GHT-SELEX demonstrates unexpectedly high intrinsic sequence specificity and complex DNA binding of many human transcription factors

Arttu Jolma[1,*], Aldo Hernandez-Corchado[2,3*], Ally W.H. Yang[1,*], Ali Fathi[4,*], Kaitlin U. Laverty[1,5*], Alexander Brechalov[1], Rozita Razavi[1], Mihai Albu[1], Hong Zheng[1], The Codebook Consortium, Ivan V. Kulakovskiy[6,7,8], Hamed S. Najafabadi[2,3**], and Timothy R. Hughes[1,4,**]

[1]Donnelly Centre, University of Toronto, Toronto, ON M5S 3E1, Canada
[2]Department of Human Genetics, McGill University, Montréal, QC H3A 0C7, Canada
[3]Victor P. Dahdaleh Institute of Genomic Medicine, Montréal, QC H3A 0G1, Canada
[4]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada
[5]Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[6]Institute of Protein Research, Russian Academy of Sciences, 142290, Pushchino, Russia
[7]Institute of Biochemistry and Genetics, Ufa Federal Research Centre of Russian Academy of Sciences, 450054, Ufa, Russia
[8]Vavilov Institute of General Genetics, Russian Academy of Sciences, 119991, Moscow, Russia

*These authors contributed equally.
**email: hamed.najafabadi@mcgill.ca and t.hughes@utoronto.ca.

The Codebook consortium

Philipp Bucher[9], Bart Deplancke[9,10], Oriol Fornes[11], Jan Grau[12], Ivo Grosse[12], Timothy R. Hughes[1,4], Arttu Jolma[1], Fedor A. Kolpakov[13,14], Ivan V. Kulakovskiy[6,7,8], Vsevolod J. Makeev[8,15,16], Mihai Albu[1], Marjan Barazandeh[1], Alexander Brechalov[1], Zhenfeng Deng[1], Ali Fathi[4], Chun Hu[1], Samuel A. Lambert[1], Kaitlin U. Laverty[1,5], 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[17,18], Antoni J. Gralak[9,10], Sachi Inukai[19], Judith F. Kribelbauer-Swietek[9,10], Marie-Luise Plescher[12], Semyon Kolmykov[13], Ivan Yevshin[20], Nikita Gryzunov[6], Ivan Kozin[6,21], Mikhail Nikonov[21], Vladimir Nozdrin[21], Arsenii Zinkevich[21], Katerina Faltejskova[22,23], Pavel Kravchenko[24], Sergey Abramov[8,25], Alexandr Boytsov[8,25], Vasilii Kamenets[8,15], Dmitry Penzar[8], Anton Vlasov[26], Ilya E. Vorontsov[8], Aldo Hernandez-Corchado[2,3], Hamed S. Najafabadi[2,3], Quaid Morris[5], Xiaoting Chen[27], Matthew T. Weirauch[27]

[9] Swiss Institute of Bioinformatics, 1015, Lausanne, Switzerland
[10] Laboratory of Systems Biology and Genetics, Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne, 1015, Lausanne, Switzerland
[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] Department of Computational Biology, Sirius University of Science and Technology, 354340, Sirius, Krasnodar region, Russia
[14] Bioinformatics Laboratory, Federal Research Center for Information and Computational Technologies, 630090, Novosibirsk, Russia
[15] Moscow Center for Advanced Studies, 123592, Moscow, Russia
[16] Current address: Cancer Research UK National Biomarker Centre, University of Manchester, M20 4BX, Manchester, UK
[17] Bioinformatics Competence Center, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
[18] Bioinformatics Competence Center, Université de Lausanne, 1015 Lausanne, Switzerland
[19] Chugai Pharmaceutical Co., Ltd, Tokyo, 103-8324, Japan
[20] Biosoft.Ru LLC, 630058, Novosibirsk, Russia
[21] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, 119991, Moscow, Russia
[22] Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, 160 00 Praha 6, Czech Republic
[23] Computer Science Institute, Faculty of Mathematics and Physics, Charles University, 118 00 Praha 1, Czech Republic
[24] Max Planck Institute of Biochemistry, 82152, Planegg, Germany
[25] Altius Institute for Biomedical Sciences, 98121, Seattle, WA, USA
[26] TUD Dresden University of Technology, Center for Molecular and Cellular Bioengineering (CMCB), Biotechnologisches Zentrum (BIOTEC), Tatzberg 47/49, 01307 Dresden, Germany
[27] Cincinnati Children's Hospital, Cincinnati, OH 45229, USA

SUMMARY

Precise identification of transcription factor (TF) binding sites is a long standing challenge in human regulatory genomics: TF binding motifs are short and degenerate, while the genome is large. Motif scans, therefore, often produce excessive binding site predictions. By surveying 179 TFs across 25 families using >1,500 cyclic *in vitro* selection experiments with fragmented, naked, and unmodified genomic DNA – a method we term GHT-SELEX (Genomic HT-SELEX) – we find that many human TFs possess much higher sequence specificity than anticipated. Moreover, genomic binding regions from GHT-SELEX are often surprisingly similar to those obtained *in vivo* (i.e., ChIP-seq peaks). Contrary to conventional wisdom, we find that high specificity can also be obtained from motif scans, but performance is highly dependent on the derivation and use of the motifs, including accounting for multiple local matches. We also observe alternative engagement of multiple DNA-binding domains within the same protein: long C2H2 zinc finger proteins often utilize modular DNA recognition, engaging different subsets of their DNA-binding domain (DBD) arrays to recognize multiple types of distinct target sites, frequently evolving via internal duplication and divergence of one or more DBDs. Thus, it is common for TFs to possess sufficient intrinsic specificity to delineate a large fraction of *in vivo* genomic targets, independently of other cellular factors.

Keywords: DNA binding specificity, Transcription factor, TF, Transcription factor binding site, Position weight matrix, PWM, ChIP-Seq, HT-SELEX, GHT-SELEX, SELEX, Modular binding, C2H2 zinc finger, C2H2, RCADEEM, MAGIX, Codebook, Gene regulation

INTRODUCTION

The DNA-binding sequence preference of a Transcription Factor (TF) is typically referred to as a motif, and is most commonly modeled as a position weight matrix (PWM), which describes the relative preference of the TF for each base in the binding site[1]. In humans, TF binding motifs are generally short and flexible; PWMs are typically 8-14 bases long[2-4], and multiple bases can be tolerated at many positions[5,6]. Thus, a typical TF PWM scan with default parameters yields over a million potential binding sites in the 3-billion-base human genome, often with multiple high-scoring matches per gene. Very few of the potential target sites are utilized in cells[7], however, and the actual number of bound sites, as measured by ChIP-seq[8-10] or other assays[11] is typically much lower than the number of motif matches.

This deficit in specificity has been resolved conceptually by the widespread cooperative binding and synergy among TFs[5,6,12,13], and evidence that the chromatin landscape generally dominates TF binding site selection, such that TF motif matches only determine binding within permissible regions[14]. In the latter model, only a special class of “pioneer” TFs can access target sequences to control the local chromatin. Indeed, some TFs have been shown to have high inherent specificity: for example, CTCF binds the majority of its strongest motif matches in the genome[15], and repositions the surrounding nucleosomes[16]. PRDM9, which controls recombination hotspots, has been reported to independently specify roughly half of its binding sites in the genome[17].

Another possible explanation for the generally low apparent specificity of TF motifs, however, is that PWMs are inaccurate, or are used inappropriately, or that the PWM model is fundamentally flawed[18]. PWMs are often derived from a non-comprehensive set of bound vs. unbound sequences, and there is ongoing controversy regarding the best methods for derivation, underlying representation, and scanning of TF motifs[1,19], as well as the impact of DNA shape[20], dependencies among base positions[18,21], multimeric binding[22,23], and lower-affinity binding sites[24]. Proteins in the CXXC zinc finger (CXXC-zf) class, for example, bind primarily to clusters of unmethylated CG dinucleotides within CpG islands[25], achieving high specificity despite an extremely simple monomeric binding site.

Many human TFs still lack binding motifs, and prominent among them are hundreds of C2H2 zinc finger (C2H2-zf) proteins[26]. These proteins recognize DNA sequences that approximate a concatenation of the three or four base specificities of their sequential constituent C2H2-zf domains[27,28]. Different C2H2-zf proteins can bind very different motifs due to both the malleability of the individual C2H2-zf domains and rearrangement of the individual C2H2-zf domains[29]. An enigmatic feature of the C2H2-zf proteins is their theoretical capacity to recognize very long sequences: the median number of C2H2-zf domains in human TFs is 11, which could contact up to 33 DNA bases, much more than would be needed to specifically recognize even a single target site in the genome, on average. Indeed, C2H2-zf proteins often use only a subset of their DBDs to contact DNA, and whether and how frequently human C2H2-zf proteins utilize different segments of the C2H2-zf domain array to bind different sequences has also been a

159 long-standing question. In a well-studied example, CTCF binding sites appear to reflect
160 a constitutive "core", bound by fingers 4-7 of the 11 C2H2-zf domain array, flanked by
161 sequences that are bound by alternative usage of upstream and/or downstream C2H2-
162 zf domains[30,31]. In another example, mouse Znf335 was found to bind to two distinct
163 motifs utilizing two separated C2H2-zf arrays[32]. Analysis of the DNA-binding of C2H2-zf
164 proteins to the genome is also complicated by the fact that they often bind repeat
165 elements such as endogenous retroelements[33], and thus the target site similarity is
166 derived both from DNA recognition and the shared ancestry of the binding sites. The
167 limited resolution of ChIP-seq (>100bp) presents a related hindrance. These
168 confounding factors, however, can be ameliorated by incorporating information about
169 the bases that are likely preferred at each position of the binding site, as predicted by a
170 C2H2-zf "recognition code" that relates the C2H2-zf amino acid sequences to their
171 binding preferences. These machine learning-based predictions can assist in identifying
172 the most plausible protein-DNA interactions in such cases, as our earlier work
173 demonstrated[34].
174
175 Methods for the study of TF sequence specificity *in vitro* are important because they
176 remove any confounding impact of cellular factors beyond the TF of interest. Some *in*
177 *vitro* approaches survey binding to randomly-generated sequences, typified by SELEX
178 (Systematic Evolution of Ligands by EXponential enrichment)[35], which uses multiple
179 cycles of affinity capture interleaved with PCR amplification of the selected DNA
180 ligands. HT-SELEX (High-throughput SELEX), in which the protein-bound DNA is
181 immobilized in microwell plates and multiplexed Illumina sequencing is performed on all
182 selection cycles, has been used to characterize DNA binding of hundreds of TFs in
183 human and other species[36,37]. SELEX has the advantage that it has a more uniform k-
184 mer distribution and potentially higher diversity than genomic DNA, but has a
185 disadvantage as increasingly long sequences are present in a decreasing proportion of
186 pool sequences. This issue is of particular concern for C2H2-zfs, due to the potential for
187 very large binding sites. Indeed, long C2H2-zf proteins have a lower success rate in HT-
188 SELEX than those with few C2H2-zf domains[36,38].
189
190 An alternative to randomized DNA is to use fragmented genomic DNA. SELEX has
191 previously been performed with *E. coli* genomic DNA with either a sequencing[39] or
192 microarray[40] readout. More recent innovations, including CAP-seq[41], Affinity-seq[17], and
193 DAP-seq[42], have employed only a single round of selection, allowing direct comparison
194 (and similar peak calling) to ChIP-seq. These approaches have revealed remarkably
195 high specificity for some TFs. They are not cost-effective when applied to the human
196 genome, however, due to its large size. To our knowledge, only two human proteins
197 (PRDM9[17] and Cfp1[41] have been studied by these methods.
198
199 Here, we describe GHT-SELEX (Genomic DNA HT-SELEX), a novel implementation of
200 the HT-SELEX[37] method that is performed with fragmented human genomic DNA, and
201 uses an associated new statistical analysis method, MAGIX (Model-based Analysis of
202 Genomic Intervals with eXponential enrichment). GHT-SELEX combines the
203 advantages of HT-SELEX (the use of barcoding, magnetic affinity beads and laboratory
204 automation makes it possible to run GHT-SELEX in parallel with hundreds of samples

over numerous cycles) with the strengths of genomic DNA selection protocols[17,42] (representation of long genomic sites such as repeat sequences, and determination of individual genomic binding sites). We developed GHT-SELEX in the context of the Codebook consortium project[43], which was aimed primarily at analysis of uncharacterized putative TFs, and provides comparison data from several other platforms for the same set of TFs (HT-SELEX, ChIP-seq, Protein Binding Microarrays[44], and SMiLE-seq[45]). We successfully applied GHT-SELEX to 139 uncharacterized TFs and 40 control TFs. For dozens of TFs, including some that are considered well-characterized, GHT-SELEX peaks correspond closely to *in vivo* binding (measured by ChIP-seq). We also derived motif models that predict genomic binding with similar accuracy in most cases. GHT-SELEX is particularly effective for C2H2-zf proteins and shows that they often use alternative subsets of their C2H2-zf domains to engage with different genomic target sites. We explore both explanations and ramifications of these observations.

RESULTS

Development and testing of GHT-SELEX

We developed GHT-SELEX to run in parallel with HT-SELEX (**Figure 1a,** see **Table S1** for oligomer designs and **Table S2** for experimental parameters) in the context of the Codebook consortium[43], a systematic effort to obtain DNA-binding motifs for 331 uncharacterized putative TFs (defined as proteins containing DBDs or having prior literature evidence but lacking known motifs[46]), as well as 61 control TFs with established motifs (**Table S3**). The Codebook clone set included both full-length proteins and DBDs, and three types of expression vectors, to enable production by three different expression systems (N-terminal eGFP fusions in both wheat germ extract and HEK293 cells, and N-terminal GST fusions in *E. coli* extracts; see **Methods**). The GHT-SELEX DNA pool used in this study was produced by nonspecific enzymatic fragmentation (Fragmentase) of HEK293 DNA to fragments with a median length of ~64 bp. We used HEK293 DNA for compatibility with ChIP-seq data generated simultaneously (see accompanying manuscript[47]), and the length of the DNA was chosen to mimic standard HT-SELEX procedures and provide relatively high resolution.

We initially tested GHT-SELEX on the Codebook control proteins. Thirty of the control TFs represented a sampling of well-studied TFs with different classes of DBDs, most of which were previously analyzed using the independent *in vitro* SMiLE-seq platform[45]. An additional 31 controls were C2H2-zf proteins for which published ChIP-seq data yielded motifs[48]. Individual mapped reads typically accumulated at sites in which all reads overlap with what appears to be a motif match (**Figure 1b**). Moreover, the GHT-SELEX data typically had a strong resemblance to ChIP-seq data, forming strong peaks found sparsely across the genome, consistent with results from CAP-seq[41], Affinity-seq[17], and DAP-seq[42]. **Figure 1c** shows raw read density for four control TFs, comparing GHT-SELEX to ChIP-seq, and to target site predictions based on existing and newly derived (see below) PWM models for the TFs. GHT-SELEX experiments showed generally good reproducibility between replicates (examples in **Figure 2a**).

Peak calling from the GHT-SELEX data with conventional algorithms is confounded by the fact that different peaks have very different enrichment ratios across the cycles. This is presumably due to varying affinity of the TF for different sites, the overall increase in motif occurrences in the pool over the successive cycles, and simultaneous reduction in pool complexity, with the strongest binding sites dominating later cycles (**Extended Data Fig. 1a**). Consequently, enrichment information is distributed across the fragments from multiple selection cycles, with weaker peaks first appearing and disappearing, and the strongest peaks dominating in the later cycles **Extended Data Fig. 1b**). To adapt to these phenomena, we developed an analytical framework "MAGIX" (Model-based Analysis of Genomic Intervals with eXponential enrichment) that capitalizes on the added information gained from multiple SELEX cycles (**Extended Data Fig. 1b**; see **Methods** for details). The approach relies on a statistical method that explicitly models the exponential growth of TF-bound genomic regions over the SELEX cycles, which leads to a progressively higher proportion of TF-bound fragments relative to genomic background. The fragment abundances, in turn, are modeled as latent variables that determine the number of observed reads through a Poisson process. This hierarchical Bayesian model enables the integration of information across different selection cycles, experiments, and batches to calculate an estimated enrichment coefficient (MAGIX score) and associated FDR (**Extended Data Fig. 1b)**.

Among the 61 control proteins, we deemed 40 as successful in at least one GHT-SELEX experiment, based on enrichment of the expected motif at the peak centre (**Figure 2b**) (see **Methods** and accompanying manuscripts[43,49] for details of how success was determined, and how PWMs were derived). The control proteins that were not successful in GHT-SELEX were mainly (15/21) from among the C2H2-zf proteins for which published ChIP-seq data yielded motifs[48]. These experiments were performed early in the study, prior to technical innovations that dramatically improved success rates with this class of proteins. The success rate for proteins for which previously-reported SMiLE-seq data yielded motifs may be a more realistic metric to gauge the success of GHT-SELEX in its current form (24/30, or 80%).

Among the 40 control TFs that were successful, 23 were represented by more than one successful experiment **(Table S4),** allowing us to assess the reproducibility of the combined GHT-SELEX/MAGIX pipeline. **Figure 2a** shows the enrichment values over genomic bins for replicates of FLI1, VDR, YY1, and CTCF. Similar plots for all successful replicates are given in **Supplementary Document S1**, and Pearson correlations for replicate pairs are in **Table S5.** We conclude that enrichment values across the genome are reproducible. MAGIX readily accommodates multiple experiments into a single analysis, and all subsequent analyses in this study utilize merged experimental data for each distinct protein.

Analysis of the data for the 40 successful controls by MAGIX resulted in between 926 and 82,045 peaks (median 9,927) with enrichment coefficient FDR less than 5% (see **Methods**). The number of strong PWM hits, on average, declines rapidly at ~50 bp from

peak centers, consistent with the utilized DNA fragment size (**Figure 2b)**; similar plots for all TFs analyzed are shown in **Supplementary Document S2**. In addition, higher PWM scores (which would, in theory, predict higher relative affinity) are clearly associated with a higher GHT-SELEX enrichment coefficient (see below).

