Published online 15 January 2015 *Nucleic Acids Research, 2015, Vol. 43, No. 3* ***1965–1984***
doi: 10.1093/nar/gku1395

NAR Breakthrough Article

A systematic survey of the Cys_2His_2 zinc finger DNA-binding landscape

Anton V. Persikov[1,†], Joshua L. Wetzel[1,2,†], Elizabeth F. Rowland[1], Benjamin L. Oakes[1], Denise J. Xu[1], Mona Singh[1,2,*] and Marcus B. Noyes[1,3,*]

[1]The Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA, [2]Department of Computer Science, Princeton University, Princeton, NJ 08544, USA and [3]Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA

Received November 06, 2014; Revised December 20, 2014; Accepted December 27, 2014

ABSTRACT

Cys_2His_2 zinc fingers (C2H2-ZFs) comprise the largest class of metazoan DNA-binding domains. Despite this domain's well-defined DNA-recognition interface, and its successful use in the design of chimeric proteins capable of targeting genomic regions of interest, much remains unknown about its DNA-binding landscape. To help bridge this gap in fundamental knowledge and to provide a resource for design-oriented applications, we screened large synthetic protein libraries to select binding C2H2-ZF domains for each possible three base pair target. The resulting data consist of >160 000 unique domain–DNA interactions and comprise the most comprehensive investigation of C2H2-ZF DNA-binding interactions to date. An integrated analysis of these independent screens yielded DNA-binding profiles for tens of thousands of domains and led to the successful design and prediction of C2H2-ZF DNA-binding specificities. Computational analyses uncovered important aspects of C2H2-ZF domain–DNA interactions, including the roles of within-finger context and domain position on base recognition. We observed the existence of numerous distinct binding strategies for each possible three base pair target and an apparent balance between affinity and specificity of binding. In sum, our comprehensive data help elucidate the complex binding landscape of C2H2-ZF domains and provide a foundation for efforts to determine, predict and engineer their DNA-binding specificities.

INTRODUCTION

The Cys_2His_2 zinc finger (C2H2-ZF) is the most frequently occurring DNA-binding domain in metazoan proteins, and is found in nearly half of human transcription factors (1,2). C2H2-ZF proteins have been implicated in a wide range of biological processes, including development (3), recombination (4) and chromatin regulation (5). Thus, a thorough understanding of how C2H2-ZF proteins specify their DNA-binding sites would be invaluable in mapping regulatory networks across a broad spectrum of eukaryotes.

An individual C2H2-ZF domain contains a well-conserved DNA-binding structural interface and specifically recognizes its DNA target via amino acids occupying four key 'canonical' positions of an alpha-helix (6–9). C2H2-ZF proteins that bind DNA typically do so via tandem arrays of multiple, closely linked C2H2-ZF domains. An individual finger binds a contiguous three-nucleotide subsequence, 3′ to 5′, along with a potential fourth, cross-strand contact that overlaps the target of the N-terminal adjacent finger (Figure 1A). Unlike other structural classes of DNA-binding domains that typically offer a limited range of specificities, C2H2-ZF domains can specify a wide range of three base pair (3bp) targets (10–15).

Due to the combination of largely modular binding and the wide range of DNA-binding specificities achievable via individual domains, C2H2-ZF arrays can, in theory, specify virtually any DNA site of interest. As such, C2H2-ZF domains serve as an attractive, general-purpose scaffold for engineering DNA-binding specificity. Indeed, efforts from the protein design community have resulted in chimeric proteins that use C2H2-ZF domains to target specific genomic locations at which to carry out particular functions. Such technology has enabled modification of transcrip-

[*]To whom correspondence should be addressed. Tel: +1 609 258 6385; Fax: +1 609 258 8020; Email: mnoyes@princeton.edu
Correspondence may also be addressed to Mona Singh. Tel: +1 609 258 7059; Fax: +1 609 258 1771; Email: mona@cs.princeton.edu
[†]These authors contributed equally to the paper as first authors.

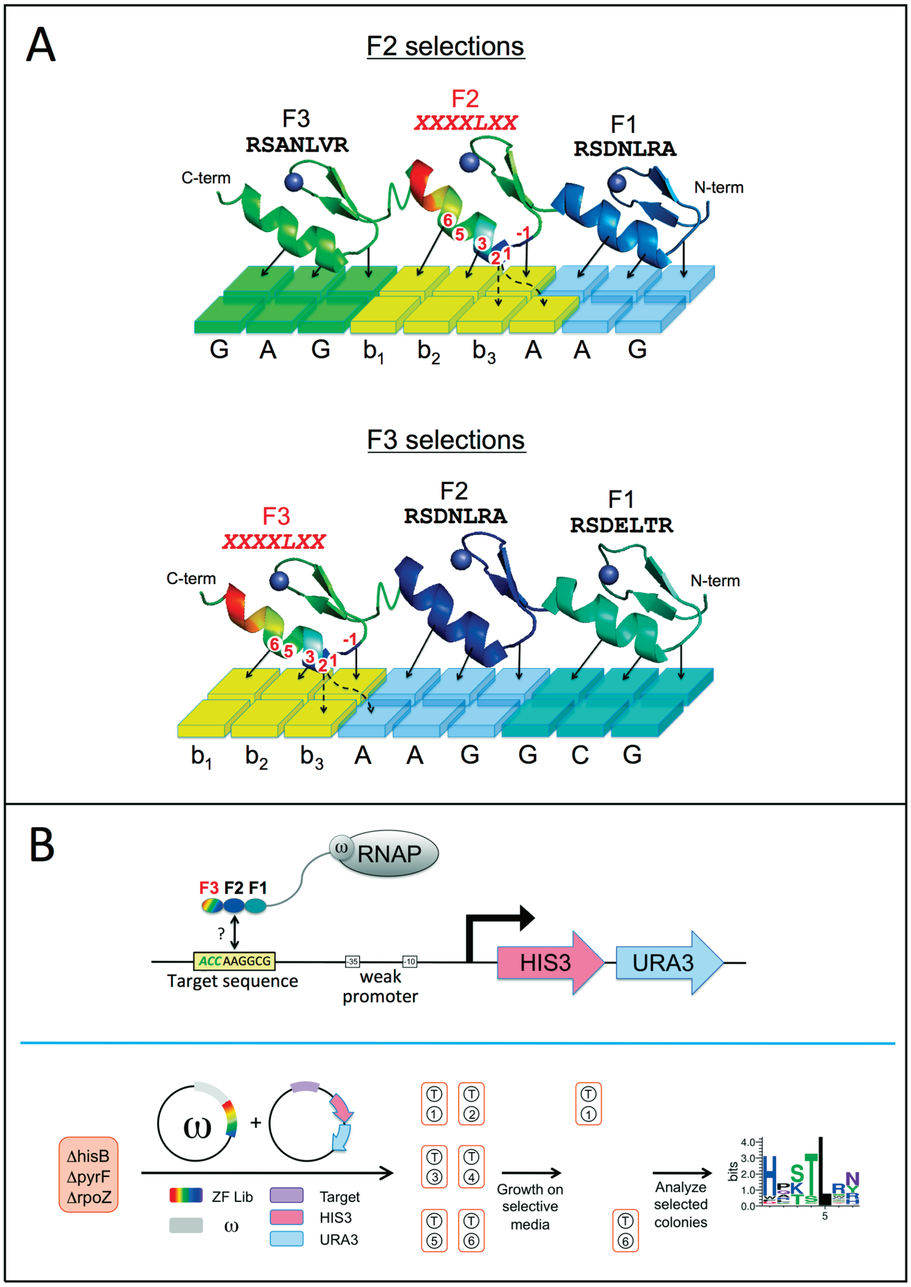


Figure 1. Schematic of bacterial one-hybrid protein selections. (**A**) Schematic of F2 (top) and F3 (bottom) protein selections. Individual C2H2-ZF domains are selected in the context of a protein containing an array of three domains. The fixed C2H2-ZF domains are shown as solid colors while the variable C2H2-ZF domain is shown as a rainbow. Primary contacts with the bases are shown with arrows. The individual selections place a unique 3bp target in the appropriate position, noted as yellow bases (b_1, b_2 and b_3), to assay the interaction of the variable C2H2-ZF domain. Underneath, the bases of the primary strand shown 5′ to 3′ are noted. Above each C2H2-ZF domain, the sequence of the recognition helix is shown N to C, with each variable position shown as a red 'X'. (**B**) Schematic of the C2H2-ZF selection. (Top) Proteins are expressed as a 3-fingered protein-direct fusion to the omega subunit of RNA polymerase. C2H2-ZF domains are selected to bind the target sequence placed 10bp upstream of the promoter that drives the reporter genes, HIS3 and URA3, as described in Supplemental Methods 1b. In the example shown, C2H2-ZF domains would be selected from the F3 library to bind the 5′-ACC-3′ (shown in green). (Bottom) Two plasmids, the protein expression vector (here shown from the F3 library) and the target reporter vector, are transformed into the bacterial strain. Double transformants are plated on selective media. DNA is recovered from the cells and the region of the library vector that codes for the variable region is sequenced. Enriched amino acid sequences are shown as a sequence logo.

tional outputs (16,17) and chromatin (18,19), as well as precise genome editing when fused with nuclease or recombinase domains (20–26).

Despite the importance of C2H2-ZF proteins for natural systems and their successful use in protein design, our knowledge about their DNA-binding landscapes remains surprisingly incomplete. Indeed, the binding specificities of most C2H2-ZFs within genomes are not known. For example, in human, specificities are known for less than a hundred of approximately 700 C2H2-ZF proteins (15). In fruit fly, specificities have been successfully determined for only ~20% of C2H2-ZF proteins, with 62% of the tested C2H2-ZFs failing characterization in a recent screen (10). Additionally, limited knowledge about context dependent effects—either between C2H2-ZF domains adjacent to one another within an array (27–29), among contacts within a single finger-DNA interface, or simply due to the position of a finger within an array (30)—has made the process of selecting, engineering and assembling synthetic C2H2-ZF proteins with desired DNA-binding specificities quite challenging (11,31). Further, while the well-defined interaction interface of C2H2-ZF domains has enabled the development of computational methods for predicting their DNA-binding specificities (32–40), the performances of these methods leave much room for improvement. Thus, a better understanding of the determinants of C2H2-ZF DNA-binding specificity would both enable highly reliable predictions of natural transcription factor binding sites and facilitate the design of engineered proteins with *de novo* binding specificities.

In order to further our understanding of C2H2-ZF DNA binding as well as to provide a resource for prediction, selection and/or design-oriented applications, here we report the results of screening all 64 possible 3bp targets for interactions with C2H2-ZF domains from multiple large protein libraries (41). This set of screens represents the most comprehensive and systematic survey of the C2H2-ZF DNA-binding landscape to date. We uncover pools of hundreds to thousands of C2H2-ZF domains capable of binding each 3bp target. Via cross-examination of these independent protein selections, we are able to simultaneously characterize binding profiles for thousands of C2H2-ZF domains and thus infer their quantitative DNA-binding specificities. For a diverse subset of the selected fingers, we confirm that these predicted specificities are highly concordant with experimentally determined specificities. We also demonstrate that the binding behavior gleaned from our large synthetic pools generalizes well to natural systems by adapting a simple nearest neighbor approach to accurately predict DNA-binding specificities for naturally occurring C2H2-ZF proteins. Furthermore, the diversity of our vast pools enables us to choose test fingers highly specific for nearly every 3bp target as well as to select three-finger combinations able to specify several challenging 9bp DNA sequences that proteins constructed via modular assembly failed to bind in previous efforts (11,31).

Additional analyses presented here elucidate the complex nature of C2H2-ZF DNA-binding landscapes. For example, we observe finger-DNA interfaces that alternately either confirm or defy previously proposed position-specific amino acid-base recognition rules for C2H2-ZFs. We explore the important role of 'within-finger' context, demonstrating that the same amino acid in a given contacting position of the recognition helix may specify up to all four different bases depending upon the context provided by the amino acids occupying other key positions in that C2H2-ZF domain. We also find that within-array domain position plays an important role in influencing base recognition, even when the neighboring finger context is the same. Lastly, we observe an apparent balance between affinity and specificity in interactions between C2H2-ZFs and DNA. Altogether, by developing an approach that integrates data from independent protein selections across all possible targets, we provide a foundational blueprint for further large-scale investigations of the DNA-binding specificity of this important domain, as well as a valuable resource for predicting and designing DNA-binding specificity for C2H2-ZFs.

MATERIALS AND METHODS

Overview of experimental approach

To systematically survey the DNA-binding landscape of the C2H2-ZF domain, we used site-directed mutagenesis to assemble diverse C2H2-ZF protein libraries with six variable amino acid positions (41), as guided by prior engineering efforts (13,42–45) and the Zif268 structure (8). These libraries allowed each of the 20 possible amino acids in the -1, 1, 2, 3, 5 and 6 positions in regard to the alpha-helix of either the middle (F2) or C-terminal (F3) positions of a model Zif268-based system (Figure 1A and Supplemental Methods 1a). The quality, diversity and uniformity of sampling within these libraries have been validated (41). This experimental system vastly expands the repertoire of C2H2-ZF domains available for selection, as most previously reported libraries either considered fewer randomized residue positions (27,28) or used a coding scheme that did not permit all amino acids (29–30,46).

