL3
CGGBP1
ZNF746
PHF21A
b
AUROC (IC-increased motif)
AUROC (original motif)
ChIP-seq
GHT-SELEX
c
Jaccard index (IC-increased motif)
Jaccard index (original motif)
ChIP-seq
GHT-SELEX

935 **Extended Data Figure 7. Motif degeneracy analysis. a,** Histogram displays the
936 maximum IC for any position within the representative PWM for all Codebook and
937 control TFs. Logos are shown for TFs at various maximum positional IC values, for
938 illustration. Red dashed line indicates an IC of 1.4. **b,** and **c,** comparison of original
939 PWMs to IC-increased PWMs for the 52 TF PWMs for which no base position exceeded
940 an IC of 1.4. **b,** AUROC values for original vs. IC-increased PWMs, discriminating ChIP-
941 seq or GHT-SELEX peaks vs. random genomic background loci. **c,** Maximum Jaccard
942 index for ChIP-seq or GHT-SELEX peak sets; using the approach described for
943 optimized TOPs in **Methods**, for original vs. IC-increased PWMs. PWM: position weight
944 matrix, IC: information content, TOP: triple-overlap regions, AUROC: area under the
945 receiver operating characteristic

Extended Data Figure 8



946 **Extended Data Figure 8. Motif activity response analysis highlights the relevance**
947 **of Codebook motifs. a,** MARA was conducted in five consecutive steps. (1) TF binding
948 motifs from Codebook and HOCOMOCO[38] (v12) were clustered, yielding 653 clusters,
949 135 of which consisted solely of Codebook motifs. (2) Motif clusters were filtered based
950 on the total $\log_2$TPM of member TFs, resulting in 632 clusters of expressed TF motifs.
951 (3) Representative motifs for each cluster were used to scan 209,374 genomic regions
952 of 260 bp (250 bp upstream and 10 bp downstream relative to the representative TSS
953 position of each promoter). (4) The sum-occupancy scores were used to perform the
954 motif activity response analysis with the MARADONER Python package; (5) Pearson
955 correlation coefficients were calculated for each representative motif by comparing the
956 total TF expression against motif activities from MARA. **b,** Correlation between the TF
957 expression ($\log_2$ TPM) and the activity of Codebook-specific (left) and known (right)
958 motifs, the top 2 activators and inhibitors are labeled. Center line: median; box limits:
959 upper and lower quartiles; whiskers: minimum and maximum values within the 1.5x
960 interquartile range; dots: values outside the 1.5x interquartile range. **c,** Principal
961 component analysis of FANTOM5 data for 142 human tissues using motif activities of
962 130 Codebook (left) and 471 known (right) representative motifs. X- and Y-axes: the
963 first two principal components. Marker color: tissue groups, marker shape: nervous
964 tissue development stage. TF: transcription factor, TPM: CAGE tags per million, CAGE:
965 cap analysis of gene expression, TSS: transcription start site


a
TIGD3
CENPB*
TIGD4
TIGD7
JRK
TIGD5
POGK
35.1/51
82.2/60
90.5/88
85/74
0.3
b
JRK motif score (relative to max)
0.0125
0.0075
0.0025
Average scaled read count
0.4
0.3
0.2
0.1
0.0
Strand
Plus
Minus
Experiments
ChIP-Seq
GHT-SELEX
Base position
0
400
800
Tigger15a
c
ZBED2 BED-zf
ZBED4 BED-zf
ZBED5 BED-zf
ZBED9/SCAND3 BED-zf
FAM200B BED-zf
GTF2IRD2 BED-zf
CGGBP1 BED-zf-like
88.6/77
74/65
80/71
30.1/55
0.3
d
PIF-Harbinger derived
MSANTD1 Myb
MSANTD4 Myb
MYPOP Myb
Mutator derivied
FLYWCH1 (5 FLYWCH-type zinc fingers)
e
5
4
3
1
2
ATCATCGAATGGAATCGAATGGAATTATCG

