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], 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]Vavilov Institute of General Genetics, Russian Academy of Sciences, 119991, Moscow, Russia and Institute of Protein Research, Russian Academy of Sciences, 142290, Pushchino, Russia
[7]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

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

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

SUMMARY

A long-standing challenge in human regulatory genomics is that transcription factor (TF) DNA-binding motifs are short and degenerate, while the genome is large. Motif scans therefore produce many false-positive 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). We find that comparable specificity can also be obtained from motif scans, but performance is highly dependent on derivation and use of the motifs, including accounting for multiple local matches in the scans. 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, contrary to conventional wisdom, it is common for TFs to possess sufficient intrinsic specificity to independently delineate cellular targets.

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 human, 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[30], such that TF motif matches only determine binding within permissible regions. 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[14], and repositions the surrounding nucleosomes[15]. PRDM9, which controls recombination hotspots, has been reported to independently specify roughly half of its binding sites in the genome[16]. 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[17]. 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[1,18], as well as the impact of DNA shape[19], dependencies among base positions[17,20], multimeric binding[21,22], and lower-affinity binding sites[23].

Many human TFs still lack binding motifs, and prominent among them are hundreds of C2H2 zinc finger (C2H2-zf) proteins[24]. These proteins recognize DNA sequences that approximate a concatenation of the three or four base specificities of their sequential constituent C2H2-zf domains[25,26]. 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[27]. 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 long-standing question. In a well-studied example, CTCF binding sites appear to reflect a constitutive "core", bound by fingers 4-7 of the 11 C2H2-zf domain array, flanked by sequences that are bound by alternative usage of upstream and/or downstream C2H2-zf domains[28,29]. Analysis of the DNA-binding of C2H2-zf proteins to the genome is also

complicated by the fact that they often bind repeat elements such as endogenous retroelements[30], and thus the target site similarity is derived both from DNA recognition and the shared ancestry of the binding sites. The limited resolution of ChIP-seq (>100bp) presents a related hindrance. These confounding factors, however, can be ameliorated by incorporating information about the bases that are likely preferred at each position of the binding site, as predicted by a C2H2-zf "recognition code" that relates the C2H2-zf amino acid sequences to their binding preferences. These machine learning-based predictions can assist in identifying the most plausible protein-DNA interactions in such cases, as our earlier work demonstrated[31].

Here, we describe GHT-SELEX (Genomic DNA HT-SELEX), a novel implementation of the HT-SELEX[32] method for identification of the sequence specificity of DNA-binding proteins. HT-SELEX is a high-throughput implementation of SELEX (Systematic Evolution of Ligands by EXponential enrichment)[33], using multi-cycle, automated affinity capture of protein-bound DNA in microwell plates, coupled to multiplexed Illumina sequencing. HT-SELEX utilizes random-sequence DNA, while GHT-SELEX is instead performed with fragmented human genomic DNA, and uses an associated new statistical analysis method, MAGIX (Model-based Analysis of Genomic Intervals with eXponential enrichment). GHT-SELEX is conceptually similar to Affinity-seq[16] and DAP-seq[34], but it incorporates multiple selection cycles, and is thus related to earlier genomic SELEX approaches that utilized Sanger sequencing[35,36]. The use of barcoding, magnetic affinity beads and laboratory automation makes it possible to run GHT-SELEX in parallel with hundreds of samples. We developed GHT-SELEX in the context of the Codebook consortium project[37], which was aimed primarily at analysis of 332 uncharacterized putative TFs (together with 61 control TFs), and provides comparison data from several other platforms for the same set of TFs (HT-SELEX, ChIP-seq, Protein Binding Microarrays[38], and SMiLE-seq[39]). We successfully applied GHT-SELEX to 179 human TFs, most of which are poorly characterized, thus providing a major expansion in the number of human TF motifs. For dozens of TFs, including some that are considered well-characterized, GHT-SELEX peaks correspond with *in vivo* binding (measured by ChIP-seq) much more accurately than current models would suggest. 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

GHT-SELEX combines the principles of previous genomic DNA selection protocols[16,34] with HT-SELEX, a method that has been applied successfully to hundreds of human TFs and is compatible with robotics[32,40]. We developed GHT-SELEX (**Figure 1A**) to run in parallel with HT-SELEX, in the context of the Codebook project. The intended purpose, initially, was to create a DNA library that contains sufficient representation of long repeat sequences that are common in the human genome (e.g. transposons and endogenous retroelements): we reasoned that the difficulty of obtaining long motifs

expected for C2H2-zf proteins may be due to the scarcity of long binding sites in a random pool, since representation of any sequence would decrease exponentially with its length. The GHT-SELEX DNA pool used in this study was produced by nonspecific enzymatic fragmentation of HEK293 DNA to fragments with a median length of ~64 bp. HEK293 DNA was chosen for compatibility with ChIP-seq data generated simultaneously (see accompanying manuscript[41]), 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 controls 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[39]. An additional 31 controls were C2H2-zf proteins for which published ChIP-seq data yielded motifs[42]. At the outset, we assumed that GHT-SELEX would yield continuous read coverage across the genome, given conventional estimates of up to a million PWM hits per TF[7], such that the data could be analyzed directly for enriched motifs among the reads. Indeed, examination of individual mapped reads revealed that they usually accumulate at sites in which all reads overlap with what appears to be a motif match (**Figure 1B**). Remarkably, it also became apparent that GHT-SELEX data typically has a strong resemblance to ChIP-seq data, forming strong peaks found sparsely across the genome. **Figure 1C** shows raw read density for four control TFs, comparing GHT-SELEX to ChIP-seq, and also to target site predictions based on existing and newly-derived (see below) PWM models for the TFs. This observation prompted us to analyze the data as peaks, instead of raw reads.

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, 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. As a consequence, enrichment information is distributed across the read cycles, with weaker peaks first appearing and disappearing, and the strongest peaks dominating in the later cycles. To adapt to these issues, we developed an analytical framework that capitalizes on the added information gained from multiple SELEX cycles (**Figure 2A**; 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 and depletion of 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 (**Figure 2B**). We refer to this approach as MAGIX (Model-based Analysis of Genomic Intervals with eXponential enrichment).

Among the 61 control proteins, 40 were deemed as successful on GHT-SELEX (see below and accompanying manuscripts[37,43] for a description of how success was determined). Analysis of the data for the 40 successful controls by MAGIX resulted in

between 13 and 137,718 peaks (median 19,400) with enrichment coefficient exceeding 5% FDR (see **Methods**). There is a clear enrichment of the motif occurrences for the corresponding TFs within the peaks, with the number of strong PWM hits, on average, declining rapidly at ~50 bp from peak centre, consistent with the DNA fragment size (**Figure 2C;** similar plots for all TFs analyzed are shown in **Document S1**). 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), suggesting that the GHT-SELEX/MAGIX is quantitative to some degree.

Application of GHT-SELEX to the Codebook TF set

We next performed GHT-SELEX and, in parallel, HT-SELEX using fragmented genomic DNA and random 40N ligands (**Table S1**), respectively, to assess DNA binding activity 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**). Several experimental variables were modulated over the course of the experiments, resulting in improvement of success rates, particularly for TFs with long C2H2-zf domain arrays (see **Methods and Table 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**).

In separate parts of the Codebook project, this same set of proteins was analyzed using ChIP-seq, Protein Binding Microarrays[38], and SMiLE-seq[39], as described in the accompanying manuscripts[37,41,44]. We gauged the success of each TF in each experiment, including the GHT-SELEX experiments, largely 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. This process produced a list of “approved” experiments, as described in an accompanying study[43]. Selection of a single PWM for each TF for subsequent analyses is described in accompanying study[37]. The PWM selections incorporated those generated from all data types. PWMs and logos are available in accompanying study[37] and online at https://codebook.ccbr.utoronto.ca, https://mex.autosome.org, and https://cisbp.ccbr.utoronto.ca[45].

In total, 139 previously uncharacterized Codebook TFs had at least one “approved” GHT-SELEX experiment (i.e., were successful in GHT-SELEX), of which 131 were also approved 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)**. In contrast, 163 of the putative TFs did not yield motifs in any of these assays, suggesting that they either do not bind DNA with sequence specificity, or require post-translational modifications or cofactors. In particular, only two of 49 proteins tested that lacked a known DBD yielded an approved experiment in GHT-

SELEX (discussed in greater detail in the accompanying studies[37]). Including the control TFs, 24 types of DBDs were present among the approved experiments (**Figure 3B**), illustrating that the method can capture motif-containing genomic target site locations of diverse TF types.