Application of GHT-SELEX to the full Codebook TF set

We next performed GHT-SELEX and, in parallel, HT-SELEX using random 40N ligands (**Table S1**) to assess DNA sequence specificity of 331 poorly characterized putative human TFs, as part of the Codebook project. We analyzed individual TFs with up to three types of constructs, and up to three protein expression strategies (two types of *in vitro* transcription–translation reactions, and expression in HEK293 cells, see **Methods**). We modulated several experimental variables over the course of the study (the protein production method, the number of washes, and digestion of single-stranded DNA with mung bean nuclease between cycles)in an effort to improve success rates. The protein production method appears to be most critical, particularly for C2H2-zf domain arrays (56%, 17% and 10% success rates using wheat germ extract, mammalian cell and *E. coli* extract-based expression systems, respectively; see **Methods, Table S2, and Extended data Fig. S2**).

For each TF, the constructs contained the full sequence of a representative isoform, or either all or a subset of its predicted DBDs. In total, we analyzed 1,315 constructs encompassing the 61 control TFs and 331 of the 332 putative TFs in the Codebook set of poorly characterized proteins. With these constructs, we performed 1,534 GHT-SELEX and 1,578 HT-SELEX experiments (see **Methods** and **Table S3**). The number of experiments per putative TF varied, due to the heterogeneity of the TFs and consequently heterogeneity in the clone collection; e.g., small proteins were represented with only a single insert, whereas larger, multi-DBD proteins were represented with both full-length and partial constructs. The vast majority (371/392, or 95%) of the proteins were tested two or more times in GHT-SELEX experiments, with a median of 4 experiments for each.

Importantly, 27% (420/1534) of the GHT-SELEX experiments were strict replicates (i.e., same construct, same expression system), covering a total of 198 potential TFs in 204 experiments (see **Methods**, **Table S3, and Table S4**), while the remainder were not strict replicates, allowing us to compare results from different expression systems. Different protein production methods could potentially yield different DNA-binding activities for the same TF, e.g., via post-translational modifications, or if the intended protein associates with interaction partners that also bind DNA. It is also possible that the isolated DBD could bind different DNA sequences from the full-length protein. To ask whether we observe such differences, we evaluated 217 pairs of experiments where the same TF was expressed by two different expression systems (e.g., wheat germ vs. mammalian cells), and a further 72 pairs that used the same expression system, by plotting enrichment of 3, 5, 7, and 9 bp k-mers in the HT-SELEX data (a total of 289 instances encompassing 62 TFs; **Supplementary Document S3, Table S6**). We identified minor technical differences, including randomly-occurring low-level enrichment

of homopolymers and concentration-dependent multimeric binding (see first page of **Supplementary Document S3** for full description), but we observed no evidence for cofactor contamination or unexpected differences in sequence specificity.

In separate parts of the Codebook project, this same set of poorly-characterized putative TFs was analyzed using ChIP-seq, Protein Binding Microarrays[44], and SMiLE-seq[45], as summarized in **Table S7** and described in the accompanying manuscripts [43,47,50]. Because the binding motif was not known in advance, we gauged the success of each protein in each experiment, including the GHT-SELEX experiments, based on whether similar DNA-binding motifs (i.e., PWMs) were obtained from different types of experiments, with all data types considered in aggregate by a team of expert curators[49]. Selection of a single PWM for each TF for subsequent analyses is also described in accompanying study[43]; PWMs logos, and information about their derivation are available in accompanying study[43] and online at https://codebook.ccbr.utoronto.ca, https://mex.autosome.org, and https://cisbp.ccbr.utoronto.ca[51].

In total, 139 of the 331 Codebook putative TFs had at least one successful GHT-SELEX experiment, of which 131 were also successful in HT-SELEX, 108 in ChIP-seq, and 102 in all three (**Figure 3a** and **Table S3**). The 139 were comprised mainly of C2H2-zf proteins, which are prevalent in the Codebook set (**Figure 3b)**. A total of 24 types of DBDs were present among the successful experiments (**Figure 3b**), however, illustrating that the method can capture motif-containing genomic target site locations of diverse TF families. An additional 163 of the putative TFs did not yield motifs in any of these three assays, however, despite the use of multiple expression systems and multiple constructs. Of these 163, only an additional nine were successful when including data from other assays used in the Codebook project (SMiLE-seq and Protein Binding Microarrays)[43] (**Table S7**). Successful and unsuccessful proteins produced as eGFP fusions yielded largely overlapping range of fluorescence signals (**Extended Data Fig 3, Table S3**), and on western blots, ~90% of sampled proteins exhibited a single prominent band at the correct molecular weight (**Table S8 and Supplementary Document S4)** for both successful and unsuccessful samples, suggesting that the majority are expressed and properly folded. It is possible that some of the putative TFs require post-translational modifications or cofactors, that our data analysis, motif derivation, and testing methods did not capture their specificity, or that they are false-negatives for some other reason. However, it is also possible that many of the tested putative TF simply do not bind DNA with sequence specificity, as in the majority of cases, they were also unsuccessful with other methods (**Table S7)**; if we assume so, then the success rate of GHT-SELEX on the Codebook proteins is 78% (139/177). Nonetheless, a central conclusion of this study is the identification of binding motifs for over 100 poorly characterized human TFs. For most of them, we produced three types of data (HT-SELEX, GHT-SELEX, and ChIP-seq) that provide differing perspectives on their sequence specificity and DNA binding.

Unexpectedly high overlap between TF binding to the genome *in vitro* and *in vivo*

We next sought to quantify the overlap between GHT-SELEX/MAGIX peaks and ChIP-seq peaks. For 137 TFs with GHT-SELEX data, encompassing 101 Codebook and 36 control TFs, ChIP-seq data in HEK293 cells were also available from a parallel Codebook study[52], which employed heterologous expression in HEK293 cells. Scoring peak overlaps requires first defining peak sets, which is typically accomplished using a threshold on enrichment values or statistical significance. GHT-SELEX, like ChIP-seq, produces peaks with a continuum of enrichment coefficient values and other associated statistics, rather than a bimodal distribution that would more readily discriminate bound from unbound loci. The overlap with ChIP-seq peaks and the enrichment of PWM scores within peaks were also generally continuous (examples in **Figure 4a**; distributions for all TFs in **Supplementary Document S2**). Moreover, the MAGIX score and associated FDR values at which the overlap of GHT-SELEX peaks with either ChIP-seq peaks or PWM matches dropped to random values varied dramatically between TFs, such that no unifying thresholds emerged that could be generally applied convincingly to all TFs.

As a workaround, we implemented a simple system to draw thresholds on the GHT-SELEX and ChIP-seq peak sets simultaneously and dynamically for each TF. This system identifies thresholds on a TF-by-TF basis by scanning through all numbers (N) of “Top N” peaks on a TF-specific basis, calculating the Jaccard statistic of overlap between the GHT-SELEX/MAGIX peaks and ChIP-seq peaks for each N, and choosing the peak number (N) with the maximum Jaccard value (**Figure 4b**). We posit that this approach accounts for unobserved TF-specific parameters in both GHT-SELEX and ChIP-seq assays, including different binding kinetics for both sequence-specific and nonspecific DNA binding, the effective concentration of the TF, and the ability of the TF to compete or cooperate with nucleosomes and other cofactors *in vivo*.

This approach yielded a very striking result, which is that for many TFs, a peak number can be identified with a surprisingly high Jaccard value (median 0.127 over all 137 TFs) (**Figure 4c,d,** and **Table S9**), meaning that almost 30% of GHT-SELEX peaks overlap with a ChIP-seq peak, and vice-versa. Peak overlap (i.e., Jaccard) is a demanding statistic, because random expectation (i.e., from choosing genomic regions at random) is near zero, as only a miniscule fraction of the genome is covered by the peaks in either data type, and both experimental variation and measurement noise will lead to fluctuation of the rank order of peaks. In permuted experiments (i.e., mismatched TFs), the Jaccard median is 0.00254 (Wilcoxon $p=1.3\times10^{-39}$) (**Figure 4c**). Using the same optimization process, the median Jaccard distance was 0.56 for ChIP-seq replicates[52]; this number presumably represents a typical upper bound. Overall, this outcome indicates that many TFs intrinsically (i.e. independently) specify many of their *in vivo* binding sites above a given TF-specific threshold. We do not expect the overlap to be perfect, due to the impact of chromatin and cofactors, but this result nonetheless contrasts with the traditional expectation that individual TF would not be able to independently specify the vast majority of their DNA targets in a genome[7]. The peak numbers yielding these high Jaccard values are often relatively low (**Figure 4d**), suggesting that the number of *in vivo* binding sites that are specified intrinsically by TFs is relatively small, and they correspond to a wide range of ChIP-seq p-value thresholds

434 and MAGIX score values (**Table S9**), such that both GHT-SELEX/MAGIX and ChIP-seq
435 data are needed to make these observations.
436
437 One potential explanation for this high overlap between GHT-SELEX and ChIP-seq
438 peaks is that the Codebook dataset is dominated by C2H2-zfs, which often have long
439 binding sites and thus intuitively high specificity. Indeed, among TFs with Jaccard > 0.1,
440 80% are C2H2-zf proteins (63 out of 79), vs 34% (20 out of 58) for those with Jaccard <
441 0.1, and overall, the median Jaccard value for C2H2-zf proteins is 0.1836, vs. 0.0559 for
442 non-C2H2-zf proteins (Wilcoxon $p=6.08x10^{-9}$). CTCF, a control TF that is known to
443 possess large number of genomic target sites, unusually high intrinsic sequence
444 specificity, and ability to control nucleosome positions[15,16], is among those with high
445 Jaccard values (0.45), although it is not the highest scoring in this dataset, and overall is
446 clearly not unique in its capacity to independently specify where it binds in living cells.
447 Counterintuitively, however, high Jaccard maxima were also obtained for a subset of
448 TFs with relatively short motifs, including TERF1, ZNF48, ZBTB8A, NACC2, and
449 several CXXC proteins, such as CXXC4 and KDM2A, that mainly bind CG
450 dinucleotides, as expected[53] (**Figure 5a**).
451
452 **Multiple explanations for the high sequence specificity observed in GHT-SELEX**
453
454 The traditional assumption that individual TFs have relatively low ability to specify their
455 genomic targets is based largely on the results of motif scans[7], which are dependent on
456 the accuracy of the PWM. To revisit this assumption, we performed a maximization of
457 the overlap between the Codebook PWM matches across the genome and the ChIP-
458 seq peaks, using the same Jaccard-based method described above (see **Methods** for
459 details). We found that the overlap between PWM predictions and ChIP-seq peaks is, in
460 fact, similar to the overlap between GHT-SELEX/MAGIX and ChIP-seq peaks (**Figure**
461 **5a, Table S9**). The number of peaks at which the maximum Jaccard was obtained is
462 also typically similar (**Figure 5b**). Thus, PWMs are often more accurate than generally
463 believed, and comparable to the GHT-SELEX data. It is important to note that we
464 selected the Codebook PWMs for their accuracy across datasets, i.e., other PWMs for
465 the same TF would display lower overlap. Therefore, the PWM derivation method,
466 including the experimental data employed, is critical for representing TF binding *in vivo*.
467
468 We also found that the motif scanning method is important to achieve the highest
469 overlap between PWM matches and ChIP-seq peaks. Consistent with a recent report[54],
470 for some TFs, the sum of predicted affinity scores over a sequence window (i.e. the sum
471 of the PWM probability scores at individual positions, rather than the log-odds that is
472 output by most PWM scanning tools) results in a considerably higher maximum Jaccard
473 value than taking the maximum or sum of log-odds PWM scores (which are generally
474 thought to represent binding energy[55,56]) (**Figure 5c**). This Sum-of-affinity scoring
475 presumably reflects the cooperation of multiple adjacent binding sites, traditionally
476 referred to as "avidity"[57]. The effect is most striking for a subset of TFs that bind short or
477 repetitive sequences, including CG dinucleotides and poly-A stretches (**Figure 5d**), but
478 it also appears to underpin the specificity of NACC2 and ZNF48, which have unique,
479 non-repetitive motifs (**Figure 5e**). Points above the diagonal in **Figure 5a**, in which

PWM prediction shows higher overlap with ChIP than the GHT-SELEX, may therefore represent the impact of TF binding sites over a larger window, which would be detected in ChIP-seq but not GHT-SELEX (ChIP-seq fragments are 100-300 bp, and the PWM scans are performed with a 200 bp window, while the GHT-SELEX fragments are only ~65 bp). For example, scanning 200-base windows with the short CG motif for CXXC4 may be better suited for the detection of CpG islands (which dominate the CXXC4 binding sites[43], in which the CG dinucleotides will be distributed over a large region (by definition $\geq$200 bp).

In contrast – and despite evaluation of hundreds to thousands of PWMs for each TF - there were still many TFs for which we could not derive a PWM that rivals GHT-SELEX data in correspondence to ChIP-seq peaks (those below the diagonal in **Figure 5a**). These TFs are almost entirely proteins with a long array of C2H2-zf domains, which we examine more closely in the next section.

Alternate usage of C2H2-zf domains within large arrays

The expansive collection of GHT-SELEX, HT-SELEX, and ChIP-seq data for C2H2-zf proteins provided an opportunity to examine the long-standing issue of the usage of individual C2H2-zf domains within large arrays. Anecdotally, we observed many instances where the motifs detected for C2H2-zf proteins were much shorter than expected based on the number of C2H2-zf domains, as well as examples in which multiple distinct motifs emerged, suggesting that the TFs might use partial subsets of their DBD array to engage DNA at different locations. Proving differential engagement of the specific C2H2-zf domains is challenging, however, due to low statistical power (there are many possible C2H2-zf domain sub-arrays, and a limited number of highly enriched peaks) and the fact that the genome is highly non-random and repeat-rich. To minimize the impact of these issues, we developed a new method that utilizes the C2H2-zf recognition code to assess which sets of C2H2-zf domains are likely to be engaged at any individual binding sites. We call this method RCADEEM (Recognition Code-Assisted Discovery of regulatory Elements by Expectation-Maximization) (see **Methods** for details). **Figure 6a** shows a schematic and the results of applying RCADEEM to CTCF, illustrating that it produces a “core” motif recognized by fingers 4-7 at all sites, and alternative usage of flanking C2H2-zf domains in a subset of sites, very similar to the differential usage of CTCF C2H2-zf domains that has been previously described[30].

We applied RCADEEM to all 120 C2H2-zf proteins for which we had successful data from GHT-SELEX (**Table S10).** We applied RCADEEM on GHT-SELEX data and separately, if available, on HT-SELEX and ChIP-seq; for GHT-SELEX and ChIP-seq, we applied it both with and without repeat sequences (i.e., removing any peaks that overlap with the UCSC RepeatMasker track). In total, we obtained RCADEEM predictions for 86 of them (**Table S10**), all of which are available via the web resources accompanying this paper (https://codebook.ccbr.utoronto.ca/). (For the remaining 34, the algorithm did not converge, suggesting that the sequence preferences of the protein do not closely follow the recognition code, and thus cannot be analyzed in this way).

Most of the 86 displayed what appears to represent alternative usage of segments of the C2H2-zf domain array on different DNA molecules (e.g., different genomic loci) within the same experiment (**Supplementary Data S2**). These patterns could not be accounted for by truncations or other potential artifacts (**Document S4**).

We manually classified the apparent C2H2-zf domain usage into the following categories, examples of which are shown in **Figure 6b-f**. **Figure 6g** provides an overview of the descriptors and other properties of each of the C2H2-zf proteins. 1) *Canonical (30 instances)* follows the baseline assumption that a TF always uses the same set of C2H2-zf domains to recognize sites that can be described with a single PWM. 2) *Core with extensions (24 instances)*, where all sites share a sequence motif bound by a subset of the C2H2-zf domains, which is supplemented by recognition of flanking sequences by adjacent C2H2-zf domains at some binding sites. 3) *Finger shift (14 instances),* where the TF recognizes a range of tiled target sites by binding with variable subsets of adjacent C2H2-zf domains. 4) *Multiple DBDs (32 instances)*, in which subsets of the C2H2-zf domain array appear to function as independent DBDs. The last three binding modes are not mutually exclusive. For example, ZNF471 displays both multiple DBDs and core with extensions with one of the DBDs (**Figure 6f**), while the long finger shift in ZNF665 (**Figure 6d**) leads effectively to multiple DBDs, as the target sites of most N-terminal and C-terminal ends do not overlap with each other. **Table S10** lists the annotations for all 86 proteins.

Evolution of C2H2-zf protein DNA-binding specificities via internal duplication

In the RCADEEM outputs, different segments of a C2H2-zf domain array (i.e., different DNA binding regions of the protein) are often predicted to bind similar yet distinct sets of sequences. For example, ZNF775 (**Figure 6e**) binds two types of sites that contain a shared GNWGAA consensus, followed by either TTT or GCA trinucleotides. RCADEEM predicts that these two sites are recognized by C2H2-zf domain arrays 1-4 and 5-8, respectively. Indeed, arrays 1-3 and 5-7, as well as 9-11, are homologous, on the basis of sequence identity (visualized at https://codebook.ccbr.utoronto.ca/details.php?TF=ZNF775), suggesting that they arose from duplications. All three arrays are present in mammals as distant as the Tasmanian devil, indicating that the duplications predate divergence from marsupials and have since been conserved. The cellular and physiological functions of this protein are unknown, to our knowledge, but this degree of sequence conservation suggests a conserved role across mammals.

Another example is ZNF721: RCADEEM indicates that it has three DNA-binding modes, with related but distinct motifs (**Figure 7a**), corresponding to homologous C2H2-zf domain arrays containing fingers 6-13, 12-16, and 18-22 (**Figure 7b**). The distinct sequence preferences of the duplicated ZNF721 arrays are supported by experimental data for partial “DBD1” and “DBD2” constructs, corresponding roughly to the first and second half of the full array, which recognize largely distinct subsets of the genomic sites bound by full length TF in GHT-SELEX (**Figure 7a**) and prefer almost entirely

distinct 10-mers in HT-SELEX (**Figure 7c**). The function of ZNF721 has not been determined, but sequences recognized by the first (6-13) and third (18-22) duplicated C2H2-zf domain arrays of ZNF721 are found in the highly numerous Alpha repeats, which are fast-evolving elements found at primate centromeres[58]. ZNF721 itself is present only in primates. ZNF721 also binds thousands of unique loci outside known repeat elements, and associates physically with TRIM28/KAP1[59], suggesting a role in gene silencing or heterochromatin formation.