966 **Extended Data Figure 9. Transcription factors co-opted from DNA transposons. a,**
967 Motif logos of human TFs that are derived from the domestication of *Tigger* and *Pogo*
968 DNA transposons and have previously known DNA binding motifs. Phylogenetic tree
969 was derived from DBD sequence alignments produced with MAFFT L-INS-I[103], and
970 produced using the same method described in the **Extended Data Fig. 2e** legend. The
971 tree is rooted on POGK, which is derived from a more distantly related family of
972 *Pogo/Tigger*-like elements relative to the other proteins[104,105]. Sequence logos are
973 Codebook-derived, except for CENPB, from the liturature[106]. **b,** average per-base read
974 count over Tigger15a TOPs in the human genome, for JRK ChIP-seq (orange) and
975 GHT-SELEX (purple), with peak sequences aligned to the Tigger15a consensus
976 sequence. JRK PWM scores at each base of the Tigger15a consensus sequence are
977 shown in black (plus strand) and grey (minus strand). **c,** Sequence logos of Codebook
978 human TFs that are derived from the domestication of DBDs from *hAT* superfamily DNA
979 transposons. Phylogenetic tree was derived from DBD sequence alignments produced
980 with MAFFT L-INS-I[103], although for ZBED4 the third of its four BED-zf domains was
981 used. The tree was produced using the same method as described in the **Extended**
982 **Data Fig. 2e** legend. **d,** Sequence logos of TFs derived from domestication of *Mutator*
983 and *PIF-Harbinger* DNA transposons. **e**, Alphafold3 structure of the five FLYWCH-type
984 zinc-fingers of FLYWCH1, and a TOP sequence from chr10:42167256-42167276 with
985 5bp flanks. Zinc finger domains are numbered according to their order in FLYWCH1
986 from the N-terminus to the C-terminus. Long linkers and terminal extensions are hidden
987 for simplicity. DBD: DNA-binding domain, PWM: position weight matrix.

a

TF	Source	Motif ID	Motif logo
ARID1	Factorbook	ENCSR491EBY	
CTCF	CisBP	M03082_3.00	
KDM5B	Factorbook	ENCSR000AQA	
YY1	CisBP	M04611_3.00	
DRAP1	Factorbook	ENCSR765MKZ	
NFYA	CisBP	M10141_3.00	
GATAD2A	Factorbook	ENCSR925BFV	
CEBPA	CisBP	M08710_3.00	
GATAD2A	Factorbook	ENCSR925BFV	
FOXO1	CisBP	M03226_3.00	
MBD1	Factorbook	ENCSR396QWK	
REST	CisBP	M08305_3.00	

b

TF	Source	Motif ID	Motif logo
GATAD2A	Factorbook	ENCSR160QYK	
GATAD2A	Codebook	PWM000707	
ZBED2	Jaspar	MA1971.1	
ZBED2	Codebook	PWM100759	
ZNF623	Jaspar	UN0210.1	
ZNF623	Codebook	PWM004913	

c

TF	Source	Motif ID	Motif logo	Notes
ZNF592	Factorbook	ENCSR701AQS		No result in codebook, validated in multiple cells
ZNF577	HOCOMOCO	ZN577.H12CORE.0.P.B		Not tested in Codebook, specific and consistent for a C2H2-zf
GATAD2B	Factorbook	ENCSR547LKC		No result in codebook, consistent with orthologs

988 **Extended Data Figure 10. Examples of evaluation of external PWMs. a,** Cases in
989 which the external PWM matches the PWM of a well-studied TF that is a frequent
990 "contaminant" motif in ChIP-seq[107]. In each example, the top sequence logo represents
991 the external PWM, and the bottom sequence logo represents a highly similar CisBP
992 PWM for a "contaminant" motif. **b,** Cases in which the external PWM (top in each
993 example) is consistent with the Codebook PWM for the same TF (bottom in each
994 example). **c,** External PWM sequence logos that cannot be explained as known
995 contaminants or artifacts, some of which are supported by multiple lines of evidence
996 and thus appear accurate. PWM: position weight matrix

997 **SUPPLEMENTARY INFORMATION**

998 **Supplementary Table 1. List of tested proteins and assay successes.** Table lists
999 the Codebook proteins and positive control TFs that were analyzed in the Codebook
1000 studies and provides metadata and information on whether they showed sequence-
1001 specific DNA binding activities in different types of experiments, together with the ID of
1002 the representative PWM selected in this study, if any.

1003 **Supplementary Table 2. Inserts used in this study.** Table provides the amino acid
1004 sequence and type (full-length or DBD) for the 716 inserts used in the Codebook
1005 studies.

1006 **Supplementary Table 3. Plasmids used in this study.** Table lists the plasmid
1007 backbone and insert for each of the 1,387 plasmids used in the Codebook studies.