A comprehensive set of protein selections was performed, wherein each of the 64 possible 3bp DNA targets was screened against our expansive C2H2-ZF libraries using an omega-based bacterial one-hybrid (B1H) system (12,21,47). Specifically, a variable finger was expressed in either the middle or C-terminal (F2 or F3) position of a three-fingered protein where the adjacent, non-varying fingers have known specificities (Supplemental Methods 1b). These three-fingered proteins were expressed as fusions to the omega subunit of RNA polymerase; omega acts as an activation domain in this hybrid assay. For each selection, the 3bp site of interest was placed in a position relative to the targets of these fixed 'anchor' fingers such that upon binding, the anchor fingers will situate the test finger in a position in proximity to the desired target (Figure 1A). Only a positive interaction between the test finger and the site of interest will lead to an omega-guided recruitment of RNA polymerase and activate the transcription of a necessary HIS3 reporter gene (Figure 1B). Therefore, when these cells are grown on minimal media that requires the activation of HIS3 transcription, only a functional protein–DNA interaction will lead to survival of the bacteria (Figure 1B). To recover these positive protein–DNA interactions, cells from each selection were pooled, DNA harvested and their C2H2-ZF constructs sequenced.

The affinity of a protein–DNA interaction has been demonstrated to relate to growth rate in the B1H system, and the level of affinity required to activate HIS3 can be modulated by changing the concentration of 3-amino triazole (3-AT, a competitive inhibitor of HIS3) in the selection media (12,48–49). All of our protein selections were performed at low (2-mM 3-AT) and high (10-mM 3-AT) levels of the inhibitor, representing low and high stringency selections, respectively. The number of sequences recovered from a given selection that correspond to a particular protein–DNA interaction (which is proportional to the size of a colony) and the recovery of that interaction at a given inhibitor concentration are both related to the affinity of the interaction. We note that the B1H system does not directly measure the affinity of particular protein–DNA interaction. Indeed, for any particular colony, other factors besides affinity and/or inhibitor concentration may influence that colony's growth rate. However, throughout this work we assume that, across a population of interactions recovered from a given selection, affinity of the protein–DNA interaction is the primary determinant of growth rate.

Details of the bacterial one-hybrid selection procedures have been described previously (12,47). Modification to these procedures and details of the libraries used in this manuscript can be found in Supplemental Methods 1a–f. To characterize the DNA-binding specificity of a particular (test) C2H2-ZF protein, the procedure is reversed, whereby binding of the test C2H2-ZF to various sequences in a randomized DNA library (Supplementary Figure S1A) leads to activation of the HIS3 reporter (Supplementary Figure S1B).

Building and selecting three-fingered C2H2-ZF libraries

The B1H system was also used to select three-fingered arrays of C2H2-ZFs that specify particular 9bp DNA targets. These selections were performed, in principle, as previously demonstrated (13,21,42,50). Briefly, we used the pools of fingers recovered from our individual zinc finger selections as polymerase chain reaction (PCR) templates to build three-fingered C2H2-ZF libraries directed at binding particular 9bp targets. These 9bp targets were chosen based on the observation that C2H2-ZF proteins built by modular assembly had failed to bind them in two separate publications (11,31). For each 9bp target, a three-fingered 'pool library' was assembled.

To create pool libraries, individual zinc finger pools corresponding to each 3bp subsite of the 9bp target were used as templates for PCR. For example, if targeting the sequence 5′-AAA-CCC-GGG-3′, the AAA, CCC and GGG pools would be used as the PCR template for each finger of the library. PCR primers were designed so that the resulting PCR pools could then be assembled by overlapping PCR into a three-fingered coding sequence of the order N-terminus-pool^{GGG}-pool^{CCC}-pool^{AAA}-C-terminus (zinc fingers bind DNA anti-parallel to the 5′-3′ sequence of DNA). This process ensures that each finger in the pools used as templates for each position of the array has already shown the ability to bind the desired 3bp subsite in the previously performed individual-finger selections. Therefore, if we estimate that each pool for a given 3bp subsite contains between 100 and 1000 C2H2-ZFs, each assembled library offers a theoretical complexity of 10^6 to10^9 three-fingered combinations from which to find compatible sets of zinc fingers that are uniquely suited for the context offered by the desired 9bp target.

The final three-fingered PCR products were digested and cloned into the B1H omega-based expression vector. The 9bp target of interest was placed 10bp upstream of the promoter that drives HIS3 expression in the B1H system and C2H2-ZF proteins were selected (as described above) from the new corresponding three-fingered library. Cells were harvested and their C2H2-ZF constructs were sequenced to find enriched protein sequences. For each 9bp target, sequenced candidates that closely resembled the enriched consensus of protein sequences were chosen and their specificities tested by B1H binding site selections as described above.

Affinity-related green fluorescent protein activation in yeast

Selected C2H2-ZFs of interest were screened for their ability to activate a green fluorescent protein (GFP) reporter in yeast as previously described (17,41). Each C2H2-ZF was cloned into the yeast genome to be expressed from an ACT1 promoter as a direct C2H2-ZF-estrogen receptor-VP16 fusion (ZEV). Binding sites to be tested were cloned into a minimal GAL1 promoter upstream of a GFP cassette on a yeast centromere (CEN) plasmid containing a URA3 cassette. These plasmids were then transformed into the appropriately constructed yeast strains in order to pair the desired ZEV-binding site combination to test. In each experiment, positive and negative controls (the original high affinity Z3EV system paired with either its optimal target or an empty vector, respectively) were also performed to control for experimental error. Next, for each sample tested, transport of the ZEV construct was induced with the addition of 100-nM β-estradiol and cultures were grown for 12 h.

The mean fluorescence of each sample was measured with a BD LSRII Multi-Laser Analyzer with High Throughput Sampler (BD Biosciences, Sparks, MD, USA). Mean fluorescence values were determined from at least 50 000 cells. Each C2H2-ZF-binding site pair was assayed in triplicate and means were normalized to the positive control. Previous work has shown that normalized GFP expression can be related to known levels of relative affinity (17,41). For this, a key is provided in our figure of normalized GFP expression when Z3EV is paired with a suite of binding sites that have known affinities relative to the Z3EV optimal target.

Processing, filtering and quality analysis of C2H2-ZF protein selection data

Following each protein selection, C2H2-ZF constructs harvested from cells were Illumina sequenced. The base-2 log of observed sequence counts were used to compute frequency distributions (per selection and considering only varied positions). Sequences with very low (<0.0001) frequencies were removed from each distribution and the resulting data were processed and filtered for quality according to an entropy-based procedure as described previously

(41). Additional details regarding the processing and filtering of protein selection data are provided in Supplemental Methods 2a. Various measures to ensure data quality and consistency were taken, as described in the Results section and further detailed in Supplemental Methods 2b.

Entropy and mutual information analysis of protein selections

For a given 3bp DNA target, we considered all protein sequences selected to bind it in the data set and computed, for each amino acid in each variable position in the alpha-helix (-1, 1, 2, 3, 5, 6), the fraction of sequences in which it was observed in that position. These were then used to derive the Shannon entropy per position as $-\Sigma_i\, p_i(\log p_i)$, where p_i is the fraction of distinct sequences with amino acid i in the position under consideration. For both the entropy and mutual information analyses, the frequency with which each protein sequence is observed within a 3bp target was ignored.

In order to examine the level of dependence between particular residue positions of the alpha-helix and particular base positions of the bound 3bp DNA regions, we performed a mutual information analysis. For each variable amino acid position *i,* we computed its distribution of amino acids A_i by calculating the fraction of times a specific amino acid was observed in this position across the data set. Similarly, for each base position j, we computed the distribution of bases B_j. We then computed the mutual information, $MI(A_i,B_j) = H(A_i) - H(A_i|B_j)$, where $H(X)$ is the Shannon entropy of the distribution of random variable X, as described above. The mutual information was then normalized to a value between 0 and 1 as $S = MI(A_i,\ B_j)/min(H(A_i),\ H(B_j))$. Using this normalization, if A_i and B_j are independent, S is zero, whereas if A_i is a deterministic function of B_j, S is 1.

In order to assess the significance of the level of normalized mutual information observed, we performed 1000 randomization experiments. Specifically, for the set of observed finger-DNA interfaces, we decoupled the interfaces by randomly permuting the DNA targets with respect to the helices that bound them. We then repeated our mutual information analysis with respect to this set of random interfaces and computed an empirical *P*-value based upon the fraction of times the normalized mutual information was higher for a pair in the randomized data than was observed in the actual data.

Core sequence representation of C2H2-ZF proteins

For each C2H2-ZF domain, we also consider its 'core sequence' representation, defined by the amino acids present in the four canonical positions of the recognition helix (i.e. -1, 2, 3 and 6). Because positions 1 and 5 can vary, each so-called 'core sequence' can correspond to multiple C2H2-ZF domains observed in our data set. Thus, when we refer to a core sequence, we are referring to all of the sequences with those amino acids occupying the -1, 2, 3 and 6 positions of the C2H2-ZF domain. The frequency of a core sequence within a specific data set (e.g., in a specific selection for uncovering domains binding a particular 3bp target or across all target selections in either F2 or F3) is defined to be the sum of the frequencies of all full-length protein sequences that share that core sequence representation.

Computing binding profiles for core sequences and computationally inferring DNA-binding specificities via lookup

For either the F2 or F3 protein selections, for each core sequence we considered the frequency with which it was found in each of the 64 possible 3bp targets. For each core sequence, we then normalized these frequencies so that they summed to 1 across the 64 3bp targets, and thereby obtained a binding profile that represents a probability distribution specifying the preference of a core sequence for each 3bp target. We denote this binding profile for a specific core sequence as $<bp_{AAA},\ bp_{AAC},\ \ldots,\ bp_{TTT}>$. To predict the DNA-binding specificity of a core sequence, we computed the probability of each nucleotide n in position b_1 as $p_{n,1} = \Sigma_{i,j}\, bp_{n,i,j}$. The predicted probabilities with which the nucleotides occur in positions b_2 and b_3 were computed analogously. We refer to this method for predicting the DNA-binding specificity for a core sequence as the 'lookup' procedure, as it is based upon finding the core sequence in question across all of the protein selections. Finally, for several analyses described below, each core sequence was assigned to its most preferred 3bp target by choosing the nucleotide in each position that has the highest inferred probability according to this lookup procedure.

Processing binding site selection data

Illumina sequencing and analysis were used to uncover binding site preferences selected by candidate C2H2-ZF proteins via B1H selection. The data were filtered for quality, searched for enriched motifs and visualized via sequence logos as described in our previous work (41). We provide further details regarding the processing and filtering of binding site selection data, motif finding, clustering and visualization in Supplemental Methods 2c.

Clustering C2H2-ZF domains within preferred targets

For the F2 and F3 protein selection data separately, we assigned each observed core sequence to the target for which it had the highest preference, as computed via the lookup procedure described above. For each 3bp target, we saw a diverse group of core sequences assigned to it in either the F2 or F3 selections. We obtained the full six amino acid sequences, including positions –1,1,2,3,5 and 6, of the corresponding zinc fingers and clustered them into 'specificity groups' of similar sequences that offer alternative binding strategies for that particular 3bp target. In particular, each 3bp target was described as a graph with all observed six amino acid sequences represented as nodes in the graph. The similarity between two sequences was computed using the BLOSUM62 matrix (51) and normalized to be between 0 and 1. Two nodes were connected with an edge if the similarity score between the two corresponding protein sequences exceeded 0.25. We used the network clustering program SPICi (52) with a minimum cluster size of six. Finally, for each cluster, we visualized the sequences within it via a sequence logo (53).

Nearest neighbor decomposition to predict C2H2-ZF binding specificity

In order to extend the predictive scope of our data to arbitrary C2H2-ZF domains that may not share a core sequence with any domain recovered in our screens, we adapted the classic nearest neighbors approach. In a typical implementation of nearest neighbors prediction, given a core sequence *C* which is not contained in our data set, we would look for all other core sequences in our data set that are most sequence-similar to *C* and predict a specificity by computing an average across the DNA-binding specificities of all such neighbors. Our approach improves on this classic paradigm by leveraging (i) prior structural knowledge about which residue positions of the core sequence are known to be most important for determining the base at a given position (54) and (ii) information about which amino acids frequently substitute for one another. Specifically, when predicting the base specificity at b_i (i.e. preferences at base position *i*) for a core sequence *C*, we first hierarchically ranked neighboring core sequences of *C* found in our data set with respect to b_i, and next took a weighted average across the specificities inferred via the lookup method over the top 25 such neighbors.