To survey the prevalence of internal duplication of C2H2-zf domains, we compared all pairs of individual human C2H2-zf domains occurring in the same protein and found that 185 human C2H2-zf proteins (~25%) contain at least one pair of C2H2-zf domains that differ by 3 or fewer edits (substitutions, deletions or insertions; **Table S10**), suggesting that they are derived from recent duplications. Furthermore, as in ZNF775 and ZNF721, there are 140 proteins with apparent internal C2H2-zf domain array duplications, defined as two (or more) adjacent C2H2-zf domains (i.e., an array) related to a second such array with two (or more) C2H2-zf domains, with 5 or fewer edits per C2H2-zf domain. Based on recognition code predictions, C2H2-zf domain arrays within internal array duplications have more diverged sequence specificities from each other than individually duplicated C2H2-zf domains (**Figure 7d, Table S11**The prevalence and diversification of internal C2H2-zf domain array duplications suggest that they are a common modality for evolution of novel functional roles for this large class of proteins.

DISCUSSION

GHT-SELEX assays direct and unassisted binding of single TFs to the unmodified and unchromatinized genome *in vitro*, revealing surprisingly specific intrinsic sequence preferences for many human TFs. The assay, and the associated MAGIX analysis pipeline, offers several technical advantages over alternatives, including smaller fragment size and compatibility with the same instrumentation used for HT-SELEX. GHT-SELEX data are often more similar to *in vivo* TF binding data than previous assessments would suggest they should be[7,26], indicating that, for an apparently large subset of TFs, chromatin and cofactors have less critical influence on where binding occurs. This same observation implies that this subset of individual TFs may have a greater ability to overcome the chromatin state than is commonly believed.

The unexpectedly high overlap between ChIP-seq, GHT-SELEX, and PWM scans could be partly explained by technical shortcomings in previous PWM-based genome scans. HT-SELEX and other *in vitro* approaches utilizing random sequences are powerful in that they are unbiased in terms of sequence composition[60], but they are inherently limited in sequence length and context that can be surveyed. ChIP-seq is invaluable because it can assay binding within cells, but it does not inherently discern direct, indirect, and non-specific binding. Thus, PWMs derived from ChIP-seq and other *in vivo* approaches are influenced by factors other than the TF, in addition to the biased sequence content of the genome. GHT-SELEX provides a powerful intermediate that can resolve ambiguities of both motif discovery and PWM scanning, and thus provides data that complements both ChIP-seq and *in vitro* assays that utilize random sequences. The PWM scanning method is of particular importance: taking the sum of predicted affinity over longer sequence windows is critical to capturing positions that are bound *in vitro* and *in vivo* by CXXC-zfs and other proteins that recognize clusters of simple sequence motifs.

GHT-SELEX is particularly effective with C2H2-zf proteins, and, together with RCADEEM, has an unprecedented ability to both obtain and dissect *in vitro* the multiple binding modes that are characteristic of this family, and inherently more difficult to represent as a single PWM. The existence of multiple binding modes also provides a potential explanation for the large number of C2H2-zf domains in each protein. In at least some cases, these large arrays derive from internal duplications of segments of the C2H2-zf domain arrays, possibly facilitating generation of evolutionary novelty via duplication and divergence.

In contrast to the C2H2-zf family, the most well-studied TFs tend to be in the TF classes such as homeodomain, bHLH, bZIP, nuclear receptor, and Sox TFs, because they are the most strongly conserved and often dictate specific biological processes (e.g., morphogenesis, body plan, lineage specification, etc.)[26]. Our study included some of these TFs (e.g., LEUTX, BATF2, RARA and SRY), and they displayed only limited overlap between GHT-SELEX and ChIP-seq peaks, indicating that many of them cannot independently specify *in vivo* binding locations and hence target genes. It has long been known that TFs controlling chromatin in yeast are largely distinct from those that

642 regulate specific pathways; we speculate that a similar division may exist in human and
643 other animals.
644
645 GHT-SELEX data, together with the larger Codebook dataset, provide an extensive new
646 dataset of TF motifs (i.e., PWMs), encompassing most putative TFs currently lacking
647 them. Accompanying papers provide a thorough analysis of the results of this project,
648 which underscore many challenges and benefits of accurate motif representations.
649 Representation of TF sequence specificity remains an open challenge, more than four
650 decades after the introduction of the standard PWM model[61]. More accurate
651 representations of multimeric binding and large and complex binding sites, in particular
652 for C2H2-zf proteins, will be useful for a variety of purposes. We propose that obtaining
653 data from GHT-SELEX for additional TFs with "known" motifs and genomic binding sites
654 from ChIP-seq will produce a more detailed view of their intrinsic DNA binding abilities
655 and how this intrinsic ability dictates TF-genome interactions in living cells.

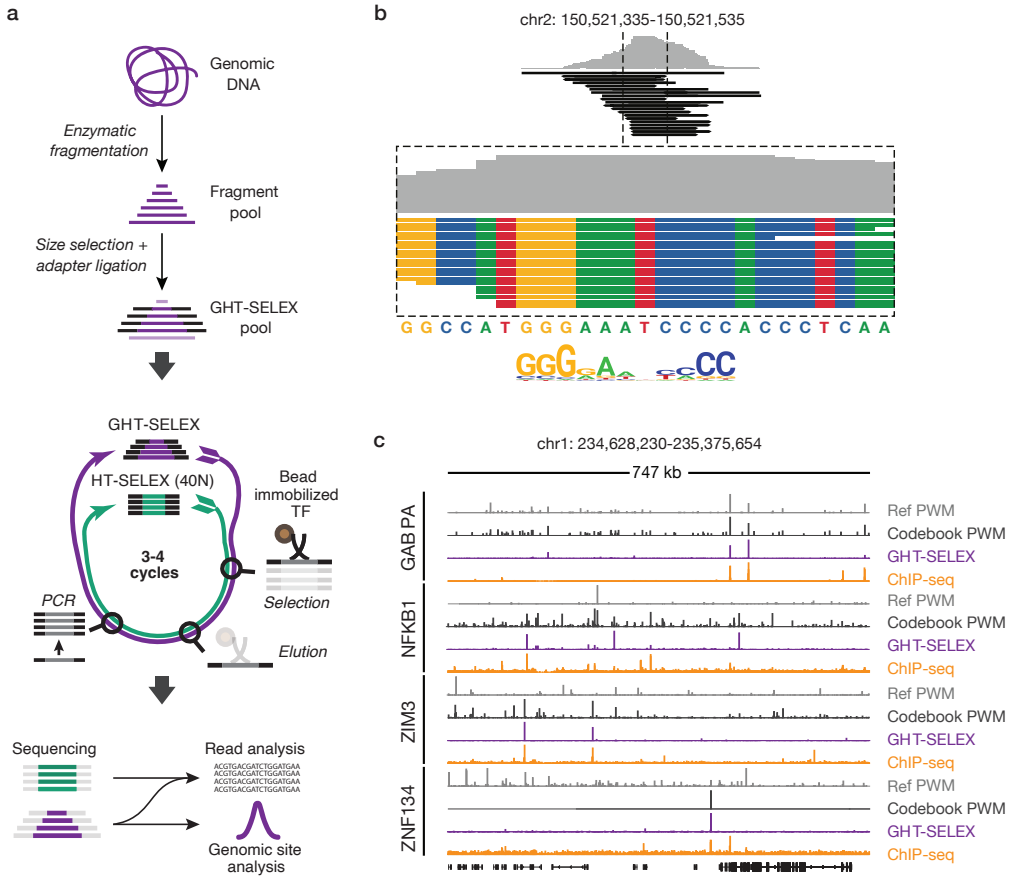

a
Genomic DNA
Enzymatic fragmentation
Fragment pool
Size selection + adapter ligation
GHT-SELEX pool
GHT-SELEX
HT-SELEX (40N)
Bead immobilized TF
3-4 cycles
PCR
Selection
Elution
Sequencing
Read analysis
Genomic site analysis
b
chr2: 150,521,335-150,521,535
G G C C A T G G G A A A T C C C C A C C C T C A A
c
chr1: 234,628,230-235,375,654
747 kb
GABPA
NFKB1
ZIM3
ZNF134
Ref PWM
Codebook PWM
GHT-SELEX
ChIP-seq

FIGURE LEGENDS

Figure 1. Overview of GHT-SELEX. a. Schematic of GHT-SELEX, showing parallels with HT-SELEX. **b.** Example of read accumulation over a TF motif match for NFKB1. **c.** Genomic binding for four positive control TFs on a genomic region showing (top to bottom) PWM scanning scores (moving average of affinity scores, from MOODS[69] scan in linear domain, using a window of size 200bp) for literature PWMs (CIS-BP Identifiers: M0839, M03448, M08312 and M02996) and Codebook PWMs, followed by read coverage signal observed in GHT-SELEX and ChIP-seq.

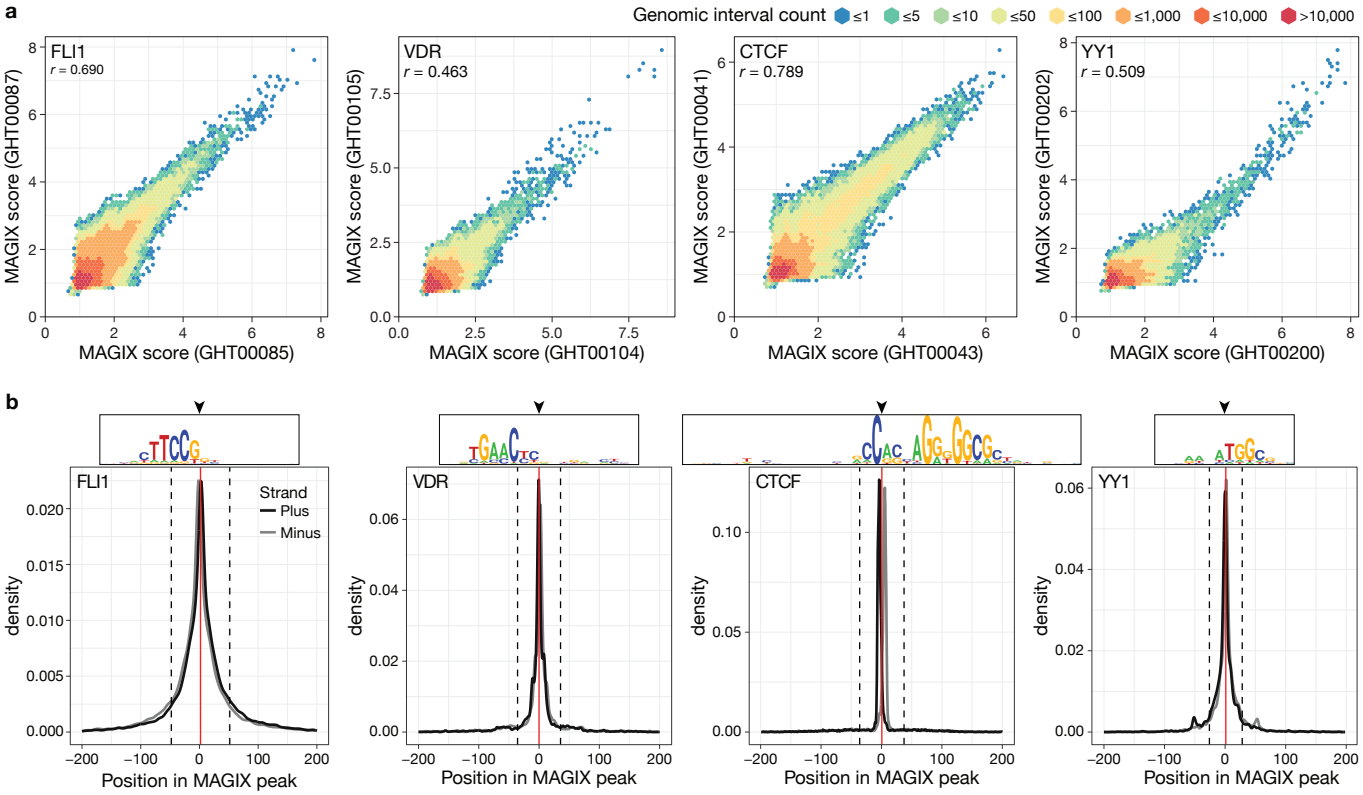

a
Genomic interval count
≤1
≤5
≤10
≤50
≤100
≤1,000
≤10,000
>10,000
FLI1
r = 0.690
MAGIX score (GHT00087)
MAGIX score (GHT00085)
VDR
r = 0.463
MAGIX score (GHT00105)
MAGIX score (GHT00104)
CTCF
r = 0.789
MAGIX score (GHT00041)
MAGIX score (GHT00043)
YY1
r = 0.509
MAGIX score (GHT00202)
MAGIX score (GHT00200)
b
FLI1
Strand
Plus
Minus
density
Position in MAGIX peak
VDR
CTCF
YY1

Figure 2. MAGIX method for interpretation of GHT-SELEX data. a. GHT-SELEX reproducibility. Scatterplots for four control TFs show MAGIX score values for all 200bp genomic bins between replicate GHT-SELEX experiments. **b.** Distribution of PWM hits for the top-ranked TF PWM (highest AUROC on GHT-peaks as determined in the accompanying study[49] within the MAGIX peaks (up to the 10,000 highest scoring peaks). PWM hits were identified with MOODS[69] ($P < 0.0001$) and the hit location is defined by the centre of the PWM, depicted above each plot. Solid red lines represent the mean PWM hit position within MAGIX peaks and dashed lines represent one standard deviation about the mean. Hits on the plus and minus strands are differentiated.

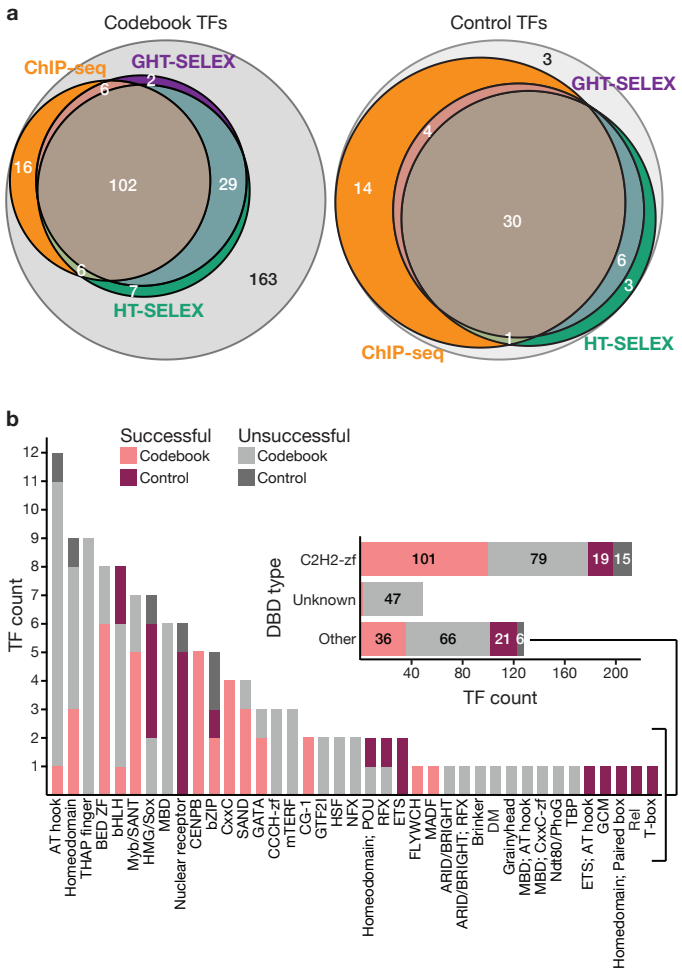

a
Codebook TFs
ChIP-seq
GHT-SELEX
HT-SELEX
16
6
2
102
29
6
7
163
Control TFs
ChIP-seq
GHT-SELEX
HT-SELEX
3
14
4
30
6
3
1
b
Successful
Codebook
Control
Unsuccessful
Codebook
Control
TF count
DBD type
C2H2-zf
Unknown
Other
101
79
19
15
47
36
66
21
6
TF count
AT hook
Homeodomain
THAP finger
BED ZF
bHLH
Myb/SANT
HMG/Sox
MBD
Nuclear receptor
CENPB
bZIP
CxxC
SAND
GATA
CCCH-zf
mTERF
CG-1
GTF2I
HSF
NFX
Homeodomain; POU
RFX
ETS
FLYWCH
MADF
ARID/BRIGHT
ARID/BRIGHT; RFX
Brinker
DM
Grainyhead
MBD; AT hook
MBD; CxxC-zf
Nd80/PhoG
TBP
ETS; AT hook
GCM
Homeodomain; Paired box
Rel
T-box

Figure 3. Analysis of 331 Codebook proteins and 61 control TFs using GHT-SELEX. a. Venn diagram displays the number of TFs with successful experiments in GHT-SELEX, HT-SELEX, and ChIP-seq for all Codebook TFs (left) and control TFs (right) assayed with GHT- and HT-SELEX. **b.** Bar chart shows the number of TFs with at least one successful GHT-SELEX experiment, categorized based by DBD type. C2H2-zf proteins and those with an unknown DBD (at the beginning of the project) are inset due to large numbers.