1008 **Supplementary Table 4. Experiments performed in this study.** Table lists the 4,805
1009 experiments performed on Codebook and control TFs, along with 20 additional GFP
1010 control experiments, as well as 43 experiments from Schmitges et al. 2016[10] and 25
1011 experiments from Isakova et al. 2017[9] that were reanalyzed. The experiment ID,
1012 experiment type, TF assayed, expert curation result, and plasmid ID are listed for each
1013 experiment. Each experiment is mapped to its ID in an accompanying manuscript[14], and
1014 9 additional experiments used only in an accompanying manuscript[14] are explicitly
1015 marked.

1016 **Supplementary Table 5. Representative PWMs.** Table displays logo representations
1017 for the PWMs that were manually selected as the representative for each of the TFs (i.e.
1018 the expert-curated motifs).

1019 **Supplementary Table 6. Re-assessment of unsuccessful proteins.** Table
1020 summarizes the re-assessment outcomes for the unsuccessful proteins. A detailed
1021 description can be found in **Supplementary Document 1.**

1022 **Supplementary Table 7. External ChIP-seq datasets**. Table lists external ChIP-seq
1023 datasets obtained from the GTRD database and ENCODE consortium for Codebook
1024 and control TFs as well as other putative TFs that were not tested in Codebook.

1025 **Supplementary Table 8. TF-wise allele-specific binding sites.** This table contains
1026 transcription factor allele-specific binding sites (ASBs) identified using Codebook data at
1027 5% FDR. Each entry represents an allele specific interaction between a TF and its DNA
1028 binding site in the vicinity of or directly encompassing the single-nucleotide variant, for
1029 which the allelic imbalance of read counts was observed between the reference and the
1030 alternative alleles.

1031 **Supplementary Table 9. Aggregated allele-specific binding sites.** This table
1032 contains the joint set of allele-specific binding sites (FDR 5%) aggregated across all
1033 Codebook data (all transcription factors and experiment types), eQTL associated
1034 phenotypes (EBI GWAS Catalog), and known ASB annotations of the respective single-

1035 nucleotide variants. Each entry corresponds to the allele-specific binding of one or more
1036 TFs. The aggregated statistics combine evidence from multiple TFs and experiments
1037 (ChIP-Seq and multiple cycles of GHT-SELEX).

1038 **Supplementary Table 10. Representative motifs for all human TFs.** Table lists
1039 representative motifs for Codebook TFs, TFs with previously known specificities, and
1040 motifs for 33 uncharacterized TFs curated from other publications (see "**Curation of**
1041 **new external motifs for previously uncharacterized TFs**"). Motif cluster membership
1042 is also indicated (**Extended Data Fig. 6**). The PWMs can be found in **Supplementary**
1043 **Data 1**.

1044 **Supplementary Table 11. MARA motif clusters and representative motifs.** Details
1045 for the 653 motif clusters generated for the motif activity response analysis using
1046 FANTOM5 data.

1047 **Supplementary Table 12. MARA results.** Table contains the MARA-predicted motif
1048 activity estimates across FANTOM5 samples for each motif cluster, their Pearson
1049 correlations with the expression of corresponding TFs in 583 samples used for the
1050 MARA analysis, ANOVA statistics, and significance estimates.

1051 **Supplementary Table 13. External PWMs**. Table lists PWM identifiers, manual
1052 curation and other metadata for external motifs available from the JASPAR,
1053 HOCOMOCO and Factorbook databases.

1054 **Supplementary Table 14. Updated census of human transcription factors and**
1055 **their motif coverage**. Table is modified from Lambert et al. 2018[1] to display an
1056 updated motif coverage of human TFs and indicates the IDs of the representative motifs
1057 (see **Supplementary Table 10**).

1058 **Supplementary Table 15. Motif derivation tools.** The table describes which methods
1059 were used for analyzing the main experiment types and summarizes the numbers of
1060 TFs and experiments that the methods were applied to, as well as the outcomes of motif
1061 analyses.

1062 **Supplementary Table 16. Common artifact motifs.** Table contains PWMs that
1063 represent artifact or false positive signals that were commonly observed in the
1064 Codebook experiments.

1065 **Supplementary Data 1. Representative motifs for all Human TFs.** The archive
1066 contains representative PWMs for all 1,421 TFs with a known binding specificity. Motif
1067 annotation is provided in **Supplementary Table 10**.