Neighbors considered for use in predicting the specificity at b_i include all core sequences in our data set that are exactly hamming distance 1 from *C* and share the same residue as *C* in the amino acid position that is known to interact with b_i in the canonical structural binding model. For example, when predicting the base at position b_1, we do not consider neighbors that vary from *C* in position a_6. This leaves us with (potentially) 57 hamming distance 1 neighbors to be hierarchically ranked. The first level of the ranking hierarchy corresponds to the order in which we allow positions of neighboring core sequences to vary with respect to *C*. This ordering is chosen based upon previous structural analysis (see Table 3 in (54)). That is, we always allow the core sequence position with the least amount of structural evidence for interacting with b_i to vary first, the second least second and so on. For example, when predicting b_1, we first look at core sequences that vary from *C* at position a_{-1}, followed by core sequences that vary at the position a_2, and finally core sequences that vary at position a_3. For b_2, the order of variation is a_2, a_{-1}, a_6, and for b_3 the order of variation is a_6, a_3, a_2. Multiple neighbors that vary in the same position with respect to *C* are sub-ranked via scores derived from a PAM30 matrix. Specifically, for a neighbor *N* of *C*, the substitution score $S = \text{PAM30}(N,C)/\text{PAM30}(C,C)$ is computed, where PAM30(*N*,*C*) is simply the sum of values from the PAM30 matrix for substituting sequence *N* for sequence *C*. Numerators of all such scores for a set of neighbors are shifted positively so that the worst possible substitution corresponds to a score of 0, and the exact match to C, if present, receives the highest possible ranking.

Once neighbors for b_i have been ranked according to the above algorithm, for each of the top 25 neighbors for b_i, the specificities inferred from the lookup procedure are computed and a weighted average is taken across the 25 neighbor's predicted specificities for position b_i, with weights corresponding to the aforementioned PAM30 substitution scores. These steps are repeated separately for base positions b_1, b_2 and b_3 to obtain the complete predicted 3bp DNA-binding specificity of core sequence *C*.

A web-form for predicting a C2H2-ZF domain's binding specificity using the nearest neighbor decomposition approach is available at http://zf.princeton.edu/b1h/.

A database of naturally occurring C2H2-ZF DNA-binding specificities

We previously gathered C2H2-ZF protein DNA-binding specificities obtained from four resources including the JASPAR database (55), the UniProbe database (56), a database of human transcription factors (57) and the FlyFactorSurvey database (58), as described in (40). This database of experimentally determined transcription factor specificities was updated with ChIP-seq data collected by the ENCODE project (59). After merging redundant protein sequences, the combined data set contains 158 proteins. We used this data set for comparing the performance of our nearest neighbor decomposition method (NN) to the performances of other state-of-the-art C2H2-ZF DNA-binding specificity prediction methods. However, substantial overlap was observed between proteins listed in this test data set and the data used to train previously published prediction methods (including those based on random forests (RFs) and support vector machines (SVMs) (39,40)). Overall, 104 out of the 158 proteins in our test data set contain at least one instance of a C2H2-ZF domain used in the training of at least one of the SVM, RF or NN methods. Thus, we compared performance of the three methods on the remaining 54 proteins of the test data set. Additional details of the processing of this test set, as well as prediction of DNA-binding specificities on the test set, are provided in Supplemental Methods 2d.

Evaluating the quality of C2H2-ZF DNA-binding specificity predictions

After predicting the DNA-binding specificity of a given C2H2-ZF (or an array of C2H2-ZFs), we evaluated agreement between the predicted and experimentally determined specificities. In the case where the correct alignment of the experimental and predicted PWMs was known *a priori* (i.e. when predicting specificity for a single domain and comparing it to a 3bp subsite selection), we compared pairs of columns (base positions) of the aligned position weight matrices (PWMs). We considered a base position to have been correctly predicted if the Pearson correlation coefficient (PCC) between the predicted and experimental columns of nucleotide frequencies for that base position is at least 0.5. When the correct alignment of the experimental and predicted PWMs was not known *a priori* (i.e. when predicting specificity for naturally occurring C2H2-ZFs), we used a previously published alignment technique, where alignment scores are based on an information content corrected version of the PCC (40). A more detailed description of our evaluation pipeline is provided in Supplemental Methods 2e.

RESULTS

Systematic screens uncover hundreds to thousands of C2H2-ZFs binding each 3bp target

Each of the 64 possible 3bp DNA targets was screened for interactions against an expansive set of C2H2-ZF libraries that varied amino acids in the recognition helix of either F2 or F3 of a three-fingered array. After subsequent sequencing and filtering of recovered C2H2-ZF domains, the resulting data consist of a vast collection of DNA-binding interfaces that arise from four primary data sets (i.e. corresponding to protein selections in either F2 or F3 and at either low or high stringency). Each filtered data set provides, for each 3bp target, a list of domains observed to interact with the corresponding target, along with the frequency with which each domain was observed in the sequencing data. We further consider three combinations of these data sets: F2 union and F3 union (each consisting of both the low and high stringency data for F2 or F3, respectively), and the F2+F3 data set (consisting of all data across the four sets of selections). Together these data consist of ∼85 000 unique F2 protein–DNA interfaces and ∼88 000 unique F3 protein–DNA interfaces (Figure 2A).

One reason for creating such large libraries and screening all possible targets was to systematically examine how amino acids occupying various positions of a C2H2-ZF domain, even those not included in the structure-based canonical binding model (6–8), might contribute to its DNA-binding specificity. Analysis of our combined F2+F3 data set shows that, within the set of fingers binding each target, the two variable amino acid positions not involved in canonical contacts (positions 1 and 5) show significantly greater variation than the other four positions (Figure 2B). This provides large-scale confirmation that the -1, 2, 3 and 6 positions of the alpha-helix, as suggested by the canonical model, are in fact the primary specificity-determining positions within C2H2-ZF domains. Therefore, in much of our ensuing analysis, for each observed protein, we also consider its core sequence representation, defined by only these four specificity-determining positions of the recognition helix.

We examined the diversity and reproducibility of the data produced by our screens. In our combined F2+F3 data set, each 3bp target was bound by hundreds to thousands of unique proteins that can be represented by hundreds of unique core sequences (Figure 2C). Within a single protein selection experiment for a DNA target, each core sequence represents on average ∼7 domains found in the filtered data set, supporting the consistency with which each core protein–DNA interface is observed. We further confirmed the reproducibility of our selections by calculating the weighted fraction of core sequences found at both high and low stringency for each 3bp target (Supplemental Methods 2b). We find excellent overlap for both the F2 and F3 selections (Supplementary Figure S2).

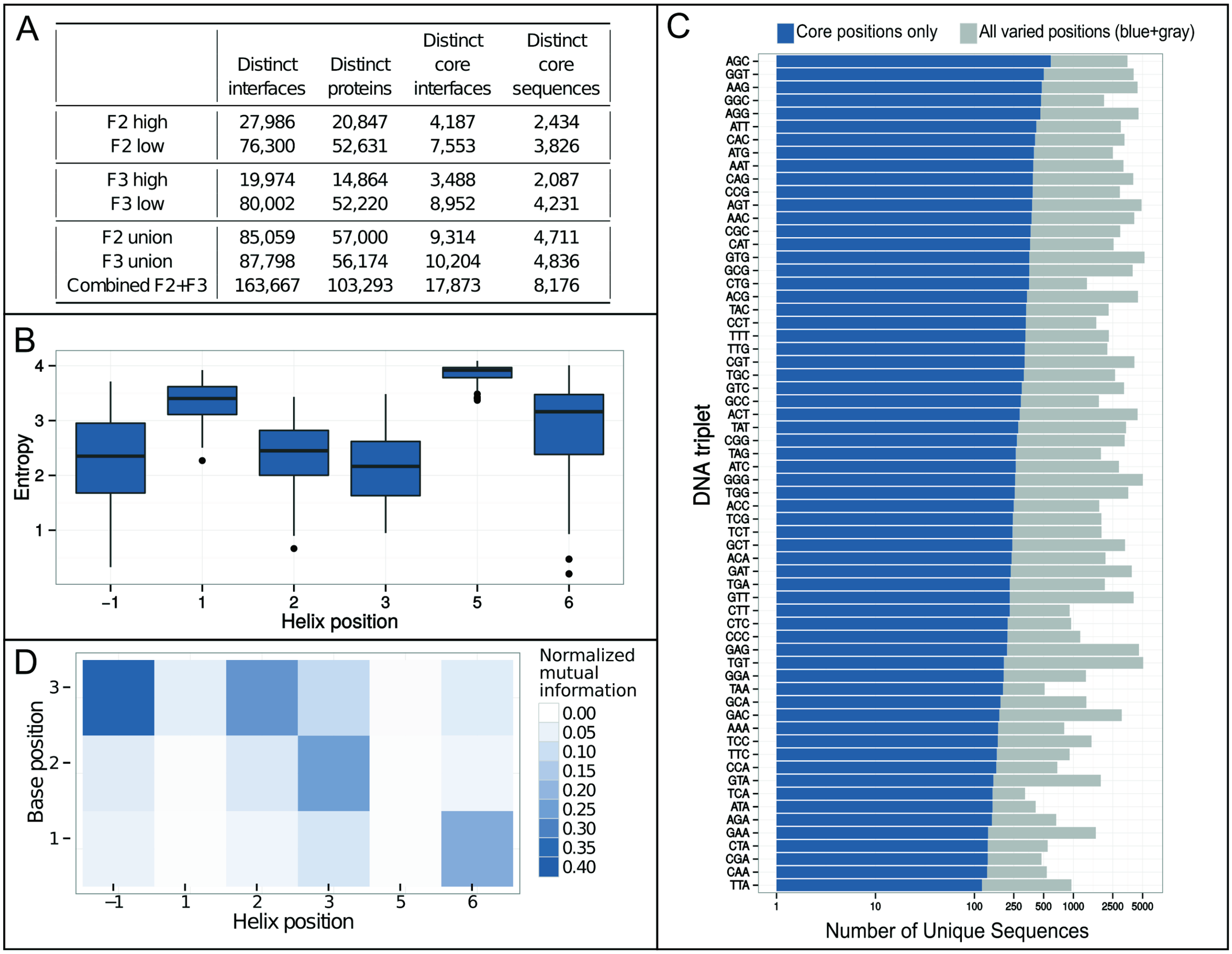
We performed a mutual information analysis on the combined F2+F3 data set to uncover the dependence between amino acid positions of the C2H2-ZF domain and base positions of the 3bp targets. The results of this analysis highlight the importance of the three within-finger canonical contacts, along with an additional fourth contact (between the amino acid at position 2 and the base at position 3 of the 3bp target) that was recently proposed to be important for intra-finger DNA-binding specificity (54) (Figure 2D). Permutation testing reveals that all amino acid-nucleotide position pairings have statistically significant normalized mutual information (empirical P-values <0.001). Although transitive correlations can confound such pairwise analyses, this result suggests that all positions within the recognition helix may, at least subtly, influence DNA-binding specificity and that multiple amino acids likely contribute to the specificity at a single base in a within-finger context-dependent manner.

C2H2-ZF domains exhibit a complex network of interactions with 3bp targets

We performed a series of quantitative analyses to gain insight into the overall trends present in C2H2-ZF DNA-binding landscapes. First, to gain a global view of the set of interactions present, we visualized our F2 data using a network representation via Cytoscape (60) (Figure 3A). Inspection of sub-networks (see, e.g. Figure 3B) revealed that the more frequently observed core sequences tend to be capable of binding multiple (usually sequence-related) targets, while the less enriched core sequences tend to bind only one unique 3bp target. In agreement with this observation, we found that in our combined F2+F3 data set, the frequency with which a core sequence bound a particular 3bp target is positively correlated with the total number of 3bp targets bound by that core sequence (Supplementary Figure S3). Since higher affinity sites yield more rapid colony growth in the B1H system (48) and more recovered sequences, this correlation suggests that there is often a trade-off between affinity and specificity for C2H2-ZF domains. We also found that targets that differ in only one nucleotide exhibit extensive overlap in the core sequences that bind them (Figure 3C, Supplementary Figure S4 and Supplemental Methods 2b); this finding is consistent with previous reports of off-target binding of engineered C2H2-ZF proteins (12,21,61–62), and highlights the challenges in engineering domains that do not bind targets similar to the intended one.