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

GHT-SELEX analyzed with MAGIX, like ChIP-seq, produces peaks with a continuum of enrichment coefficient values and other associated statistics. Across both Codebook TFs and controls, there are typically a relatively small number of peaks with exceptionally high MAGIX enrichment coefficient values (hundreds to thousands), but we did not observe bimodal distributions that would imply a natural threshold which could be used to discriminate “bound” from “unbound” loci (examples in **Figure 4A**; distributions for all TFs in **Document S1**). We also examined the correspondence between GHT-SELEX/MAGIX peaks, ChIP-seq peaks, and PWM scores, focusing on the 137 TFs for which both ChIP-seq and GHT-SELEX data were available (101 Codebook TFs and 36 controls). In most cases, there was a much higher overlap with ChIP-seq peaks and high PWM scores among the highest-scoring GHT-SELEX/MAGIX peaks (examples are shown in **Figure 4A**, and plots for all TFs in **Document S1**). We did not, however, identify a specific peak enrichment coefficient or significance value across all experiments that uniformly corresponds to high enrichment of PWM hits, or the probability of overlap with ChIP-seq peaks.

Lack of a universal enrichment coefficient threshold across all experiments could be accounted for by 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 TFs, and the ability of the TFs to compete or cooperate with nucleosomes and other cofactors *in vivo*. Given that these parameters are unobserved and difficult to estimate from the data available, we implemented a simple scheme to draw thresholds on both peak sets: by sequentially taking equal numbers of highest scoring peaks on a TF-specific basis, we identified the peak number that maximizes the Jaccard statistic of overlap between the GHT-SELEX/MAGIX peaks and ChIP-seq peaks (**Figure 4B**).

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 (Jaccard median 0.1117) (**Figure 4C,D** and **Table S4**), indicating that the TF intrinsically (i.e. independently) specifies many of the *in vivo* binding sites above the threshold selected. Peak overlap 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 noise in generation of peaks will lead to fluctuation of the rank order of peaks, even for replicates. Indeed, this result is not obtained from permuted peak positions, or permuted experiments (i.e. mismatched TFs) (after permutation, Jaccard median 0.0073; Wilcoxon $p=2.6\times10^{-38}$) (**Figure 4D**). The peak numbers yielding these high Jaccard values are often relatively

low, and correspond to a wide range of ChIP-seq p-value thresholds and MAGIX enrichment coefficient values (**Table S4)**.

Overall, this outcome contrasts with traditional expectation, which is that individual TF would normally not be able to independently specify their DNA targets in the genome[7]. We note that many of the TFs with highest Jaccard maxima are uncharacterized C2H2-zf proteins with long (and intuitively specific) motifs: Among those with Jaccard > 0.1, 78% are C2H2-zf proteins (57 out of 73), vs 42% (27 out of 64) for those with Jaccard below 0.1, and overall, the median Jaccard value for C2H2-zf proteins is 0.1582, vs. 0.0616 for non-C2H2-zf proteins (Wilcoxon $p=5.14x10^{-8}$). CTCF, a control protein that is known to possess large number of genomic target sites, unusually high specificity and ability to control nucleosome positions[14,15], is among those with high Jaccard values, although it is not the highest scoring in this dataset. Counterintuitively, high Jaccard maxima were also obtained for a subset of TFs with relatively short motifs, including NFKB1, GABPA, NACC2, and several CXXC proteins, such as CXXC4 and KDM2A, that mainly bind CG dinucleotides, as expected[46] (**Figure 5A**).

Multiple explanations for high sequence specificity observed in GHT-SELEX

We next asked whether PWM predictions across the genome could achieve a level of correspondence to ChIP-seq that we obtained with GHT-SELEX/MAGIX. To do this, we performed a similar maximization of the overlap score (Jaccard) as described above for GHT-SELEX and ChIP-seq, here sweeping through PWM scores (i.e. using PWMs to predict and score “peaks” in the genome; see **Methods** for details). Remarkably, on average, the overlap between PWM predictions and ChIP-seq peaks is similar to that for GHT-SELEX/MAGIX and ChIP-seq peaks (**Figure 5A, Table S4**), and the numbers of peaks at which the maximum Jaccard was obtained is also typically similar (**Figure 5B**). The slightly higher Jaccard for PWMs in some cases may be due to the simple PWM models smoothing experimental noise in the GHT-SELEX. In some cases, however, this explanation seems implausible; for example, in several instances, very small PWMs (e.g. that of CXXC4, which is mainly a single CG dinucleotide) yielded higher overlap with ChIP-seq peak locations than GHT-SELEX did.

To our knowledge, such strong ability of PWMs to predict *in vivo* binding sites, over a large set of TFs, is unprecedented. We attribute two main sources. First, the PWMs used in these analyses were selected from a panel of hundreds to thousands of candidate PWMs, specifically choosing those that performed best across numerous test statistics and several data types. The Jaccard statistics against ChIP-seq and GHT-SELEX were among the selection criteria. Thus, lower maximal Jaccard scores – often vastly lower - are obtained from virtually all other PWMs. Hence, in addition to optimizing the thresholds, part of the explanation for the high Jaccard values we obtained lies in the derivation of the PWM itself.

The second apparent source of performance increase is the PWM scanning and scoring method. For some TFs, scoring a DNA fragment using the sum of predicted affinity scores over a sequence window (i.e. the sum of the PWM probability scores at

individual positions, rather than the log-odds that is output by most PWM scanning tools) results in considerably higher maximum Jaccard value than taking the maximum or sum of log-odds PWM scores (which are generally thought to represent binding energy[47,48]) (**Figure 5C**). Sum-of-affinity scoring presumably reflects the cooperation of multiple adjacent binding sites, traditionally referred to as "avidity"[49]. The effect is most striking for a subset of TFs that bind short or repetitive sequences, including CG dinucleotides and poly-A stretches (**Figure 5D**, but it also appears to underpin the specificity of NACC2 and ZNF48, which have unique, non-repetitive motifs (**Figure 5E**). Points above the diagonal in **Figure 5A**, where PWM prediction shows higher overlap with ChIP than the GHT-SELEX, may therefore represent the impact of TF binding sites over a larger window influencing ChIP-seq but not GHT-SELEX (we performed the PWM scans 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 detection of CpG islands (which dominate the CXXC4 binding sites[37], in which the CG dinucleotides will be distributed over a large region (by definition $\geq$200 bp).

These analyses indicate that PWMs can often predict *in vivo* TF binding sites as effectively as actual measurements of binding to the genome made with GHT-SELEX. **Figure 1C** illustrates the increase in correspondence between PWM predictions and ChIP-seq peaks that can be achieved with carefully-selected PWMs and improved scanning procedures. There are, however, many TFs in which no PWM could be derived that rivals GHT-SELEX data in correspondence of 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 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-zfs domains that has been previously described[28].

We applied RCADEEM to all 120 C2H2-zf proteins for which we had approved data from GHT-SELEX (**Table S4**). 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 S4**), 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. We manually classified the apparent C2H2-zf domain usage into the following categories, examples of which are shown in **Figure 6B-F**, while **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 S4** 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 1-4 and 5-8, 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[50]. ZNF721 itself is present only in primates. ZNF721 also binds thousands of unique loci outside known repeat elements, and associates physically with TRIM28/KAP1[51], 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 S6**), 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 S6**). 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 a 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 ChIP-seq data than conventional wisdom would suggest it should be[7,24], 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 greater ability to overcome the chromatin state than is commonly believed. This apparent discrepancy with expectation can be explained partly by technical shortcomings in previous PWM-based genome scans, which are based on PWMs derived from other methods. HT-SELEX and other *in vitro* approaches utilizing random sequence are powerful in that they are unbiased in terms of sequence composition[52], 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.

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 uniquely 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.)[24]. 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 regulate specific pathways[24]; we speculate that a similar division may exist in human and other animals.