1068 **Supplementary Document 1. Case-by-case re-assessment of unsuccessful**
1069 **Codebook proteins.** This document contains detailed case-by-case manual
1070 assessments for the proteins that did not yield successful experiments. The proteins are
1071 organized by DNA binding domain composition and the result of the re-assessment.
1072 The objective of the re-assessment was to determine whether manual examination of

existing knowledge, and in some cases the Codebook data itself, suggest that they are more likely than not to be potential sequence specific TFs (i.e., false negatives in the scope of the Codebook project), or whether the available evidence instead supports them being true negatives..

Supplementary Discussion. Additional discussion of undercharacterized DNA-binding domains (particularly the SKI/SNO/DAC and C-Clamp domains), potential reasons for the failure of experiments for some proteins to yield motifs, and transposon-derived TFs.

Supplementary Methods. Expanded methods sections related to the allele-specific binding (ASB) analysis and Motif Activity Response Analysis (MARA).

1082 **REFERENCES**
1083

1084 1. Lambert, S.A. *et al.* The Human Transcription Factors. *Cell* **175**, 598–599 (2018).
1085 2. Wasserman, W.W. & Sandelin, A. Applied bioinformatics for the identification of
1086 regulatory elements. *Nat Rev Genet* **5**, 276–87 (2004).
1087 3. Srivastava, D. & Mahony, S. Sequence and chromatin determinants of transcription
1088 factor binding and the establishment of cell type-specific binding patterns. *Biochim*
1089 *Biophys Acta Gene Regul Mech* **1863**, 194443 (2020).
1090 4. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast
1091 genomes. *Genome Res* **15**, 1034–50 (2005).
1092 5. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome.
1093 *Nature* **489**, 57–74 (2012).
1094 6. Stormo, G.D. & Zhao, Y. Determining the specificity of protein-DNA interactions. *Nat Rev*
1095 *Genet* **11**, 751–60 (2010).
1096 7. Stormo, G.D. Consensus patterns in DNA. *Methods Enzymol* **183**, 211–21 (1990).
1097 8. Schneider, T.D. & Stephens, R.M. Sequence logos: a new way to display consensus
1098 sequences. *Nucleic Acids Res* **18**, 6097–100 (1990).
1099 9. Isakova, A. *et al.* SMiLE-seq identifies binding motifs of single and dimeric transcription
1100 factors. *Nat Methods* **14**, 316–322 (2017).
1101 10. Schmitges, F.W. *et al.* Multiparameter functional diversity of human C2H2 zinc finger
1102 proteins. *Genome Res* **26**, 1742–1752 (2016).
1103 11. Kuiper, M. *et al.* The gene regulation knowledge commons: the action area of GREEKC.
1104 *Biochim Biophys Acta Gene Regul Mech* **1865**, 194768 (2022).
1105 12. Jolma, A. *et al.* GHT-SELEX demonstrates unexpectedly high intrinsic sequence specificity
1106 and complex DNA binding of many human transcription factors. *bioRxiv*,
1107 2024.11.11.618478 (2024).
1108 13. Razavi, R. *et al.* Extensive binding of uncharacterized human transcription factors to
1109 genomic dark matter. *bioRxiv*, 2024.11.11.622123 (2024).
1110 14. Vorontsov, I.E. *et al.* Cross-platform motif discovery and benchmarking to explore
1111 binding specificities of poorly studied human transcription factors. *Commun Biol* **8**, 1545
1112 (2025).
1113 15. Gralak, A.J. *et al.* Identification of methylation-sensitive human transcription factors
1114 using meSMiLE-seq. *bioRxiv*, 2024.11.11.619598 (2024).
1115 16. Brayer, K.J., Kulshreshtha, S. & Segal, D.J. The protein-binding potential of C2H2 zinc
1116 finger domains. *Cell Biochem Biophys* **51**, 9–19 (2008).
1117 17. Bird, A.J., Gordon, M., Eide, D.J. & Winge, D.R. Repression of ADH1 and ADH3 during zinc
1118 deficiency by Zap1-induced intergenic RNA transcripts. *EMBO J* **25**, 5726–34 (2006).
1119 18. Font, J. & Mackay, J.P. Beyond DNA: zinc finger domains as RNA-binding modules.
1120 *Methods Mol Biol* **649**, 479–91 (2010).
1121 19. Zhou, J. *et al.* Attenuation of Forkhead signaling by the retinal determination factor
1122 DACH1. *Proc Natl Acad Sci U S A* **107**, 6864–9 (2010).
1123 20. Ikeda, K., Watanabe, Y., Ohto, H. & Kawakami, K. Molecular interaction and synergistic
1124 activation of a promoter by Six, Eya, and Dach proteins mediated through CREB binding
1125 protein. *Mol Cell Biol* **22**, 6759–66 (2002).