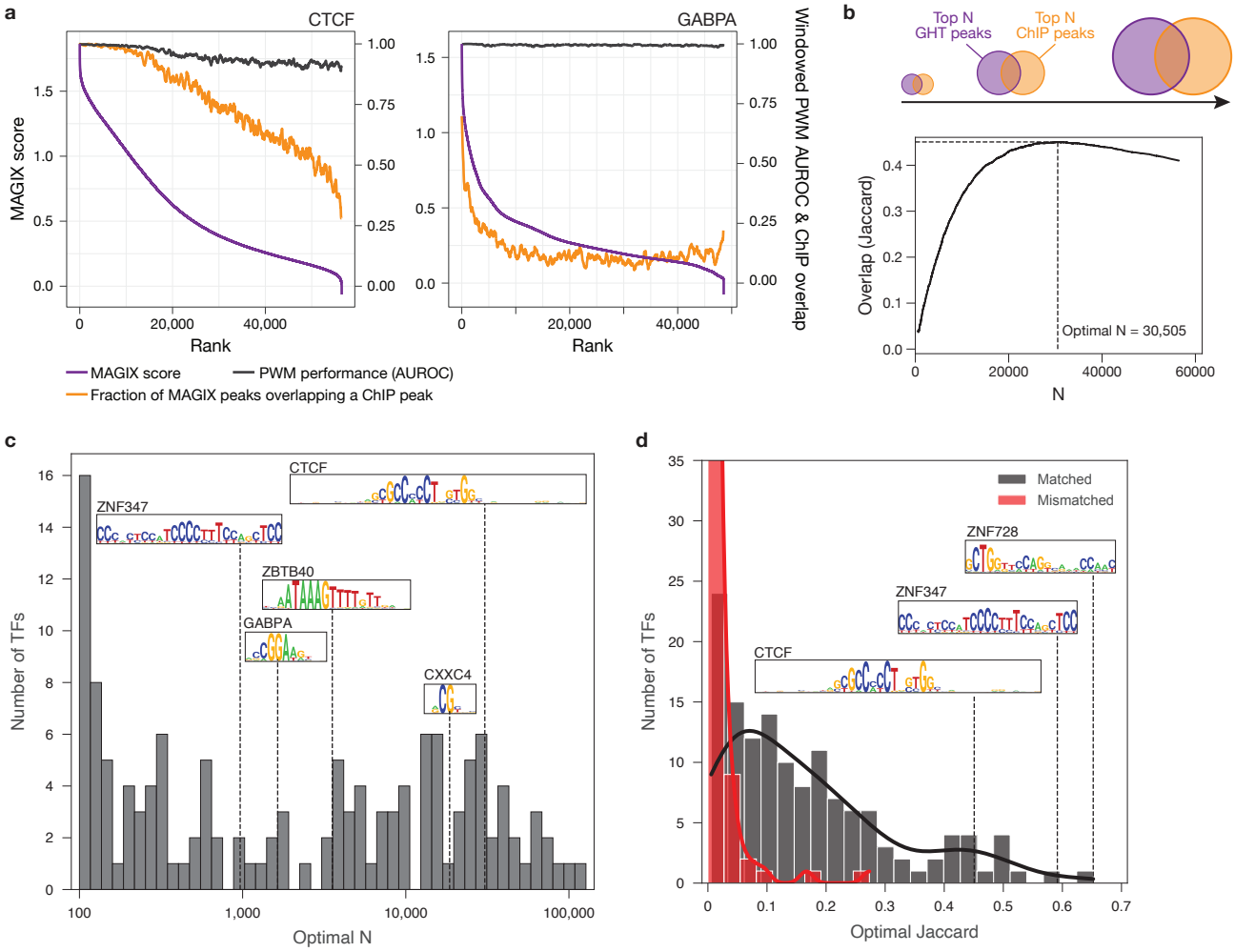

a
CTCF
GABPA
MAGIX score
Windowed PWM AUROC & ChIP overlap
Rank
MAGIX score
PWM performance (AUROC)
Fraction of MAGIX peaks overlapping a ChIP peak
b
Top N GHT peaks
Top N ChIP peaks
Overlap (Jaccard)
Optimal N = 30,505
N
c
CTCF
ZNF347
ZBTB40
GABPA
CXXC4
Number of TFs
Optimal N
d
Matched
Mismatched
ZNF728
ZNF347
CTCF
Number of TFs
Optimal Jaccard

Figure 4. Correspondence between GHT-SELEX and ChIP-seq peaks. a. Enrichment of ChIP-seq peaks and PWM hits within MAGIX peaks, for two example control TFs. MAGIX peaks are sorted by their MAGIX score (purple, *left y-axis*). Orange line shows the proportion of peaks (in a sliding window of 500 peaks over the ranked peaks, with a step size of 50) that overlap with a ChIP-seq peak (at MACS threshold $P < 0.001$). Black line shows the AUROC for PWM affinity scores (calculated by AffiMx[52]) of MAGIX peaks in the same window vs. 500 random genomic sites. **b.** Illustration of peak number optimization (for CTCF as an example). **c.** Histogram of the optimal values of N (peak count) for the 137 TFs that have both GHT-SELEX and ChIP-seq peaks. **d.** Histogram of optimal Jaccard values, compared to the maximum Jaccard for mismatched TFs (i.e. between GHT-SELEX for one TF and ChIP-seq for a randomly selected TF).

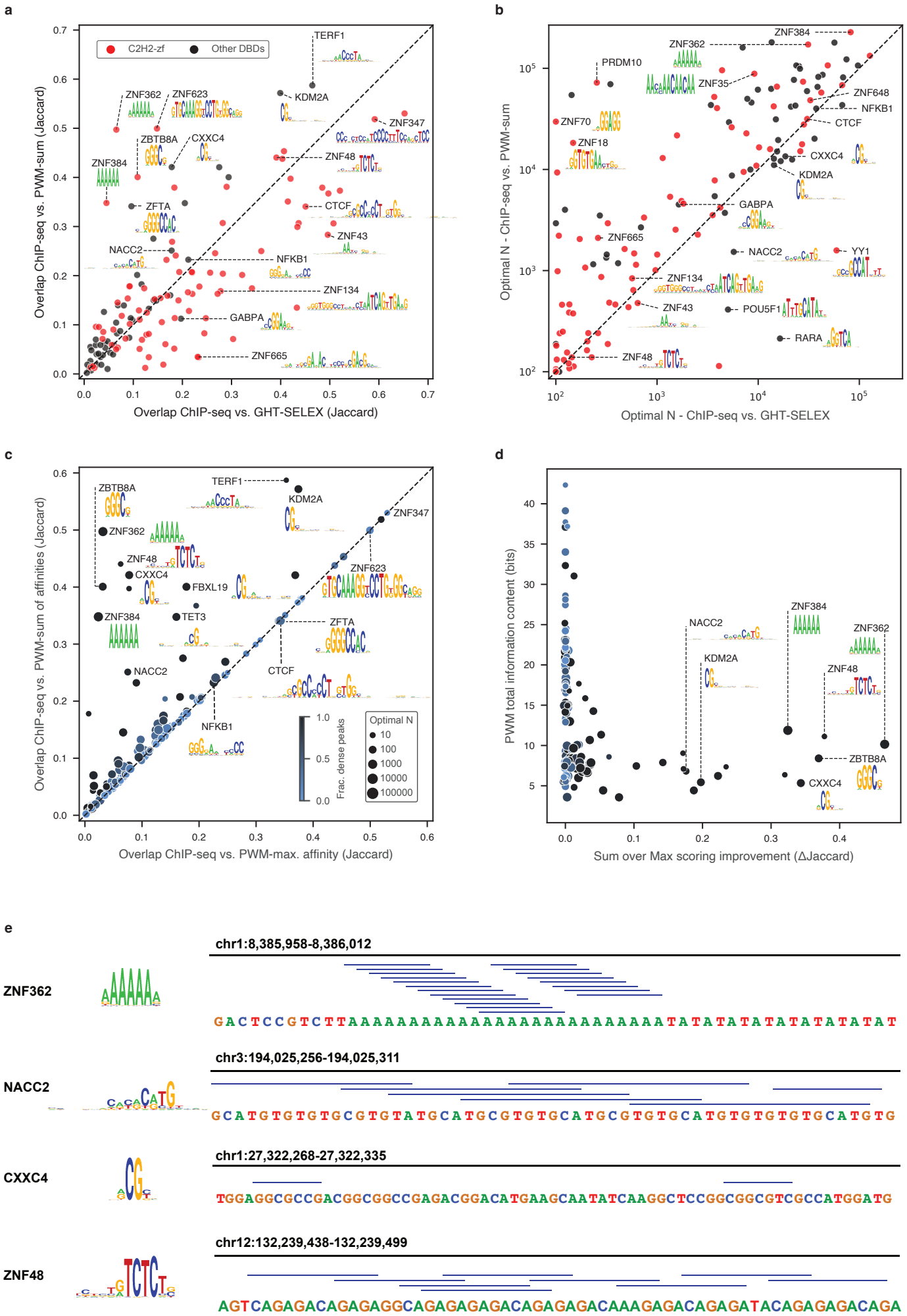

a
C2H2-zf
Other DBDs
Overlap ChIP-seq vs. PWM-sum (Jaccard)
Overlap ChIP-seq vs. GHT-SELEX (Jaccard)
TERF1
KDM2A
ZNF362
ZNF623
ZNF347
ZBTB8A
CXXC4
ZNF48
ZNF384
ZFTA
CTCF
ZNF43
NACC2
NFKB1
ZNF134
GABPA
ZNF665
b
Optimal N - ChIP-seq vs. PWM-sum
Optimal N - ChIP-seq vs. GHT-SELEX
ZNF384
ZNF362
PRDM10
ZNF35
ZNF648
NFKB1
CTCF
ZNF70
ZNF18
CXXC4
KDM2A
GABPA
ZNF665
NACC2
YY1
ZNF134
ZNF43
POU5F1
RARA
ZNF48
c
Overlap ChIP-seq vs. PWM-sum of affinities (Jaccard)
Overlap ChIP-seq vs. PWM-max. affinity (Jaccard)
ZBTB8A
TERF1
KDM2A
ZNF347
ZNF362
ZNF48
CXXC4
ZNF623
FBXL19
ZNF384
TET3
ZFTA
NACC2
CTCF
NFKB1
Frac. dense peaks
Optimal N
10
100
1000
10000
100000
d
PWM total information content (bits)
Sum over Max scoring improvement (ΔJaccard)
ZNF384
NACC2
ZNF362
KDM2A
ZNF48
ZBTB8A
CXXC4
e
chr1:8,385,958-8,386,012
ZNF362
GACTCCGTCTTAAAAAAAAAAAAAAAAAAAAAAAAAATATATATATATATATATAT
chr3:194,025,256-194,025,311
NACC2
GCATGTGTGTGCGTGTATGCATGCGTGTGCATGCGTGTGCATGTGTGTGTGCATGTG
chr1:27,322,268-27,322,335
CXXC4
TGGAGGCGCCGACGGCGGCCGAGACGGACATGAAGCAATATCAAGGCTCCGGCGGCGTCGCCATGGATG
chr12:132,239,438-132,239,499
ZNF48
AGTCAGAGACAGAGAGGCAGAGAGAGACAGAGAGACAAAGAGACAGAGATACAGAGAGACAGA

Figure 5. High quality PWMs often predict *in vivo* binding sites as effectively as GHT-SELEX peaks. a. Scatter plot of optimal Jaccard value between GHT-SELEX peaks and ChIP-seq peaks (x-axis) vs. optimal Jaccard value between PWM-predicted sites and ChIP-seq peaks (y-axis), for all 137 TFs (dots). **b.** Scatter plot of optimal N (peak number) for the same peak set comparisons shown in (**a**). **c.** Scatter plot showing optimal Jaccard value between PWM-predicted sites and ChIP-seq peaks, for maximum-affinity PWM scoring and sum-of-affinities PWM scoring. Points (TFs) are scaled based on the optimal number of peaks (in the sum scoring), and the color reflects the fraction of binding sites comprised of multiple PWM hits. **d.** Scatter plot of the improvement in the optimal Jaccard value associated with sum-of-affinities PWM scoring vs. information content of the PWM. Points' size and color are the same as panel (**c**). **e.** Examples of four TFs with multiple motif matches within a single ChIP-seq peak.

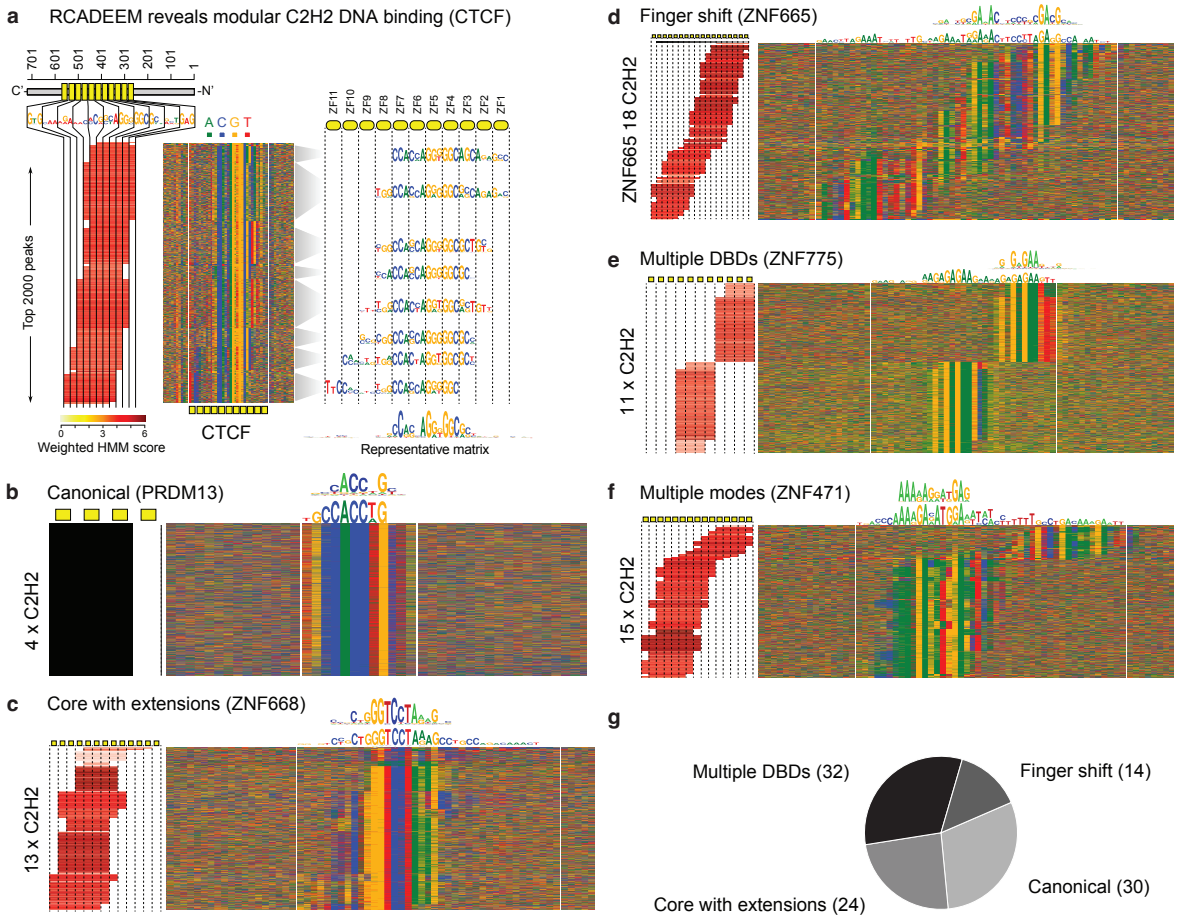

a
RCADEEM reveals modular C2H2 DNA binding (CTCF)
701
601
501
401
301
201
101
1
C'
-N'
A C G T
ZF11
ZF10
ZF9
ZF8
ZF7
ZF6
ZF5
ZF4
ZF3
ZF2
ZF1
Top 2000 peaks
0
3
6
Weighted HMM score
CTCF
Representative matrix
b
Canonical (PRDM13)
4 x C2H2
c
Core with extensions (ZNF668)
13 x C2H2
d
Finger shift (ZNF665)
ZNF665 18 C2H2
e
Multiple DBDs (ZNF775)
11 x C2H2
f
Multiple modes (ZNF471)
15 x C2H2
g
Multiple DBDs (32)
Finger shift (14)
Canonical (30)
Core with extensions (24)

Figure 6. Alternative engagement of individual C2H2-zf domains at genomic binding sites inferred from the recognition code. a. RCADEEM applied to CTCF. *Middle* panel displays the top 2,000 nonrepetitive GHT-SELEX peaks. White vertical bars indicate the region that is expected to contact the DNA based on the assumption that each of the C2H2-zf domains define three contiguous bases. *Left* panel indicates which C2H2-zf domains are inferred to engage each DNA sequence, which is used to determine the row order in the figure. *Right* panel shows motifs for the major sub-sites, derived from base frequencies in the sequence alignment. **b-f,** Top 2,000 non-repeat peak sequences, as in (**a**), for representative TFs with different binding modes, as described in the main text. Above each is shown the sequence logo for the single representative Codebook PWM (*top*) and a motif generated by RCADEEM that represents all of the observed sequences (*bottom*). **g.** Number of occurrences of each category among all 86 C2H2-zf proteins for which RCADEEM yielded a significant outcome; note that a TF might appear in multiple or no categories.