GHT-SELEX data, together with the larger Codebook dataset, provides an extensive new dataset of TF motifs (i.e. PWMs), encompassing most putative TFs currently

lacking them. Accompanying papers provide a thorough analysis of the results of this project, which underscore many challenges and benefits of accurate motif representations. Representation of TF sequence specificity remains an open challenge, more than four decades after the introduction of the standard PWM model[53]. More accurate representations of large and complex binding sites, in particular for C2H2-zf proteins, could be useful for a variety of purposes, including attributing deep learning filters to individual TFs. Finally, we propose that obtaining data from GHT-SELEX for additional TFs with “known” motifs and genomic binding sites from ChIP-seq will produce a more detailed view of their intrinsic DNA binding abilities, and how this intrinsic ability dictates TF-genome interactions in living cells.

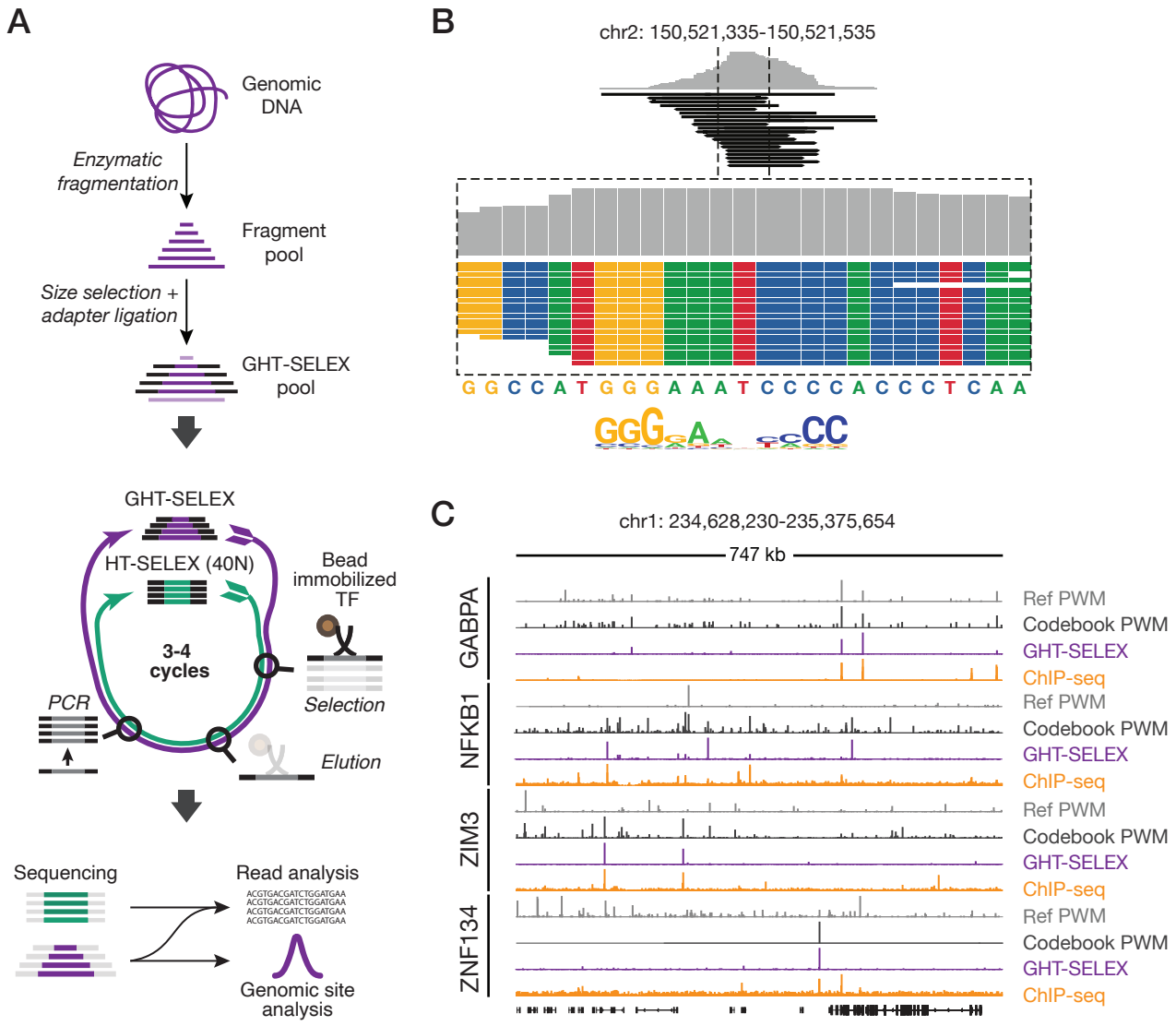


Figure 1. Overview of GHT-SELEX. A. Schematic of GHT-SELEX, showing parallels with HT-SELEX. **B.** Example of read accumulation over a TF motif match for NFKB1. **C.** Genomic binding for four positive control TFs on a genomic region showing (top to bottom) PWM scanning scores (moving average of affinity scores, from MOODS[60] scan in linear domain, using a window of size 200bp) for literature (Ref) PWMs and Codebook PWMs, followed by read coverage signal observed in GHT-SELEX and ChIP-seq.

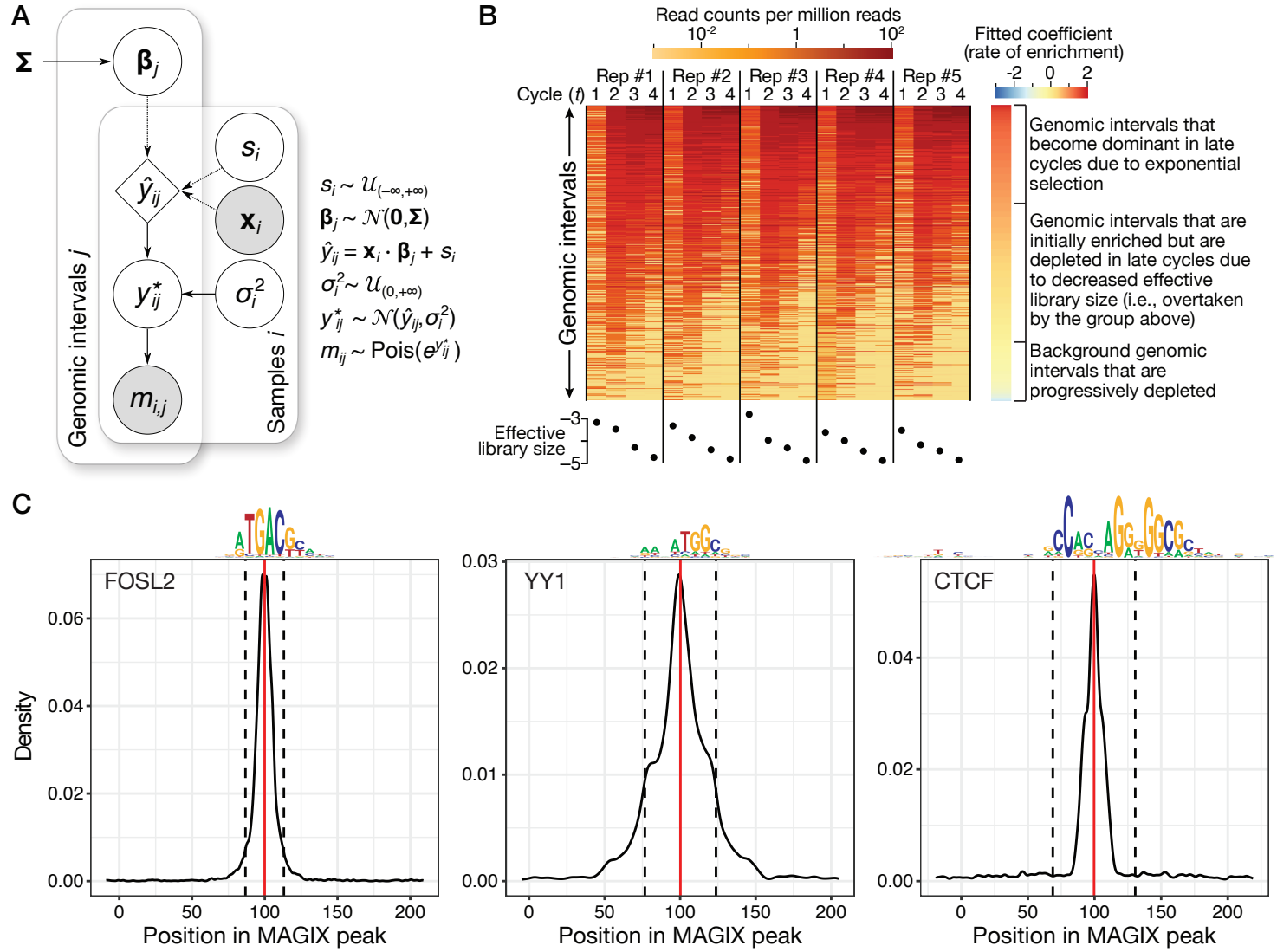


Figure 2. MAGIX method for interpretation of GHT-SELEX data. A. A brief overview of the statistical framework of the generative model of MAGIX. Open circles, closed circles, and the diamonds represent latent variables, observed variables, and deterministic computations, respectively. s_i: library size 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. **B.** 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). **C.** Distribution of PWM hits for the top-ranked TF PWM (highest AUROC on GHT-peaks as determined by accompanying study[43]) within the 5,000 highest scoring MAGIX peaks. PWM hits were identified with MOODS[60] ($P < 0.0001$). Solid red lines represent the mean PWM hit position within MAGIX peaks and dashed lines represent one standard deviation about the mean.