1126 21. Rass, M. *et al.* The Drosophila functional Smad suppressing element fuss, a homologue
1127 of the human Skor genes, retains pro-oncogenic properties of the Ski/Sno family. *PLoS*
1128 *One* **17**, e0262360 (2022).
1129 22. Wolfe, S.A., Nekludova, L. & Pabo, C.O. DNA recognition by Cys2His2 zinc finger
1130 proteins. *Annu Rev Biophys Biomol Struct* **29**, 183–212 (2000).
1131 23. Consortium, E.P. *et al.* Perspectives on ENCODE. *Nature* **583**, 693–698 (2020).
1132 24. Partridge, E.C. *et al.* Occupancy maps of 208 chromatin-associated proteins in one
1133 human cell type. *Nature* **583**, 720–728 (2020).
1134 25. Lai, W.K.M. *et al.* A ChIP-exo screen of 887 Protein Capture Reagents Program
1135 transcription factor antibodies in human cells. *Genome Res* **31**, 1663–1679 (2021).
1136 26. Deplancke, B., Alpern, D. & Gardeux, V. The Genetics of Transcription Factor DNA
1137 Binding Variation. *Cell* **166**, 538–554 (2016).
1138 27. Buyan, A. *et al.* Statistical framework for calling allelic imbalance in high-throughput
1139 sequencing data. *Nat Commun* **16**, 1739 (2025).
1140 28. Abramov, S. *et al.* Landscape of allele-specific transcription factor binding in the human
1141 genome. *Nat Commun* **12**, 2751 (2021).
1142 29. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association
1143 studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005–
1144 D1012 (2019).
1145 30. Consortium, G.T. The GTEx Consortium atlas of genetic regulatory effects across human
1146 tissues. *Science* **369**, 1318–1330 (2020).
1147 31. Lambert, S.A., Albu, M., Hughes, T.R. & Najafabadi, H.S. Motif comparison based on
1148 similarity of binding affinity profiles. *Bioinformatics* **32**, 3504–3506 (2016).
1149 32. Emerson, R.O. & Thomas, J.H. Adaptive evolution in zinc finger transcription factors.
1150 *PLoS Genet* **5**, e1000325 (2009).
1151 33. Illingworth, R.S. *et al.* Orphan CpG islands identify numerous conserved promoters in
1152 the mammalian genome. *PLoS Genet* **6**, e1001134 (2010).
1153 34. Zhao, Y. & Stormo, G.D. Quantitative analysis demonstrates most transcription factors
1154 require only simple models of specificity. *Nat Biotechnol* **29**, 480–3 (2011).
1155 35. Ruan, S., Swamidass, S.J. & Stormo, G.D. BEESEM: estimation of binding energy models
1156 using HT-SELEX data. *Bioinformatics* **33**, 2288–2295 (2017).
1157 36. Weirauch, M.T. *et al.* Evaluation of methods for modeling transcription factor sequence
1158 specificity. *Nat Biotechnol* **31**, 126–34 (2013).
1159 37. Kotelnikova, E.A., Makeev, V.J. & Gelfand, M.S. Evolution of transcription factor DNA
1160 binding sites. *Gene* **347**, 255–63 (2005).
1161 38. Vorontsov, I.E. *et al.* HOCOMOCO in 2024: a rebuild of the curated collection of binding
1162 models for human and mouse transcription factors. *Nucleic Acids Res* **52**, D154–D163
1163 (2024).
1164 39. Moses, A.M., Chiang, D.Y., Pollard, D.A., Iyer, V.N. & Eisen, M.B. MONKEY: identifying
1165 conserved transcription-factor binding sites in multiple alignments using a binding site-
1166 specific evolutionary model. *Genome Biol* **5**, R98 (2004).
1167 40. Deaton, A.M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev* **25**,
1168 1010–22 (2011).