Integration of comprehensive selections allows accurate inference and design of C2H2-ZF DNA-binding specificities

Our set of systematic screens allowed us to uncover the binding preference for each selected finger across all 3bp targets. In particular, for each core sequence, we aggregated the frequencies with which it was found in each target selection and then used these frequencies to construct a binding profile. This binding profile corresponds to a probability distribution specifying the preference of a core sequence for each 3bp target. We observed that core sequences found in both high and low stringencies for either the F2 or F3 data tend to have similar binding profiles (Supplementary Figure S5 and Supplemental Methods 2b); this supports the consistency of our selections and suggests that, in aggregate, these selections can be used to infer binding specificities. Thus, for each core sequence observed in F2 or F3, we predicted its



	Distinct interfaces	Distinct proteins	Distinct core interfaces	Distinct core sequences
F2 high	27,986	20,847	4,187	2,434
F2 low	76,300	52,631	7,553	3,826
F3 high	19,974	14,864	3,488	2,087
F3 low	80,002	52,220	8,952	4,231
F2 union	85,059	57,000	9,314	4,711
F3 union	87,798	56,174	10,204	4,836
Combined F2+F3	163,667	103,293	17,873	8,176

Figure 2. Comprehensive protein selections across all 3bp targets. (**A**) The total numbers of distinct interfaces (i.e. protein-target pairs), proteins, core interfaces (i.e. core sequence-target pairs) and core sequences are listed. These are further separated into sequences recovered per finger position (F2 or F3) and stringency of selection (low or high). Three combinations of these primary data sets are also considered: F2 union, F3 union and the combined F2+F3 data sets. (**B**) For each variable amino acid position of the selected proteins, we display a boxplot of the Shannon entropy of the distribution of amino acids selected, computed individually across each of the 64 possible DNA targets in the F2+F3 data set. Shown are the median and the interquartile range, with whiskers on the top and bottom representing the maximum and minimum data points within 1.5 times the interquartile range. The higher entropy of the two amino acid positions not included in the canonical binding model (1 and 5) indicates that these positions are more variable than the core positions of the recognition helix (-1,2,3,6) with respect to a particular 3bp DNA target and suggests that amino acids in these non-canonical positions are less important for DNA-binding specificity. Entropy scores are significantly higher for positions 1 and 5 than for position 6, as judged by one-tailed Mann–Whitney U-tests ($P < 0.003$, $P < 2.2\text{e-}16$, respectively). (**C**) The total number of distinct sequences recovered from protein selections for each 3bp target, shown 5′ to 3′, for the combined F2+F3 data set. Blue is with respect to only the core positions of the C2H2-ZF domain, while blue plus gray is with respect to all six varied amino acid positions. (**D**) The normalized mutual information between base and amino acid positions computed on the set of distinct domain-DNA interfaces in the F2+F3 data set.

putative binding specificity by ‘looking up’ its constructed binding profile and using this profile to weight each 3bp target; that is, a PWM was computed by taking a weighted average over the 3bp targets.

We tested the performance of this ‘lookup’ procedure for inferring binding specificities for C2H2-ZF domains observed in the selections. In particular, from among the set of C2H2-ZFs observed in the F2 protein selections, we chose 166 with a range of predicted binding specificities and characterized their specificities experimentally. Each of these C2H2-ZF domains was tested as F2 in our three-fingered, Zif268-based system (Supplemental Methods 1c and 1d). Briefly, for each test finger, binding sites were selected from a randomized 28bp library via the B1H assay ((12); Supplementary Figure S1) and followed by deep sequencing and motif finding (Supplemental Methods 2c).

Experimentally determined binding specificities exhibited exceptional concordance with those computationally inferred from the F2 selections: across the entire set of 166 C2H2-ZF domains, 83% of the computationally inferred per-position base specificities match their experimentally determined counterparts, as judged by a PCC between the actual and predicted nucleotide distributions $> = 0.5$ (Supplementary Figure S6). A wide spectrum of experimental

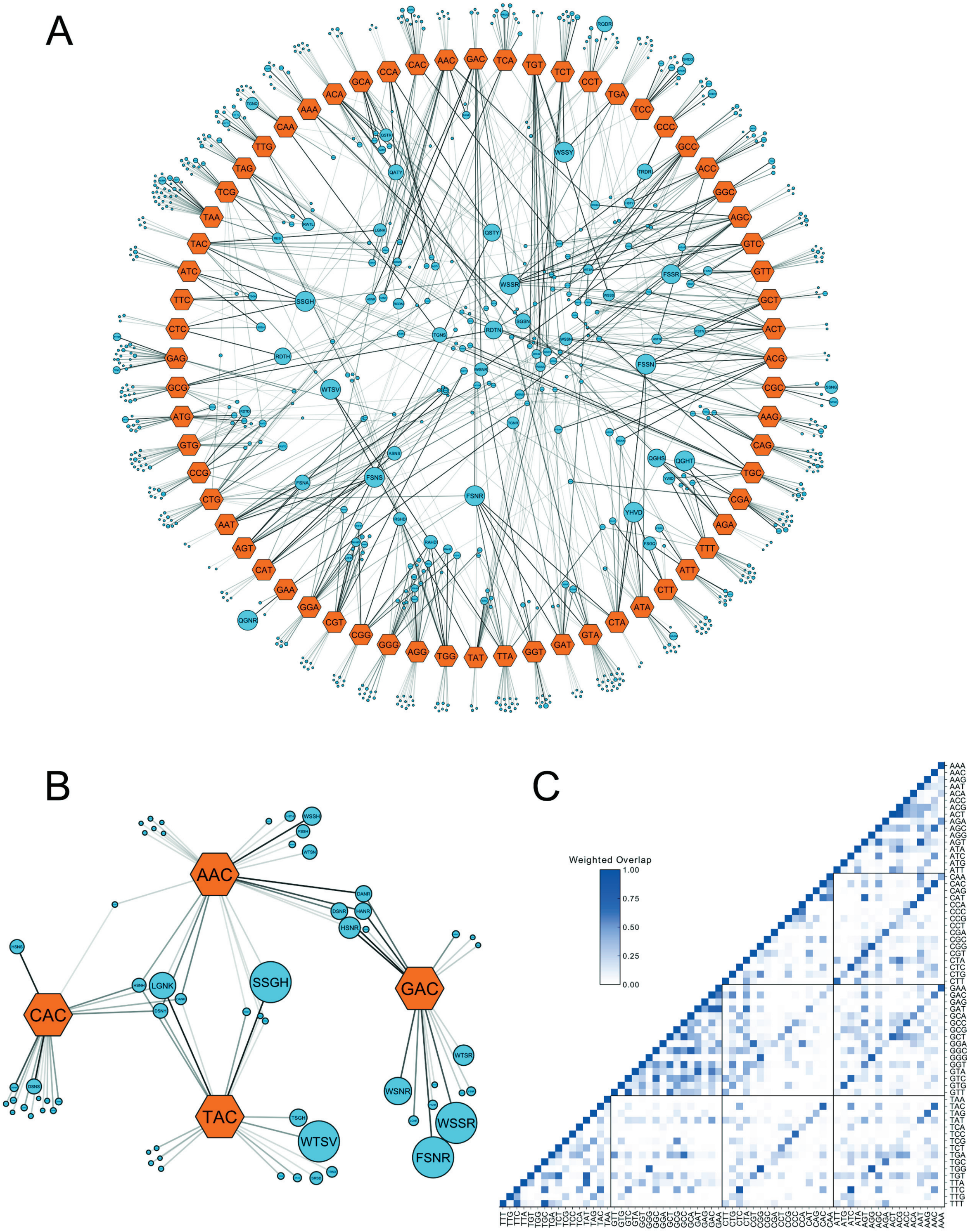


Figure 3. Visualization of C2H2-ZF domain–DNA interactions. (**A**) The F2 union data set in a network representation. Each core sequence is shown as a blue circle, with the size of the circle proportional to the frequency with which this sequence appeared across the data set. Each 3bp target is shown as an orange hexagon, with a connection between a target and each core sequence that is bound to it. Core sequences that bound multiple targets are shown on the inside of the circle. For visualization purposes, only core sequences that occur with frequencies greater than 1% are shown. The transparency level of an edge between a core sequence and a 3bp target corresponds to the frequency with which that core sequence was observed in the selection for that 3bp target. (**B**) The induced sub-network consisting of the targets AAC, CAC, GAC and TAC, and the core sequences bound to them. (**C**) The frequency weighted overlap (Supplemental Methods 2b) of the core sequences binding each pair of targets for the F2 union data set shown as a heat map, with evident patterns illustrating higher levels of overlap between targets that differ by one nucleotide. A high-resolution version of this figure can be viewed in the Supplementary Material online.

binding specificities was obtained, including, for most 3bp targets, at least one C2H2-ZF domain capable of binding that target. Indeed, for ~73% of the 64 targets, we observed at least one tested domain whose most frequently observed nucleotides match that target in all three base positions (Figure 4, left logo of each table entry). Experimentally determined specificities for this subset of domains are in excellent agreement with their computationally determined counterparts (Figure 4, right logo of each table entry). Thus, an integrative analysis of our separately performed large-scale protein selections enabled the successful design and prediction of individual C2H2-ZFs specific for most 3bp targets.

To confirm C2H2-ZF specificities and determine affinities outside of the B1H system, we screened a set of eight C2H2-ZFs, representative of core sequences recovered in the CAA and ATG selections, for their abilities to bind each of a suite of targets (Figure 5) in an affinity-based reporter system in yeast (17,41). We find that each C2H2-ZF tested activates the suite of GFP reporters in a pattern strikingly similar to the specificities uncovered via the B1H system. In particular, by comparing GFP intensities for a particular C2H2-ZF paired with either its optimal target or a target with a single-base substitution from the optimal target (i.e. row of Figure 5), we observe that base-substitutions in positions highly specific for the optimal base generally result in markedly decreased GFP output. Meanwhile, base-substitutions suggested as tolerable according to the experimentally obtained specificities result in relatively less impact on GFP output. Further, comparison of the on-target GFP activity across the various C2H2-ZFs when paired with the same target (i.e. a column of Figure 5) demonstrates that we have recovered distinct proteins that each offer a preference for their selected target, but with a wide range of affinities. These results underscore the dynamic affinity range of the B1H selection system, as while activation of the bacterial reporter is based on affinity, a wide range of affinities is functional. Therefore, we are likely to have recovered protein sequences in each of our B1H selections that may be missed by other methods requiring higher levels of affinity.

Distinct groups of C2H2-ZF core sequences can specify each 3bp target

We performed multiple computational analyses to elucidate patterns in the sets of C2H2-ZFs that prefer each 3bp target. To begin with, for both the F2 and F3 protein selection data, we assigned each observed protein sequence to its most preferred target based on the binding profile of its core sequence (Supplementary Figure S7). For each 3bp target, a diverse group of sequences was assigned to it, and subsequent sequence clustering analysis revealed that for each 3bp target, multiple distinct strategies are possible for binding that target. In particular, despite the recognition sequences having been assigned to targets individually, distinct groups of similar sequences were observed within each target and between similar targets (Figure 6). For each core sequence, on average ~5% of the other core sequences assigned to the same target were identical in three of the four core sequence positions. A significantly lower percent was found when looking across unrelated targets (Supplementary Figure S8). Conversely, highly dissimilar protein sequences were also assigned to each of the 3bp targets (Supplementary Figure S9), highlighting that there are diverse ways with which each DNA target may be specified.

Our set of 166 domains for which experimentally determined DNA-binding specificities were obtained confirmed the above-mentioned trends. We observed striking cases where pairs of core sequences having no amino acids in common specify the same 3bp target (e.g. Figure 7A). Conversely, we observed similar but distinct protein sequences with binding specificities that differ, with those differences corresponding well to the divergent core residues (Supplementary Figure S10, box). Thus, we observe notable complexity in the binding landscape of C2H2-ZFs: C2H2-ZF domains with completely different sequences can have highly similar DNA-binding specificities, while fingers with very similar sequences can diverge in their base preferences.

Within-finger context affects nucleotide-binding preferences of core amino acid residues

Multiple, simple one-to-one 'codes' of C2H2-ZF DNA-binding specificity have been proposed (6,10,63–64). These have been based upon contacts found in solved crystal structures and amino acid-nucleotide pairings inferred from tested C2H2-ZF–DNA interactions (e.g. Supplementary Figure S11). While these codes agree in the most important one-to-one preferences, such as arginine selecting for guanine, it was previously noted that they also disagree on several amino acid-nucleotide pairings (33).

Our set of 166 experimentally determined binding specificities has also yielded examples of C2H2-ZF domains that alternately either support or contradict previous codes of specificity (Figure 7B). We observed that the same amino acid in a given core position may specify different bases in the corresponding contacting nucleotide position, depending on the entire within-finger amino acid context. In some extreme examples, we observed identical amino acids in a given core position specifying three or four different bases at the corresponding nucleotide position (Figure 7C). Additionally, consistent with the mutual information analysis (Figure 2D), a within-finger cross-strand contact that is not part of the original canonical model is a strong determinant of DNA-binding specificity (Supplementary Figure S12). Together our data give compelling evidence that amino acids at all core positions have the flexibility to offer different base preferences depending upon the internal context of the C2H2-ZF domain itself, with all of the residues within the core positions of the C2H2-ZF domain presenting a unique environment that contributes to the full 3bp preference.