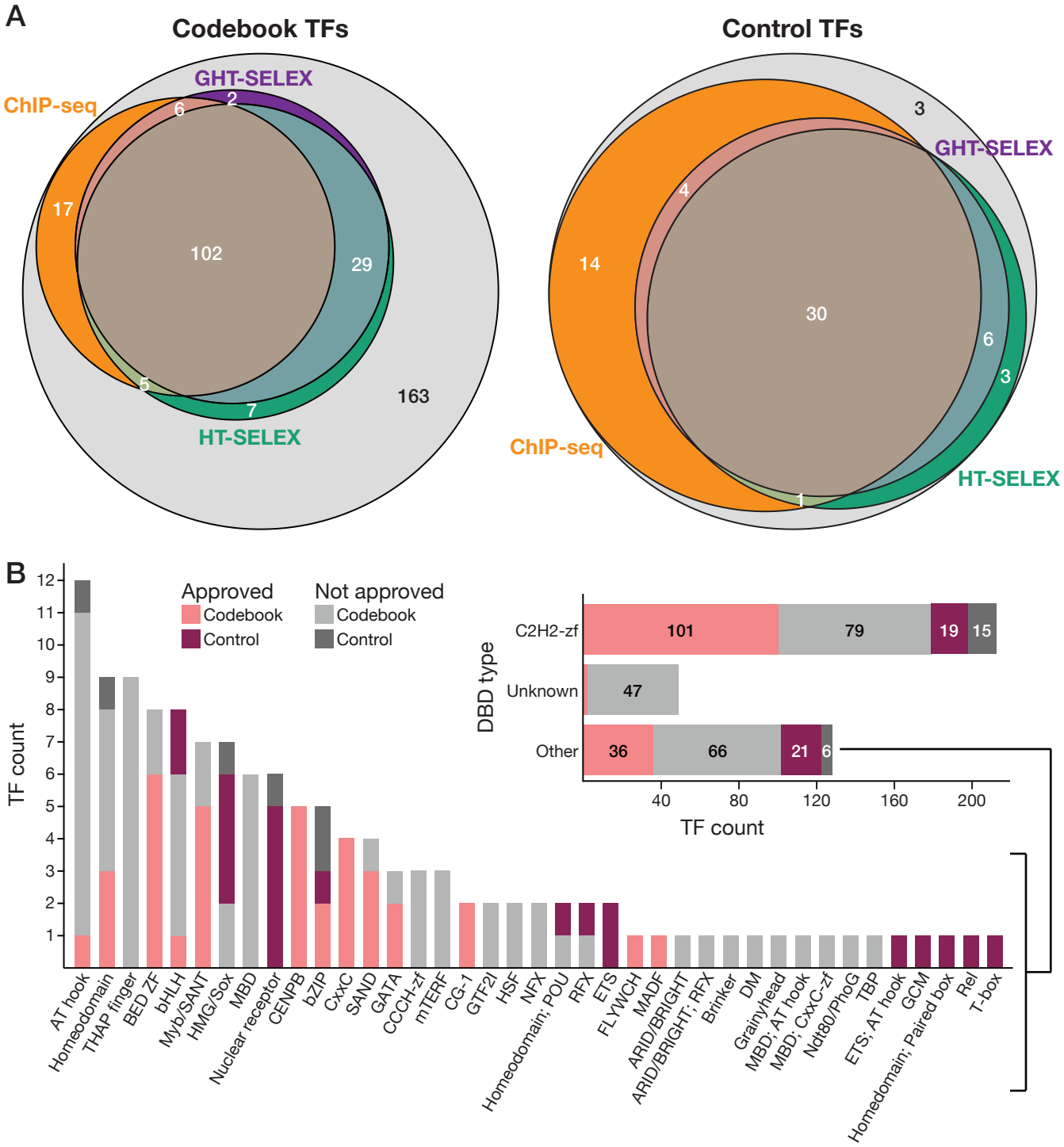


Figure 3. Analysis of 331 Codebook proteins and 61 control TFs using GHT-SELEX. A. Venn diagram displays the number of TFs with approved experiments in GHT-SELEX, HT-SELEX, and ChIP-seq for all Codebook TFs (left) and control TFs (right) assayed with GHT- and HT-SELEX. **B.** Bar chart shows the number of TFs with at least one approved 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.