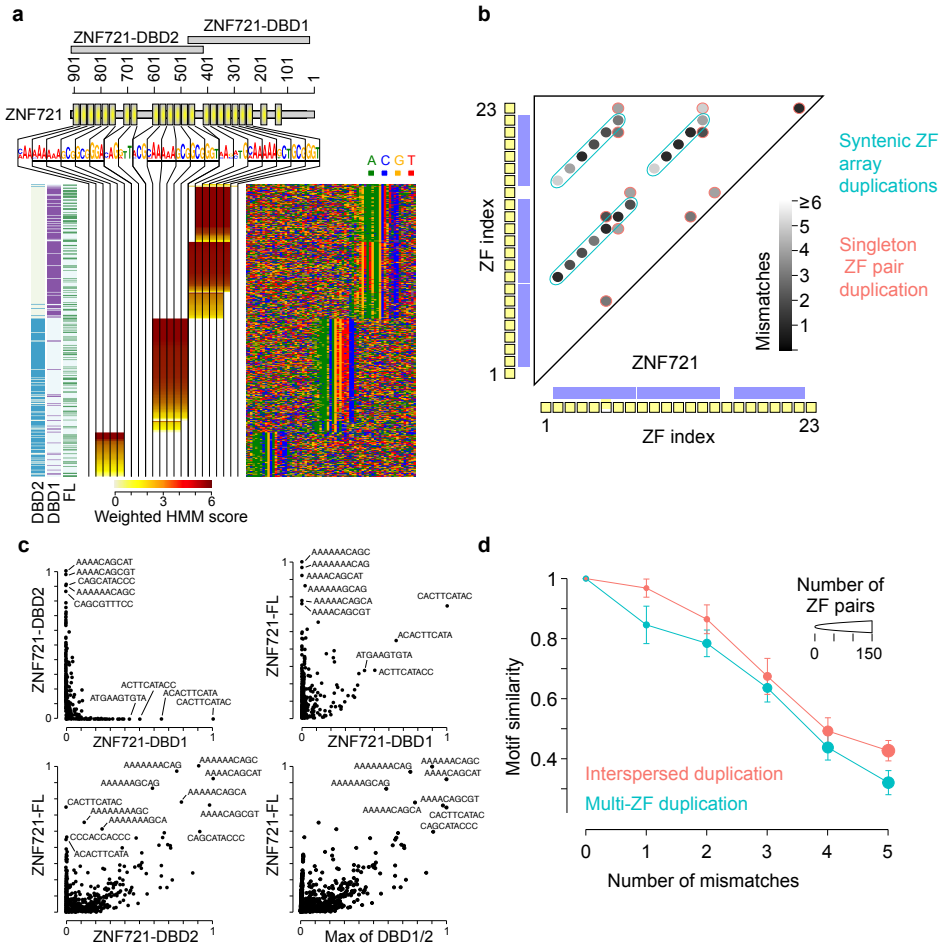

a
ZNF721-DBD2
ZNF721-DBD1
901
801
701
601
501
401
301
201
101
1
ZNF721
A C G T
DBD2
DBD1
FL
0
3
6
Weighted HMM score
b
23
1
ZF index
Syntenic ZF array duplications
Singleton ZF pair duplication
≥6
5
4
3
2
1
Mismatches
ZNF721
1
ZF index
23
c
ZNF721-DBD2
ZNF721-DBD1
ZNF721-FL
Max of DBD1/2
d
Motif similarity
Number of ZF pairs
0
150
Interspersed duplication
Multi-ZF duplication
Number of mismatches

Figure 7. Evolution of C2H2-zf protein DNA-binding specificities through internal duplication of DBDs and DBD arrays. a. RCADEEM results for the pool of top 500 peaks from full-length ZNF721 and two DBD constructs, after removing peaks that overlap repeats. The construct in which the peak was observed is indicated on the left. **b.** Similarity of C2H2-zf domains of ZNF721 (based on the number of mismatches in the global alignment). Apparent duplicated arrays are encircled by blue border (i.e. syntenic duplications), while single pairs that may also be duplicates are circled in red. **c.** Scatterplots of the HT-SELEX *k*-mer scores[65] (relative counts) across the three ZNF721 constructs **d.** Comparison of average per-base similarity (correlation of nucleotide frequency) in PWMs predicted by the recognition code, for those present in duplicated arrays vs. those duplicated as individual C2H2-zf domains, with duplicated C2H2-zf domains taken as pairs that are separated from each other by 5 or less edits. DBD pairs have been filtered to contain only combinations where both DBDs are likely to have retained their ability to bind DNA (have DNA binding functionality score[68] >0.5). Error bars show standard error of the mean.

METHODS

TFs and constructs. Selection of TFs, design of constructs for gene synthesis, and expression vectors are described in the accompanying study[43]. Sequences and other information are available as described below in Data Availability.

Protein production and quality control. We used three protein expression systems, which we refer to in **Table S2** and below as *Lysate*, *IVT*, and *eGFP-IVT*, respectively. The *Lysate* system employed recombinant HEK293 cells, created in the accompanying study[47], and in a previous study[48], which express eGFP-tagged full-length proteins from a Tet-inducible promoter (plasmid backbones pTH13195[48] and pTH12027). We induced expression by Doxycycline treatment for 24 hours prior to harvest, and confirmed via fluorescent microscopy. Whole cell lysates were then harvested from a 10cm plate (~10 million cells) for each line using 1 ml of lysis buffer (50 mM Tris-Cl at pH 7.4 containing 150 mM NaCl and 1% Triton X-100), supplemented with protease-inhibitor cocktail (Roche cOmplete mini, 04693159001), as described previously[37]. Each of the SELEX cycles used 50 μl of lysate. *IVT* used an *in vitro* transcription-translation reaction (PURExpress In Vitro Protein Synthesis Kit, NEB, Cat# E6800L) to express T7-driven, GST-tagged proteins (either full-length or DBDs) (plasmid backbone pTH6838[51]). *eGFP-IVT* employs the TNT SP6 High-Yield Wheat Germ Protein Expression System (Promega, Cat# L3260) to express SP6-driven, eGFP-tagged proteins (either full-length or DBDs) (plasmid backbone pTH16505, an SP6-promoter driven, N-terminal eGFP-tagged bacterial expression vector, modified from pF3A–eGFP [45] to contain AscI and SbfI restriction sites after the eGFP. For *IVT* and *eGFP-IVT* production systems, we performed reactions according to kit instructions, but using a smaller volume: 7.5 μl of *IVT* or 5 μl of *eGFP-IVT* reaction sample was used in each binding reaction of each SELEX cycle. Fluorescence levels shown in **Table S3** and **Extended Data Fig. S3** were measured using a BIOTEK Synergy Neo microplate reader from 10 μl volume of lysate or wheat germ extracts on a conical transparent 96-well plate using bottom optics and with a gain setting of 80. A subset of these proteins was also analyzed on Western blots (**Table S7**, **Supplementary Document S4)** with the same anti-GFP antibody (ab290, Abcam) as used in SELEX, and a secondary horseradish peroxidase fusion antibody (Cat Nr. Anti-Rabbit IgG HRP Conjugate, Cat W401B, Promega) and Immobilon Western Chemiluminescent HRP Substrate (WBKLS0050, Millipore Sigma). Blots were imaged with Odyssey Fc Imager (LICORbio).

GHT-SELEX and HT-SELEX library preparation. We fragmented HEK293 genomic DNA (Genscript, USA; Cat. No. M00094) for 45 minutes using NEBNext dsDNA Fragmentase enzyme mix (NEB, M0348S), and then performed a size selection step to reduce the amounts of fragments larger than 200 bp. In the size selection we added 0.9X volume of bead suspension (magnetic SPRI beads, supplied with the kit, NEB, E7103S) to the fragmented DNA, mixed the reaction for a minute, and then removed the large DNA fragment bound beads with a magnet, after which we diluted the supernatant 5X with water, followed by purification with a PCR purification kit (NEB, T1030S), to recover fragments as small as 25 bp. Next, the fragments were converted to an Illumina

sequencing compatible library using NEBNext® Ultra™ II DNA Library Prep kit (NEB: E7103S) and NEB E7350 adapters. After adapter ligation, we purified the library with a PCR purification kit (NEB, T1030S) and then amplified it for five PCR cycles to convert the partially single-stranded adapter flanks to fully double-stranded DNA, to increase the amount of the product and reduce the amount of methylated cytosine residues in the initial library. The ninety-six (96) HT-SELEX ligands were prepared as described[62], with the exception that the reverse primer was replaced with a primer (5'-CTGGAGTTCAGACGTGTGCTCTTCCGATCT-3') that does not contain a T7 promoter sequence. HT-SELEX ligands differ from each other by containing a well-specific variable region that flanks the randomized 40 bases indicated in the name of the experiments (e.g., AA40NCCAGTG contains 40 bases flanked by AA and CCAGTG sequences and partial Illumina adapter sequences). All primers and library preparation schemes are given in **Table S1**.

HT-SELEX and GHT-SELEX. We modified protocols from a previously-described HT-SELEX procedure[37]. HT-SELEX and the GHT-SELEX ligands contain the same flanking constant regions and thus there were no differences in the selections or sequencing library preparations. We conducted the magnetic bead washing operations below using a Biotek 405TS plate washer fitted with a magnetic carrier. We performed 21 different batches of SELEX, which varied in some technical respects in order to accommodate the three protein production systems and to implement improvements developed during the study (See **Table S2** for description of conditions used in each experimental batch). Protein immobilization was carried out in buffers based either on Lysis buffer (150 mM NaCl and 1% Triton X100 in Tris-Cl, pH 8) or Low stringency binding buffer (LSBB)(140 mM KCl, 5 mM NaCl, 1 mM K2HPO4, 2 mM MgSO4, 100 μM EGTA, 1 mM ZnSO4, and 0.1% Tween20 in 20 mM HEPES-HCl (pH 7). All DNA-protein reactions used LSBB. For GST-tagged proteins, we used glutathione magnetic beads (Sigma-Aldrich G0924-1ML), and for GFP-protein immobilization, we used GFP-Trap Magnetic Agarose" (Chromotek, gtma-100) for initial batches, and Anti-GFP antibody (ab290, Abcam) immobilized to Protein G Mag Sepharose® Xtra (Cytiva, 28-9670-70) for later batches, as the latter showed a higher success rate. All selections used 1μl of the magnetic bead slurry, a volume that in majority of the cases, according to manufacturers' information, contains excess protein binding capacity but is still visible in microwell plates, allowing quality control of the washing steps.

SELEX process: All of the protocols (described in **Table S2)** followed these general steps: 1) Affinity beads and 96-well plates were blocked with BSA for 15 minutes; 2) Beads and plates were washed to remove unbound BSA; 3) Protein was immobilized into beads for 1h on a shaker; 4) Beads were washed to remove nonspecific proteins and carryover DNA; 5) Protein coated beads were incubated with DNA ligand for 1h to allow the proteins to bind their target sites; 6) Unbound and weakly bound DNA ligands were removed with extensive washing; 7) DNA ligands were eluted by suspending the beads into heat elution buffer (0.4 μM forward and reverse primers, 1 mM EDTA and 1% Tween 20 in 10 mM Tris-Cl, pH 8) transferring the suspension into a conical PCR plate and heat treating it in a PCR machine using a program that cycled between temperatures of 98 and 60°C, in order to denature the proteins and DNA, use convection to drive the DNA into the solution, and to hybridize DNA to the amplification

748 primers; 8) Bead suspension obtained from heat elution was used as template in PCR
749 and qPCR reactions; 9) An additional DNA amplification cycle was performed with 2X
750 more primers and dNTPs to ensure that majority of the ligands are in fully double-
751 stranded state and 10) In batch YWO and an associated batch YWN, we tested a set of
752 same proteins with (YWO) or without (YWN) mung bean nuclease treatment to digest
753 partially single-stranded ligands from the experiments, in order to reduce enrichment of
754 DNA with artifactual sites, such as potential aptamers. This strategy proved effective in
755 HT-SELEX experiments and was used in the following batches (P, Q, R, and S). For
756 these last four batches, the TFs were mainly in alphabetical order. In each mung bean
757 nuclease reaction, the pH of the solution (PCR reaction) was first lowered by the
758 addition of 1:10 volume of 100 mM acetic acid, followed by addition of 1μl (0.75 units)
759 of the enzyme and incubation for one hour at 37°C.
760
761 **Sequencing**. Samples were prepared for sequencing by performing a PCR reaction
762 that indexes each sample and its selection cycle with a unique combination of i7 and i5
763 barcodes, followed by a double-stranding reaction with primers that target regions of
764 DNA outside indices (**Table S1**). Following this step, DNA libraries were pooled, purified
765 with a PCR purification kit (NEB, T1030S), and then subjected to Illumina sequencing
766 with 60bp reads at ~3M reads per sample (Donnelly Centre sequencing core facility).
767
768 **HT- and GHT-SELEX read processing and mapping.** HT-SELEX reads were filtered
769 by Phred quality score (Q >= 30 in at least 90% of bases). GHT-SELEX reads were
770 parsed with Trimmomatic[63] to remove the constant regions from genomic fragments that
771 were shorter than the sequencing read length (options:
772 ILLUMINACLIP:CustomAdapters.fa:2:5:5, LEADING:3, TRAILING:3 MINLEN:25). The
773 custom adapters in the fasta file were AGATCGGAAGAGCACACGTCTGAACTCCAG
774 and AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTTA. For GHT-SELEX, we
775 mapped trimmed reads to the human genome build hg38 with bowtie2 (options: --very-
776 sensitive, --no-unal). The mapped reads were further filtered using Samtools (options: -
777 F 1548, version 1.20)[64]. Data were processed further for MAGIX-based analyses: each
778 cycle of individual experiments was filtered through, sorting mapped fragments by read
779 name, deduplicating them by position (samtools command: markdup, option: -r), and
780 keeping only properly paired reads (option: –f 2).
781
782 **HT-SELEX k-mer enrichment-based comparison of protein production methods**.
783 To identify binding specificity differences conferred by either protein production method
784 or using a different clone type that contains all predicted DBDs (DBD or FL), we
785 identified 62 TFs for which we had at least one pair of successful experiments
786 performed with a lysate HT-SELEX experiment and another with either of the IVT
787 systems. For each successful experiment performed for these 62 TFs, we counted
788 occurrences of 3-, 5-, 7-, and 9-mers in the unique reads contained in the final two
789 cycles, as well as in the corresponding DNA library (cycle 0) using jellyfish (v2.3.1)[65].
790 For all possible experiment comparisons, $\log_2$ fold change was compared for each k-
791 mer size.
792

MAGIX statistical framework. At the core of MAGIX is a generative model that explicitly connects the enrichment of TF-bound genomic intervals to the fragment counts observed across GHT-SELEX cycles. MAGIX models how TF-bound intervals progressively occupy a higher proportion of the selected fragments pool in each cycle relative to the genomic background. These fragment proportions, in turn, are treated as latent variables in the model that, together with a sample-specific library size factor, determine the number of observed reads through a Poisson process. Consider the genomic interval $j \in [1, G]$, where G is the total number of unique genomic intervals that we are modeling. Assume that fragments originating from interval j have a starting abundance of a_j in the library. We also assume an exponential enrichment for the fragments, that in each cycle of SELEX, the abundance of these fragments changes by a factor of e^{b_j}, where b_j is the log fold-change in abundance per cycle (referred to as *enrichment coefficient*), conceptually associated with biophysical parameters such as binding energies. Therefore, at cycle t, the abundance of the fragment originating from interval j is given by:

$$f_j^t = a_j \left(e^{b_j}\right)^t = a_j e^{t b_j}$$

For convenience, we work with the logarithm of abundance, $y_j = \log f_j$, transforming the exponential equation above to a linear equation as follows:

$$E\left(\mathrm{y}_{tj}\right) = log\left(f_j^t\right) = \log a_j + t b_j,$$

which can be seen as the linear multiplication of a feature vector $\mathbf{x}_i = [1\ t_i]$ for all the samples i (and across different cycles) corresponding to the TF of interest, and the interval-specific parameters $\boldsymbol{\beta}_j = [\log a_j\ \ b_j]$.
We note that to accurately model the enrichment of each fragment per cycle, other factors also need to be taken into consideration, such as background or batch effects, and therefore, the linear equation above needs to be fitted not only to the samples that correspond to the TF of interest, but also to samples from other experiments. We embed these dependencies in a design matrix $\boldsymbol{X} \in \mathbb{R}^{N \times K}$, where N is the total number of samples and K is the number of variables to consider, including an intercept term (whose coefficient will correspond to $\log a_j$ above), a term for the SELEX cycle t (variable for the samples corresponding to a same TF), and other terms for batch and background effects. In addition to the variables included in $\boldsymbol{X}$, the abundance of each fragment in each sample depends on a sample-specific scaling factor that is often referred to as the library size. Assume that this library effect, for each sample $i \in [1, N]$, is the scaling factor s_i (in logarithmic scale). Therefore:

$$E\left(\mathrm{y}_{ij}\right) = \hat{\mathrm{y}}_{ij} = \mathbf{x}_i \cdot \boldsymbol{\beta}_j + s_i$$

Here, $\hat{y}_{ij}$ corresponds to the expected logarithm of the abundance of interval j in sample i, $\mathbf{x}_i \in \mathbb{R}^K$ is a vector representing the i'th row of the design matrix $\mathbf{X}$ (i.e., the sample-level variables for sample i), and $\boldsymbol{\beta}_j \in \mathbb{R}^K$ is an interval-specific vector of coefficients for the K variables included in the model.

We note that the equation above does not have a unique solution. For example, any $\Delta\boldsymbol{\beta}$ can be added to $\boldsymbol{\beta}_j$, followed by subtraction of $\mathbf{X}\Delta\boldsymbol{\beta}$ from $\mathbf{s}$, without any change in $\hat{\mathbf{y}}_j$:

$$\mathbf{x}_i \cdot \boldsymbol{\beta}_j + \mathbf{s} = \mathbf{x}_i \cdot \left(\Delta\boldsymbol{\beta} + \boldsymbol{\beta}_j\right) + (s_i - \mathbf{x}_i \cdot \Delta\boldsymbol{\beta})$$

Therefore, to make the model identifiable, we limit $\boldsymbol{\beta}_j$ so that $\Sigma_{j\in[1,G]}\boldsymbol{\beta}_j=\mathbf{0}$, where $\mathbf{0}$ is the zero vector of length K. This constraint is also useful since it means that, across all G intervals, the mean of each coefficient in $\boldsymbol{\beta}$, including the coefficient for the SELEX cycle, is zero; in other words, the enrichment per cycle for each interval is calculated relative to the mean of all G intervals.
To incorporate the experimental noise in the logarithm of the abundance of interval j in sample i (i.e., building the real distribution of y_{ij}), we modeled it as a Gaussian random variable whose mean is given by $\hat{y}_{ij}$ (the linear model above) with a sample-specific variance σ_i^2:

$$y_{ij} \sim \mathcal{N}\left(\hat{y}_{ij}, \sigma_i^2\right)$$

To complete the Bayesian framework, we also assume a multivariate Gaussian prior for $\boldsymbol{\beta}_j$:

$$\boldsymbol{\beta}_j \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

Here, $\boldsymbol{\Sigma}$, the covariance matrix of the prior distribution, is shared across all intervals.