Within-array position can affect a C2H2-ZF's DNA-binding specificity, even with a fixed neighboring finger

It has been previously observed that adjacent C2H2-ZFs within an array may affect one another's binding specificities (13,21,27–28,44,65). Indeed, sets of two-finger modules that take this effect into account have been recently selected for and designed (28–29,66). However, all current approaches to modular assembly operate on the principle

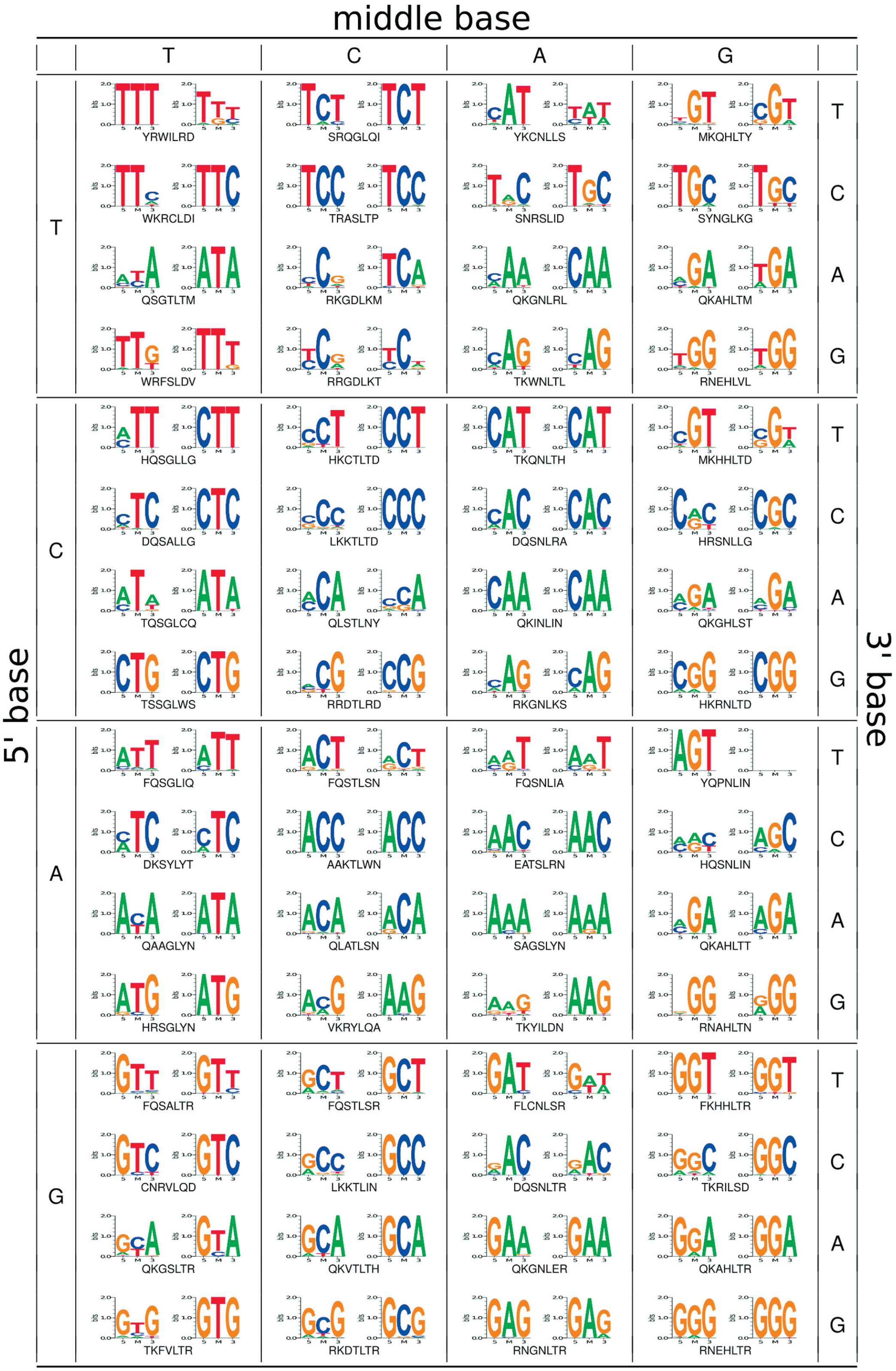


Figure 4. C2H2-ZF domains designed to bind nearly every 3bp target. Experimentally determined (left) and computationally inferred (right) DNA-binding specificities for 64 designed C2H2-ZF domains, visualized as sequence logos (Supplemental Methods 2c). The amino acid sequence of each tested finger is given below each pair of logos. Experimental specificities were determined for C2H2-ZFs as F2 of a *Zif268*-based construct (Supplementary Figure S1). Specificities were computationally inferred from binding profiles using our lookup procedure.

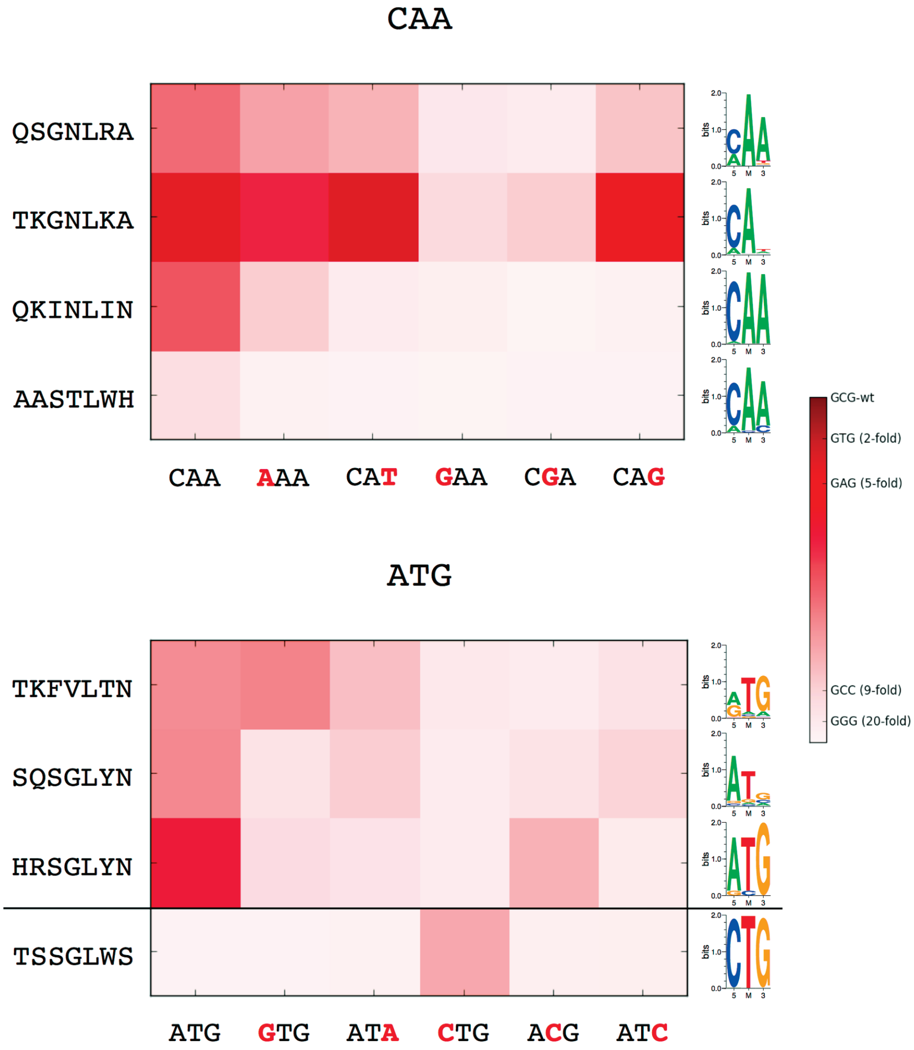


Figure 5. C2H2-ZF activity in yeast. C2H2-ZFs chosen from the CAA and ATG selections were expressed as F2 in yeast and challenged to activate an affinity-related GFP reporter using test binding sites placed at a critical position within the promoter (as described in (17)). (Top) Test C2H2-ZF sequences, listed to the left of each row, were chosen from the CAA selections and challenged to bind the six 3bp targets noted below each column of the chart. Alterations to the preferred target (CAA) are noted by bold, red letters. GFP expression for each protein-3bp target combination is normalized to the expression of the positive control and the data are shown as a heat map. The key denotes normalized GFP expression for protein–DNA interaction as compared to known affinity measures relative to the positive control Z3EV (*Zif268)* paired with its optimal target. A comparison of each protein–DNA interaction to the key provides an approximation of relative affinity. The B1H produced specificity of each zinc finger domain tested as F2 is displayed (as a sequence logo) to the right of each row. (Bottom) Test C2H2-ZF sequences, listed to the left of each row, were chosen from the ATG (top three rows) and CTG (bottom row) selections and challenged to bind the six 3bp targets noted below each column of the chart. A heat map of normalized GFP expressions for each protein-3bp target combination is shown as in the CAA chart above and B1H produced sequence logos are listed to the right of each row.

that array position alone (i.e. within the same inter-finger context) should not change a zinc finger's binding potential. Although this seems like a reasonable assumption, to our knowledge it has never been interrogated in a systematic fashion. Thus our F2 and F3 selections were designed to uncover how a domain's position within an array may affect its binding specificity, even when maintaining an identical neighboring finger (Figure 1A).

For each 3bp target, we compared the C2H2-ZFs recovered in either the F2 or F3 selections and found a significantly lower degree of overlap than when comparing domains recovered at high or low stringency selections within the same finger position (Figure 8A and Supplemental Methods 2b). We further tested this by comparing the binding profiles of a subset of C2H2-ZFs for which we were most confident with respect to the results of our selections. This subset consists of those core sequences recovered in both finger positions that had similar low and high stringency binding profiles in each individual positional context. A surprisingly large fraction of this high confidence subset of core sequences displays starkly differing binding profiles for the F2 versus F3 finger positions (Supplementary Figure S13); this is in contrast to the similarity observed between binding profiles derived from low and high stringencies at the same finger position (Supplementary Figure S5). This suggests that a C2H2-ZF's position within a DNA-binding array may influence its binding geometry and thus its target preferences. Based on this subset of F2 and F3 binding pro-

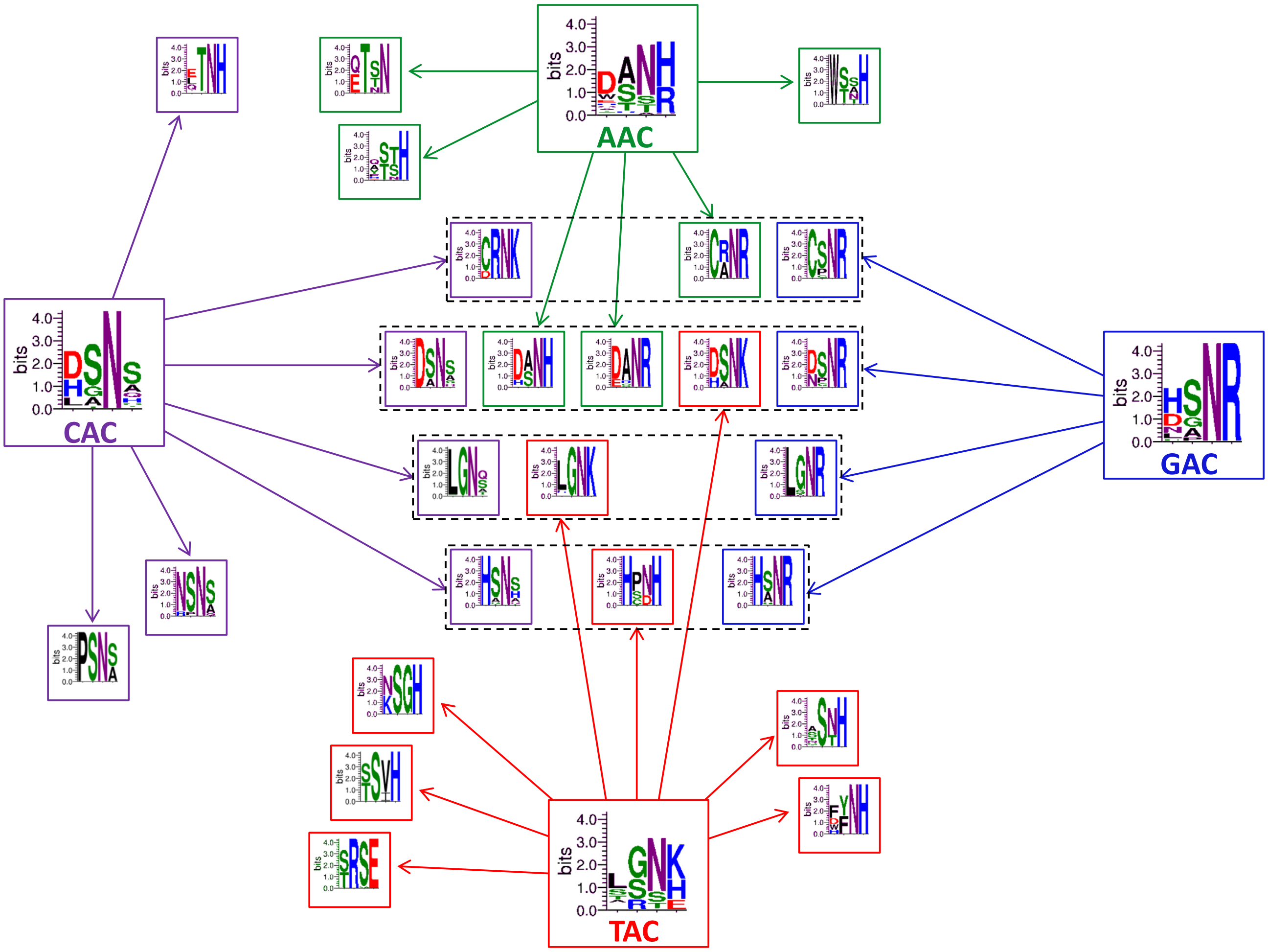


Figure 6. Common versus specialized solutions for binding similar 3bp targets. We assigned each C2H2-ZF sequence recovered from a set of protein selections to its most preferred 3bp target (according to the lookup procedure described in the Materials and Methods section) and subsequently clustered the set of sequences assigned to each 3bp target. Sequence logo representations of the full set of sequences assigned to each of the four similar 3bp targets, CAC, AAC, GAC and TAC (based on data obtained from the F2 union protein selections) are displayed in large boxes and labeled by the target. Logos for clusters of similar sequences assigned to a given 3bp target are pointed to by arrows originating from the corresponding boxed logo. Similar clusters derived from sequences assigned to different targets (shown in the center) suggest 'common solutions' for specifying nAC targets. However, 'specialized solutions' to each of the individual targets (grouped near the corresponding boxed logo) are also apparent.