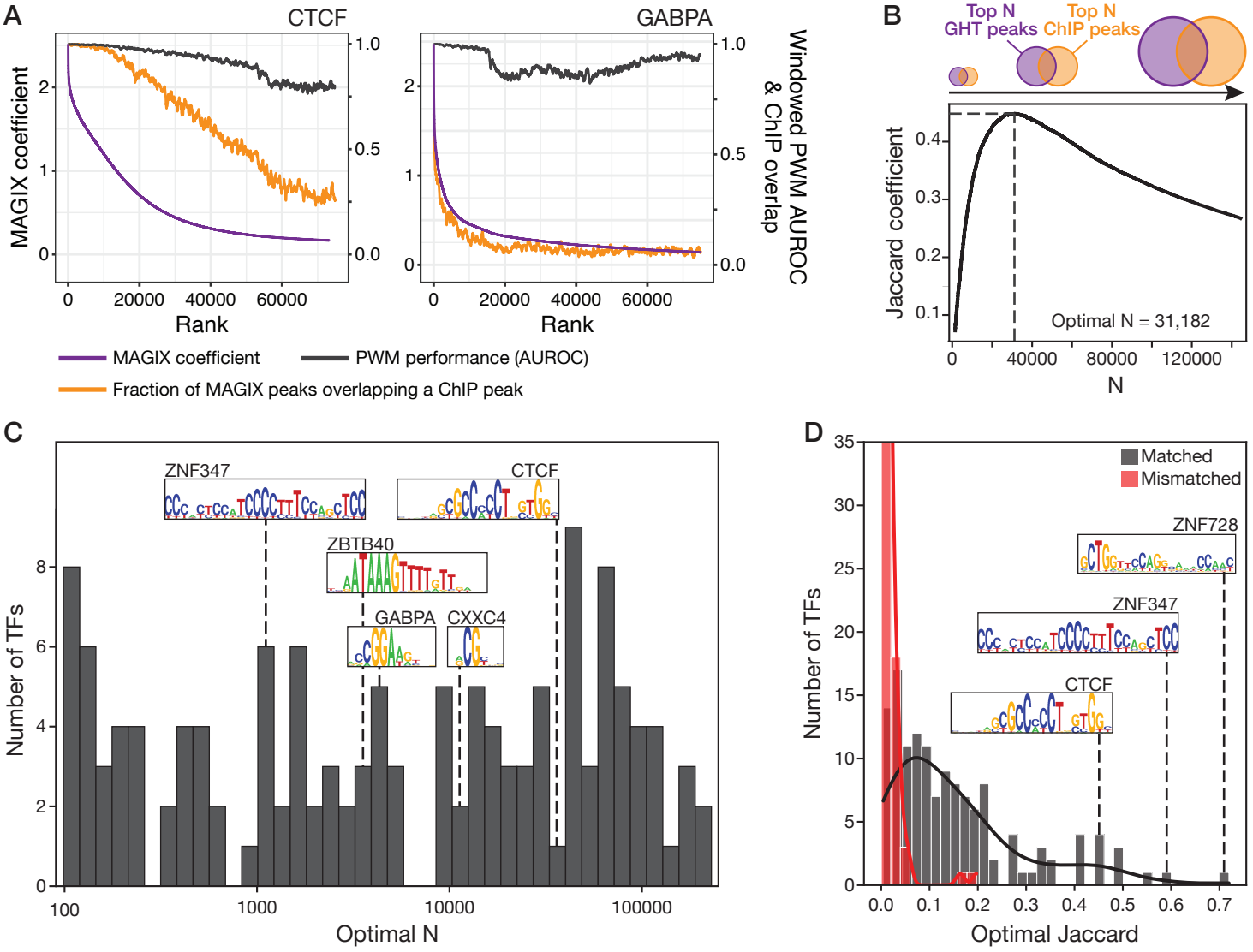
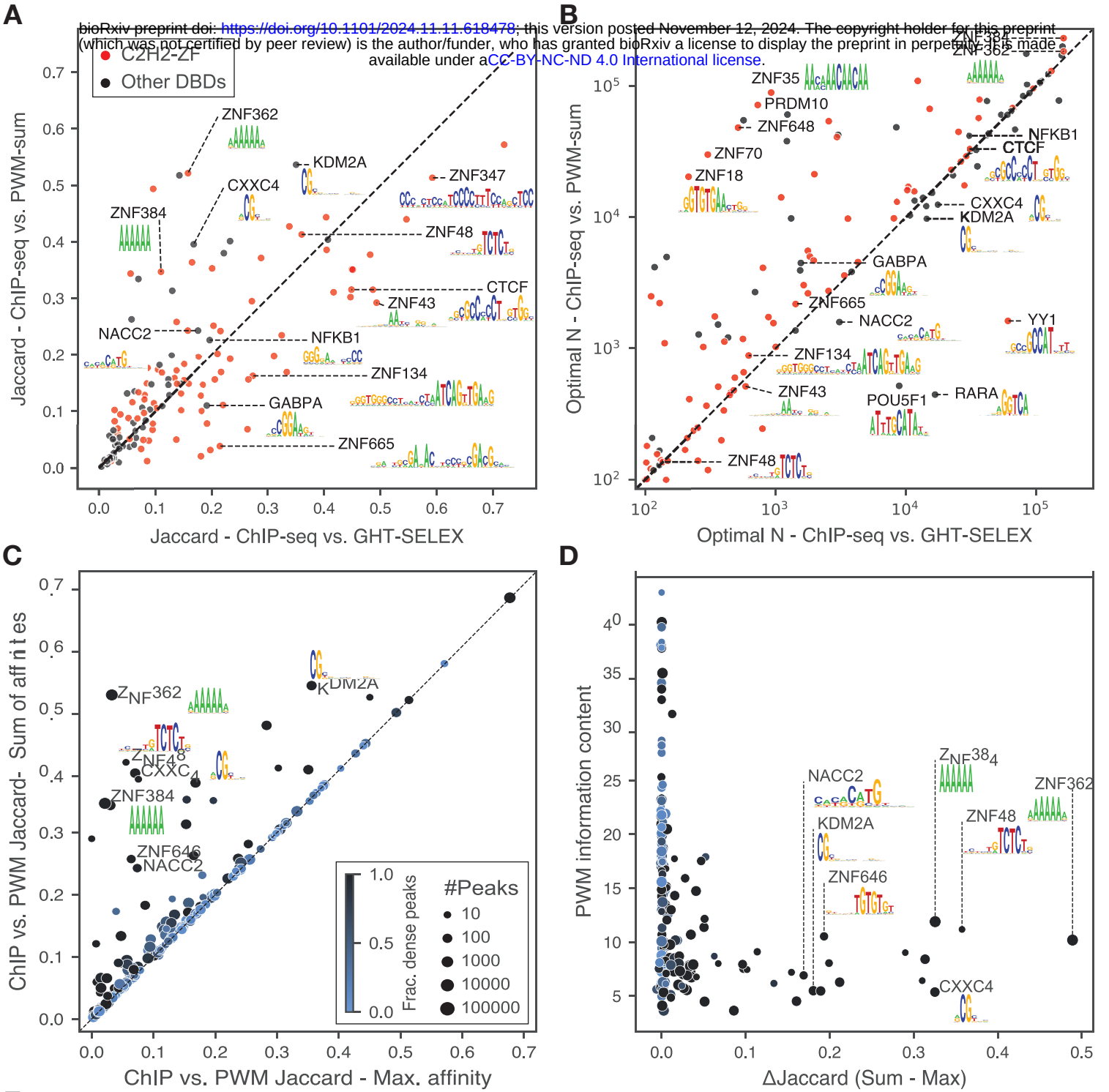


Figure 4. Correspondence between GHT-SELEX and ChIP-seq peaks. A. Enrichment of ChIP-seq peaks and PWM hits within MAGIX peaks, for two example control TFs. The top 75,000 MAGIX peaks are sorted by their MAGIX enrichment coefficient (purple, left y-axis). Orange line shows the proportion of peaks (in 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).



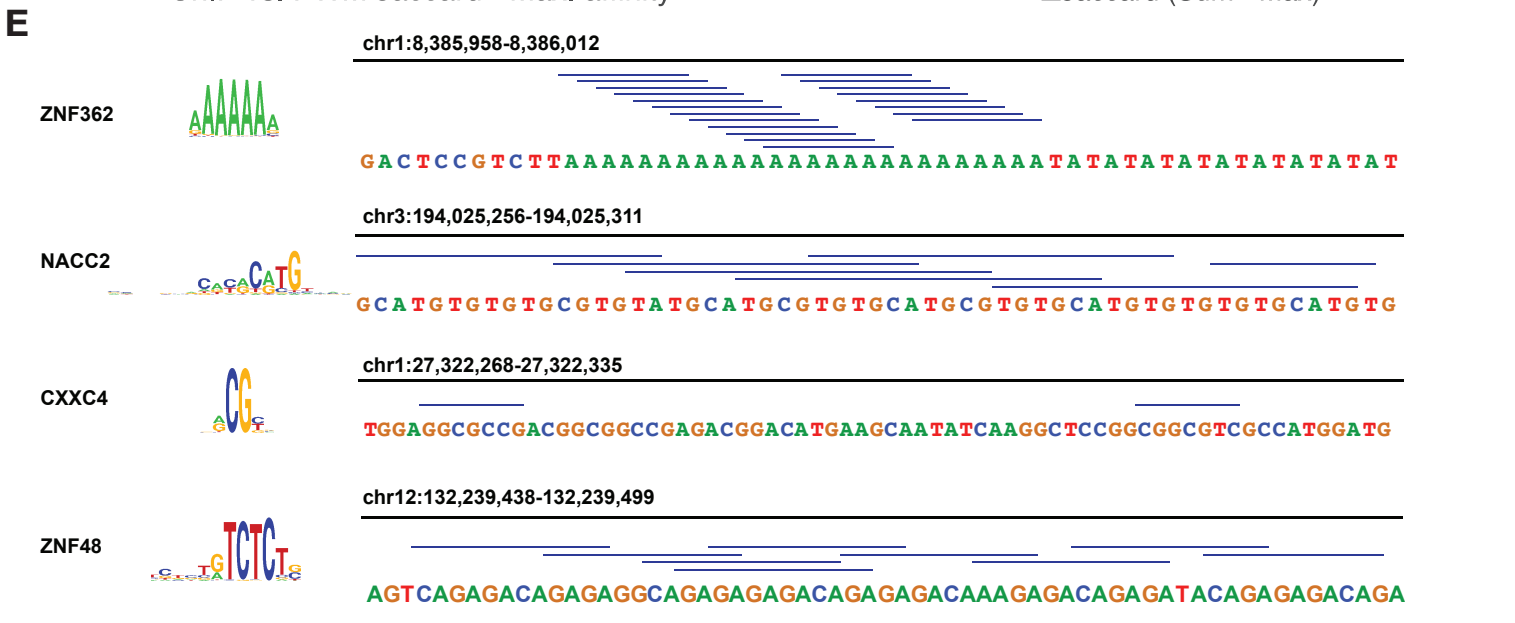
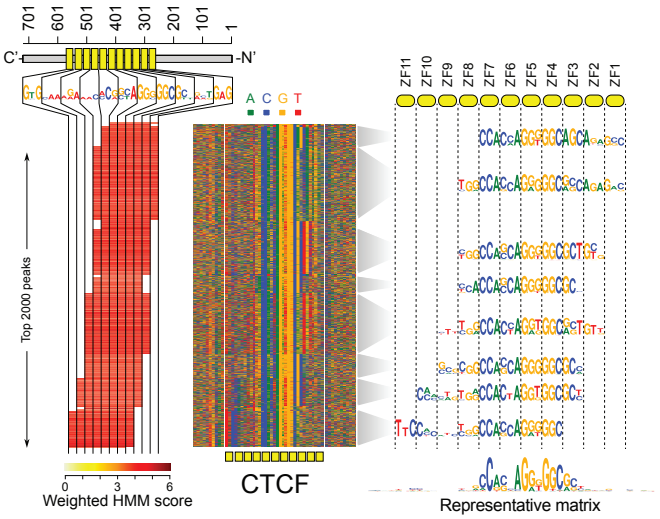
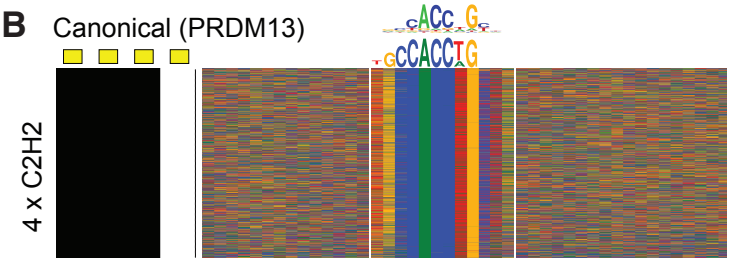