Altogether, the equations above form the following Bayesian model:

$$\boldsymbol{\beta}_j \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$
$$s_i \sim \mathrm{U}(-\infty, +\infty)$$
$$\sigma_i^2 \sim \mathrm{U}(0, +\infty)$$
$$\hat{y}_{ij} = \mathbf{x}_i \cdot \boldsymbol{\beta}_j + s_i$$
$$y_{ij} \sim \mathcal{N}\left(\hat{y}_{ij}, \sigma_i^2\right)$$

Assuming that the values for y_{ij} are directly observed, we can obtain the maximum a posteriori (MAP) estimates of the parameters $\boldsymbol{\beta}_j$ (for $j\in[1,G]$), $\mathbf{s}$, and σ_i^2 (for $i\in[1,N]$) through a block coordinate descent algorithm, as previously described[48].
The prior covariance matrix $\boldsymbol{\Sigma}$ is a hyperparameter that is obtained using an empirical Bayes approach. More specifically, we first obtain the maximum likelihood estimate (MLE) of the parameters $\boldsymbol{\beta}_j$, s_i, and σ_i^2 without assuming any prior on $\boldsymbol{\beta}_j$, and then estimate the values of $\sigma_{\beta k}^2$, the variance of each element k in $\boldsymbol{\beta}_j$, using the MLE solutions of all $\boldsymbol{\beta}_j$ coefficients. The covariance matrix $\boldsymbol{\Sigma}$ is then constructed as $\boldsymbol{\Sigma}=\mathrm{diag}(\sigma_{\beta 1}^2,\ldots, \sigma_{\beta K}^2)$.

We note, however, that the log-abundance values y_{ij}'s are not directly observable in GHT-SELEX data. Instead, we observe m_{ij}, the count of reads mapping to interval j in sample i (see **HT- and GHT-SELEX read processing and mapping**). This parameter adds another step to the framework, leading to the following hierarchical Bayesian model:

$$\boldsymbol{\beta}_j \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

$$s_i \sim \mathrm{U}(-\infty, +\infty)$$
$$\sigma_i^2 \sim \mathrm{U}(0, +\infty)$$
$$\hat{y}_{ij} = \mathbf{x}_i \cdot \boldsymbol{\beta}_j + s_i$$
$$y_{ij}^* \sim \mathcal{N}\left(\hat{y}_{ij}, \sigma_i^2\right)$$
$$m_{ij} \sim \mathrm{Pois}\left(e^{y_{ij}^*}\right)$$

Here, y^*_{ij} is the logarithm of the true abundance of fragment j in sample i, which is latent. We obtain the MAP estimates of the parameters $\boldsymbol{\beta}_j$ (for $j \in [1,G]$), s_i, and σ_i^2 (for $i \in [1,N]$) using an expectation maximization (EM) algorithm, in which at each E-step we obtain the expected value of each y^*_{ij} given the observed read count m_{ij} and the current model parameters, followed by re-estimation of the model parameters in the M-step, similar to a previous method established for EM optimization of Poisson-lognormal models[48].

Identifying GHT-SELEX peaks with MAGIX. The statistical framework described above calculates the rate of enrichment across GHT-SELEX cycles for a set of given genomic intervals, e.g., a set of candidate peak regions. Below, we describe how we identify such candidate peaks.

For each TF, an aggregate BAM file is created across replicates and cycles. For each genomic region with continuous non-zero coverage in this aggregate BAM file, the maximum coverage is calculated. Regions with a maximum coverage smaller than a threshold are discarded, with the threshold being set to the total fragment count in the aggregate BAM file divided by 2×10^6. Then, per region, the coordinate with the highest coverage is defined as the summit. The candidate peaks are defined as the 200 bp regions centered on the summits.

Once the candidate peaks are identified, their count profiles are calculated from the original (unaggregated) BAM files (per cycle and replicate) and used as input to MAGIX to calculate MAGIX scores (enrichment coefficients), using prefixed library sizes that are estimated by fitting a similar MAGIX model to count profiles of 200-bp non-overlapping genomic intervals (a total of ~13M bins). For each candidate peak, we also calculate a P-value, representing the statistical significance of the enrichment coefficient (the null hypothesis is that the enrichment coefficient is zero). To do so, we obtain a maximum likelihood estimate of the model coefficients and perform a likelihood ratio test (LRT) against a reduced model in which the enrichment coefficient is restricted to zero. The source code for MAGIX is available at https://github.com/csglab/MAGIX.

MAGIX peak sets analyzed throughout these papers and provided at codebook.ccbr.utoronto.ca are derived from a merge of all successful experiments that contain full set of TF's DBDs (marked with "DBD" or "FL") prior to running MAGIX. Experiments performed with specific subsets of DBDs (Marked with "DBD1", "DBD2", or "DBD3") are analyzed separately.

Selection of thresholds for peak sets. We sorted the GHT-SELEX peaks by their MAGIX score, or as named in the peaks BED files, *coefficient.br*, which estimates cycle enrichment. Similarly, we sorted the merged ChIP-seq peaks by P-value. Then, for different values of N (between 100 and the total number of peaks), we took the top N peaks for both peaks sets and calculated the Jaccard index (= O/(2N-O), in which O is the intersection of peaks). To eliminate the error in the cases when one peak in a set overlaps multiple peaks in another set, we used the average of the overlaps for the intersection (i.e., O=(O1+O2)/2, in which O1 is the number of peaks in set1 overlapping with any peaks in set2 and vice versa). The value of N that yielded the maximum Jaccard value was identified, and the threshold for each peak set was taken as that which yielded this maximum N. The same process was applied to compare PWM-predicted binding sites and ChIP-seq peaks.

GHT-SELEX experiment overlap comparison. To investigate the reproducibility of GHT-SELEX profiles, we focused on experiment pairs with the same TF construct and protein production approach. Per replicate, we fitted genome-wide MAGIX models to 200-bp non-overlapping genomic intervals (~13M genomic intervals), resulting in a model coefficient per genomic bin. The exponential of this coefficient represents the fold-enrichment per cycle, which we visualize as scatterplots and use to calculate the Pearson correlation between replicates.

Motif derivation and assessment of experiment success.
We evaluated the success of GHT- and HT-SELEX experiments systematically alongside other Codebook datasets generated with ChIP-seq, protein binding microarray, and SMiLE-seq experiments (See **Table S7** for a summary of all experiments), by performing motif discovery with nine different programs and then testing their performance in all experiments for the same TF. Experiments were classified by expert curation as successful if they produced motifs that scored well in data generated by other methods (preferably, and in most cases) and, in rare cases, in replicate data from the same experimental method, when other methods failed. All motifs can be browsed at https://mex.autosome.org. See accompanying article for detailed description of the process[66].

Representative motif selection.
For each TF, a list of up to 60 motifs was automatically assembled based on auROC, auPRC, and motif centrality metrics performance across all successful experiments[66]. These motifs were then assessed manually by an assembly of authors to select a representative motif. Besides scoring across all data sets, we considered whether the motif appears likely to describe the inherent specificity of the TF, rather than, for example, a larger sequence that is partially derived from a genomic repeat element (a scenario common with KRAB zinc fingers). In the cases where several motifs displayed highly similar performance, we prioritized motifs with higher information content. See[66] for further details and **Supplementary Data S1** for the representative motifs.

Comparing PWM scoring methods. To create *in silico* predicted binding sites for a TF, we first scanned the genome using the generated PWM (see the Codebook overview

manuscript for the details on PWM selection), using MOODS[67] with a p-value threshold of 0.0001. We then merged the clusters of PWM hits with a distance less than 200bp between neighboring hits, since this is the median length of ChIP-seq fragments, and the task is predicting in vivo binding sites. Singleton PWM hits and boundary hits were also expanded to have a width of at least 200bp. The clusters of PWM hits were re-scored using sum-of-affinity (i.e., with PWM log-odds scores at each base converted to linear/probability space, prior to calculation of the sum) and maximum-affinity methods, by either applying a sum or maximum, respectively, over the PWM scores of the cluster members. The resulting sites were sorted by their new score and processed through the same optimization procedure described above for peaks, to maximize their overlap with ChIP-seq peaks.

Modeling alternative C2H2-zf binding modes with RCADEEM. RCADEEM uses a hidden Markov model (HMM) to represent multiple, alternative DNA-binding motifs, each corresponding to the binding preference of a C2H2-zf array. Briefly, the DNA sequences (e.g., GHT-SELEX peaks) are modeled as sequences generated from a discrete Markov process with hidden states that include a background state (S_0) and M motif states S_m ($m \in [1,M]$). The background state, with marginal probability π_0, emits each nucleotide n with probability $b_0(n)$ ($\Sigma_n b_0(n)=1$). The background state can transition to itself (i.e., consecutive DNA nucleotides can be generated from the background state) with probability $a_{0,0}$, or to each motif state S_m ($m \in [1,M]$) with probability $a_{0,m}$ ($a_{0,0}+\Sigma_m a_{0,m}=1$). Each motif m, with marginal probability π_m, generates a sequence of length l_m, with each nucleotide n at position i emitted with probability $b_{m,i}(n)$ ($\Sigma_n b_{m,i}(n)=1$ $\forall m \in [1,M]$, $i \in [1,l_m]$). Note that, for each $m \in [1,M]$, the values $b_{m,i}(n)$ form a position-specific frequency matrix (PFM, i.e. the exponential of the classical log-odds PWM) with width l_m, which is fixed to be 3 times the number of zinc finger domains in the array represented by motif m, as each zinc finger domain binds to three nucleotides. Finally, each motif state S_m transitions to the background state with probability $a_{m,0}=1$.

We start the model by including the motifs representing all possible consecutive zinc finger domain arrays[50]. We initialize the emission probabilities $b_{m,i}(n)$ for each motif m using the PFM predicted for the associated zinc finger array by a previously created C2H2-zf recognition code[51]—this recognition code is a machine learning model that, given the sequence of a zinc finger array, predicts the expected binding preference. The HMM parameters, including all marginal state probabilities, state transition probabilities, and emission probabilities, are then optimized via expectation maximization using Baum–Welch algorithm. Then, each of the optimized PFMs is tested for (*i*) enrichment of the motif in actual sequences compared to dinucleotide-shuffled sequences, and (*ii*) similarity to the original recognition code-predicted PFM. To achieve (*i*), for each position x in each DNA sequence k, we calculate $\gamma_{k,x}(S_m)$, the probability that it was generated from motif state S_m, using the forward-backward algorithm. The motif score for DNA sequence k is then calculated as $\Sigma_x \gamma_{k,x}(S_m)/l_m$, representing the expected number of times the state S_m is seen in sequence k. For each motif m, these scores are calculated both for actual GHT-SELEX peak sequences and their dinucleotide-shuffled version. Then, the top 100 sequences with the largest scores for each motif are tested to see whether they are enriched in the motif compared to shuffled sequences (Fisher's

exact test, FDR≤0.01). Motifs that do not pass this cutoff are removed from the model. To achieve (*ii*), each HMM-optimized PFM is first converted to log-scale (representing a PWM), followed by the calculation of the Pearson correlation of the PWM entries with those predicted by the recognition code. Pearson correlations are then converted using Fisher transformation in order to calculate a P-value, followed by removal of motifs that do not pass the FDR cut-off ≤0.01. The remaining motifs are then used to reconstruct a smaller HMM, similar to the procedure described above, followed by another round of EM optimization. This procedure is repeated until all motifs pass the cut-offs for enrichment in GHT-SELEX sequences while maintaining significant similarity to the original recognition code-predicted sequences.

To visualize the binding modes predicted by RCADEEM, the resulting PWMs are used to identify their best match in each of the input sequences using AffiMx[52]. Then, for each sequence, the PWM with the highest weighted HMM score on the best match is kept as the predicted binding mode. To align the sequences, offsets are calculated based on the corresponding C2H2-zf domains (**Figures 6a-f**). C2H2-zf proteins were categorized based on their alternative usage of C2H2-zf domains (i.e., Multiple DBDs, Finger shift, Canonical, and Core with extensions; **Figure 6**) through an expert-curated evaluation (**Table S10**). To make a motif model for each binding mode, we manually selected representative peaks corresponding to each binding mode over the 2000 GHT-SELEX peaks with the highest enrichment coefficient. The sequence (already aligned by RCADEEM) and C2H2-zf domain array coordinates of these peaks were used to create PFMs. The resulting PFMs for those C2H2-zf TFs are available in **Supplementary Data S2** and online at https://cisbp.ccbr.utoronto.ca[51]. The logos, coordinates, selected sequences, annotated sequence heatmaps, and associated metadata are available online at https://codebook.ccbr.utoronto.ca. The source code for RCADEEM is available at https://github.com/csglab/RCADEEM.

Comparison of C2H2 DBDs. C2H2 DBD similarities were compared by pairwise alignment with Needleman-Wunsch algorithm, as implemented in the R-package Biostrings and counting substitutions, insertions, and unmatched flanking bases as edits. DNA-binding functionality scores and predicted motif similarity for the DBDs were analyzed as described previously[68].

DATA AVAILABILITY

The sequencing raw data for the HT-SELEX and GHT-SELEX experiments have been deposited into the SRA database under identifiers PRJEB61115 (HT-SELEX) and PRJEB76622 (GHT-SELEX). Additionally, genomic interval information generated for the GHT-SELEX has been deposited into GEO under accession GSE278858. The entire Codebook data structure, with many accessory files and browsable results at is available at https://codebook.ccbr.utoronto.ca. Larger collection of motifs generated for these experiments in an accompanying study[49] can be browsed at https://mex.autosome.org.

CODE AVAILABILITY

Source code for MAGIX and RCADEEM is available in GitHub (https://github.com/csglab/MAGIX and https://github.com/csglab/RCADEEM).

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.
-
- Russian Science Foundation grant 24-14-20031 to F.A.K.
- MSHERF grant 075-15-2025-014 (prev. 075-15-2024-666)
- State assignment 125091010189-3 (FFRW-2025-010)
- Swiss National Science Foundation grant (no. 310030_197082) to B.D.
- Marie Skłodowska-Curie (no. 895426) and EMBO long-term (1139-2019) fellowships to J.F.K.
- NIH grants R01HG013328 and U24HG013078 to M.T.W., T.R.H., and Q.M.
- NIH grants R01AR073228, P30AR070549, and R01AI173314 to M.T.W.
- NIH grant P30CA008748 partially supported Q.M.
- Canada Research Chairs funded by CIHR to T.R.H. and H.S.N.
- Ontario Graduate Scholarships to K.U.L and I.Y.
- A.J. was supported by Vetenskapsrådet (Swedish Research Council) Postdoctoral Fellowship (2016-00158)
- The Billes Chair of Medical Research at the University of Toronto to T.R.H.
- EPFL Center for Imaging
- Institutional funding from EPFL
- Resource allocations from the Digital Research Alliance of Canada

Author contributions

A.J. and T.R.H. designed the study. A.J. and A.W.H.Y. performed the SELEX experiments with assistance from H.Z., R.R., and A.B. A.H.C. and H.S.N. developed and performed MAGIX and RCADEEM. A.H.C., A.F., K.U.L. and A.J. performed sequence and data analyses with assistance from A.B.; A.J., K.U.L and A.H.C. prepared the illustrations. M.A. and I.V.K. orchestrated the data and motif repositories. A.J, A.H.C, A.F, K.U.L, H.S.N and T.R.H. wrote the paper. All authors contributed to data analysis and reviewed the manuscript.

SUPPLEMENTARY TABLES AND DOCUMENTS

Extended Data Fig. 1: a. Example of actual read count data for CTCF over five replicates of four cycles, illustrating enrichment patterns, fitted coefficients (right), and estimated library sizes (bottom). **b.** A brief overview of the statistical framework of the generative model of MAGIX. Open circles, closed circles, and the diamonds represent latent variables, observed variables, and deterministic computations, respectively. s_i: library size for sample i; $\mathbf{x}_i$: vector of sample-level variables for sample i, including an intercept term and a term for the SELEX cycle, in addition to other terms for batch and background effects; $\boldsymbol{\beta}_j$: vector of model coefficients for interval j; m_{ij}: number of observed reads mapping to interval j in sample i. See **Methods** for description of other variables..

Extended Data Fig. 2: TF success rate by protein production method and structural class. Figure displays bar graphs of success rates for putative TFs based on their structural class showing a bar for TFs based on success on if the protein was successful in any of the production methods and then individually based on protein production methods. Number of TFs tested in each category is listed with the name of the structural class.

Extended Data Fig. 3: TF expression levels are plotted for four categories based on experiment success and protein production. Data is shown for the full range of fluorescence (*top)* and for the region with most of the signal (*bottom).* Note that even though the successful samples have generally higher expression levels in both protein production systems it is not an effective predictor of experiment success.

Table S1. HT- and GHT-SELEX ligand sequences and descriptions. Table lists the oligonucleotide sequences used in the assay and describes how they anneal with each other on the synthesis and amplification steps.

Table S2. Experimental batch specific protocol details. Table lists the reagents and experimental conditions that varied between different experimental batches.

Table S3 GHT-SELEX-experiment metadata. Table lists all GHT- and HT-SELEX experiments performed in this study indicating: unique experiment identifier; human readable identifier; plasmid identifiers; HNGC symbol; experimental batch; construct type; protein production approach, position in the 96-well; sequencing strategy; number of selection cycles; and whether the experiment was successful or not. Note that GFP control experiments (i.e. empty plasmids) are also included in the table (5 GHT-SELEX and 7 HT-SELEX).

Table S4: Summary data of GHT-SELEX replicates. Table shows counts of successful and unsuccessful experiments for all analyzed TFs, listing both direct experiments, that were performed with the same construct and expression system and indirect replicates based on TF identity, regardless of protein production method or insert.