1169 41. Blackledge, N.P. & Klose, R. CpG island chromatin: a platform for gene regulation.
1170 *Epigenetics* **6**, 147–52 (2011).
1171 42. Cohen, N.M., Kenigsberg, E. & Tanay, A. Primate CpG islands are maintained by
1172 heterogeneous evolutionary regimes involving minimal selection. *Cell* **145**, 773–86
1173 (2011).
1174 43. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target
1175 genes in GeneCards. *Database (Oxford)* **2017**(2017).
1176 44. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–66
1177 (2015).
1178 45. Consortium, F. *et al.* The transcriptional network that controls growth arrest and
1179 differentiation in a human myeloid leukemia cell line. *Nat Genet* **41**, 553–62 (2009).
1180 46. Balwierz, P.J. *et al.* ISMARA: automated modeling of genomic signals as a democracy of
1181 regulatory motifs. *Genome Res* **24**, 869–84 (2014).
1182 47. Consortium, F. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–70
1183 (2014).
1184 48. Mance, L. *et al.* Dynamic BTB-domain filaments promote clustering of ZBTB proteins.
1185 *Mol Cell* **84**, 2490–2510 e9 (2024).
1186 49. Arabzade, A. *et al.* Synthetic ZFTA fusions pinpoint disordered protein domain
1187 acquisition as a mechanism of brain tumorigenesis. *Nat Cell Biol* **27**, 1496–1509 (2025).
1188 50. Donaldson, N.S. *et al.* Kaiso regulates Znf131-mediated transcriptional activation. *Exp*
1189 *Cell Res* **316**, 1692–705 (2010).
1190 51. Vuong-Brender, T.T., Flynn, S., Vallis, Y., Sonmez, S.E. & de Bono, M. Neuronal
1191 calmodulin levels are controlled by CAMTA transcription factors. *Elife* **10**(2021).
1192 52. Kaushik, K. *et al.* Requirements for the neurodevelopmental disorder-associated gene
1193 ZNF292 in human cortical interneuron development and function. *Cell Rep* **44**, 115597
1194 (2025).
1195 53. Knock, E. *et al.* The methyl binding domain 3/nucleosome remodelling and deacetylase
1196 complex regulates neural cell fate determination and terminal differentiation in the
1197 cerebral cortex. *Neural Dev* **10**, 13 (2015).
1198 54. Mondal, B. *et al.* The histone deacetylase complex MiDAC regulates a
1199 neurodevelopmental gene expression program to control neurite outgrowth. *Elife*
1200 **9**(2020).
1201 55. Emery, B. *et al.* Myelin gene regulatory factor is a critical transcriptional regulator
1202 required for CNS myelination. *Cell* **138**, 172–85 (2009).
1203 56. Sander, E.E. & Grummt, I. Oligomerization of the transcription termination factor TTF-I:
1204 implications for the structural organization of ribosomal transcription units. *Nucleic*
1205 *Acids Res* **25**, 1142–7 (1997).
1206 57. Wang, M. *et al.* TAF1A and ZBTB41 serve as novel key genes in cervical cancer identified
1207 by integrated approaches. *Cancer Gene Ther* **28**, 1298–1311 (2021).
1208 58. Koh, D.I. *et al.* A novel POK family transcription factor, ZBTB5, represses transcription of
1209 p21CIP1 gene. *J Biol Chem* **284**, 19856–66 (2009).
1210 59. Najafabadi, H.S. *et al.* C2H2 zinc finger proteins greatly expand the human regulatory
1211 lexicon. *Nat Biotechnol* (2015).

