Why Do Long Zinc Finger Proteins have Short Motifs?

A case study of ZFY and CTCF reveals non-independent recognition of tandem zinc finger proteins.

Zheng Zuo[1*], Timothy Billings[6], Michael Walker[6], Petko Petkov[6], Polly Fordyce[1, 2, 3, 4], Gary D. Stormo[5*]

1. Department of Genetics, Stanford University, CA, USA
2. Chan Zuckerberg Biohub, San Francisco, CA, USA
3. Department of Bioengineering, Stanford University, CA, USA
4. Stanford CheM-H Institute, Stanford University, CA, USA
5. Department of Genetics, Washington University in St. Louis, MO, USA
6. The Jackson Laboratory, ME, USA

Correspondence: zzuo@stanford.edu

Summary

The human genome has more than 800 C2H2 Zinc Finger-containing genes, and many of them are composed of long tandem arrays of zinc fingers. Current Zinc Finger Protein (ZFP) motif prediction models assume longer finger arrays correspond to longer DNA-binding motifs and higher specificity. However, recent experimental efforts to identify ZFP binding sites *in vivo* contradict this assumption with many having short reported motifs. Using Zinc Finger Y (ZFY), which has 13 ZFs, we quantitatively characterize its DNA binding specificity with several complementary methods, including Affinity-seq, HT-SELEX, Spec-seq and fluorescence anisotropy. Besides the previously identified core motif GGCCT recognized by fingers 12-13, we find a novel secondary irregular motif recognized by accessory fingers. Via high-throughput energy measurements and two-color anisotropy, we establish that this secondary motif contributes to binding and recognition in a non-independent manner, increasing overall affinity only in the presence of the core recognition site. Through additional experimental and iterative computational analysis of CTCF and ZNF343, we further establish that this non-independent recognition between core and secondary motifs could be a general mechanism for tandem zinc finger proteins. These results establish that better motif discovery methods that consider the intrinsic properties of tandem zinc fingers including irregular motif structure, variable spacing and non-independent recognition are essential to improve prediction of ZFP recognition, occupancies, and effects on downstream gene expression *in vivo.*

Keywords: zinc finger proteins, protein-DNA interactions specificity, ZFY, CTCF, Spec-seq

Introduction

The zinc finger domain was first described in the TFIIIA protein of Xenopus which contains an array of 9 zinc fingers involved in the recognition of RNA Polymerase III promoters [1, 2]. With the advent of whole genome sequencing it has been discovered that ZFPs occur in all eukaryotic species but have expanded enormously in the vertebrate lineage [3-5] where they are the most abundant class of transcription factors (TFs) [6-8]. While ZFPs can have other roles, such as binding to RNA and in protein-protein interactions [9-11], it is generally thought that most of them function as TFs [12, 13].

The most common form of ZFPs are those with a C2H2 motif in which two cysteine and two histidine residues coordinate a zinc atom in a beta-beta-alpha protein structure [14, 15]. Some proteins contain only one C2H2 domain, but it is much more common for TFs to contain multiple C2H2 domains in an array, or sometimes in multiple, separated arrays [12]. The total number of C2H2 domains has also greatly expanded in vertebrates, with some ZFPs containing more than 30 (Figure 1A-E) [13].

The initial structure determination of a ZFP bound to DNA showed that each finger could interact with 3-4 base pairs with adjacent fingers binding to adjacent DNA triplets [15]. That structural model and information

about the binding sites for several ZFPs led to simple models of DNA recognition which could be used to design ZFPs to recognize specific sequences [14, 16-20]. Those early models only provided predictions of the preferred binding site sequence, whereas more comprehensive predictions of the specificity of the ZFPs is available in probabilistic models [21-23]. Based largely on the development of zinc finger nucleases for genome editing, specificities have been determined for large collections of short zinc finger proteins [12, 24-26] and from those improved methods for motif prediction have been developed [27-29]. But all models assume that each finger of the ZFP interacts with 3bp of DNA, concatenated together to form the complete binding site of length 3n for a ZFP with n fingers. While the majority of ZFPs with many fingers have unknown specificity, recent advances have led to the identification of motifs for many more [30, 31]. For some ZFPs with only a few ZFs shown in Fig. 1F, the observed motifs are longer than expected probably due to binding in vivo with another factor[9, 10, 32], but it is evident that the motifs are usually much shorter than expected based on the number of fingers.

There are several examples of well-studied ZFPs with long arrays of zinc fingers, including CTCF [33], Gli1 [34], and PRDM9 [35, 36], in which significant portions of their DNA binding domains are apparently not engaged in motif recognition but still evolutionarily conserved. This “many fingers but short motif” paradox could be explained in several ways. It is possible that the additional fingers may contribute to DNA binding specificity, but in ways that are missed by current methods for motif discovery. It may be that the additional fingers have roles other than DNA binding, such as being involved in protein-protein or protein-RNA interactions in vivo. It may also be that the additional fingers contribute to binding affinity but not specificity, for example by contributing favorable binding energy through interactions with the DNA backbone. Of course, these alternative functions for the additional zinc fingers are not mutually exclusive, and different proteins may employ different functions.

One notable case is Zinc Finger Y (ZFY). As the only zinc finger gene on human Y chromosome, ZFY was initially perceived as the putative regulator for sex determination until the discovery of the SRY gene [37, 38]. In mice, its close homologs mZFY1 and mZFY2 (Figure 2A) are required for the meiotic sex chromosome inactivation (MSCI) process [39]. ZFY consists of an N-terminal activation domain followed by thirteen tandem fingers, therefore motif prediction models suggest it should have a ~39bp long consensus binding site (Fig. 2B). Taylor-Harris et al [40] performed SELEX on mZFY1 and found it preferentially binds to sequences containing GGCCT and Grants et al [41] showed that fingers 11-13 were sufficient to recognize a motif of RGGCCT, both of which are consistent with the model from zinc finger motif prediction algorithms (Fig. 2B) [29]. Since ZFY genes share close homology with their X-chromosome counterpart ZFX, we list all the contact residues of ZFY and ZFX for both human and mouse versions in Fig. 2A. ZFY and ZFX have the same contact residues composition in Finger 12 and 13, so not surprisingly previous PBM [42] and Chip-seq [43] experiments of ZFX produced the same 5 base-pair motif (Fig. 2C, D).

We employ four complementary experimental approaches to determine the specificity of ZFY, and we use different truncated versions of the protein to assess the contribution of different fingers in DNA binding affinity. We find that the additional fingers do contribute to specificity, with a weak, irregular motif not consistent with the motif prediction programs. We also find that the additional fingers contribute significantly to binding affinity only in the presence of the core site, increasing the half-life of the complex considerably, which could be important for fulfilling ZFY’s *in vivo* biological functions [39]. Similar experiments with CTCF show the recognition of an upstream motif depends on the presence of core motif with variable spacing. Given this, it is more effective to use a sequential or iterative motif discovery method to analyze and extract more motif information of tandem ZFPs, as we show for ZNF343 from high-throughput genomic data.

Results

Short motifs are obtained for hZFY and mZFY1 using Affinity-seq.

Affinity-seq is a method for *in vitro* selection of fragmented genomic DNA followed by MEME motif analysis [35], which can determine all potential binding sites of a given TF over the entire genome sequence. We performed Affinity-seq on human ZFY and mouse ZFY1. For hZFY and mZFY1, we obtained 90,084 and 50,170 peaks at p value < 0.01, respectively, from which we found motifs very similar to those previously reported (Fig. 2E-F) and no secondary motifs were reported by MEME.

High-throughput SELEX (HT-SELEX) identifies an extended consensus site and Spec-seq confirms this irregular secondary motif by accessory fingers of ZFY

Spec-seq is a high-throughput sequencing-based method to quantitatively characterize the energetic landscape of TF-DNA interactions with resolution down to 0.1k_BT [44, 45]. However, one practical limitation of Spec-seq is that one can only obtain the relative binding energy of a few thousands of variants in one assay, so having prior knowledge about the TF consensus site is preferable. To test whether ZFY has any extended motif beyond GGCCT, we adopted High-throughput SELEX (HT-SELEX) first to infer an extended consensus site, similar to our previous approach of following SELEX with a more quantitative analysis [46], then performed Spec-seq. We chose randomized dsDNA libraries with prefixed GGCCT in the flanking region for HT-SELEX [47, 48] analysis by full-length human ZFY (Fig. 3A). After two rounds of bound DNA selection by EMSA separation and amplification, we sequenced the enriched DNA pool and found the most enriched site to be GGCCTAGGCGTTG. We then fixed that extended consensus site, extended the randomized dsDNA region, and redid the SELEX assay. This returned the most enriched site of GGCCTAGGCGTTATTTT (Fig. 3A).

Given this SELEX-enriched site, we constructed tandemly non-overlapping dsDNA libraries for test runs and noticed that GGCCTAGTCGTTTTTG has slightly higher affinity than the SELEX-enriched site, thus it was chosen as the reference site for Spec-seq library design [44, 49]. The four dsDNA libraries (Rand 9, 10, 11, 12), each with four randomized positions (a total of 1,021 sequences), were designed as in Fig. 3B. Spec-seq experiments were performed with mZFY1 (F7-F13), hZFY-full (F1-F13), and truncated versions of hZFY containing different subset of ZFs (F11-F13, F9-F13, F7-F13, F5-F13, F1-F11). Consistent with our previous work, we observed ~0.2*kT* measurement variation for individual sequence between multiple replicate runs, so we draw conclusions only with energy deviations above 0.2*kT*.

From the Spec-seq results, energy logos were created from the reference site and all single base variants. Figure 3C shows the logos for hZFY(F11-F13), hZFY(F7-F13) and mZFY(F7-F13). The logos for the longer proteins are essentially equivalent to the F7-F13 versions, indicating that those fingers are sufficient for the determining the specificity of the proteins. Comparing those logos, it is clear that Finger 11 can only recognize a 2nt long TA motif in positions 7-8, as opposed to the predicted 3nt TCA motif for positions 7-9 in Fig. 2B. The remaining GTN motif in positions 9-11 is likely recognized by Finger 10, which is missing in our (F11-F13) construct. While this kind of compressed or irregular motif has been reported before, this has only been for cases with a known crystal structure [34] or irregular linkers between fingers [50, 51].