Table S5: Reproducibility of GHT-SELEX experiments. Table lists 102 experiment pairs with Pearson correlation of coefficient BR values observed in genomic bins.

Table S6: Annotations of experiment pairs shown in Document S3, production method derived differences in target specificity data. Table shows Pearson correlations of k-mer enrichment between pairs of replicate experiments.

Table S7: Summary of Codebook experiments and motifs. Table lists all GHT-SELEX, HT-SELEX, Protein binding microarray, SMiLE-seq and ChIP-seq experiments performed in the Codebook consortium and indicates experiments that were used to derive the representative PFMs.

Table S8: List and annotations of western blot analyses for lysates and wheat germ extracts. Each row describes one of the 97 western blot wells in the associated **Document S4** and provides information on; Associated experiment; HNGC identifier; Construct type of the tested TF productions; structural class of the DBD; GHT-SELEX result; Gel and well; All wells with this TF; Protein expression method; Gel examination based annotation of the protein production success; Whether the western blot shows partial bands; If the experiment is paired comparison of different construct for the same TF (FL vs. DBD); whether a successful experiment for C2H2 TF displayed modular binding activities based on RCADEEM analysis; Is a IVT vs lysate production comparison; eGFP fluorescence measured for the protein, protein construct amino-acid count and estimated molecular weight.

Table S9: Genomic region overlap of GHT-SELEX and ChIP-seq peaks and PWM-predicted target regions. Table shows the overlap of optimal ChIP-seq peaks with GHT-SELEX/MAGIX and PWM based predictions for each of the TFs where both datasets were available. Columns show the highest Jaccard coefficient between each pair of datasets and the number of peaks that yielded it.

Table S10: C2H2-zf protein DNA-binding mode annotation. Table lists the 86 C2H2 TFs for which RCADEEM result was obtained (out of 120 total C2H2-zf TFs with GHT-SELEX data available) with information of: Total number of C2H2 zinc finger domains; amino acid gaps between these DBDs; number of distinct motifs bound by the TF; modular binding activity annotated for it; whether the protein is likely to contain zinc fingers obtained from internal duplications and whether data was obtained from experiments that expressed different subsets of the TFs C2H2-zf domains.

Table S11: Intra-protein C2H2-zf domain duplication dataset. Table displays all pairs of human C2H2 DBDs that are separated from each other by five or less edits.

Document S1: GHT-SELEX reproducibility. Document shows scatterplots of coefficient-BR scores for all possible 15 million non-overlapping 200bp genomic bins for

pairs of successful experiments that have been expressed with the same protein production system. See associated **Table S5** for details and summary data.

Document S2: Motif centrality and enrichment in GHT-SELEX/MAGIX peaks and its correspondence with ChIP-seq peaks. Same plots as in **Figure 2c** and **Figure 4c** for all the TFs and DBD constructs in this study with successful GHT-SELEX experiments. *Top*, top-ranked TF PWM (highest AUROC on GHT-peaks as determined by [49]). *Middle*, Distribution of PWM hits within the 5,000 highest scoring MAGIX peaks. Solid red lines represent the mean PWM hit position within MAGIX peaks and dashed lines represent one standard deviation about the mean. *Bottom*, Enrichment of ChIP-seq peaks and PWM hits within MAGIX peaks. Orange line shows the proportion of peaks (in a sliding window of 500 peaks over the ranked peaks, with a step size of 50) that overlap with a ChIP-seq peak (at MACS threshold $P < 0.001$). Black line shows the AUROC for PWM affinity scores of MAGIX peaks in the same window vs. 500 random genomic sites.

Document S3: Analysis of k-mer enrichment in HT-SELEX experiment pairs. (*First page*) Description of likely mechanisms that cause systematic differences between replicate pairs between. (*Pages 2-290*) Enrichment of 3,5,7 and 9bp k-mers are shown for 289 combinations of HT-SELEX experiments, see **Table S6** for details.

Document S4: Western blot analyses for lysate and wheat germ extract based proteins. (*First page*) Description of the document and considerations of effects of partial protein products on modular C2H2 binding patterns. **(***Second page, Panel A)* Summary of blot results. (*Second page, Panel B*) Modular C2H2-zfs shown in Fig6; (*Second page, Panel C*) C2H2-zfs ZNF233, ZNF497 and ZNF853, show sites in RCADEEM analysis that could have been affected by C-terminal truncations. Note, only ZNF853 show smaller than expected size fragments. (*Pages 2-6*) All western blot analyses: Wells have been annotated to display HNGC symbol of the TF, its expression construct along with gel and well references for all positions with the same TF; Whether the expression was deemed successful (black text) or unsuccessful (red text); If this protein expression yielded a successful experiment in GHT-SELEX (green bar) or not (red bar); If the TF is a C2H2 protein (black bar on top); If the C2H2 TF displayed modular binding specificity (blue bar) and the expected size of the fusion protein (asterisk over the blot).

Data S1: Representative motifs for TFs analyzed in Codebook Consortium. PFMs for all of the representative motifs chosen in the Codebook project.

Data S2: PFMs of C2H2-zf proteins with alternative DNA binding modes. PFMs representing the different binding modes of C2H2-zf proteins.

1245 **REFERENCES**

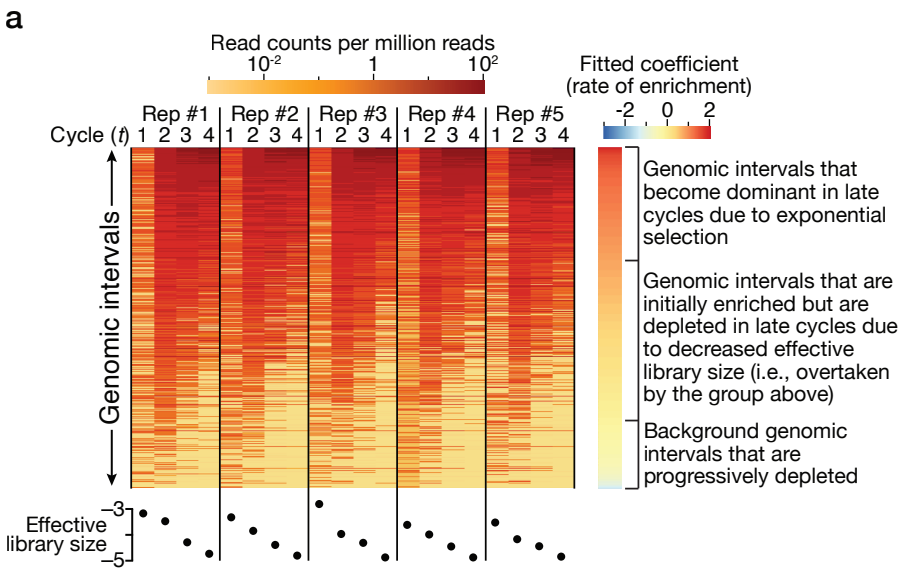
1246 1. Stormo, G.D. & Zhao, Y. Determining the specificity of protein-DNA interactions. *Nat Rev*
1247 *Genet* **11**, 751-60 (2010).
1248 2. Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor
1249 binding profiles and its web framework. *Nucleic Acids Res* **46**, D1284 (2018).
1250 3. Bernard, B., Thorsson, V., Rovira, H. & Shmulevich, I. Increasing coverage of transcription
1251 factor position weight matrices through domain-level homology. *PLoS One* **7**, e42779
1252 (2012).
1253 4. Lambert, S.A. *et al.* Similarity regression predicts evolution of transcription factor
1254 sequence specificity. *Nat Genet* **51**, 981-989 (2019).
1255 5. Wunderlich, Z. & Mirny, L.A. Different gene regulation strategies revealed by analysis of
1256 binding motifs. *Trends Genet* **25**, 434-40 (2009).
1257 6. Patel, Z.M. & Hughes, T.R. Global properties of regulatory sequences are predicted by
1258 transcription factor recognition mechanisms. *Genome Biol* **22**, 285 (2021).
1259 7. Wasserman, W.W. & Sandelin, A. Applied bioinformatics for the identification of
1260 regulatory elements. *Nat Rev Genet* **5**, 276-87 (2004).
1261 8. Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on
1262 ChIP-Seq data. *Nat Methods* **5**, 829-34 (2008).
1263 9. Marinov, G.K., Kundaje, A., Park, P.J. & Wold, B.J. Large-Scale Quality Analysis of
1264 Published ChIP-seq Data. *G3 (Bethesda)* (2013).
1265 10. Consortium, E.P. *et al.* Expanded encyclopaedias of DNA elements in the human and
1266 mouse genomes. *Nature* **583**, 699-710 (2020).
1267 11. Rhee, H.S. & Pugh, B.F. Comprehensive Genome-wide Protein-DNA Interactions
1268 Detected at Single-Nucleotide Resolution. *Cell* **147**, 1408-19 (2011).
1269 12. Long, H.K., Prescott, S.L. & Wysocka, J. Ever-Changing Landscapes: Transcriptional
1270 Enhancers in Development and Evolution. *Cell* **167**, 1170-1187 (2016).
1271 13. Zinzen, R.P., Girardot, C., Gagneur, J., Braun, M. & Furlong, E.E. Combinatorial binding
1272 predicts spatio-temporal cis-regulatory activity. *Nature* **462**, 65-70 (2009).
1273 14. Liu, X., Lee, C.K., Granek, J.A., Clarke, N.D. & Lieb, J.D. Whole-genome comparison of
1274 Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in
1275 target site selection. *Genome Res* **16**, 1517-28 (2006).
1276 15. Kim, T.H. *et al.* Analysis of the vertebrate insulator protein CTCF-binding sites in the
1277 human genome. *Cell* **128**, 1231-45 (2007).
1278 16. Fu, Y., Sinha, M., Peterson, C.L. & Weng, Z. The insulator binding protein CTCF positions
1279 20 nucleosomes around its binding sites across the human genome. *PLoS Genet* **4**,
1280 e1000138 (2008).
1281 17. Walker, M. *et al.* Affinity-seq detects genome-wide PRDM9 binding sites and reveals the
1282 impact of prior chromatin modifications on mammalian recombination hotspot usage.
1283 *Epigenetics Chromatin* **8**, 31 (2015).
1284 18. Morgunova, E. *et al.* Two distinct DNA sequences recognized by transcription factors
1285 represent enthalpy and entropy optima. *Elife* **7**(2018).
1286 19. Stormo, G.D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16-23
1287 (2000).

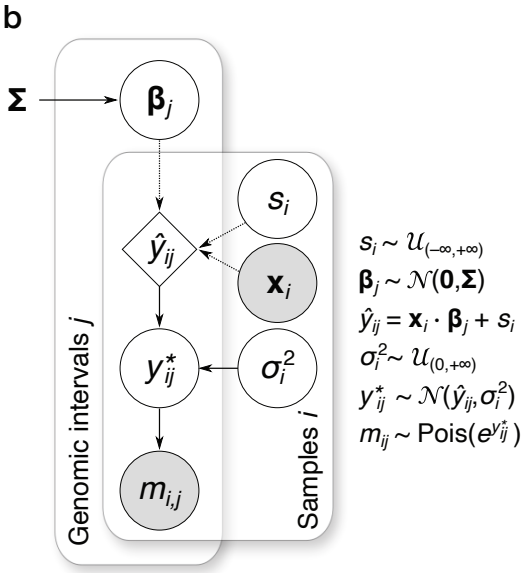
1288 20. Rohs, R. *et al.* The role of DNA shape in protein-DNA recognition. *Nature* **461**, 1248-53
1289 (2009).
1290 21. Zhao, Y., Ruan, S., Pandey, M. & Stormo, G.D. Improved models for transcription factor
1291 binding site identification using nonindependent interactions. *Genetics* **191**, 781-90
1292 (2012).
1293 22. Jolma, A. *et al.* DNA-dependent formation of transcription factor pairs alters their
1294 binding specificity. *Nature* **527**, 384-8 (2015).
1295 23. Horton, C.A. *et al.* Short tandem repeats bind transcription factors to tune eukaryotic
1296 gene expression. *Science* **381**, eadd1250 (2023).
1297 24. Kribelbauer, J.F., Rastogi, C., Bussemaker, H.J. & Mann, R.S. Low-Affinity Binding Sites and
1298 the Transcription Factor Specificity Paradox in Eukaryotes. *Annu Rev Cell Dev Biol* **35**,
1299 357-379 (2019).
1300 25. Thomson, J.P. *et al.* CpG islands influence chromatin structure via the CpG-binding
1301 protein Cfp1. *Nature* **464**, 1082-6 (2010).
1302 26. Lambert, S.A. *et al.* The Human Transcription Factors. *Cell* **175**, 598-599 (2018).
1303 27. Wolfe, S.A., Nekludova, L. & Pabo, C.O. DNA recognition by Cys2His2 zinc finger proteins.
1304 *Annu Rev Biophys Biomol Struct* **29**, 183-212 (2000).
1305 28. Klug, A. The discovery of zinc fingers and their applications in gene regulation and
1306 genome manipulation. *Annu Rev Biochem* **79**, 213-31 (2010).
1307 29. Stubbs, L., Sun, Y. & Caetano-Anolles, D. Function and Evolution of C2H2 Zinc Finger
1308 Arrays. *Subcell Biochem* **52**, 75-94 (2011).
1309 30. Nakahashi, H. *et al.* A genome-wide map of CTCF multivalency redefines the CTCF code.
1310 *Cell Rep* **3**, 1678-1689 (2013).
1311 31. Kieffer-Kwon, K.R. *et al.* Interactome maps of mouse gene regulatory domains reveal
1312 basic principles of transcriptional regulation. *Cell* **155**, 1507-20 (2013).
1313 32. Han, B.Y., Foo, C.S., Wu, S. & Cyster, J.G. The C2H2-ZF transcription factor Zfp335
1314 recognizes two consensus motifs using separate zinc finger arrays. *Genes Dev* **30**, 1509-
1315 14 (2016).
1316 33. Najafabadi, H.S. *et al.* C2H2 zinc finger proteins greatly expand the human regulatory
1317 lexicon. *Nat Biotechnol* (2015).
1318 34. Najafabadi, H.S., Albu, M. & Hughes, T.R. Identification of C2H2-ZF binding preferences
1319 from ChIP-seq data using RCADE. *Bioinformatics* **31**, 2879-81 (2015).
1320 35. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA
1321 ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505-10 (1990).
1322 36. Jolma, A. *et al.* DNA-Binding Specificities of Human Transcription Factors. *Cell* **152**, 327-
1323 39 (2013).
1324 37. Jolma, A. *et al.* Multiplexed massively parallel SELEX for characterization of human
1325 transcription factor binding specificities. *Genome Res* **20**, 861-73 (2010).
1326 38. Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human
1327 transcription factors. *Science* **356**(2017).
1328 39. Reiss, D.J. & Mobley, H.L. Determination of target sequence bound by PapX, repressor of
1329 bacterial motility, in flhD promoter using systematic evolution of ligands by exponential
1330 enrichment (SELEX) and high throughput sequencing. *J Biol Chem* **286**, 44726-38 (2011).

1331 40. Ishihama, A., Shimada, T. & Yamazaki, Y. Transcription profile of Escherichia coli: genomic
1332 SELEX search for regulatory targets of transcription factors. *Nucleic Acids Res* **44**, 2058-
1333 74 (2016).
1334 41. Illingworth, R.S. *et al.* Orphan CpG islands identify numerous conserved promoters in the
1335 mammalian genome. *PLoS Genet* **6**, e1001134 (2010).
1336 42. O'Malley, R.C. *et al.* Cistrome and Epicistrome Features Shape the Regulatory DNA
1337 Landscape. *Cell* **165**, 1280-1292 (2016).
1338 43. Jolma, A. *et al.* Perspectives on Codebook: sequence specificity of uncharacterized
1339 human transcription factors. *bioRxiv*, 2024.11.11.622097 (2024).
1340 44. Berger, M.F. *et al.* Compact, universal DNA microarrays to comprehensively determine
1341 transcription-factor binding site specificities. *Nat Biotechnol* **24**, 1429-35 (2006).
1342 45. Isakova, A. *et al.* SMiLE-seq identifies binding motifs of single and dimeric transcription
1343 factors. *Nat Methods* **14**, 316-322 (2017).
1344 46. Lambert, S.A. *et al.* The Human Transcription Factors. *Cell* **172**, 650-665 (2018).
1345 47. Razavi, R. *et al.* Extensive binding of uncharacterized human transcription factors to
1346 genomic dark matter. *bioRxiv*, 2024.11.11.622123 (2024).
1347 48. Schmitges, F.W. *et al.* Multiparameter functional diversity of human C2H2 zinc finger
1348 proteins. *Genome Res* **26**, 1742-1752 (2016).
1349 49. Vorontsov, I.E. *et al.* Cross-platform DNA motif discovery and benchmarking to explore
1350 binding specificities of poorly studied human transcription factors. *bioRxiv*,
1351 2024.11.11.619379 (2024).
1352 50. Gralak, A.J. *et al.* Identification of methylation-sensitive human transcription factors
1353 using meSMiLE-seq. *bioRxiv*, 2024.11.11.619598 (2024).
1354 51. Weirauch, M.T. *et al.* Determination and inference of eukaryotic transcription factor
1355 sequence specificity. *Cell* **158**, 1431-43 (2014).
1356 52. Razavi, R. *et al.* Extensive binding of uncharacterized human transcription factors to
1357 genomic dark matter. *bioRxiv*, 2024.11.11.622123 (2024).
1358 53. Birke, M. *et al.* The MT domain of the proto-oncoprotein MLL binds to CpG-containing
1359 DNA and discriminates against methylation. *Nucleic Acids Res* **30**, 958-65 (2002).
1360 54. Khetan, S., Carroll, B.S. & Bulyk, M.L. Multiple overlapping binding sites determine
1361 transcription factor occupancy. *Nature* (2025).
1362 55. Stormo, G.D. & Fields, D.S. Specificity, free energy and information content in protein-
1363 DNA interactions. *Trends Biochem Sci* **23**, 109-13 (1998).
1364 56. Weirauch, M.T. *et al.* Evaluation of methods for modeling transcription factor sequence
1365 specificity. *Nat Biotechnol* **31**, 126-34 (2013).
1366 57. Kuznetsov, V.A. Mathematical Modeling of Avidity Distribution and Estimating General
1367 Binding Properties of Transcription Factors from Genome-Wide Binding Profiles.
1368 *Methods Mol Biol* **1613**, 193-276 (2017).
1369 58. Alexandrov, I., Kazakov, A., Tumeneva, I., Shepelev, V. & Yurov, Y. Alpha-satellite DNA of
1370 primates: old and new families. *Chromosoma* **110**, 253-66 (2001).
1371 59. Huttlin, E.L. *et al.* Architecture of the human interactome defines protein communities
1372 and disease networks. *Nature* **545**, 505-509 (2017).
1373 60. de Boer, C.G. & Taipale, J. Hold out the genome: a roadmap to solving the cis-regulatory
1374 code. *Nature* **625**, 41-50 (2024).