1212 60. Barazandeh, M., Lambert, S.A., Albu, M. & Hughes, T.R. Comparison of ChIP-Seq Data
1213 and a Reference Motif Set for Human KRAB C2H2 Zinc Finger Proteins. *G3 (Bethesda)* **8**,
1214 219–229 (2018).
1215 61. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nat Rev*
1216 *Genet* **9**, 397–405 (2008).
1217 62. Singh, U. & Westermark, B. CGGBP1--an indispensable protein with ubiquitous
1218 cytoprotective functions. *Ups J Med Sci* **120**, 219–32 (2015).
1219 63. Hayward, A., Ghazal, A., Andersson, G., Andersson, L. & Jern, P. ZBED evolution:
1220 repeated utilization of DNA transposons as regulators of diverse host functions. *PLoS*
1221 *One* **8**, e59940 (2013).
1222 64. Smit, A.F. & Riggs, A.D. Tiggers and DNA transposon fossils in the human genome. *Proc*
1223 *Natl Acad Sci U S A* **93**, 1443–8 (1996).
1224 65. Etchegaray, E., Baas, D., Naville, M., Haftek-Terreau, Z. & Volff, J.N. The
1225 neurodevelopmental gene MSANTD2 belongs to a gene family formed by recurrent
1226 molecular domestication of Harbinger transposons at the base of vertebrates. *Mol Biol*
1227 *Evol* **39**(2022).
1228 66. Marquez, C.P. & Pritham, E.J. Phantom, a new subclass of Mutator DNA transposons
1229 found in insect viruses and widely distributed in animals. *Genetics* **185**, 1507–17 (2010).
1230 67. Harrison, P.W. *et al.* Ensembl 2024. *Nucleic Acids Res* **52**, D891–D899 (2024).
1231 68. Toth, M., Grimsby, J., Buzsaki, G. & Donovan, G.P. Epileptic seizures caused by
1232 inactivation of a novel gene, jerky, related to centromere binding protein-B in transgenic
1233 mice. *Nat Genet* **11**, 71–5 (1995).
1234 69. Pace, J.K., 2nd & Feschotte, C. The evolutionary history of human DNA transposons:
1235 evidence for intense activity in the primate lineage. *Genome Res* **17**, 422–32 (2007).
1236 70. Cordaux, R., Udit, S., Batzer, M.A. & Feschotte, C. Birth of a chimeric primate gene by
1237 capture of the transposase gene from a mobile element. *Proc Natl Acad Sci U S A* **103**,
1238 8101–6 (2006).
1239 71. Stielow, B. *et al.* The SAM domain-containing protein 1 (SAMD1) acts as a repressive
1240 chromatin regulator at unmethylated CpG islands. *Sci Adv* **7**(2021).
1241 72. Li, H. *et al.* Polycomb-like proteins link the PRC2 complex to CpG islands. *Nature* **549**,
1242 287–291 (2017).
1243 73. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with
1244 AlphaFold 3. *Nature* (2024).
1245 74. Ambrosini, G. *et al.* Insights gained from a comprehensive all-against-all transcription
1246 factor binding motif benchmarking study. *Genome Biol* **21**, 114 (2020).
1247 75. Weirauch, M.T. *et al.* Determination and inference of eukaryotic transcription factor
1248 sequence specificity. *Cell* **158**, 1431–43 (2014).
1249 76. Jolma, A. *et al.* GHT-SELEX demonstrates unexpectedly high intrinsic sequence specificity
1250 and complex DNA binding of many human transcription factors. *bioRxiv*,
1251 2024.11.11.618478 (2024).
1252 77. Razavi, R. *et al.* Extensive binding of uncharacterized human transcription factors to
1253 genomic dark matter. *bioRxiv*, 2024.11.11.622123 (2024).
1254 78. Narasimhan, K. *et al.* Mapping and analysis of Caenorhabditis elegans transcription
1255 factor sequence specificities. *Elife* **4**(2015).