Binding energy data by Spec-seq shows the dependence of accessory motif on the core site

The energetic landscape of ZFY-DNA interactions can be visualized by mapping the assayed binding sites according to their energy values and highlighting their mismatches to the reference site with red color (Fig.

3D). For hZFY(F7-F13), a few variants from the RAND11 library, covering the CGTT portion of the reference sequence, have slightly higher binding affinity (lower energy) than the reference. But most of the variants of the RAND11 and RNAD12 libraries have somewhat lower affinity, up to about 1*kT* higher binding energy. For the RAND10 library, covering the TAGT portion of the reference sequence, most variants have between 1-2.5 *kT* higher energy, and some of them have increases beyond 2.5*kT*. All of the variants from RAND9, covering the core region of the motif, have binding energies >2.5*kT* above the reference. These results confirm the existence of an irregular extended motif containing ----TAGT--TTT---. The greatly reduced affinity of those core variants explains why it is observed by methods like ChiP-seq or PBM but the weaker extended motif is not. To determine why Affinity-seq didn't detect this extended motif, we did a correlation analysis between the occurrences of all GGCCTNNN sites near Affinity-seq peak summits and their corresponding relative binding affinity by Spec-seq. For mZFY1, the most abundant site, GGCCTAGT, does have the highest binding affinity (Fig. 3E), nonetheless the overall correlation coefficient is not particularly high (r^2=0.224), probably because significant portions of Affinity-seq peaks (58,541/90,084 for hZFY and 24,801/50,170 for mZFY1) consist of multiple GGCC sites (at least 2) and therefore are not solely determined by the binding affinity of individual ones. That makes the enrichment too weak to be detected by MEME although it can be observed with quantitative binding assays.

To further explore the contribution of the extended motif, we included two extra DNA libraries R9N and R9NN (Fig 3B) and re-ran Spec-seq assay using multiple differently truncated versions of ZFY proteins. Library R9N randomized the core positions, 3-6, in the context of non-preferred bases in positions 8-12. Library R9NN also contains non-preferred bases in positions 15-18. If we symbolically divide the whole binding site to three regions, i.e., core site, extended motif part I, and part II, then many binding sites can be classified and given a short acronym depending on the sequence feature, either specific (S) or non-specific (N). For example, the acronym S-N-S refers to the site that has the specific (preferred) sequence in positions 3-6 and 15-18, but the non-specific sequence in positions 8-12. Figure 3F summarizes our measurement results for some characteristic sites. In a conventional motif prediction model, each finger recognizes its own motif and operates in independent, additive fashion. Under this additive model, the energy gap between a site with all specific subsequences (S-S-S) and sites with one or two non-specific subsequences (S-N-S and S-N-N) should be the same as the difference between N-S-S and the sites N-N-S and N-N-N. As expected, when the core site is present (S-x-x), the two other regions increase overall affinity additively (2.1*kT* for middle region, 0.7*kT* for last region, and 2.7*kT* for their sum for hZFY(F7-F13)). However, in the absence of the core (N-x-x), there is essentially no difference in binding specificity with or without those two regions, suggesting that the proper recognition of adjacent secondary motif depends on the presence of the core GGCC site. Note that this analysis does not allow us to determine relative affinities for the same sites between different proteins, rather each protein's binding energy is determined relative to the reference sequence.

Two-color fluorescence anisotropy reveals multi-state DNA binding for ZFY

Besides sequencing, we sought an alternative quantitative method to validate and further investigate differences in ZFY-DNA binding affinities. We recently developed an orthogonal method, two-color competitive fluorescence anisotropy (2C-CFA, Fig. 4A)[45], to quantify the relative binding affinity of a protein of interest to different DNA sequences. Two-color competitive anisotropy overcomes the noise inherent to conventional single-color anisotropy by measuring anisotropy values of reference and competitor probes in the same protein and buffer environment simultaneously. Jantz and Berg [52] demonstrated that a difference in anisotropy is detectable only if the fluorophore is proximal to the actual TF binding site within a few nucleotides, which makes it an ideal tool to probe the local properties of TF binding on DNA.

We designed five representative ZFY competitor probes, each with FAM labeled on either the left or right side, as in Fig. 4B. As expected, titration of increasing amounts of ZFY(F7-F13) protein into the "SSS vs. SSS" binding

reactions yielded consistent linear correspondence between FAM- and TAMRA- probes on either side, establishing that the placement of attached fluorophores has no effect on binding. For “SNS vs. SSS”, “NSS vs. SSS”, and “NNN vs. SSS” experiments on both sides, we obtained the same energy values matching the Spec-seq results well (Fig. 4E). Clearly for the NSS and NNN cases, because of lack of a core GGCC site to form a stable complex, both sequences showed almost the same affinities that were at least 3*kT* worse than the SSS reference site (Fig. 4C, D). However, interestingly, while the “FAM-SSN vs. TAMRA-SSS” experiment detected ~0.7*kT* energy difference, the reciprocal “SSN-FAM vs. SSS-TAMRA” case reported a significantly larger value (1.3*kT*). This discrepancy simply cannot be explained by any one-step, all-or-none DNA-binding model, and one plausible explanation is the existence of some intermediate protein-DNA complex state in which only part of the DNA probe is bound by ZFY. Given this, we proposed a multi-state, DNA binding and motif recognition model to explain our observation (Fig. 4F). Assuming for each DNA site, there is some intermediate state that only the core and its closely adjacent regions (positions 1-12) are bound by ZFY, denoted as state 2, 5, 8 respectively in Fig. 4F, then due to the local sensing property of anisotropy, for “FAM-SSN vs. TAMRA-SSS” experiment, we are detecting the binding occupancy difference between ensemble state 2+3 and ensemble state 5+6, whereas for “SSN-FAM vs. SSS-TAMRA” run, we are only monitoring the difference between state 3 and 6. Since state 2 and 5 are in the same energy level, overall smaller binding occupancy difference is observed by fluorophores on the left side than ones on the right side. On the other hand, for “SNS vs. SSS” runs, state 5 and state 8 already carry different sequences and energy values, so we wouldn’t get such asymmetric results. Additionally, we carefully performed conventional single-color anisotropy titrated by increasing ZFY(F7-F13) and observed its dissociation constant in nanomolar range (Fig. S4), which corresponds to at least 16 *kT* binding energy, thus the ~4 *kT* specific binding energy only contributes to small part of the overall protein-DNA interactions. Combining Spec-seq and anisotropy data, we are able to construct the energy landscapes for multiple sites including some intermediate states (Fig. 4G). The ~2 *kT* difference between fully occupied SSS and partially occupied SSS is derived using “FAM-SSS vs. TAMRA-SS” data with detail in Supplemental Info. Thermodynamically, many protein-DNA complex conformations should exist at room temperature, but most of them fall in the non-specific range. It is usually referred as multi valent binding in other areas [53], but has not been explored for ZFPs before.

ZFY accessory fingers incrementally increase the stability of the protein-DNA complex

Besides binding energy measurement, we directly measured the intrinsic dissociation rate for various ZFY truncation constructs interacting with different DNA sequences using fluorescence anisotropy (Fig. 5A). Each experiment assessed binding to one of the four versions of hZFY containing different subsets of fingers (F9-F13, F7-F13, F5-F13 and F1-F13 (full-length)) interacting with one of the four dsDNA probes of different lengths and sequences in the presence of a long consensus unlabeled ‘competitor’ probe. After reactions reach equilibrium, a 200-fold excess of the unlabeled competitor DNA is added and the anisotropy values of the FAM-DNA probes are monitored over time to visualize dissociation processes (Fig. 5B and 5C). The changes in mean lifetime for full-length hZFY to the different sequences are consistent with the Spec-seq measurement, for example SSS and SSN have 0.7*kT* energy difference measured by Spec-seq, corresponding to 1,176s and 656s of mean life time respectively. This means that binding affinity differences between different forms of protein-DNA complex are primarily driven by difference in dissociation rates (Fig. 5C, E). The full length ZFY protein displayed the longest mean lifetime of about 50mins, with each additional truncation of the protein, the mean lifetime are reduced down to about 15 minutes for fingers 9-13 (Fig. 5B and 5D). When the shorter SSS probe is bound by hZFY-full, the mean lifetime also drops to below 20mins (Fig. 5C and 5D). These results show that the additional fingers of hZFY contribute to binding affinity by stabilizing the complex, even fingers that do not appear to contribute to specificity.

CTCF fingers 10-11 recognize an upstream motif conditionally with variable spacing to the core sequence

Besides ZFY, do other tandem ZFPs recognize their motif in such a non-independent way? The CTCF insulator protein is composed of 11 tandem zinc fingers that have identical contact residues in humans and mice. Previous ChIP-chip and ChIP-seq work [31, 54, 55] identified a 14nt core motif CCnnnAGGGGGCGC, recognized by fingers 7 to 3, as in Fig. 6A, B. Recent work [56] using ChIP-seq, ChIP-exo, and DNaseI-seq combined with site-directed mutagenesis reported an extra "upstream motif" and "downstream motif" with a variable 5-6 nt distance to the core sequence (Fig. 6C, downstream motif not included). According to their analysis, among 48,137 detected ChIP-seq peaks, only 31,474 peaks contain the core sequences alone with no flanking motifs, and ~ 6,000 peaks have the upstream motif 5 or 6nt away from the core sites. We hypothesized that this flanking upstream motif may also represent a 'conditional motif' that depends on the proper recognition of CTCF to its core site and can enhance the overall protein-DNA complex binding.

To test this, we designed 4 Spec-seq libraries (R1, R2, R3 and R2L (Fig. 6D)) to assay the contributions of this upstream motif CTCF's specificity by testing the effects of the presence or absence of this motif at 5 and 6 nt spacings in the presence and absence of variations within the core at positions 2 and 6 (Y=C or T; M = A or C). Consistent with previous results, we identify an optimal upstream motif with preferred sequence TGCAGTACCC and a preferred spacing of 6 nt from the core motif (Fig. 6E, F). Fig. 6H shows that when the core sequence is intact (state C), the upstream motif can further enhance the DNA binding by ~1.7k_BT (U5C vs. N5C, U6C vs. N6C), but for mutated core sites, the energy differences are negligible (U5N vs. N5N, U6N vs. N6N). We also did Spec-seq on truncated mouse CTCF(F1-F9) and found no effect of the upstream motif, proving that fingers 10 and 11 are necessary for the upstream recognition.