Figure 5. High quality PWMs often predict *in vivo* binding sites as effectively as GHT-SELEX peaks. A. Scatter plot of optimal Jaccard value between GHT-SELEX peaks and ChIP-seq peaks (x-axis) vs. optimal Jaccard value between PWM-predicted sites and ChIP-seq peaks (y-axis), for all 137 TFs (dots). **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.

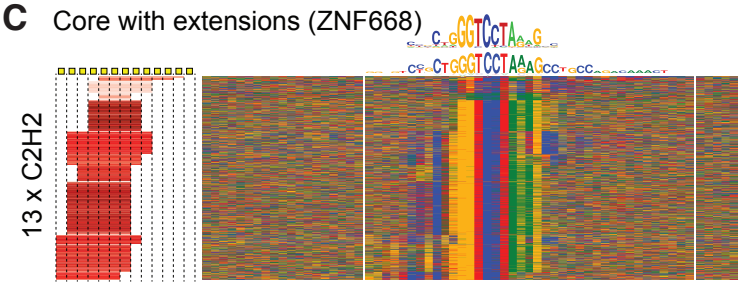
A RCADEEM reveals modular C2H2 DNA binding (CTCF)



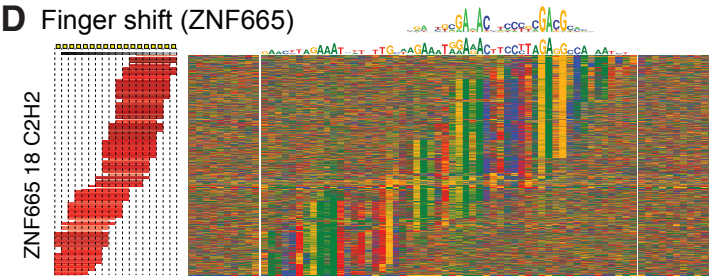
B Canonical (PRDM13)



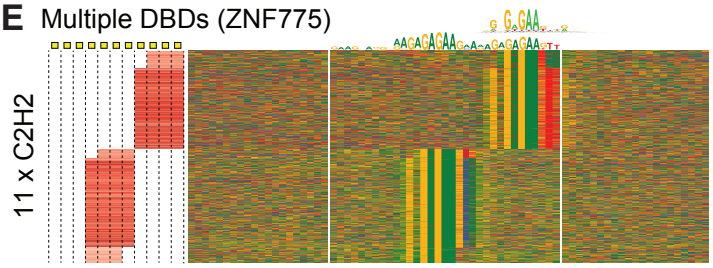
C Core with extensions (ZNF668)



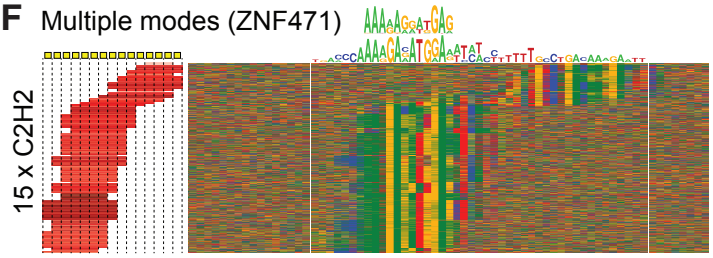
D Finger shift (ZNF665)



E Multiple DBDs (ZNF775)



F Multiple modes (ZNF471)



G

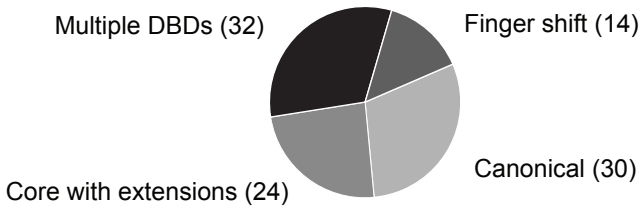


Figure 6. Alternative engagement of individual C2H2-zf domains at genomic binding sites inferred from the recognition code. A. RCADEEM applied to CTCF. *Middle* panel displays the top 2,000 nonrepetitive GHT-SELEX peaks. White vertical bars indicate the region that is expected to contact the DNA based 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 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.