1256 79. Vorontsov, I.E., Kulakovskiy, I.V. & Makeev, V.J. Jaccard index based similarity measure
1257 to compare transcription factor binding site models. *Algorithms Mol Biol* **8**, 23 (2013).
1258 80. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic
1259 features. *Bioinformatics* **26**, 841–2 (2010).
1260 81. Satopaa, V., Albrecht, J., Irwin, D. & Raghavan, B. Finding a" kneedle" in a haystack:
1261 Detecting knee points in system behavior. in *2011 31st international conference on*
1262 *distributed computing systems workshops* 166–171 (IEEE, 2011).
1263 82. Korhonen, J., Martinmaki, P., Pizzi, C., Rastas, P. & Ukkonen, E. MOODS: fast search for
1264 position weight matrix matches in DNA sequences. *Bioinformatics* **25**, 3181–2 (2009).
1265 83. Weirauch, M.T. *et al.* Determination and inference of eukaryotic transcription factor
1266 sequence specificity. *Cell* **158**, 1431–1443 (2014).
1267 84. Lovmar, L., Ahlford, A., Jonsson, M. & Syvanen, A.C. Silhouette scores for assessment of
1268 SNP genotype clusters. *BMC Genomics* **6**, 35 (2005).
1269 85. Agarwal, V. *et al.* Massively parallel characterization of transcriptional regulatory
1270 elements. *Nature* **639**, 411–420 (2025).
1271 86. Kolmykov, S. *et al.* GTRD: an integrated view of transcription regulation. *Nucleic Acids*
1272 *Res* **49**, D104–D111 (2021).
1273 87. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data
1274 portal. *Nucleic Acids Res* **48**, D882–D889 (2020).
1275 88. Rauluseviciute, I. *et al.* JASPAR 2024: 20th anniversary of the open-access database of
1276 transcription factor binding profiles. *Nucleic Acids Res* **52**, D174–D182 (2024).
1277 89. Pratt, H.E. *et al.* Factorbook: an updated catalog of transcription factor motifs and
1278 candidate regulatory motif sites. *Nucleic Acids Res* **50**, D141–D149 (2022).
1279 90. Orenstein, Y. & Shamir, R. A comparative analysis of transcription factor binding models
1280 learned from PBM, HT-SELEX and ChIP data. *Nucleic Acids Res* **42**, e63 (2014).
1281 91. Armstrong, J. *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-
1282 genome era. *Nature* **587**, 246–251 (2020).
1283 92. Amemiya, H.M., Kundaje, A. & Boyle, A.P. The ENCODE Blacklist: Identification of
1284 Problematic Regions of the Genome. *Sci Rep* **9**, 9354 (2019).
1285 93. Frankish, A. *et al.* GENCODE: reference annotation for the human and mouse genomes
1286 in 2023. *Nucleic Acids Res* **51**, D942–D949 (2023).
1287 94. Nassar, L.R. *et al.* The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res*
1288 **51**, D1188–D1195 (2023).
1289 95. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping
1290 and population genetical parameter estimation from sequencing data. *Bioinformatics*
1291 **27**, 2987–93 (2011).
1292 96. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J.K. WASP: allele-specific software
1293 for robust molecular quantitative trait locus discovery. *Nat Methods* **12**, 1061–3 (2015).
1294 97. Selected abstracts of Bioinformatics: from Algorithms to Applications 2021 Conference.
1295 *BMC Bioinformatics* **22**, 591 (2021).
1296 98. Vorontsov, I.E., Kulakovskiy, I.V., Khimulya, G., Nikolaeva, D.D. & Makeev, V.J.
1297 PERFECTOS-APE-predicting regulatory functional effect of SNPs by approximate P-value
1298 estimation. in *International Conference on Bioinformatics Models, Methods and*
1299 *Algorithms* Vol. 2 102–108 (SCITEPRESS, 2015).

1300 99. Lambert, S.A. *et al.* Similarity regression predicts evolution of transcription factor
1301 sequence specificity. *Nat Genet* **51**, 981–989 (2019).
1302 100. De Maio, N., Ly-Trong, N., Martin, S., Minh, B.Q. & Goldman, N. Assessing phylogenetic
1303 confidence at pandemic scales. *Nature* **647**, 472–478 (2025).
1304 101. Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human
1305 transcription factors. *Science* **356**(2017).
1306 102. Liu, Z., Yang, X., Tan, F., Cullion, K. & Thiele, C.J. Molecular cloning and characterization
1307 of human Castor, a novel human gene upregulated during cell differentiation. *Biochem*
1308 *Biophys Res Commun* **344**, 834–44 (2006).
1309 103. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of
1310 multiple sequence alignment. *Nucleic Acids Res* **33**, 511–8 (2005).
1311 104. Dupeyron, M., Baril, T., Bass, C. & Hayward, A. Phylogenetic analysis of the Tc1/mariner
1312 superfamily reveals the unexplored diversity of pogo-like elements. *Mob DNA* **11**, 21
1313 (2020).
1314 105. Gao, B. *et al.* Evolution of pogo, a separate superfamily of IS630-Tc1-mariner
1315 transposons, revealing recurrent domestication events in vertebrates. *Mob DNA* **11**, 25
1316 (2020).
1317 106. Jolma, A. *et al.* DNA-Binding Specificities of Human Transcription Factors. *Cell* **152**, 327–
1318 39 (2013).
1319 107. Worsley Hunt, R. & Wasserman, W.W. Non-targeted transcription factors motifs are a
1320 systemic component of ChIP-seq datasets. *Genome Biol* **15**, 412 (2014).
1321