Next, we asked whether the presence of the secondary motif could be detected by Affinity-seq. We profiled binding of the cloned mouse CTCF ZF array (which is identical to the human ZF array) to mouse genomic DNA. In total, we detected 192,670 CTCF peaks in the mouse genome at $p<0.01$. Analysis of all data with MEME reveals only the core motif. Because previous work reported that the secondary motif was found in only a fraction of the ChIP-exo data [56], we subdivided the 35,370 top activity peaks into 249 bins of decreasing peak intensity and ran MEME on each bin separately. Interestingly, the top 60 bins containing 4518 peaks showed strong match to the core motif only. The secondary motif was found predominantly in peaks with moderate peak intensities and of those, 4856 peaks, or 13.85%, showed secondary motif presence. These results suggest that secondary 'conditional motifs' may provide a method by which cells can compensate for core mutations to maintain overall high levels of binding.

Iterative analysis of Chip-exo data with prefixed core site reveals extended motif of human ZNF343

Our specificity analysis for ZFY and CTCF suggests many long ZFPs that are currently found to have short motifs may actually have extended motifs underrepresented in current motif discovery program. For instance, human ZNF343 is a 12-finger long KRAB ZFP, yet RCADE analysis of published ChIP-exo data reveal only a 6-nt long consensus site (GAAGCG) [30, 57], as in Fig. 7C. The motif prediction model (Fig. 7B) suggests this hexamer motif is most likely recognized by fingers 2-1. With this prior knowledge, we identified 3,237 GAAGCG sites within the reported 4,532 Chip-exo peaks. Using these sites as the "anchor point", we then aligned, mapped, and counted all the ChIP-exo reads near these core sites (Fig. 7F). There is a significant reverse peak signal at position 7, confirming that this hexamer is indeed recognized by the N-terminal fingers 1-2, but for the forward ChIP-exo signals, the peaks are quite broad and ambiguous. Ideally, the ChIP-exo read count at each site should be proportional to the binding occupancy of that site; if the *in vivo* ZNF343 protein concentration is low enough, reads should further be proportional to the binding affinity. Based on these assumptions, we calculated the negative logarithmic ratio of ChIP-exo read counts to yield relative binding

energies for data regression and motif analysis (Fig. 7D). This analysis revealed an extended motif in positions -8 to 0 that is very similar to the published result of HT-SELEX [58] (Fig 7E). We also group each binding site based on its number of mismatches to a putative consensus sequence GCCNNGGTGAAGCG and count the Chip-exo reads distribution for each group (Fig. 7G). The consensus site has the highest Chip-exo signal, consistent with the case for the longer motif.

Discussion

To tackle this "long fingers but short motifs" paradox, we must first ask, is it true that long ZFPs really have short motifs? There are some technical reasons that hinder long motif discovery. First, most existing techniques like PBM, ChIP-seq, and Affinity-seq have limited resolving power for weak binding sites. As suggested by one reviewer, for ChIP-seq work on human genomic DNA, each 15-mer will show up only once on average, which makes the discovery of longer motifs increasingly difficult. Second, the irregular nature of extended motifs like those seen for ZFY, where each finger does not appear to interact with the 'expected' three bases, can also inhibit motif detection when searches are guided by motif prediction methods. The CTCF upstream motif, which is outside the border of the predicted motif, was recently shown to be conferred by some irregular structural configuration [59]. A visual comparison between the predicted motifs and ChIP-seq results for 131 human ZFPs with reported motifs suggests that such irregular motifs may be quite common: for example, ZNF140, ZNF324, and ZNF449 all likely contain some irregular motifs with fewer than three base-pairs for some internal fingers (Supplemental Figure S3). Lastly, for ZFY and CTCF, the secondary motif is harder to detect when it is conditioned on the presence of primary motif. In the CTCF case, the upstream motif has significantly smaller energy amplitude ($\pm 0.6kT$/position) than the core motif ($\pm 2kT$/position), and the variable spacing between the two motifs also makes it harder for conventional motif programs to discover the upstream one.

Given those reasons, we think better motif discovery algorithms should take into account these intrinsic properties of long ZFPs, including irregular motif structure, variable spacing and non-independent recognition between sub-motifs. Because the non-additive dependence of secondary motif on the core site, it is computationally more effective to identify the core motif first and use it as anchor site to analyze flanking regions to extract more motif information for many other ZFPs, as we did for ZNF343. Current ZFP motif prediction and ChiP-seq motif discovery programs all assume independent specificity contribution from each finger, thus we expect some iterative motif discovery program in the future on high-quality data will yield more specificity information.

If some long ZFPs indeed have short motifs compared to their finger numbers, what would be the biological functions of those extra fingers? Very likely many of them are engaged in protein-protein and protein-RNA interactions. Alternatively, we speculate that the tandem array ZF architecture can modulate the TF-DNA complex stability, so that once the TF is bound to its regulatory elements, it can remain there long enough to fulfill its activation or repression function. Only very limited work has been done on this topic previously [60-62]. Here we used fluorescence anisotropy to monitor the *in vitro* dissociation process of ZFY protein-DNA complex over time, and found that those accessory fingers (F1-F11) incrementally increase the half-lives of the protein-DNA complex, which could be important for fulfilling ZFY's *in vivo* biological functions [39]. Alternatively, these secondary motifs make it possible to maintain binding in the face of evolutionary changes to the core motif. When we look at the overall correlation between *in vitro* binding predictions and *in vivo* occupancies, the correlations just really aren't all that satisfactory. At least in some cases, this could be because we're predicting binding based on the consensus site – but cells have a lot of ways to 'tune' binding. Here, we are showing ZFPs can increase overall affinity after a hit to the main binding site by just adding an additional upstream site.

Experimental and Data analysis Procedures

Data census for C2H2-containing proteins in model organisms

All raw protein sequences data were downloaded from InterPro database (IPR013087). For each model organisms, Cys-$X_{2,4}$-Cys-X_{12}-His-$X_{3,4,5}$-His was used to search and match individual finger in each protein sequence, and only protein sequences with unique name were included for final census (e.g. PRDM9 was only counted once for human). Processed data was deposited to Mendeley Data.

Construction and expression of recombinant proteins

The coding sequences for human ZFY (Uniprot P08048:408-768) and mouse ZFY1 (Unirprot P10925:390-782) were codon optimized for E. coli expression and synthesized by IDT gBlock service, whereas mouse CTCF (Uniprot Q61164-1:241-583) was cloned from mouse cDNA libraries. After In-Fusion cloning into NEB DHFR control vector with N-terminal hisSUMO tag, the expression and purification procedures were essentially the same as our previous work [45] except that extra heparin column purification is used for anisotropy experiment. The plasmid of hisSUMO-mCTCF(F1-F9) was reused from our previous work is readily available from Addgene(#102859). All constructs, including truncated versions, are listed in Supplemental Table S1.

Affinity-seq procedures

Affinity-seq was essentially done as in [36] with minor adjustments. A ZF array of the protein of interest is amplified then cloned into a universal Affinity-seq vector by recombineering. The resulting construct expresses a fused protein containing 6HisHALO–the 412-511 aa fragment of PRDM9– ZF array of interest. The fused protein is expressed in Rosetta 2 cells at 15^{o}C for 24 h, and partially purified by ion exchange chromatography on SP-sepharose. The purified protein is mixed with genomic DNA sheared to ~200 bp on a Covaris ultrasonicator, and allowed to bind overnight. The protein-DNA complexes are then isolated on HisPur Ni-NTA Resin (Thermo Scientific) preincubated with partially purified prep of the empty tag to reduce the background. DNA is then eluted and used to prepare genomic libraries using TruSeq ChIP Library Prep Kit (Illumina). The libraries are sequenced on HiSeq2500 or NextSeq platform ensuring ~50 mln. reads per library. Data are analyzed using a custom pipeline as described before [36]. Motif analysis was done using MEME software package.

HT-SELEX procedures

For the first round of HT-SELEX, ~200ng dsDNA libraries containing randomized sequence CAGGCCTNNNNNNNN was used for EMSA shift with hisSUMO-hZFY titrated from low to high concentration. Each time only the lane containing lowest amount protein was chosen and the bound portion of DNA (no more than 20% of total DNA) was cut, amplified for next round of SELEX selection enrichment. Since in the first round of HT-SELEX, the most enriched site turned out to be CAGGCCTAGGCGTTG, the DNA library was redesigned as CAGGCCTAGGCGTNNNNNNNN for further HT-SELEX by EMSA separation. Again, each time no more than 20% of total DNA should be in bound state for selection and enrichment analysis.

Spec-seq procedures

The experimental procedures are essentially the same as our previous work, with all binding reactions set up at 1X NEBuffer 4, room temperature, all EMSA performed at 9% Tris-glycine gel, cold room, 200V, 30mins. For ZFY, given the putative consensus site, we designed tandemly randomized dsDNA libraries Rand1-8 for pilot run, and it turned out GGCCTAGTCGTTTTTG has slightly higher affinity than the SELEX-enriched site, so we redesigned dsDNA libraries Rand9-12 including 9N, 9NN to cover some partially non-specific sites listed in Table 1 (main text). We noticed that for some ZFY proteins, particularly ZFY(F11-F13), when the protein concentration is too high, the shifted DNA fragments easily to form protein oligomers or aggregate near the EMSA well, so generally low concentration of protein (<100uM) was used and only monomer ZFY-DNA complex was cut for Spec-seq analysis. By default, all energy matrices are derived by data regression of the

binding energy of reference site and all its single variants. For ZFY and CTCF upstream motif, the reference sites are GGCCTAGTCGTTTTTG and TGCAGTNCCN respectively.

Two-color fluorescence anisotropy

All assays are set at 1X NEBuffer 4, room temperature, 40nM FAM- or TAMRA-labeled DNA probes. Nickel and heparin column purified hisSUMO-ZFY (F7-F13) was used for anisotropy experiment. Experimental procedures are same as described in previous work[45].