files, we predicted putative binding specificities for each core sequence in both finger positions via the lookup procedure. The lowest similarity between F2 and F3 putative binding specificities is observed at the 5′ base (Supplementary Figure S14). It is worth noting that, in the case of F3, the 5′ base is at the 5′ edge of the entire 3-fingered binding site and, in the case of F2, the 5′ base can additionally contact the adjacent F3 finger according to the canonical binding model.

Based on the F3 protein selections, we chose 69 C2H2-ZFs to characterize further by experimentally determining their DNA-binding specificities when tested as F3 in a Zif268-based construct with fixed N-terminal fingers (Supplemental Methods 1c and 1d). To compare positional influence, 26 of these fingers tested in the F3 position had identical residues in the core positions of their recognition helices as fingers tested in the F2 position. Predicted specificities computed using either the F2 or F3 protein selection data were similar for approximately half of these (Supplementary Figure S15). Experimentally determined DNA-binding specificities confirmed that 16 core sequences share similar specificities in both finger positions (Figure 8B, left). Further, fingers predicted to behave differently in the F2 and F3 positions (Supplementary Figure S15) show distinct experimental binding specificities (Figure 8B, right, and Figure 8C), with either no detectable DNA binding or very weak nucleotide preferences in one of the two positions. Overall, our experimental binding specificities for fingers placed in the F3 position display higher variability per base-position than experimental specificities for fingers placed in F2 (Figure 8D), suggesting that the C2H2-ZFs placed in the F3 position may not be interacting with DNA as tightly as those placed in the F2 position.

Thus, while many C2H2-ZF domains maintain their binding specificities across different positional contexts, some conversely display drastically different binding poten-

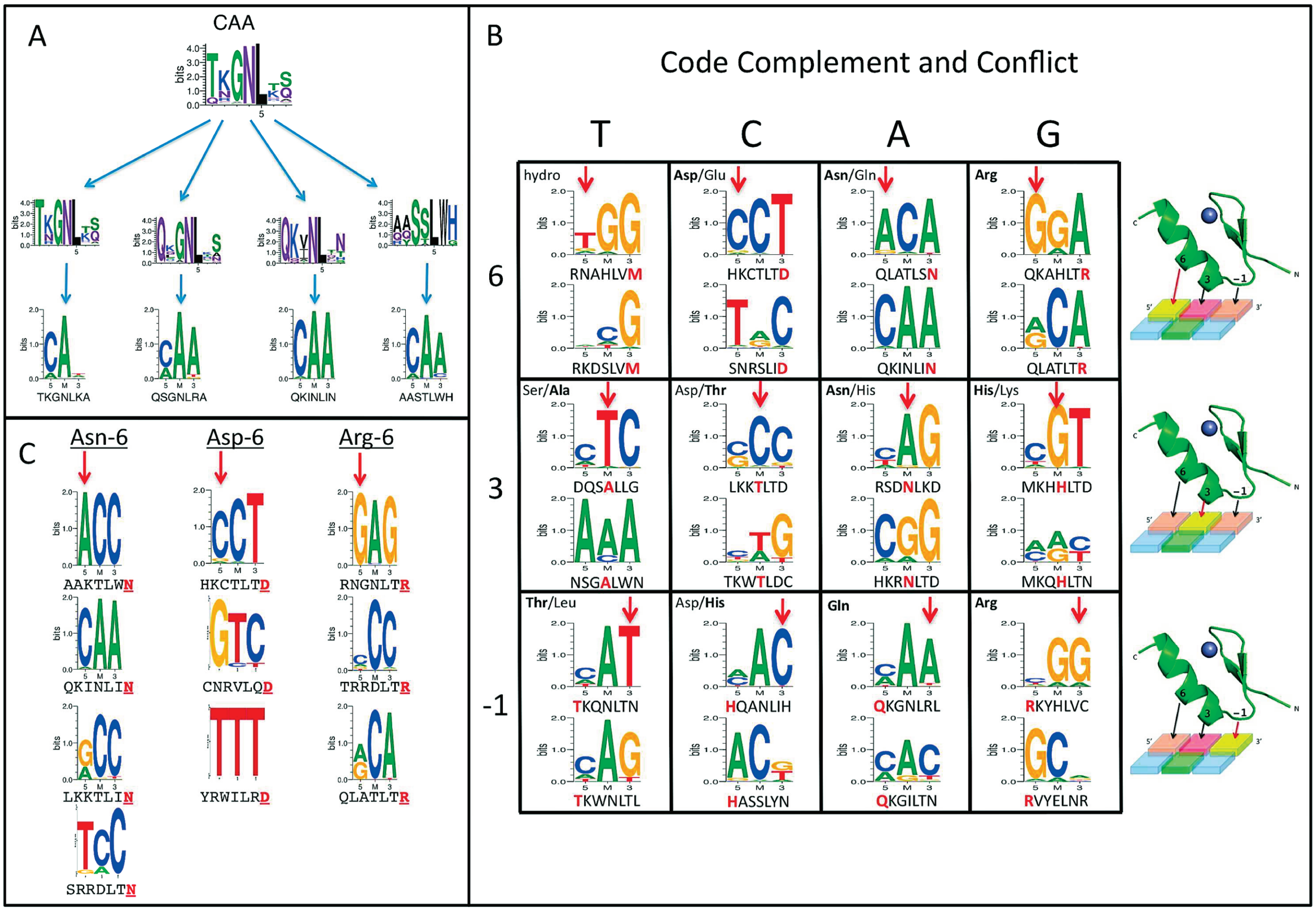


Figure 7. Variation in amino acid-base pairings for C2H2-ZF domain–DNA interactions. (**A**) All C2H2-ZF domains inferred to prefer CAA (based on F2 union protein selections) represented in sequence logo format (top). Protein sequences were clustered into distinct groups of similar proteins (middle). DNA-binding specificities were experimentally determined for a representative protein from each shown cluster, protein sequence noted below (bottom). (**B**) A simple code of specificity has been described based on C2H2-ZF selection and structural data (Supplementary Figure S11; (6)). Selected C2H2-ZFs complement and contradict this code. Each row represents an alpha-helix sequence position and each column a predicted base preference. In each box, the residue(s) thought to give the desired base preference is noted in the upper left. The sequence of the finger tested as F2 is noted below each logo with the critical amino acid shown in red. The top sequence of each box is consistent with the simple code. The bottom sequence contradicts the code. The cartoon to the right highlights the row-specific contact with a red arrow and yellow base. (**C**) DNA-binding specificities were determined for fingers that offer a shared amino acid at position 6 of the alpha-helix. Despite the conserved residue, the complementary base (noted with a red arrow) differs depending upon other residues of the test finger. Asn6, Asp6 and Arg6 examples are shown.

tials, even when a neighboring finger is fixed. This could complicate attempts to engineer or predict specificities under the widely implemented 'modular' paradigm. However, this is not to say that designing for specificity with C2H2-ZF proteins is not possible. To the contrary, it is possible (indeed likely) that future systematic screens of large libraries in different positional and neighboring finger contexts, in combination with good computational methods to integrate the data obtained, can lead to zinc fingers with stronger sequence specificity.

Diverse pools allow selection of C2H2-ZF arrays that specify challenging targets

We have described C2H2-ZFs able to bind every 3bp target, many with exceptional specificity. However, as noted above, position and interaction with neighboring fingers can influence the 3bp specificity of an individual finger within a multi-finger array. Therefore, assembly of multiple zinc fingers into a functional array is a more complicated problem. The common failure of zinc fingers assembled as modules to activate a reporter drives home this point; that is, a C2H2-ZF domain with known specificity may exhibit a range of functionality when placed in new contexts (11,31). As a result, zinc finger arrays have been more successfully produced either when neighboring fingers are selected from pools to find the individual monomers that are most compatible with one another (13,21,27,29–30,42,50), or through selection and design that is based on pairs of zinc fingers that have been previously selected from pools (28–29,66). However, any such approach is fundamentally limited by the size and diversity of the available pools.

To demonstrate the advantage of having a complete set of pools (one for each 3bp target), most of which offer hundreds to thousands of unique C2H2-ZF domains, we created three-fingered array libraries using the diverse pools as templates. We then selected zinc finger arrays from these