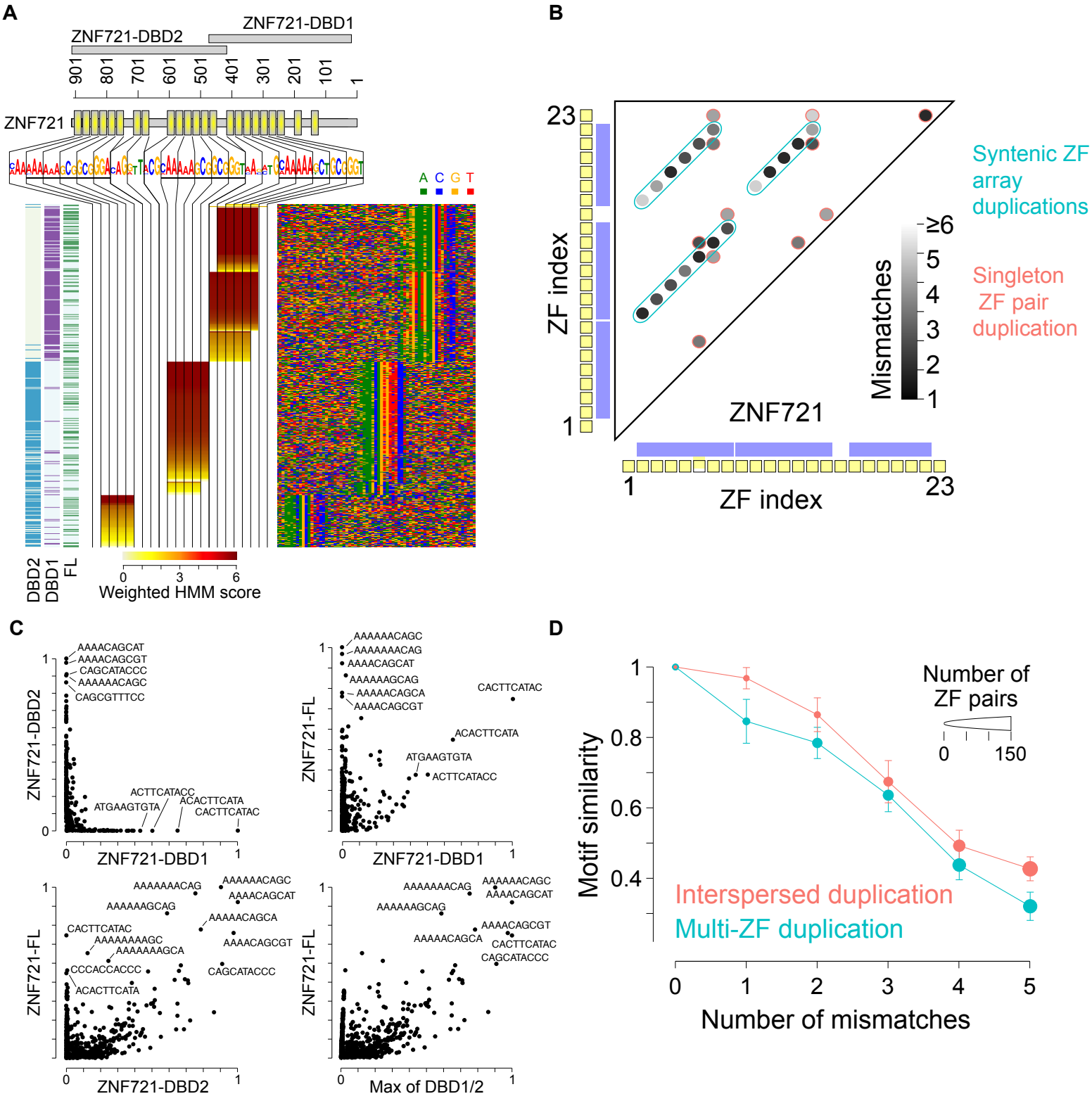


Figure 7. Evolution of C2H2-zf protein DNA-binding specificities through internal duplication of DBDs and DBD arrays. A. RCADEEM results for the pool of top 500 peaks from full-length ZNF721 and two DBD constructs, after removing peaks that overlap repeats. The construct in which the peak was observed 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[61] (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[59] >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 accompanying study[37]. Sequences and other information are available as described below in Data Availability.

Protein production. We used three protein expression systems, which we refer to in **Table S1** and below as *Lysate*, *IVT*, and *eGFP-IVT*, respectively. The *Lysate* system employed recombinant HEK293 cells, created in the accompanying study[41], and in a previous study[42], which express eGFP-tagged full-length proteins from a Tet-inducible promoter (plasmid backbones pTH13195[42] and pTH12027). We induced expression by Doxycycline treatment for 24 hours prior to harvest, 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[32]. Each of the SELEX cycles used 50 ul 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[45]). *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 [39] 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: 7.5ul of *IVT* or 5ul of *eGFP-IVT* reaction sample was used in each binding reaction of each SELEX cycle.

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 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[54], with the exception that the reverse primer was replaced with a primer (5’

CTGGAGTTCAGACGTGTGCTCTTCCGATCT 3'), that does not contain a T7 promoter sequence, and that 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 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[32]. 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 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 primers; 8) Bead suspension obtained from heat elution was used as template in PCR and qPCR reactions; 9) An additional DNA amplification cycle was performed with 2X more primers and dNTPs to ensure that majority of the ligands are in fully double-stranded state and 10) For batches YWO through YWS, we performed an additional step in which the double-stranded ligands were treated with mung bean nuclease to digest single stranded DNA such as primers or unpaired bases within selection ligands. In each mung bean nuclease reaction, the pH of the solution (PCR reaction) was first

lowered by addition of 1:10 volume of 100 mM acetic acid, followed by addition of 1ul (0.75 units) of the enzyme and incubation for one hour at 37°C.

Sequencing. Samples were prepared for sequencing by performing a PCR reaction that indexes each sample and its selection cycle with a unique combination of i7 and i5 barcodes, followed by a double stranding reaction with primers that target regions of DNA outside indices (**Table S1**). Following this step, DNA libraries were pooled, purified with a PCR purification kit (NEB, T1030S), and then subjected to Illumina sequencing with 60bp reads at 3M reads per sample (Donnelly Centre sequencing core facility).

HT- and GHT-SELEX read processing and mapping. HT-SELEX reads were filtered by Phred quality score (Q >= 30 in at least 90% of bases). GHT-SELEX reads were parsed with Trimmomatic[55] to remove the constant regions from genomic fragments that were shorter than the sequencing read length (options: ILLUMINACLIP:CustomAdapters.fa:2:5:5, LEADING:3, TRAILING:3 MINLEN:25). The custom adapters in the fasta file were AGATCGGAAGAGCACACGTCTGAACTCCAG and AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTTA. For GHT-SELEX, we mapped trimmed reads to the human genome build hg38 with bowtie2 (options: --very-sensitive, --no-unal). The mapped reads were further filtered using Samtools (options: -F 1548, version 1.20)[56].

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 selected fragments pool in each cycle relative to 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^{bj}, 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:

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

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:

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

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 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:

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

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 $\hat{\mathbf{y}}_j$:

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

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:

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

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:

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

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 $\mathbf{\Sigma}$ is a hyper-parameter 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 $\mathbf{\Sigma}$ is then constructed as $\mathbf{\Sigma}=\text{diag}(\sigma_{\beta 1}^2,\ldots,\sigma_{\beta K}^2)$.

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:

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

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, for each genomic interval, the rate of enrichment across GHT-SELEX cycles. To identify GHT-SELEX peaks, we used this framework along with the procedure described below to systematically examine all genomic intervals and identify regions with the highest signal (peaks). First, we binned the entire human genome (build hg38) into ~13M non-overlapping intervals of 200 bp and then, for each TF, calculated the read count profiles of these intervals across the GHT-SELEX cycles and replicates. Counts were obtained using bedtools multicov (version 2.30.0)[57]. These read count profiles were used as the input to MAGIX to obtain an enrichment coefficient for each 200 bp interval (while controlling for batch effects by including batch-specific pooled controls and variables in the design matrix). Next, the top 200,000 regions with the highest enrichment coefficients were selected as candidate intervals for peak refinement.

To refine the peak coordinates, we first merged any adjacent candidate intervals, and then calculated the base pair-resolution read count coverage profile across each merged interval (and sum of all GHT-SELEX cycles). The position with the highest read coverage was selected as the candidate peak summit. The read counts overlapping the ±100 bp around the summits were computed, which were used as input to the MAGIX

statistical inference component (above) to recalculate enrichment coefficients, while reusing the library sizes and the empirical hyperparameters estimated from the analysis of all 13M genomic intervals, without re-optimization. For each candidate peak, we also calculated a P-value, representing the statistical significance of the enrichment coefficient (null hypothesis is that the enrichment coefficient is zero). To do so, we obtained maximum likelihood estimate of the model coefficients for each coefficient (i.e., ignoring the prior distributions), and performed a likelihood ratio test (LRT) against a reduced model in which the enrichment coefficient was restricted to zero.

To calculate empirical FDRs for the peaks, we first obtained negative peaks by repeating the procedure described above but inverting the cycle labels. In other words, we obtained depleted peaks, relative to the pool, instead of enriched peaks. Then, we calculate the FDR as the fraction of depleted peaks relative to enriched peaks for each coefficient value. The source code for MAGIX is available at https://github.com/csglab/MAGIX.

Selection of thresholds for peak sets. We sorted the GHT-SELEX peaks by their MAGIX score (enrichment coefficient, 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 with 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 the set2 and vice versa). The value of N that yielded the maximum Jaccard value was identified, and the threshold for each peak set taken as that which yielded this maximum N. The same process was applied to compare PWM-predicted binding sites and ChIP-seq peaks.

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[58] 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; this length is also consistent with the MAGIX bin size. 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 are 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 calculation of 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 S5**). To make a motif model for each binding mode, we manually selected representative peaks corresponding to each biding 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 **Document S2** and online at https://cisbp.ccbr.utoronto.ca[45]. 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 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[59].

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[43] can be browsed at mex.autosome.org. Source codes for MAGIX and RCADEEM are available from 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 20-74-10075 to I.V.K.
- Russian Science Foundation grant 24-14-20031 to F.A.K.
- Swiss National Science Foundation grant (no. 310030_197082) to B.D.
- Marie Skłodowska-Curie (no. 895426) and EMBO long-term (1139-2019) fellowships to J.F.K.
- NIH grants R01HG013328 and U24HG013078 to M.T.W., T.R.H., and Q.M.
- NIH grants R01AR073228, P30AR070549, and R01AI173314 to M.T.W.
- NIH grant P30CA008748 partially supported Q.M.
- Canada Research Chairs funded by CIHR to T.R.H. and H.S.N.
- Ontario Graduate Scholarships to K.U.L and I.Y.
- A.J. was supported by Vetenskapsrådet (Swedish Research Council) Postdoctoral Fellowship (2016-00158)
- The Billes Chair of Medical Research at the University of Toronto to T.R.H.
- EPFL Center for Imaging
- Institutional funding from EPFL
- Resource allocations from the Digital Research Alliance of Canada

SUPPLEMENTARY TABLES AND DOCUMENTS

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 approved 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: 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 S5: 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 S6: 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: Motif centrality and enrichment in GHT-SELEX/MAGIX peaks and its correspondence with ChIP-seq peaks. Same plots as in **Figure 2C** and **Figure 4A** for all the TFs and DBD constructs in this study with approved GHT-SELEX experiments. *Top*, top-ranked TF PWM (highest AUROC on GHT-peaks as determined by [43]). *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.