Dissociation kinetics assay by fluorescence anisotropy

All binding assays are all set in 1X NEBuffer 4, 37° C with 30nM FAM-DNA probe, and in this condition the basal value for FAM-DNA probe without protein is ~15mA. With saturating concentration of ZFY protein added, the anisotropy values can go above 100mA, but in our cases, we titrated low volume of protein (<4% v/v) and the initial values in equilibrium state are slightly above 40mA, therefore in such low occupancy state (<20%), the DNA probe is more likely bound by the protein in assumed specific conformations. After we injected highly concentrated unlabeled competitor DNA (500pM/uL X 2uL) into 100uL binding reactions, the molar ratio between FAM probe and competitor DNA should go below 1:200, after that we measured the anisotropy values with 20s or 40s time interval for up to 90mins.

To measure the intrinsic dissociation rate of protein-DNA complex, we did titration experiments first with different molar ratio of unlabeled competitor DNA into the binding reactions, as in Fig. S1. When the competition ratio is below 1:100, the observed dissociation rate reaches some plateau and cannot increase further, thus we assume it is appropriate to use 1:200 competitor ratio curve to estimate the intrinsic dissociation rate or mean lifetime of the protein-DNA complex.

After setting up the binding reaction for at least 20mins, we assumed the equilibrium state is reached. Slightly to our surprise, with no competitor DNA added, we still observed the anisotropy value slowly decrease over time, probably owing to the steady inactivation or degradation of ZFY protein at 37 °C. To exclude the possibility that the measured dissociation rates are differentially biased by different protein inactivation rates, we measured this inactivation process alone for different proteins, and they all showed very similar inactivation rates, which are significantly slower than our observed dissociation rates (Fig. S2).

To estimate the dissociation rate k_{off} or mean lifetime τ, we fit our data using exponential decay model with following equation:

$$\mathrm{FAM(t)} = range \times e^{-t/\tau} + base$$

The base value is usually in the range 15 to 17, and the range parameter depends on the first measured anisotropy value in each experiment, which should not affect the mean lifetime. Each experiment was repeated at least three times to calculate the mean values and standard deviations (as in Table S2). All experiments were performed using TECAN Safire2 instrument, 490nm excitation/525nm emission wavelength.

Motif analysis of human ZNF343 using published Chip-exo data

For human ZNF343, to derive the extended motif beyond the previously identified core motif GAAGCA, we searched this hexamer and found 3,237 intact binding sites within all those 4,532 binding peaks called by Trono lab (NCBI GEO GSE78099). Using this hexamer as the prefixed core (position 1-6), all flanking sequences within ±40bp ranges were extracted and aligned accordingly (from -39 to 46). For each particular site, all those raw Chip-exo reads (NCBI SRA SRX2512772) falling in its neighboring range were mapped based on the their starting positions, either in forward or reverse direction, thus the total Chip-exo reads for this site was calculated as the forward exo reads count (position -39 to +5) plus the reverse exo reads count (position +2 to 46). Ideally the Chip-exo reads count for each site should be proportional to the binding probability of that

site, or approximately its binding affinity if the *in vivo* ZNF343 protein occupancy is low enough, therefore the negative logarithmic ratio of its Chip-exo reads was used as the relative binding energy for data regression and motif analysis. All processed data including binding site reads and box plot analysis was deposited to Mendeley Data.

Supplemental Information

Supplemental information includes experimental conditions and procedures, anisotropy measurement descriptions and record, and irregular motif predictions. All raw and processed sequencing data were deposited to NCBI GEO database (Affinity-seq #GSE111772, ZFY Spec-seq #GSE109098, CTCF Spec-seq #GSE110155). C2H2 proteins census, Chip-exo data analysis for ZNF343, and all related PWM/Energy matrices were deposited to Mendeley Data.

Author Contributions

Z.Z. designed and performed the HT-SELEX, Spec-seq and anisotropy dissociation experiments in G.D.S. lab, performed two-color anisotropy in P.F. lab, wrote the first draft of manuscript. P.M.P., T.B., and M.W. performed and analyzed the Affinity-seq experiments. G.D.S. and P.F. supervised and revised this work.

Acknowledgement and Funding

We thank Dr. Rafael Casellas for providing the CTCF PWM data for motif comparison purposes. We also thank an anonymous reviewer for helpful suggestions to improve the presentation. This work was supported by NIH grants HG000249 (GDS) and GM078452 (PMP).

References

1. Klug, A. and D. Rhodes, *'Zinc fingers': a novel protein motif for nucleic acid recognition.* Trends in Biochemical Sciences, 1987. **12**: p. 464-469.
2. Miller, J., A.D. McLachlan, and A. Klug, *Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes.* EMBO J, 1985. **4**(6): p. 1609-14.
3. Hamilton, A.T., et al., *Lineage-specific expansion of KRAB zinc-finger transcription factor genes: implications for the evolution of vertebrate regulatory networks.* Cold Spring Harb Symp Quant Biol, 2003. **68**: p. 131-40.
4. Huntley, S., et al., *A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors.* Genome Res, 2006. **16**(5): p. 669-77.
5. Tadepally, H.D., G. Burger, and M. Aubry, *Evolution of C2H2-zinc finger genes and subfamilies in mammals: species-specific duplication and loss of clusters, genes and effector domains.* BMC Evol Biol, 2008. **8**: p. 176.
6. Vaquerizas, J.M., et al., *A census of human transcription factors: function, expression and evolution.* Nat Rev Genet, 2009. **10**(4): p. 252-63.

7. Kapopoulou, A., et al., *The evolution of gene expression and binding specificity of the largest transcription factor family in primates.* Evolution, 2016. **70**(1): p. 167-80.
8. Ecco, G., M. Imbeault, and D. Trono, *KRAB zinc finger proteins.* Development, 2017. **144**(15): p. 2719-2729.
9. Brayer, K.J. and D.J. Segal, *Keep your fingers off my DNA: protein-protein interactions mediated by C2H2 zinc finger domains.* Cell Biochem Biophys, 2008. **50**(3): p. 111-31.
10. Gamsjaeger, R., et al., *Sticky fingers: zinc-fingers as protein-recognition motifs.* Trends Biochem Sci, 2007. **32**(2): p. 63-70.
11. Hall, T.M., *Multiple modes of RNA recognition by zinc finger proteins.* Curr Opin Struct Biol, 2005. **15**(3): p. 367-73.
12. Najafabadi, H.S., et al., *C2H2 zinc finger proteins greatly expand the human regulatory lexicon.* Nature biotechnology, 2015. **33**(5): p. 555-562.
13. Lambert, S.A., et al., *The Human Transcription Factors.* Cell, 2018. **172**(4): p. 650-665.
14. Wolfe, S.A., L. Nekludova, and C.O. Pabo, *DNA recognition by Cys2His2 zinc finger proteins.* Annu Rev Biophys Biomol Struct, 2000. **29**: p. 183-212.
15. Pavletich, N.P. and C.O. Pabo, *Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A.* Science, 1991. **252**(5007): p. 809-17.
16. Desjarlais, J.R. and J.M. Berg, *Use of a zinc-finger consensus sequence framework and specificity rules to design specific DNA binding proteins.* Proceedings of the National Academy of Sciences, 1993. **90**(6): p. 2256-2260.
17. Desjarlais, J.R. and J.M. Berg, *Toward rules relating zinc finger protein sequences and DNA binding site preferences.* Proc Natl Acad Sci U S A, 1992. **89**(16): p. 7345-9.
18. Choo, Y. and A. Klug, *Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage.* Proc Natl Acad Sci U S A, 1994. **91**(23): p. 11163-7.
19. Nardelli, J., T. Gibson, and P. Charnay, *Zinc finger-DNA recognition: analysis of base specificity by site-directed mutagenesis.* Nucleic Acids Res, 1992. **20**(16): p. 4137-44.
20. Pomerantz, J.L., S.A. Wolfe, and C.O. Pabo, *Structure-based design of a dimeric zinc finger protein.* Biochemistry, 1998. **37**(4): p. 965-70.
21. Benos, P.V., A.S. Lapedes, and G.D. Stormo, *Probabilistic code for DNA recognition by proteins of the EGR family.* J Mol Biol, 2002. **323**(4): p. 701-27.

22. Benos, P.V., A.S. Lapedes, and G.D. Stormo, *Is there a code for protein-DNA recognition? Probab(ilistical)ly.* Bioessays, 2002. **24**(5): p. 466-75.
23. Kaplan, T., N. Friedman, and H. Margalit, *Ab initio prediction of transcription factor targets using structural knowledge.* PLoS Comput Biol, 2005. **1**(1): p. e1.
24. Fu, F. and D.F. Voytas, *Zinc Finger Database (ZiFDB) v2.0: a comprehensive database of C(2)H(2) zinc fingers and engineered zinc finger arrays.* Nucleic Acids Res, 2013. **41**(Database issue): p. D452-5.
25. Gupta, A., et al., *An optimized two-finger archive for ZFN-mediated gene targeting.* Nat Methods, 2012. **9**(6): p. 588-90.
26. Persikov, A.V., et al., *A systematic survey of the Cys2His2 zinc finger DNA-binding landscape.* Nucleic Acids Res, 2015. **43**(3): p. 1965-84.
27. Gupta, A., et al., *An improved predictive recognition model for Cys(2)-His(2) zinc finger proteins.* Nucleic Acids Res, 2014. **42**(8): p. 4800-12.
28. Najafabadi, H.S., M. Albu, and T.R. Hughes, *Identification of C2H2-ZF binding preferences from ChIP-seq data using RCADE.* Bioinformatics, 2015. **31**(17): p. 2879-81.
29. Persikov, A.V. and M. Singh, *De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins.* Nucleic acids research, 2013. **42**(1): p. 97-108.
30. Imbeault, M., P.Y. Helleboid, and D. Trono, *KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks.* Nature, 2017. **543**(7646): p. 550-554.
31. Schmitges, F.W., et al., *Multiparameter functional diversity of human C2H2 zinc finger proteins.* Genome Res, 2016. **26**(12): p. 1742-1752.
32. Edelstein, L.C. and T. Collins, *The SCAN domain family of zinc finger transcription factors.* Gene, 2005. **359**: p. 1-17.
33. Hashimoto, H., et al., *Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA.* Molecular Cell, 2017. **66**(5): p. 711-720. e3.
34. Pavletich, N.P. and C.O. Pabo, *Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers.* Science, 1993. **261**(5129): p. 1701-1707.
35. Billings, T., et al., *DNA binding specificities of the long zinc-finger recombination protein PRDM9.* Genome biology, 2013. **14**(4): p. R35.
36. 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, 2015. **8**: p. 31.