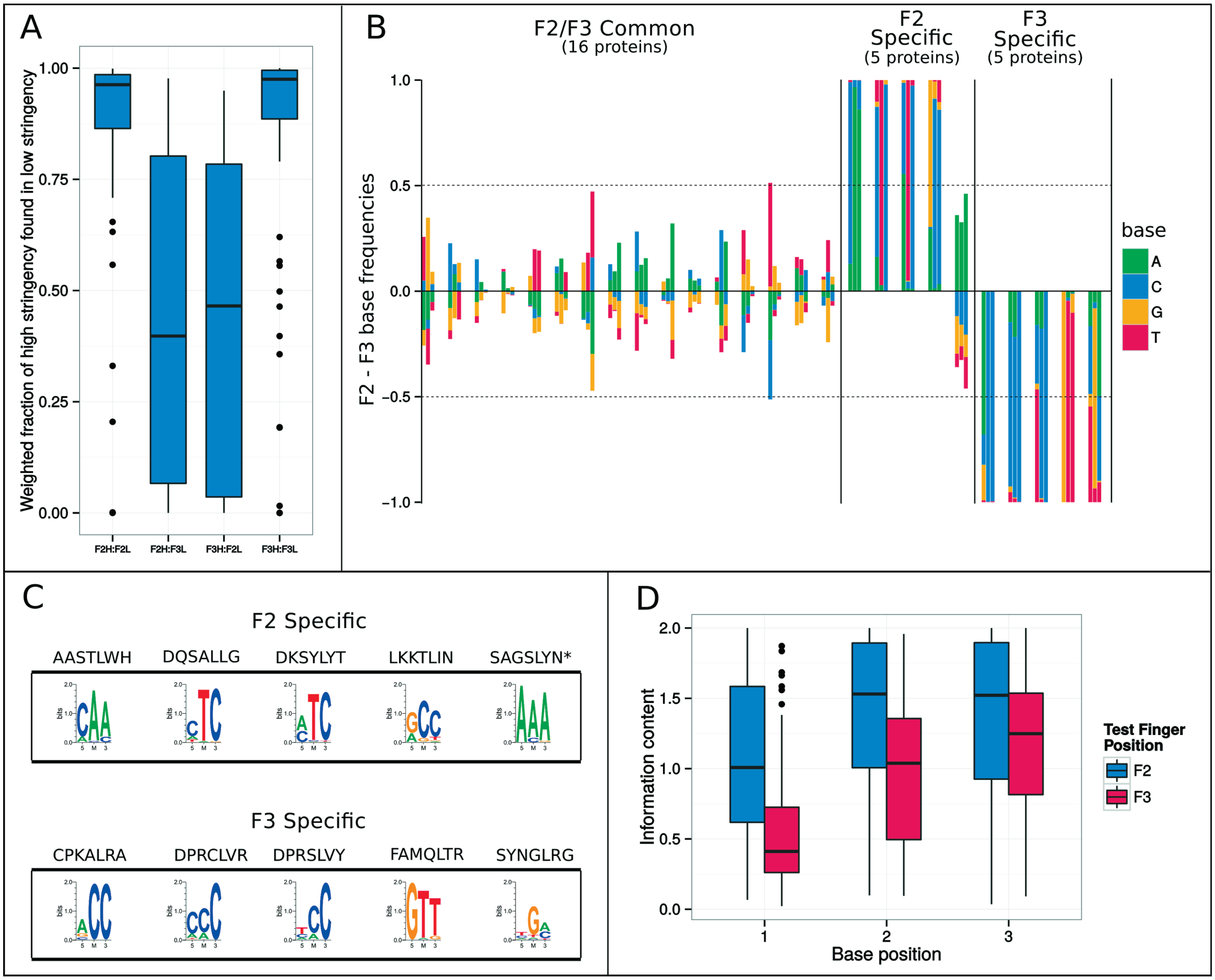


Figure 8. Positional context in domain–DNA interactions. (**A**) Weighted fractions of core sequences found in F2 or F3 high stringency selections that are also found in F2 or F3 low stringency selections (Supplemental Methods 2b). Shown left to right are: F2 high in F2 low, F2 high in F3 low, F3 high in F2 low and F3 high in F3 low. For each of these, weighted fractions are computed for each 3bp target and are depicted as boxplots. (**B**) The DNA-binding specificities of 26 core sequences were tested in both the F2 and F3 positions. Each set of three bars along the x-axis represents the 3bp specificity (5′ to 3′) in both positions for one core sequence. The y-axis represents the difference in the frequency with which a base is observed when comparing the F2 and F3 specificities of that same core sequence. If the base is more commonly observed in the F2 position, the bar is above the x-axis (base indicated by the color key, right). If the base is more commonly observed in the F3 position, the bar is below the axis. The closer this difference is to zero, the more similar the specificities are of the given core sequence in F2 and F3. The first group of core sequences exhibited similar DNA-binding specificities when tested in F2 and F3. The second group exhibited DNA binding when tested in F2 but either no detectable binding or extremely weak binding when tested in F3. The third group exhibited DNA binding when tested in F3 but not when tested in F2. (**C**) Sequence logos of fingers that function in either F2 (top) or F3 (bottom), with no colony growth or weak (as depicted by a star) DNA-binding specificity observed in the other position. (**D**) For each of the 166 binding site selections performed in the F2 context and 69 binding site selections performed in the F3 context, we computed the information content (IC) of each experimentally determined base position. Specifically, for each base position, we compute the Shannon entropy of the distribution of bases to uncover its variability and subtract this value from the maximum possible value (2 bits) to obtain its IC. For each of these base positions, we depict a side-by-side boxplot of the distribution of ICs across C2H2-ZF sequences tested in F2 and F3. Shown in each boxplot are the median and the interquartile range, with whiskers on the top and bottom representing the maximum and minimum data points within 1.5 times the interquartile range. For each position, IC is significantly lower for F3 binding site selections than for F2 binding site selections (red and blue boxes, respectively) as judged by a Mann–Whitney U-test ($P < 10^{-8}$, $P < 10^{-7}$ and $P < .02$ for base positions 1, 2 and 3, respectively).

'pool-assembled' libraries to bind six targets that arrays designed via modular assembly had failed to specify according to two independent projects (11,31). Moreover, four of these six targets contain at least one 3bp sub-target for which no other known pool is available, including in the OPEN resource (30). From our comprehensive pools, we were able to select arrays that specifically bind five out of six 9bp targets (Supplementary Figure S16). Further, we tested the activity of these fingers outside of the B1H system and demonstrated that they activate a GFP reporter in yeast, but only when paired with their respective target sequences. While zinc finger assembly remains challenging, our deep pools provide a resource where multiple strategies are available for binding each 3bp target. In many cases, these pools will provide at least one solution able to bind a 3bp sub-target in a desired context, even when that context has proven challenging for prior methods.

A nearest neighbor extension incorporating within-finger context expands predictive scope of the protein selections

Given the success of our `lookup' approach for predicting the experimental specificities of individual zinc fingers selected and tested in the same positional and neighboring finger contexts, we set out to determine how well the data generalize to predict DNA-binding specificities of C2H2-ZF domains with any core sequence (including those not appearing in our protein selections) and in differing positional and/or neighboring finger contexts. To make such an analysis possible, we extended a standard nearest neighbor approach as described above in the Materials and Methods section.

To gain baseline knowledge about how well our nearest neighbor decomposition (NN) approach performs under near-ideal circumstances, we first tested how accurately it predicts the experimental specificities of the 166 F2 and 69 F3 tested fingers when using the F2 or F3 protein selection data sets, respectively. For each prediction, the closest neighbor (where all amino acids matched in the core sequence) was removed from the list of neighbors. While predictions for individual core sequences differ (Supplementary Figure S17), in aggregate, >86% of the base positions within the binding sites were predicted well (Figure 9A). This is an improvement over our simpler lookup procedure, where we previously noted ∼83% concordance for fingers tested as F2. This result demonstrates that our NN approach performs well within a given positional and neighboring finger context, even in the absence of an exact match to the core sequence being predicted.

As noted above, we observed significant differences between selections performed in different positional contexts (i.e. F2 versus F3 protein and binding site selections). Thus, as a basic first test of our algorithm's ability to effectively generalize the data to different contexts, we examined how well predictions based on protein selections performed in one position correspond to experimental specificities obtained from domains tested in the other position. Specifically, for each domain whose specificity was experimentally tested in the F2 position, we predicted its specificity using the F3 protein selection data, and vice versa. As expected, we observed a drop in performance when compared to predicting within the same positional context, with ∼71–78% of base positions within binding sites predicted correctly (Figure 9A). A corresponding decline in performance was also observed when considering the fraction of C2H2-ZFs with well-matching predictions in all three base positions of the subsite (Figure 9B).

We compared these results to those obtained via prediction of the same fingers' specificities using other state-of-the-art C2H2-ZF DNA-binding site predictors including those based on RFs and SVMs (39,40). We expected our algorithm to perform favorably in this comparison, as the experimental specificities were produced with similar neighboring fingers and in the same B1H selection system as the data used for making predictions, albeit in different finger positions. Meanwhile, the data used to train the other methods were obtained across a variety of systems and neighboring finger contexts. Our simple NN method does indeed outperform the other more sophisticated machine learning methods according to the fraction of C2H2-ZF domains for which all three base positions are predicted correctly (Supplementary Figure S18 and Supplemental Methods 2e). It is also worth noting that, regardless of whether predicting within or across positional contexts, NN generally outperforms our simpler lookup approach (Supplementary Figure S19). This indicates that observing the trends across multiple binding profiles of similar core sequences is more effective than simply looking at the binding profile of a single core sequence of interest.

Finally, we compared NN's performance to the performances of the other methods in predicting binding sites for a set of naturally occurring C2H2-ZF proteins whose sequences were not used for training any of these three predictors. We observe that the performance of our simple NN approach is similar to that of the other more sophisticated methods. Agreement between aligned experimental and predicted PWMs is significant at the $P = 0.05$ level for ∼60% of these proteins using any of the three methods. At a more rigorous statistical threshold of $P = 0.01$, our method reports the highest percent of proteins that have statistically significant alignments (∼46%) among the three methods (Supplementary Figure S20A and Supplemental Methods 2d and 2e). Furthermore, about half of the proteins tested have at least 60% of their PWM columns predicted correctly by all three of the prediction methods (Supplementary Figure S20B). These results show that the DNA-binding behaviors observed in our large-scale synthetic C2H2-ZF protein selection screens apply quite well to naturally occurring C2H2-ZF proteins. In fact, the volume of quality data produced by our screens has allowed even a simple nearest neighbor method to perform comparably to models that are based on more statistically rigorous machine learning techniques but rely on less data. Thus, further development of within-finger context-dependent predictive approaches based upon our data holds great promise for yielding even better predictions.

DISCUSSION

We screened large randomized C2H2-ZF libraries with the B1H system to recover vast pools of C2H2-ZFs capable of binding each of the 64 possible 3bp targets in two differ-

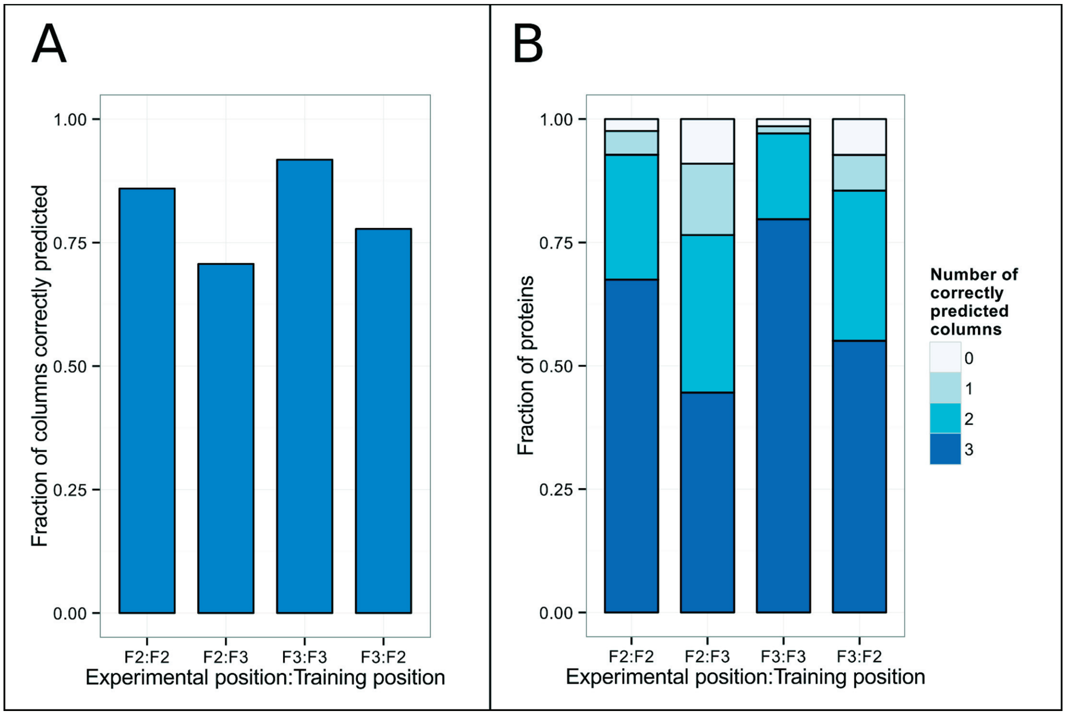


Figure 9. Performance of the nearest neighbor decomposition (NN) approach within and across positional contexts. Accuracy of predictions using nearest neighbor decomposition based upon either F2 or F3 protein selection data (training sets) when predicting the specificities of C2H2-ZFs experimentally tested in either the F2 or F3 positions (test sets). (**A**) Fraction of correctly predicted per-nucleotide base preferences, as judged by a Pearson correlation coefficient > = 0.5. (**B**) The fraction of predicted 3bp binding specificities that have 0, 1, 2 or 3 base preferences correctly predicted. For both (A) and (B), shown left to right are performances in predicting DNA-binding specificities tested as: F2 when nearest neighbor uses F2 union protein selection data; F2 when nearest neighbor uses F3 union protein selection data; F3 when nearest neighbor uses F3 union protein selection data; and F3 when nearest neighbor uses F2 union protein selection data.

ent positional contexts. One clear advantage of using large synthetic zinc finger libraries in a systematic screen of targets is that it enables the concurrent characterization of putative binding specificities for thousands of C2H2-ZF domains in a relatively small number of experiments. Using this approach, we have generated the largest collection of distinct C2H2-ZF protein–DNA interfaces to date. Computational analyses of the entirety of these data confirm the importance of the four canonical positions within the C2H2-ZF domain in determining DNA-binding specificity, and yet also provide strong evidence for the role of an alternate predicted (54) amino acid-nucleotide contact (Figures 1A and 2D) that is not currently included in the widely accepted binding model. Further, we show that data arising from these synthetic screens have great value in predicting the binding specificities of naturally occurring C2H2-ZF domains, as even a simple nearest neighbor approach performs comparably to other more sophisticated state-of-the-art C2H2-ZF DNA-binding specificity prediction algorithms.

While C2H2-ZF protein selections have been performed in the past, they have typically focused on a small number of targets at a time and many of these studies predate the era of deep sequencing (13–14,21–22,27,30,42,44,46,67). As a result, those approaches recovered a relatively small number of domains for each target that are likely biased toward the very highest affinity solutions and not necessarily the most specific solutions. In fact, in this work, we often observe an apparent balance between affinity and specificity. By taking a comprehensive approach and using deep sequencing, we were able to compute the binding profiles for all C2H2-ZFs observed in our protein selections and leverage these profiles to infer their DNA-binding specificities. Independent DNA-binding site selections within the B1H system confirm the overall accuracy of our inferred DNA-binding specificities for 235 domains (166 in F2 and 69 in F3). Further, we show that specificities derived via the B1H system correspond well to the affinity-related activation of a GFP reporter in a yeast-based system. In the future, we expect that more approaches for uncovering proteins with desired DNA-binding specificities will do so through exhaustive sampling across all possible targets, as we have done here.