988 **Document S2: PFMs of C2H2-zf proteins with alternative binding modes.** PFMs
989 representing the different binding modes of C2H2-zf proteins.
990
991

REFERENCES

1. Stormo, G.D. & Zhao, Y. Determining the specificity of protein-DNA interactions. *Nat Rev Genet* **11**, 751-60 (2010).
2. Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res* **46**, D1284 (2018).
3. Bernard, B., Thorsson, V., Rovira, H. & Shmulevich, I. Increasing coverage of transcription factor position weight matrices through domain-level homology. *PLoS One* **7**, e42779 (2012).
4. Lambert, S.A. *et al.* Similarity regression predicts evolution of transcription factor sequence specificity. *Nat Genet* **51**, 981-989 (2019).
5. Wunderlich, Z. & Mirny, L.A. Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet* **25**, 434-40 (2009).
6. Patel, Z.M. & Hughes, T.R. Global properties of regulatory sequences are predicted by transcription factor recognition mechanisms. *Genome Biol* **22**, 285 (2021).
7. Wasserman, W.W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* **5**, 276-87 (2004).
8. Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* **5**, 829-34 (2008).
9. Marinov, G.K., Kundaje, A., Park, P.J. & Wold, B.J. Large-Scale Quality Analysis of Published ChIP-seq Data. *G3 (Bethesda)* (2013).
10. Consortium, E.P. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699-710 (2020).
11. Rhee, H.S. & Pugh, B.F. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell* **147**, 1408-19 (2011).
12. Long, H.K., Prescott, S.L. & Wysocka, J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**, 1170-1187 (2016).
13. Zinzen, R.P., Girardot, C., Gagneur, J., Braun, M. & Furlong, E.E. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* **462**, 65-70 (2009).
14. Kim, T.H. *et al.* Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**, 1231-45 (2007).
15. Fu, Y., Sinha, M., Peterson, C.L. & Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet* **4**, e1000138 (2008).
16. Walker, M. *et al.* Affinity-seq detects genome-wide PRDM9 binding sites and reveals the impact of prior chromatin modifications on mammalian recombination hotspot usage. *Epigenetics Chromatin* **8**, 31 (2015).
17. Morgunova, E. *et al.* Two distinct DNA sequences recognized by transcription factors represent enthalpy and entropy optima. *Elife* **7**(2018).
18. Stormo, G.D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16-23 (2000).
19. Rohs, R. *et al.* The role of DNA shape in protein-DNA recognition. *Nature* **461**, 1248-53 (2009).

20. Zhao, Y., Ruan, S., Pandey, M. & Stormo, G.D. Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics* **191**, 781-90 (2012).
21. Jolma, A. *et al.* DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384-8 (2015).
22. Horton, C.A. *et al.* Short tandem repeats bind transcription factors to tune eukaryotic gene expression. *Science* **381**, eadd1250 (2023).
23. Kribelbauer, J.F., Rastogi, C., Bussemaker, H.J. & Mann, R.S. Low-Affinity Binding Sites and the Transcription Factor Specificity Paradox in Eukaryotes. *Annu Rev Cell Dev Biol* **35**, 357-379 (2019).
24. Lambert, S.A. *et al.* The Human Transcription Factors. *Cell* **175**, 598-599 (2018).
25. Wolfe, S.A., Nekludova, L. & Pabo, C.O. DNA recognition by Cys2His2 zinc finger proteins. *Annu Rev Biophys Biomol Struct* **29**, 183-212 (2000).
26. Klug, A. The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annu Rev Biochem* **79**, 213-31 (2010).
27. Stubbs, L., Sun, Y. & Caetano-Anolles, D. Function and Evolution of C2H2 Zinc Finger Arrays. *Subcell Biochem* **52**, 75-94 (2011).
28. Nakahashi, H. *et al.* A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep* **3**, 1678-1689 (2013).
29. Kieffer-Kwon, K.R. *et al.* Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell* **155**, 1507-20 (2013).
30. Najafabadi, H.S. *et al.* C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat Biotechnol* (2015).
31. Najafabadi, H.S., Albu, M. & Hughes, T.R. Identification of C2H2-ZF binding preferences from ChIP-seq data using RCADE. *Bioinformatics* **31**, 2879-81 (2015).
32. Jolma, A. *et al.* Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res* **20**, 861-73 (2010).
33. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505-10 (1990).
34. O'Malley, R.C. *et al.* Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* **165**, 1280-1292 (2016).
35. Singer, B.S., Shtatland, T., Brown, D. & Gold, L. Libraries for genomic SELEX. *Nucleic Acids Res* **25**, 781-6 (1997).
36. Zimmermann, B., Bilusic, I., Lorenz, C. & Schroeder, R. Genomic SELEX: a discovery tool for genomic aptamers. *Methods* **52**, 125-32 (2010).
37. Jolma, A. *et al.* Perspectives on Codebook: sequence specificity of uncharacterized human transcription factors. *bioRxiv*, 2024.11.11.622097 (2024).
38. Berger, M.F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* **24**, 1429-35 (2006).
39. Isakova, A. *et al.* SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nat Methods* **14**, 316-322 (2017).
40. Jolma, A. *et al.* DNA-Binding Specificities of Human Transcription Factors. *Cell* **152**, 327-39 (2013).

41. Razavi, R. *et al.* Extensive binding of uncharacterized human transcription factors to genomic dark matter. *bioRxiv*, 2024.11.11.622123 (2024).
42. Schmitges, F.W. *et al.* Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Res* **26**, 1742-1752 (2016).
43. Vorontsov, I.E. *et al.* Cross-platform DNA motif discovery and benchmarking to explore binding specificities of poorly studied human transcription factors. *bioRxiv*, 2024.11.11.619379 (2024).
44. Gralak, A.J. *et al.* Identification of methylation-sensitive human transcription factors using meSMiLE-seq. *bioRxiv*, 2024.11.11.619598 (2024).
45. Weirauch, M.T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-43 (2014).
46. Birke, M. *et al.* The MT domain of the proto-oncoprotein MLL binds to CpG-containing DNA and discriminates against methylation. *Nucleic Acids Res* **30**, 958-65 (2002).
47. Stormo, G.D. & Fields, D.S. Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem Sci* **23**, 109-13 (1998).
48. Weirauch, M.T. *et al.* Evaluation of methods for modeling transcription factor sequence specificity. *Nat Biotechnol* **31**, 126-34 (2013).
49. Kuznetsov, V.A. Mathematical Modeling of Avidity Distribution and Estimating General Binding Properties of Transcription Factors from Genome-Wide Binding Profiles. *Methods Mol Biol* **1613**, 193-276 (2017).
50. Alexandrov, I., Kazakov, A., Tumeneva, I., Shepelev, V. & Yurov, Y. Alpha-satellite DNA of primates: old and new families. *Chromosoma* **110**, 253-66 (2001).
51. Huttlin, E.L. *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505-509 (2017).
52. de Boer, C.G. & Taipale, J. Hold out the genome: a roadmap to solving the cis-regulatory code. *Nature* **625**, 41-50 (2024).
53. Stormo, G.D., Schneider, T.D., Gold, L. & Ehrenfeucht, A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res* **10**, 2997-3011 (1982).
54. Laverty, K.U. *et al.* PRIESSTESS: interpretable, high-performing models of the sequence and structure preferences of RNA-binding proteins. *Nucleic Acids Res* **50**, e111 (2022).
55. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-20 (2014).
56. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
57. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-2 (2010).
58. Korhonen, J., Martinmaki, P., Pizzi, C., Rastas, P. & Ukkonen, E. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics* **25**, 3181-2 (2009).
59. Najafabadi, H.S. *et al.* Non-base-contacting residues enable kaleidoscopic evolution of metazoan C2H2 zinc finger DNA binding. *Genome Biol* **18**, 167 (2017).
60. Korhonen, J.H., Palin, K., Taipale, J. & Ukkonen, E. Fast motif matching revisited: high-order PWMs, SNPs and indels. *Bioinformatics* **33**, 514-521 (2017).

61. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-70 (2011).