37. Palmer, M., et al., *Genetic evidence that ZFY is not the testis-determining factor.* Nature, 1989. **342**(6252): p. 937-939.
38. Berta, P., et al., *Genetic evidence equating SRY and the testis-determining factor.* Nature, 1990. **348**(6300): p. 448-450.
39. Vernet, N., et al., *Zfy genes are required for efficient meiotic sex chromosome inactivation (MSCI) in spermatocytes.* Human molecular genetics, 2016. **25**(24): p. 5300-5310.
40. Taylor-Harris, P., S. Swift, and A. Ashworth, *Zfy1 encodes a nuclear sequence-specific DNA binding protein.* FEBS letters, 1995. **360**(3): p. 315-319.
41. Grants, J., et al., *Characterization of the DNA binding activity of the ZFY zinc finger domain.* Biochemistry, 2010. **49**(4): p. 679-686.
42. Weirauch, M.T., et al., *Evaluation of methods for modeling transcription factor sequence specificity.* Nature biotechnology, 2013. **31**(2): p. 126-134.
43. Chen, X., et al., *Integration of external signaling pathways with the core transcriptional network in embryonic stem cells.* Cell, 2008. **133**(6): p. 1106-1117.
44. Zuo, Z. and G.D. Stormo, *High-resolution specificity from DNA sequencing highlights alternative modes of Lac repressor binding.* Genetics, 2014. **198**(3): p. 1329-43.
45. Zuo, Z., et al., *Measuring quantitative effects of methylation on transcription factor-DNA binding affinity.* Sci Adv, 2017. **3**(11): p. eaao1799.
46. Liu, J. and G.D. Stormo, *Combining SELEX with quantitative assays to rapidly obtain accurate models of protein-DNA interactions.* Nucleic Acids Res, 2005. **33**(17): p. e141.
47. Jolma, A., et al., *Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities.* Genome Res, 2010. **20**(6): p. 861-73.
48. Zhao, Y., D. Granas, and G.D. Stormo, *Inferring binding energies from selected binding sites.* PLoS Comput Biol, 2009. **5**(12): p. e1000590.
49. Stormo, G.D., Z. Zuo, and Y.K. Chang, *Spec-seq: determining protein-DNA-binding specificity by sequencing.* Brief Funct Genomics, 2015. **14**(1): p. 30-8.
50. Nolte, R.T., et al., *Differing roles for zinc fingers in DNA recognition: structure of a six-finger transcription factor IIIA complex.* Proc Natl Acad Sci U S A, 1998. **95**(6): p. 2938-43.

51. Garton, M., et al., *A structural approach reveals how neighbouring C2H2 zinc fingers influence DNA binding specificity.* Nucleic Acids Res, 2015. **43**(19): p. 9147-57.
52. Jantz, D. and J.M. Berg, *Probing the DNA-binding affinity and specificity of designed zinc finger proteins.* Biophysical journal, 2010. **98**(5): p. 852-860.
53. Markin, C.J., W. Xiao, and L. Spyracopoulos, *Mechanism for recognition of polyubiquitin chains: balancing affinity through interplay between multivalent binding and dynamics.* Journal of the American Chemical Society, 2010. **132**(32): p. 11247-11258.
54. Kim, T.H., et al., *Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome.* Cell, 2007. **128**(6): p. 1231-1245.
55. Jothi, R., et al., *Genome-wide identification of in vivo protein–DNA binding sites from ChIP-Seq data.* Nucleic acids research, 2008. **36**(16): p. 5221.
56. Nakahashi, H., et al., *A genome-wide map of CTCF multivalency redefines the CTCF code.* Cell Rep, 2013. **3**(5): p. 1678-1689.
57. Barazandeh, M., et al., *Comparison of ChIP-Seq Data and a Reference Motif Set for Human KRAB C2H2 Zinc Finger Proteins.* G3: Genes, Genomes, Genetics, 2018. **8**(1): p. 219-229.
58. Yin, Y., et al., *Impact of cytosine methylation on DNA binding specificities of human transcription factors.* Science, 2017. **356**(6337): p. eaaj2239.
59. Yin, M., et al., *Molecular mechanism of directional CTCF recognition of a diverse range of genomic sites.* Cell research, 2017. **27**(11): p. 1365.
60. Elf, J., G.-W. Li, and X.S. Xie, *Probing transcription factor dynamics at the single-molecule level in a living cell.* Science, 2007. **316**(5828): p. 1191-1194.
61. Scholes, C., A.H. DePace, and Á. Sánchez, *Combinatorial Gene Regulation through Kinetic Control of the Transcription Cycle.* Cell systems, 2017. **4**(1): p. 97-108. e9.
62. Riggs, A.D., S. Bourgeois, and M. Cohn, *The lac represser-operator interaction: III. Kinetic studies.* Journal of molecular biology, 1970. **53**(3): p. 401-417.

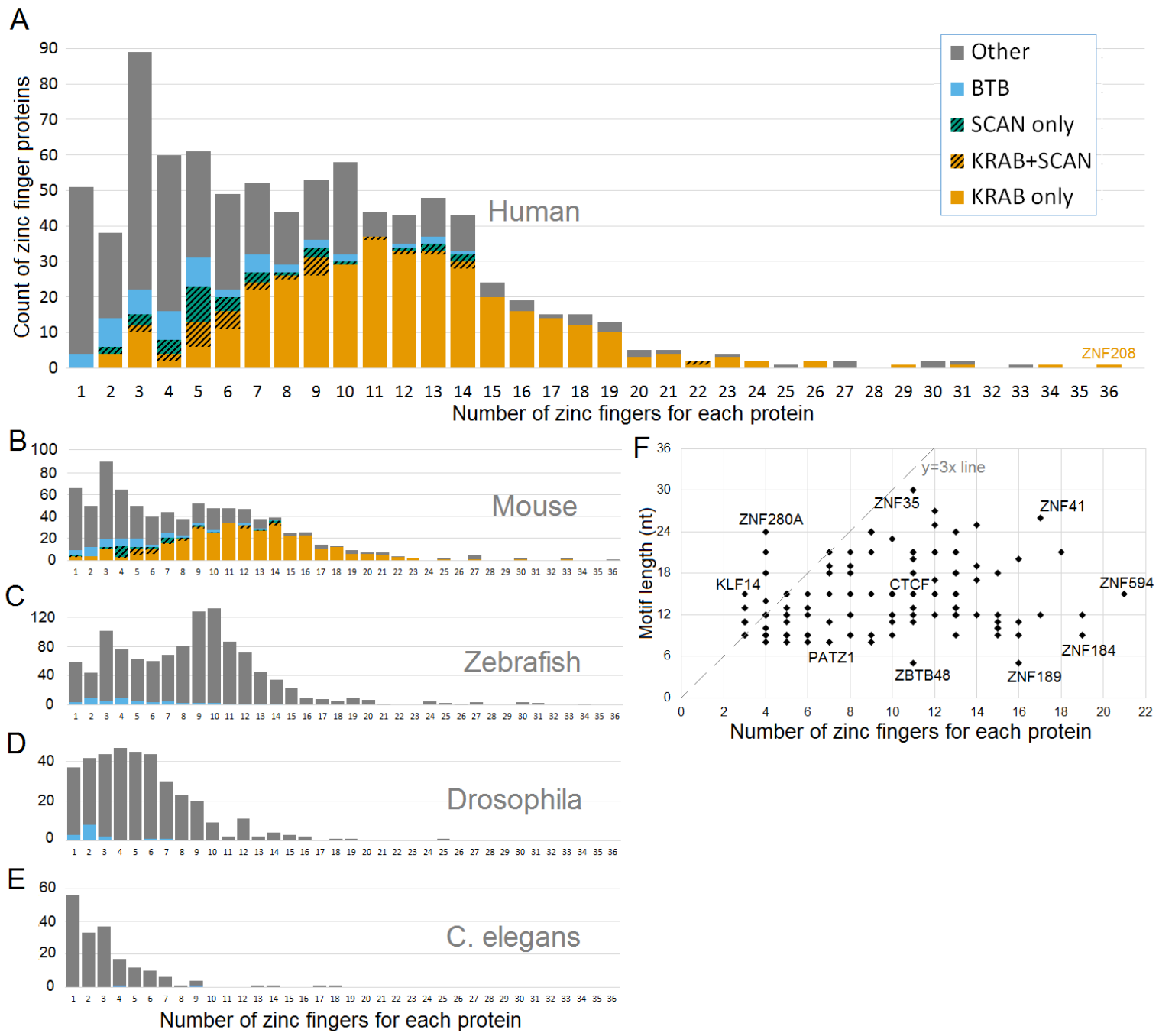


Figure 1. A) A census for all those C2H2-ZF containing proteins in human genome. Only ZFPs with unique names are counted; each gene is classified based on its effector domain, i.e., KRAB, SCAN, or BTB; B-E) Census for Mouse, Zebrafish, Drosophila, and C. elegans respectively; F) For 131 human ZFPs, the finger numbers and motif lengths according to recent Chip-seq results are displayed [31].