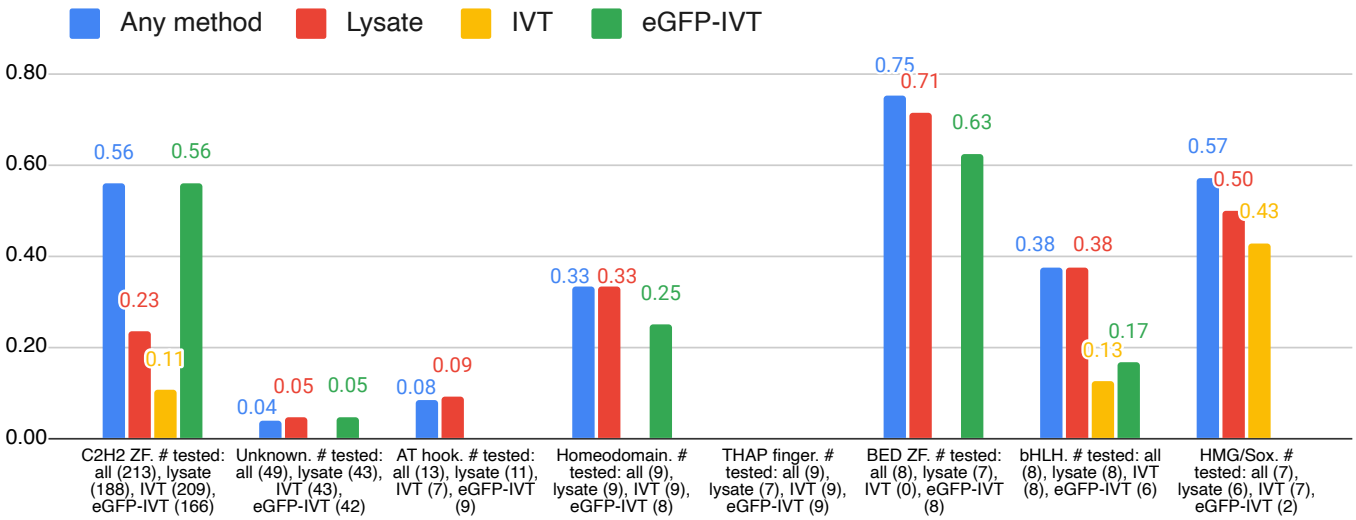
1375 61. Stormo, G.D., Schneider, T.D., Gold, L. & Ehrenfeucht, A. Use of the 'Perceptron'
1376 algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res* **10**, 2997-
1377 3011 (1982).
1378 62. Laverty, K.U. *et al.* PRIESSTESS: interpretable, high-performing models of the sequence
1379 and structure preferences of RNA-binding proteins. *Nucleic Acids Res* **50**, e111 (2022).
1380 63. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
1381 sequence data. *Bioinformatics* **30**, 2114-20 (2014).
1382 64. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
1383 2078-9 (2009).
1384 65. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of
1385 occurrences of k-mers. *Bioinformatics* **27**, 764-70 (2011).
1386 66. Vorontsov, I.E. *et al.* Cross-platform DNA motif discovery and benchmarking to explore
1387 binding specificities of poorly studied human transcription factors. *bioRxiv*,
1388 2024.11.11.619379 (2024).
1389 67. Korhonen, J., Martinmaki, P., Pizzi, C., Rastas, P. & Ukkonen, E. MOODS: fast search for
1390 position weight matrix matches in DNA sequences. *Bioinformatics* **25**, 3181-2 (2009).
1391 68. Najafabadi, H.S. *et al.* Non-base-contacting residues enable kaleidoscopic evolution of
1392 metazoan C2H2 zinc finger DNA binding. *Genome Biol* **18**, 167 (2017).
1393 69. Korhonen, J.H., Palin, K., Taipale, J. & Ukkonen, E. Fast motif matching revisited: high-
1394 order PWMs, SNPs and indels. *Bioinformatics* **33**, 514-521 (2017).

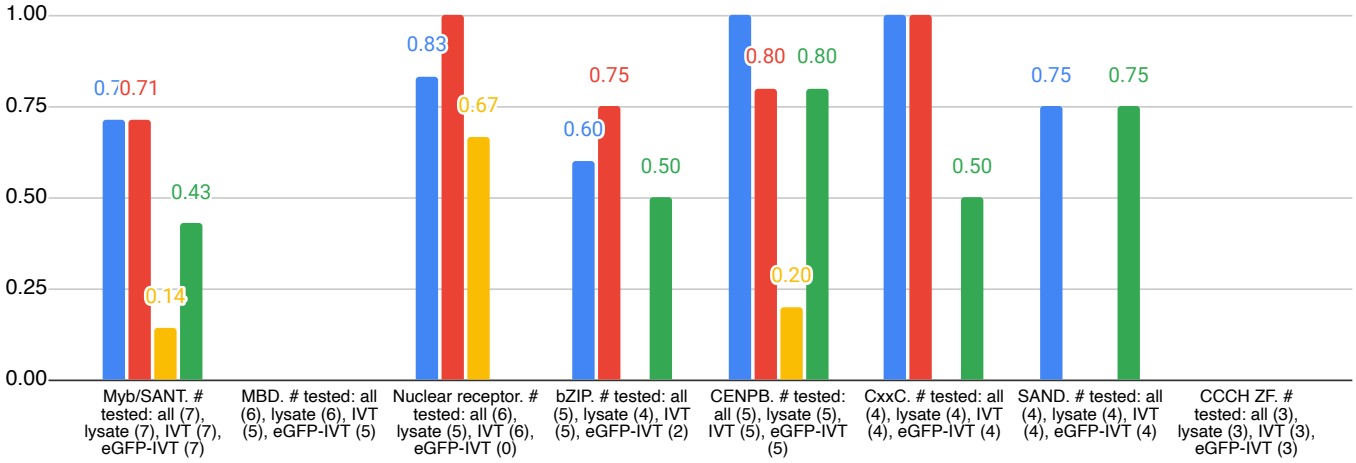
1395

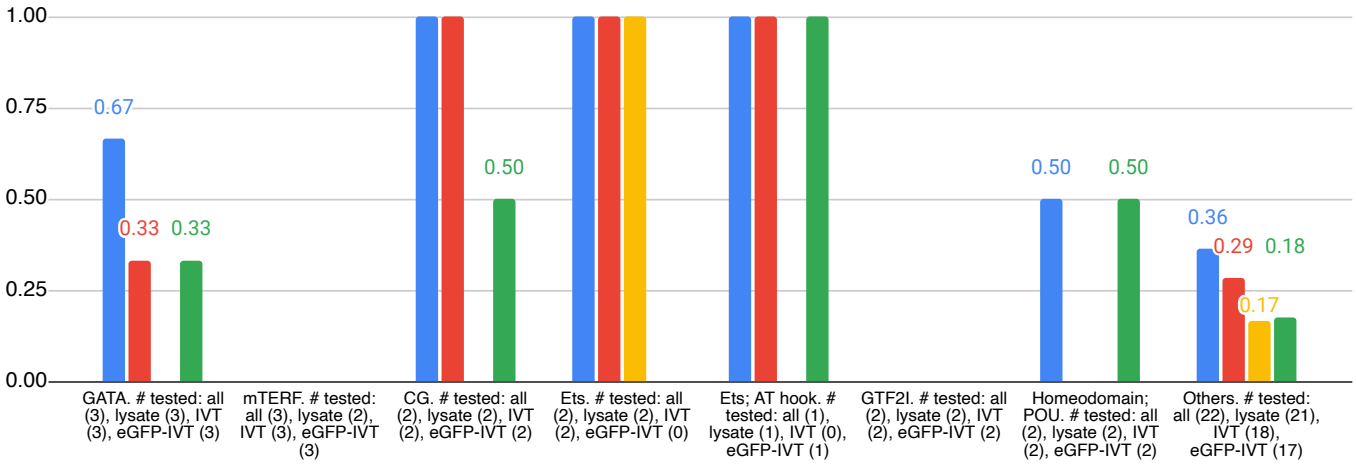

a
Read counts per million reads
10-2
1
102
Fitted coefficient
(rate of enrichment)
-2
0
2
Rep #1
Rep #2
Rep #3
Rep #4
Rep #5
Cycle (t)
1 2 3 4
Genomic intervals
Genomic intervals that become dominant in late cycles due to exponential selection
Genomic intervals that are initially enriched but are depleted in late cycles due to decreased effective library size (i.e., overtaken by the group above)
Background genomic intervals that are progressively depleted
Effective library size
−3
−5


b
Σ
βj
ŷij
si
xi
yij*
σi2
mi,j
Genomic intervals j
Samples i
si ~ 𝒰(−∞,+∞)
βj ~ 𝒩(0,Σ)
ŷij = xi · βj + si
σi2 ~ 𝒰(0,+∞)
yij* ~ 𝒩(ŷij, σi2)
mij ~ Pois(eyij*)

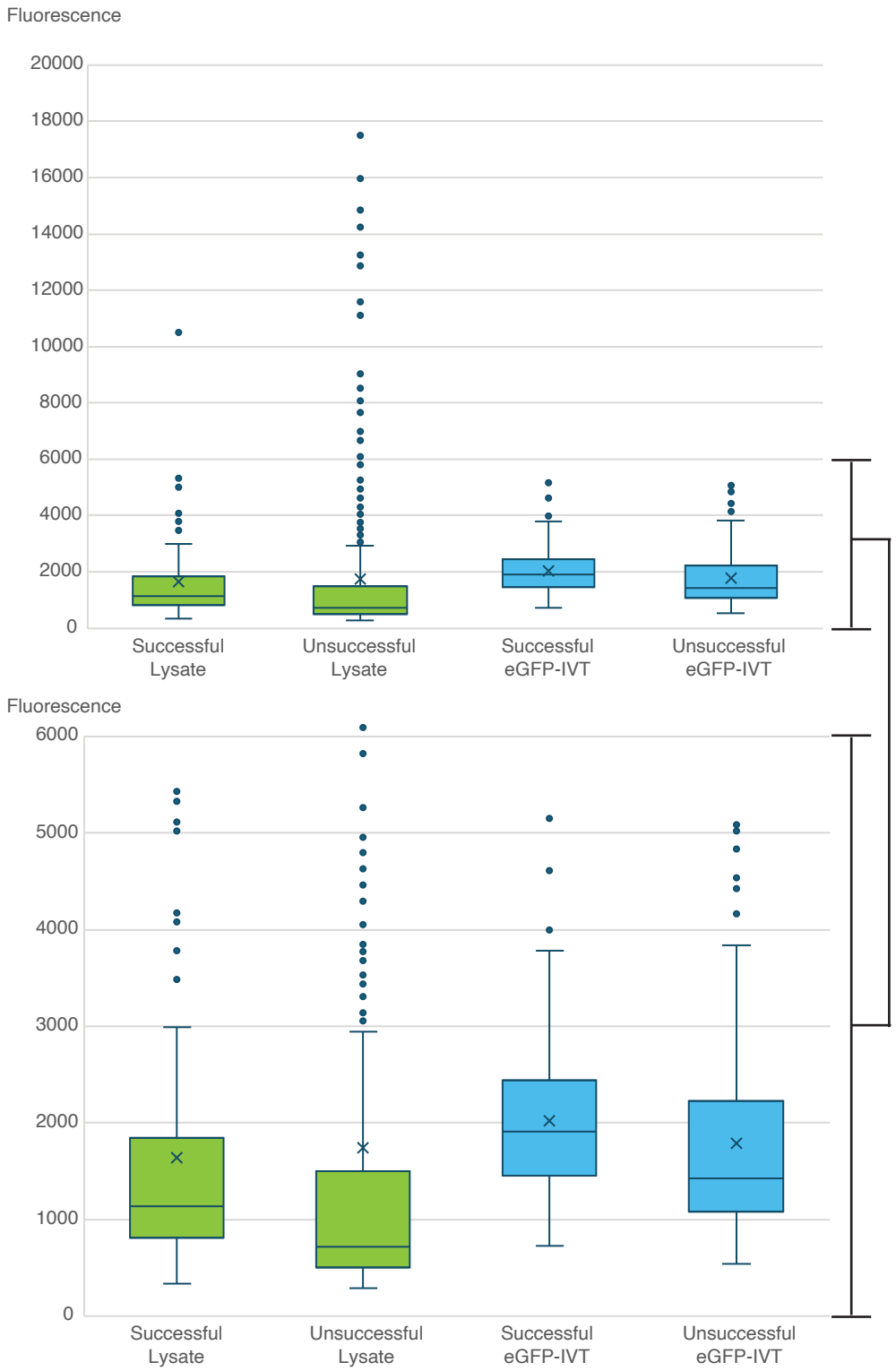
Extended Data Fig. 1: a. Example of actual read count data for CTCF over five replicates of four cycles, illustrating enrichment patterns, fitted coefficients (right), and estimated library sizes (bottom). **b.** A brief overview of the statistical framework of the generative model of MAGIX. Open circles, closed circles, and the diamonds represent latent variables, observed variables, and deterministic computations, respectively. s_i: library size for sample i; $\mathbf{x}_i$: vector of sample-level variables for sample i, including an intercept term and a term for the SELEX cycle, in addition to other terms for batch and background effects; $\boldsymbol{\beta}_j$: vector of model coefficients for interval j; m_{ij}: number of observed reads mapping to interval j in sample i. See **Methods** for description of other variables..


Any method
Lysate
IVT
eGFP-IVT
0.80
0.60
0.40
0.20
0.00
0.56
0.23
0.11
0.56
0.04
0.05
0.05
0.08
0.09
0.33
0.33
0.25
0.75
0.71
0.63
0.38
0.38
0.13
0.17
0.57
0.50
0.43
C2H2 ZF. # tested: all (213), lysate (188), IVT (209), eGFP-IVT (166)
Unknown. # tested: all (49), lysate (43), IVT (43), eGFP-IVT (42)
AT hook. # tested: all (13), lysate (11), IVT (7), eGFP-IVT (9)
Homeodomain. # tested: all (9), lysate (9), IVT (9), eGFP-IVT (8)
THAP finger. # tested: all (9), lysate (7), IVT (9), eGFP-IVT (9)
BED ZF. # tested: all (8), lysate (7), IVT (0), eGFP-IVT (8)
bHLH. # tested: all (8), lysate (8), IVT (8), eGFP-IVT (6)
HMG/Sox. # tested: all (7), lysate (6), IVT (7), eGFP-IVT (2)


1.00
0.75
0.50
0.25
0.00
0.70
0.71
0.14
0.43
0.83
0.67
0.60
0.75
0.50
0.80
0.80
0.20
0.50
0.75
0.75
Myb/SANT. # tested: all (7), lysate (7), IVT (7), eGFP-IVT (7)
MBD. # tested: all (6), lysate (6), IVT (5), eGFP-IVT (5)
Nuclear receptor. # tested: all (6), lysate (5), IVT (6), eGFP-IVT (0)
bZIP. # tested: all (5), lysate (4), IVT (5), eGFP-IVT (2)
CENPB. # tested: all (5), lysate (5), IVT (5), eGFP-IVT (5)
CxxC. # tested: all (4), lysate (4), IVT (4), eGFP-IVT (4)
SAND. # tested: all (4), lysate (4), IVT (4), eGFP-IVT (4)
CCCH ZF. # tested: all (3), lysate (3), IVT (3), eGFP-IVT (3)


1.00
0.75
0.50
0.25
0.00
0.67
0.33
0.33
0.50
0.50
0.50
0.36
0.29
0.18
0.17
GATA. # tested: all (3), lysate (3), IVT (3), eGFP-IVT (3)
mTERF. # tested: all (3), lysate (2), IVT (3), eGFP-IVT (3)
CG. # tested: all (2), lysate (2), IVT (2), eGFP-IVT (2)
Ets. # tested: all (2), lysate (2), IVT (2), eGFP-IVT (0)
Ets; AT hook. # tested: all (1), lysate (1), IVT (0), eGFP-IVT (1)
GTF2I. # tested: all (2), lysate (2), IVT (2), eGFP-IVT (2)
Homeodomain; POU. # tested: all (2), lysate (2), IVT (2), eGFP-IVT (2)
Others. # tested: all (22), lysate (21), IVT (18), eGFP-IVT (17)

Extended Data Fig. 2: TF success rate by protein production method and structural class. Figure displays bar graphs of success rates for putative TFs based on their structural class showing a bar for TFs based on success on if the protein was successful in any of the production methods and then individually based on protein production methods. Number of TFs tested in each category is listed with the name of the structural class.


Fluorescence
20000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
Successful Lysate
Unsuccessful Lysate
Successful eGFP-IVT
Unsuccessful eGFP-IVT
Fluorescence
6000
5000
4000
3000
2000
1000
0
Successful Lysate
Unsuccessful Lysate
Successful eGFP-IVT
Unsuccessful eGFP-IVT

Extended Data Fig. 3: TF expression levels are plotted for four categories based on experiment success and protein production. Data is shown for the full range of fluorescence (*top)* and for the region with most of the signal (*bottom).* Note that even though the successful samples have generally higher expression levels in both protein production systems it is not an effective predictor of experiment success.