A

Finger number	13	12	11	10	9	8	7	6	5	4	3	2	1
Critical residues human ZFY	RSEQ	DSGR	QNEK	DFVR	RSEK	NSDR	DKEQ	DSNT	HSEK	EGLR	HGAT	KISN	SGFR
mouse ZFY1			QCET	DFLH					HSAK	ETLH	HGTT		TRFR
mouse ZFY2			QCET	DFLH					HSAK	ETLH	missing		TRFR
human/mouse ZFX			QSEK										

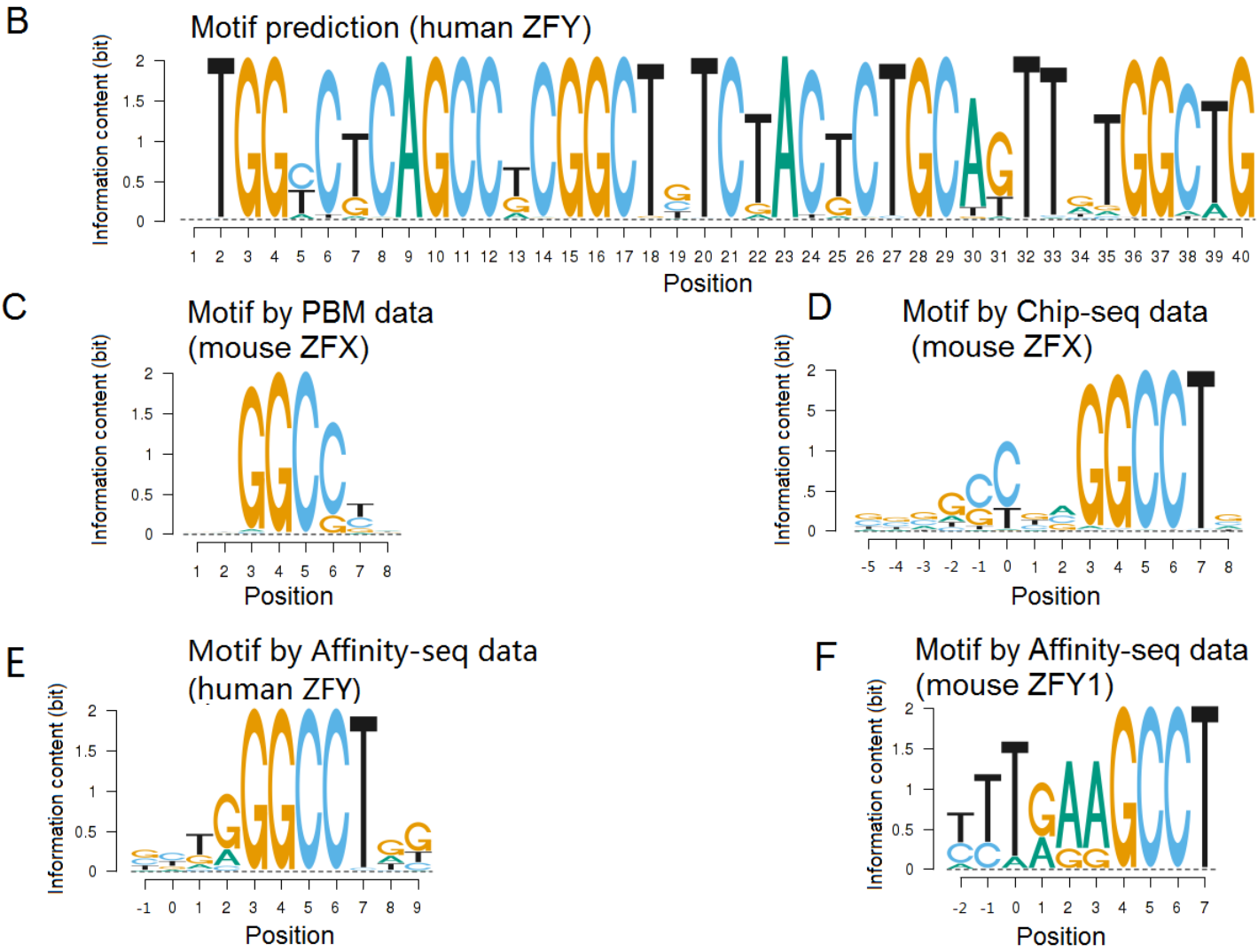


Figure 2 ZFY/ZFX protein architectures and motifs made by various methods.

A) Critical residues composition for human ZFY, mouse ZFY1, mouse ZFY2 and human/mouse ZFX. Human ZFY was used as reference so for other proteins only those different portions were shown; B) Motif prediction for human ZFY[29], in which positions 1-39 are recognized by Finger 13 to 1 and aligned with finger pattern in Fig. 1A; C) Motif from mouse ZFX PBM data [41]; D) Motif from mouse ZFX Chip-seq data [42]; E) Human ZFY motif from Affinity-seq data; F) Mouse ZFY1 motif from Affinity-seq data. MEME was used as the default motif-finding program.

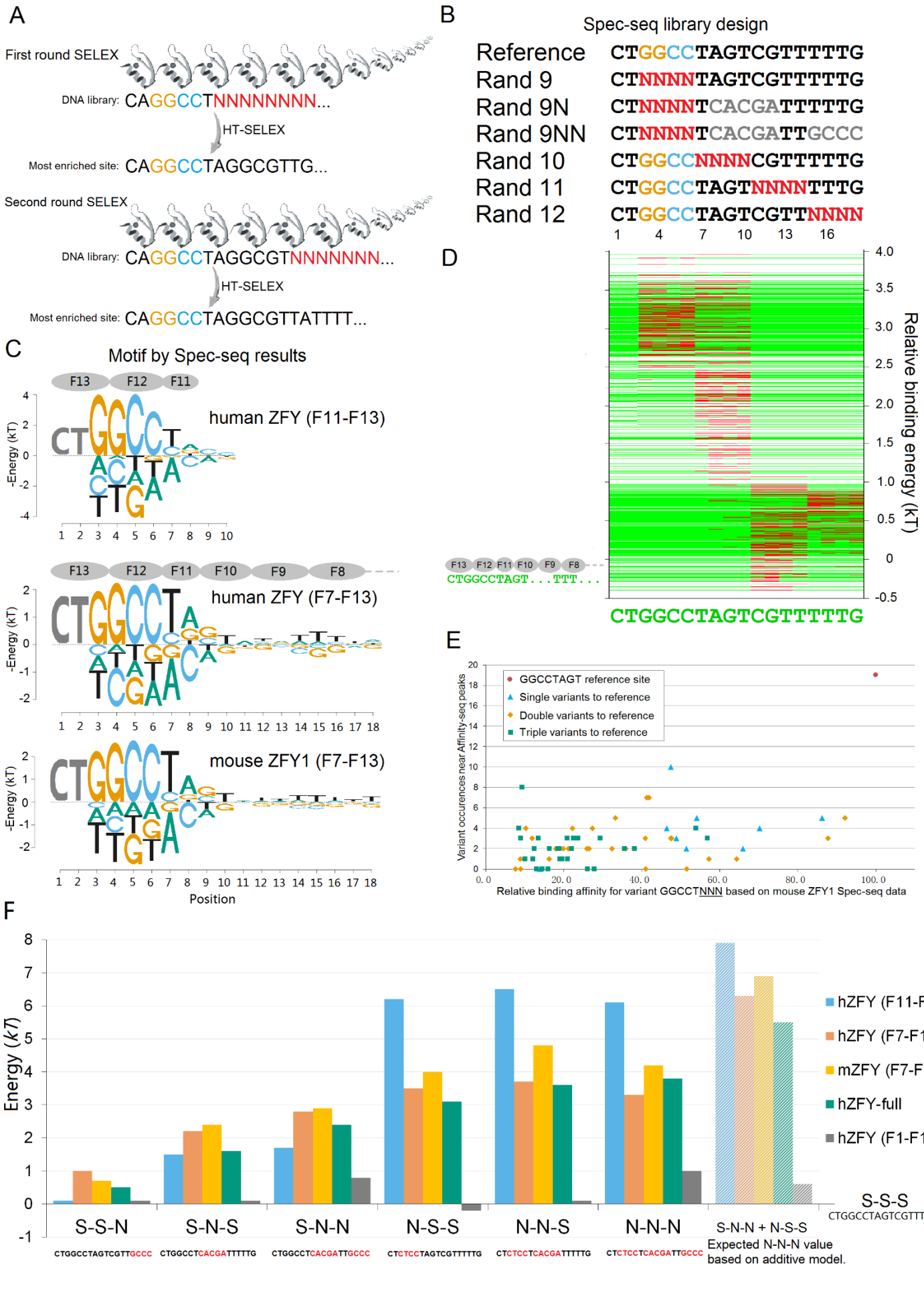


Figure 3. High-throughput SELEX discovered extended motif, which can be further quantified by Spec-seq analysis. A) Schematics for two successive rounds of HT-SELEX, in which the GGCCT core sequence was prefixed for recognition by F13-F12; B) Given the preferred binding site identified by HT-SELEX, tandemly non-overlapping randomized dsDNA libraries were designed for Spec-seq analysis; C) Energy logos produced by Spec-seq data for truncated human F11-F13, human F7-F13, and mouse F7-F13 respectively; D) All variants in libraries Rand 9-12 were ranked and displayed according to their measured binding energy values for hZFY(F7-F13). The segments of each sequence that match and mismatch the reference site are highlighted in green and red respectively; E) Correlation analysis between measured binding affinity by mZFY1(F7-F13) Spec-seq and the observed sequence occurrences near mZFY1 Affinity-seq peak summits for all GGCCTNNN variants; F) Spec-seq measurement of different ZFY constructs to different representative DNA sequences. S-S-S (CTGGCCTAGTCGTTTTTG) is chosen as reference site with relative energy 0. For each sequence mismatches to the reference site are highlighted in red. Expected N-N-N value based additive model is also shown.

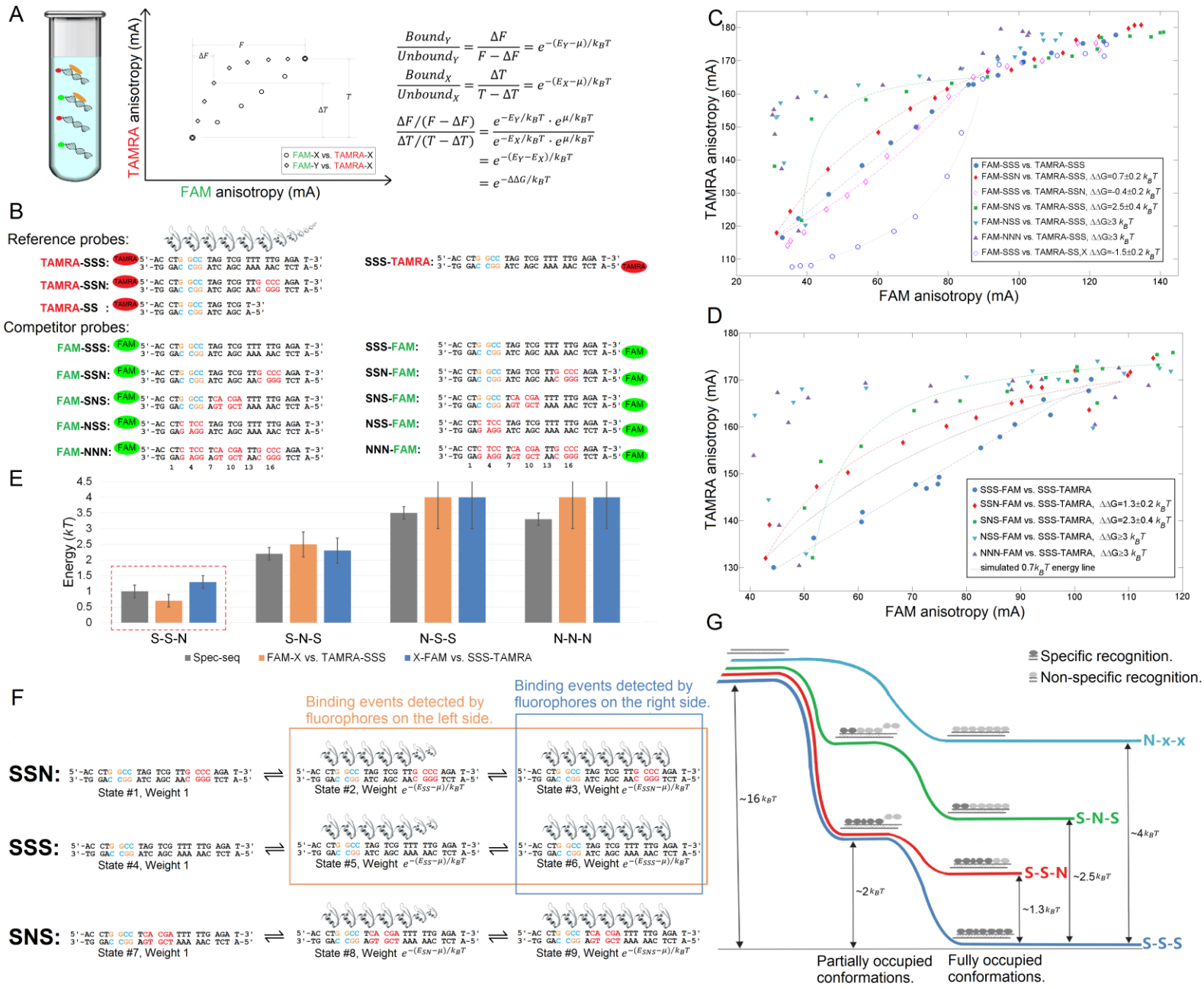


Figure 4. A multi-state DNA binding model of ZFY(F7-F13) revealed by Two-color fluorescence anisotropy.

A) Schematics for measuring the relative affinity or energy difference between two binding sites X and Y to protein of interest. The energy difference is calculated based on simple all-or-none, two-state binding model; B) Design of TAMRA- labeled reference probes and FAM- labeled competitor probes. Each probe was labeled either from left or right side; C) Two-color anisotropy value correspondence for each pair of sequences with fluorophores on the left side; D) Two-color anisotropy value correspondence for each pair of sequences with fluorophores on the right side; Energy difference between reference and competitor was calculated by curve fitting of equations in 4A; Simulated 0.7kT energy line was drawn as comparison with observed SSN data; E) Comparisons of Spec-seq and Two-color anisotropy measurement results for ZFY(F7-F13); F) Proposed multi-state DNA-binding model for sequence SSN, SSS, and SNS. For intermediate state 2, 5, 8, only part of DNA is bound by ZFY(F7-F13); G) Energy landscapes for different sequences modeled on Spec-seq and fluorescence anisotropy data. Specific and Non-specific recognitions are labeled in different shades; The 2*kT* energy gap between fully and partially bound conformations are derived by equations in Supplemental Info.

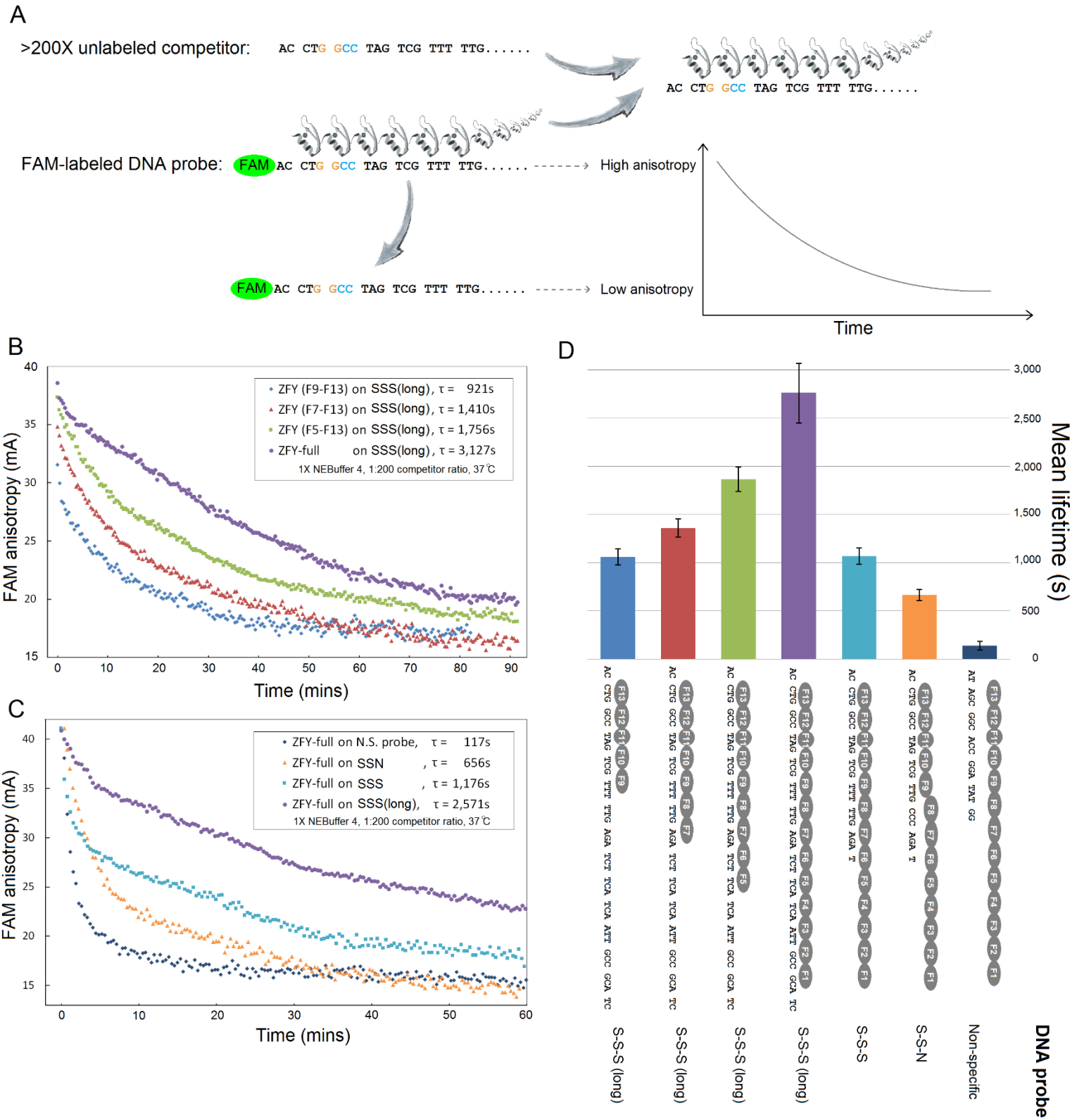


Figure 5. Dissociation kinetics of ZFY proteins by fluorescence anisotropy.

A) Schematics for measuring the dissociation rate/mean lifetime of ZFY protein-DNA complex. dsDNA probes with different sequences and length were labeled with FAM fluorophores on the 5' end for tracing fluorescence anisotropy values, whereas the competitor DNA probe has no FAM label. After 200-fold competitor DNA was added into binding reactions, human ZFY and its differentially truncated proteins exhibit differential dissociation curves on various probes; B) Dissociation curves for different truncated ZFY proteins on long probe; C) Dissociation curves for full-length human ZFY on different DNA probes; D) Average values and standard deviations of the mean lifetime for different ZFY protein-DNA

complexes. Each measurement was repeated at least three times to calculate average and standard deviation (Supplemental Table S2), where 5B and 5C only displayed one particular curve for each sample.

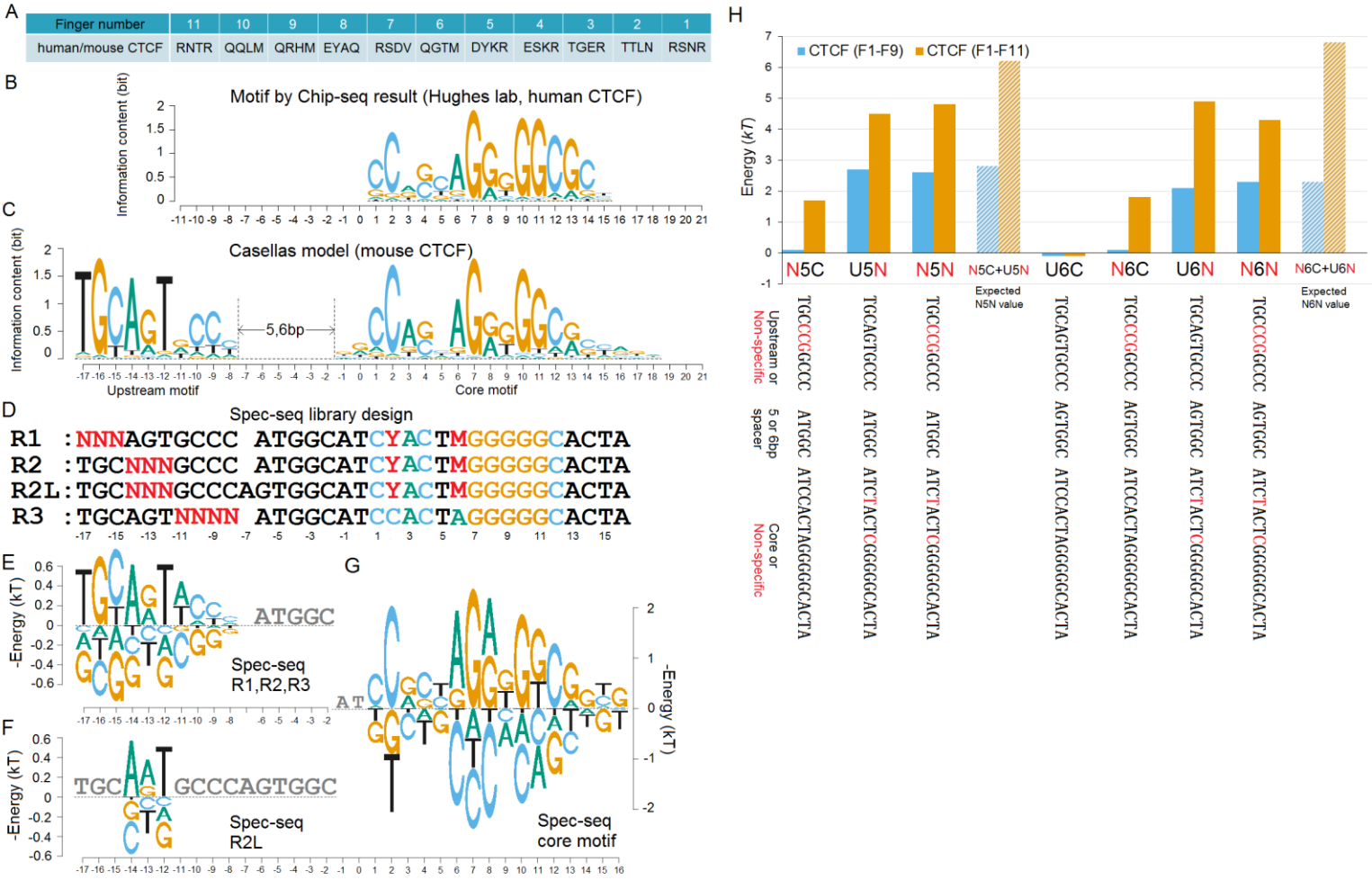


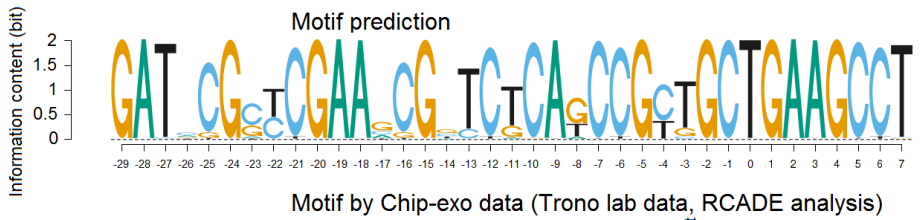
Figure 6. CTCF protein architecture and recognition motifs made by different methods

A)The contact residues composition for human and mouse CTCF proteins; B) Recognition motif of human CTCF produced by Chip-seq experiment[31]; C) Recognition motif model of mouse CTCF based on Chip-exo experiment, including the upstream motif with variable distance to the canonical core motif [43]; D) Spec-seq library design; Y for C/T, M for A/C, and N for A/C/G/T; E) Spec-seq energy logo made by data regression of upstream reference site TGCAGTNCCN and all those single variants from library R1,R2, and R3; F) Spec-seq energy logo made by regression of upstream site TGCAGTGCCC and all its single variants in library R2L; G) Spec-seq energy logo of CTCF core motif using mouse CTCF(F1-F9), from previously published data [45]; H) Relative binding energy for characteristic binding sites with mouse CTCF(F1-F9) or CTCF(F1-F11) by Spec-seq. TGCAGTGCCC ATGGC ATCCACTAGGGGGCACTA was chosen as reference site with energy zero and designated as “Upstream-5bp Gap-Core”, or “U5C” state, whereas for other sites, the mismatched portions of sequence are highlighted in red. All energy number are in k_BT units. Expected values of N5N and N6N based on additive model are also shown.

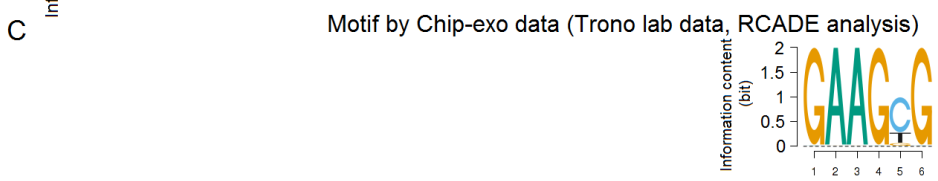
A

Finger number	12	11	10	9	8	7	6	5	4	3	2	1
human ZNF343	HSNR	RSLV	DSTV	QSNR	RSLV	DSTI	QSDK	DSTK	ESSR	SSIR	QSNR	DSTR

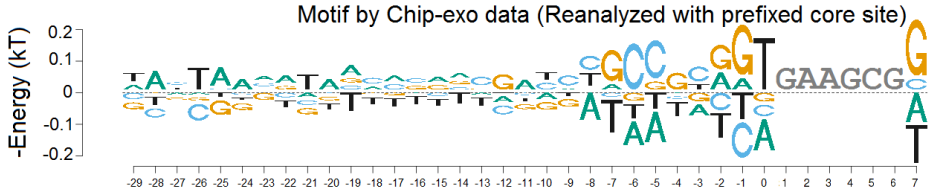
B



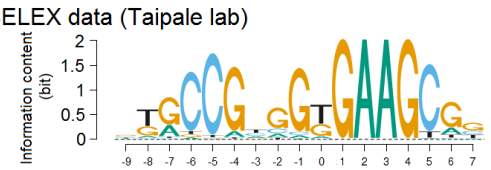
C



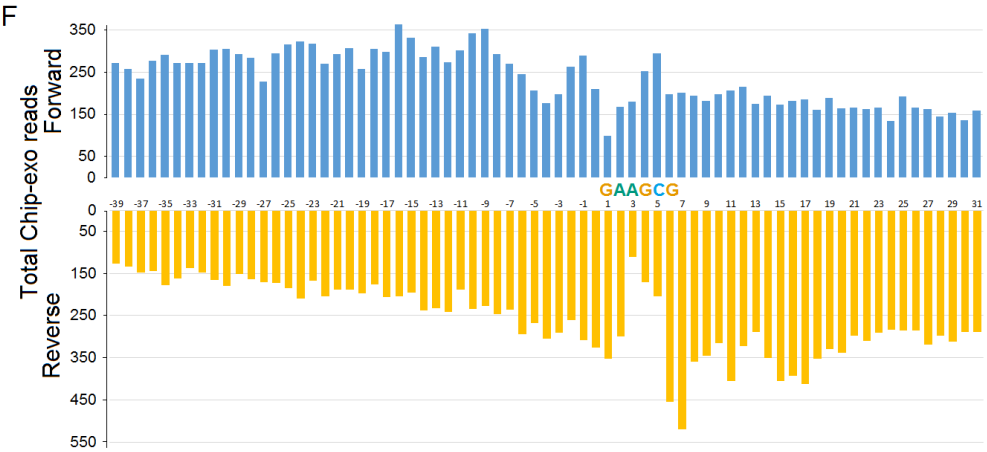
D



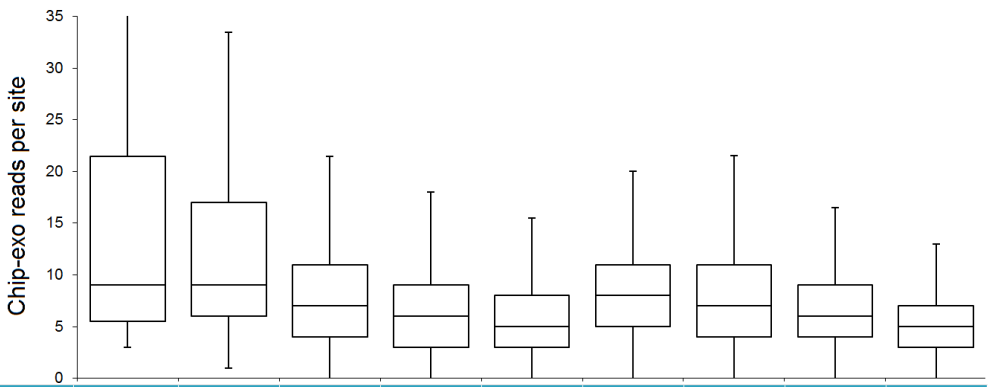
E



F



G



Group	1	2	3	4	5	6	7	8	9
Reference	GCCNNGGTGAAGCG		GCCNNNNNGAAGCG			GGTGAAGCG			
Mismatches	0	0	1	2	3	0	1	2	3
Number of sites	7	208	774	1,485	770	91	857	1,630	659
Mean value	20.7	13.6	9.1	7.3	6.8	9.9	9.6	7.7	6.6

Figure 7 Specificity analysis of human ZNF343

A)The contact residues composition for human ZNF343; B) Motif prediction for human ZNF343; C) Motif made by RCADE analysis of ZNF343 Chip-exo peaks[28]; D) Energy logo made by reanalysis of Chip-exo data with prefixed core site GAAGCG at position 1-6, detailed in Supplemental Method section; E) Motif produced by Taipale lab HT-SELEX results[56]; F) Accumulative Chip-exo reads near the GAAGCG core sites within Chip-exo peaks; Each read was mapped by its starting position, either in forward or reverse direction; G) Box plots with maximum whiskers at 1.5 IQR showing the Chip-exo reads distribution. Each site is classified by its mismatches to underlined portions of the reference sequences; Group 1 contains all reference sites matching GCCNNGGTGAAGCG; group 2-5 contain reference site GCCNNNNNGAAGCG and its single, double, triple variants respectively; group 6-9 contain reference site GGTGAAGCG and its single, double, triple variants respectively.