Our protein selections uncovered many C2H2-ZF domains that have similar DNA-binding specificities despite having dissimilar core residues. These multiple, distinct solutions to specify a target may be especially useful in engineering C2H2-ZF proteins with desired specificities, as the binding specificity of each finger may be influenced or altered by the inter-finger context provided by neighboring C2H2-ZFs within the same array. We were able to take advantage of our deep, diverse pools to enable selection of three-fingered C2H2-ZF ‘solutions’ for binding 9bp DNA sequences that arrays generated by modular assembly had previously failed to target. Interestingly, in each example, the solution found by our selection used an amino acid at a core position that could not be coded in previous pools that utilized a VNS coding scheme (Supplementary Figure S16, red letters). These selections demonstrate that, beyond significantly increasing the number of 3bp targets for which zinc finger pools are available, the size and diversity of our pools provide solutions that have been overlooked by less comprehensive approaches. Further large-scale selections with alternate neighboring fingers may yield other solutions that were not functional in the selection context utilized here. Comparison of such selections for each 3bp target may guide us toward a set of 'universal' C2H2-ZFs whose specificities are maintained across various contexts.

The number of distinct C2H2-ZFs observed in our data far exceeds those found in any single genome, and indeed 89% of the core sequences binding DNA in the combined F2+F3 data are not present in the human genome. As many C2H2-ZFs undergo positive selection in their DNA-binding positions (68), the large set of C2H2-ZFs recovered in our screens may provide numerous evolutionary trajectories that maintain DNA-binding function, with a subset additionally maintaining or only gradually changing DNA-binding specificities.

Conversely, our data have high coverage of natural C2H2-ZFs: for example, ∼25% of human C2H2-ZFs are identical in their core sequence positions to at least one finger in our combined F2+F3 protein selection data, and ∼95% share at least three of the four amino acids in the core sequence. This current level of coverage enables predictions for most natural proteins via nearest neighbor decomposition, but also has implications regarding the fraction of natural C2H2-ZFs that are likely to bind DNA. In particular, some natural and engineered C2H2-ZFs bind protein or RNA instead of (or in addition to) DNA (69–71), and the ∼5% of human C2H2-ZFs that are not similar to any core sequence in our data set are the best candidates for such functionality. Further, in most cases, a single amino acid substitution in a DNA-binding position will not abolish DNA-binding activity, and thus it is likely that most human C2H2-ZF domains can bind DNA, though with varying levels of affinity and specificity.

In conclusion, our systematic and integrated analysis of synthetic protein selections to bind an exhaustive range of DNA targets lays a foundation for a 'bottom-up' approach to exploring DNA-binding specificity for an important regulatory domain—one that has proven difficult to experimentally characterize using state-of-the-art methods. Here we have focused on a single neighboring-finger context at high resolution, but we expect that future approaches will extend our blueprint, allowing for comparative analyses of systematic selections across a variety of contexts. We believe that such efforts will significantly increase our understanding of protein–DNA interactions for the C2H2-ZF domain, thereby facilitating the construction of more complete transcriptional regulatory networks and enabling the design of proteins that can specify any DNA-binding site—even those sites that could not be targeted by alternate methodology.

DATA AVAILABILITY

Our data set is available for download at http://zf.princeton.edu/b1h/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank the staffs of the Sequencing Core Facility and the Flow Cytometry Resource Facility at Princeton University for their advice and guidance on this project.

FUNDING

NIH [R01 GM076275 to A.V.P., J.W., M.S., in part]; NSF [ABI 1062371 to A.V.P., J.W., M.S., in part]; Graduate Research Fellowship, NSF [DGE 1148900 to J.W.]; Lewis-Sigler Fellow Program [to E.F.R., B.L.O., D.J.X., M.B.N.]. Funding for open access charge: NIH [R01 GM076275]; Lewis-Sigler Fellow Program.
Conflict of interest statement. None declared.

REFERENCES

1. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.
2. Tupler,R., Perini,G. and Green,M.R. (2001) Expressing the human genome. *Nature*, **409**, 832–833.
3. Sommer,R.J., Retzlaff,M., Goerlich,K., Sander,K. and Tautz,D. (1992) Evolutionary conservation pattern of zinc-finger domains of Drosophila segmentation genes. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 10782–10786.
4. Myers,S., Bowden,R., Tumian,A., Bontrop,R.E., Freeman,C., MacFie,T.S., McVean,G. and Donnelly,P. (2010) Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science*, **327**, 876–879.
5. Phillips,J.E. and Corces,V.G. (2009) CTCF: master weaver of the genome. *Cell*, **137**, 1194–1211.
6. Wolfe,S.A., Nekludova,L. and Pabo,C.O. (2000) DNA recognition by Cys2His2 zinc finger proteins. *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 183–212.
7. Klug,A. (2010) The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annu. Rev. Biochem.*, **79**, 213–231.
8. Pavletich,N.P. and Pabo,C.O. (1991) Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science*, **252**, 809–817.
9. Pavletich,N.P. and Pabo,C.O. (1993) Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. *Science*, **261**, 1701–1707.
10. Enuameh,M.S., Asriyan,Y., Richards,A., Christensen,R.G., Hall,V.L., Kazemian,M., Zhu,C., Pham,H., Cheng,Q., Blatti,C. *et al.* (2013) Global analysis of Drosophila Cys(2)-His(2) zinc finger proteins reveals a multitude of novel recognition motifs and binding determinants. *Genome Res.*, **23**, 928–940.
11. Lam,K.N., van Bakel,H., Cote,A.G., van der Ven,A. and Hughes,T.R. (2011) Sequence specificity is obtained from the majority of modular C2H2 zinc-finger arrays. *Nucleic Acids Res.*, **39**, 4680–4690.
12. Noyes,M.B., Meng,X., Wakabayashi,A., Sinha,S., Brodsky,M.H. and Wolfe,S.A. (2008) A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. *Nucleic Acids Res.*, **36**, 2547–2560.
13. Joung,J.K., Ramm,E.I. and Pabo,C.O. (2000) A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 7382–7387.
14. Segal,D.J., Dreier,B., Beerli,R.R. and Barbas,C.F. III (1999) Toward controlling gene expression at will: selection and design of zinc finger domains recognizing each of the 5′-GNN-3′ DNA target sequences. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 2758–2763.

15. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
16. Khalil,A.S., Lu,T.K., Bashor,C.J., Ramirez,C.L., Pyenson,N.C., Joung,J.K. and Collins,J.J. (2012) A synthetic biology framework for programming eukaryotic transcription functions. *Cell*, **150**, 647–658.
17. McIsaac,R.S., Oakes,B.L., Wang,X., Dummit,K.A., Botstein,D. and Noyes,M.B. (2013) Synthetic gene expression perturbation systems with rapid, tunable, single-gene specificity in yeast. *Nucleic Acids Res.*, **41**, e57.
18. Snowden,A.W., Gregory,P.D., Case,C.C. and Pabo,C.O. (2002) Gene-specific targeting of H3K9 methylation is sufficient for initiating repression in vivo. *Curr. Biol.*, **12**, 2159–2166.
19. Carvin,C.D., Parr,R.D. and Kladde,M.P. (2003) Site-selective in vivo targeting of cytosine-5 DNA methylation by zinc-finger proteins. *Nucleic Acids Res.*, **31**, 6493–6501.
20. Bhakta,M.S., Henry,I.M., Ousterout,D.G., Das,K.T., Lockwood,S.H., Meckler,J.F., Wallen,M.C., Zykovich,A., Yu,Y., Leo,H. *et al.* (2013) Highly active zinc-finger nucleases by extended modular assembly. *Genome Res.*, **23**, 530–538.
21. Meng,X., Noyes,M.B., Zhu,L.J., Lawson,N.D. and Wolfe,S.A. (2008) Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases. *Nat. Biotechnol.*, **26**, 695–701.
22. Urnov,F.D., Miller,J.C., Lee,Y.L., Beausejour,C.M., Rock,J.M., Augustus,S., Jamieson,A.C., Porteus,M.H., Gregory,P.D. and Holmes,M.C. (2005) Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature*, **435**, 646–651.
23. Gersbach,C.A., Gaj,T., Gordley,R.M. and Barbas,C.F. III (2010) Directed evolution of recombinase specificity by split gene reassembly. *Nucleic Acids Res.*, **38**, 4198–4206.
24. Proudfoot,C., McPherson,A.L., Kolb,A.F. and Stark,W.M. (2011) Zinc finger recombinases with adaptable DNA sequence specificity. *PLoS One*, **6**, e19537.
25. Kim,Y.G., Cha,J. and Chandrasegaran,S. (1996) Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 1156–1160.
26. Morton,J., Davis,M.W., Jorgensen,E.M. and Carroll,D. (2006) Induction and repair of zinc-finger nuclease-targeted double-strand breaks in Caenorhabditis elegans somatic cells. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 16370–16375.
27. Isalan,M., Choo,Y. and Klug,A. (1997) Synergy between adjacent zinc fingers in sequence-specific DNA recognition. *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 5617–5621.
28. Gupta,A., Christensen,R.G., Rayla,A.L., Lakshmanan,A., Stormo,G.D. and Wolfe,S.A. (2012) An optimized two-finger archive for ZFN-mediated gene targeting. *Nat. Methods*, **9**, 588–590.
29. Sander,J.D., Dahlborg,E.J., Goodwin,M.J., Cade,L., Zhang,F., Cifuentes,D., Curtin,S.J., Blackburn,J.S., Thibodeau-Beganny,S., Qi,Y. *et al.* (2011) Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA). *Nat. Methods*, **8**, 67–69.
30. Maeder,M.L., Thibodeau-Beganny,S., Osiak,A., Wright,D.A., Anthony,R.M., Eichtinger,M., Jiang,T., Foley,J.E., Winfrey,R.J., Townsend,J.A. *et al.* (2008) Rapid 'open-source' engineering of customized zinc-finger nucleases for highly efficient gene modification. *Mol. Cell*, **31**, 294–301.
31. Ramirez,C.L., Foley,J.E., Wright,D.A., Muller-Lerch,F., Rahman,S.H., Cornu,T.I., Winfrey,R.J., Sander,J.D., Fu,F., Townsend,J.A. *et al.* (2008) Unexpected failure rates for modular assembly of engineered zinc fingers. *Nat. Methods*, **5**, 374–375.
32. Mandel-Gutfreund,Y. and Margalit,H. (1998) Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. *Nucleic Acids Res.*, **26**, 2306–2312.
33. Benos,P.V., Lapedes,A.S. and Stormo,G.D. (2002) Probabilistic code for DNA recognition by proteins of the EGR family. *J. Mol. Biol.*, **323**, 701–727.
34. Kaplan,T., Friedman,N. and Margalit,H. (2005) Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput. Biol.*, **1**, e1.
35. Persikov,A.V., Osada,R. and Singh,M. (2009) Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics*, **25**, 22–29.
36. Liu,J. and Stormo,G.D. (2008) Context-dependent DNA recognition code for C2H2 zinc-finger transcription factors. *Bioinformatics*, **24**, 1850–1857.
37. Cho,S.Y., Chung,M., Park,M., Park,S. and Lee,Y.S. (2008) ZIFIBI: Prediction of DNA binding sites for zinc finger proteins. *Biochem. Biophys. Res. Commun.*, **369**, 845–848.
38. Yanover,C. and Bradley,P. (2011) Extensive protein and DNA backbone sampling improves structure-based specificity prediction for C2H2 zinc fingers. *Nucleic Acids Res.*, **39**, 4564–4576.
39. Gupta,A., Christensen,R.G., Bell,H.A., Goodwin,M., Patel,R.Y., Pandey,M., Enuameh,M.S., Rayla,A.L., Zhu,C., Thibodeau-Beganny,S. *et al.* (2014) An improved predictive recognition model for Cys(2)-His(2) zinc finger proteins. *Nucleic Acids Res.*, **42**, 4800–4812.
40. Persikov,A.V. and Singh,M. (2014) De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. *Nucleic Acids Res.*, **42**, 97–108.
41. Persikov,A.V., Rowland,E.F., Oakes,B.L., Singh,M. and Noyes,M.B. (2014) Deep sequencing of large library selections allows computational discovery of diverse sets of zinc fingers that bind common targets. *Nucleic Acids Res.*, **42**, 1497–1508.
42. Greisman,H.A. and Pabo,C.O. (1997) A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science*, **275**, 657–661.
43. Dreier,B., Segal,D.J. and Barbas,C.F. III (2000) Insights into the molecular recognition of the 5′-GNN-3′ family of DNA sequences by zinc finger domains. *J. Mol. Biol.*, **303**, 489–502.
44. Rebar,E.J. and Pabo,C.O. (1994) Zinc finger phage: affinity selection of fingers with new DNA-binding specificities. *Science*, **263**, 671–673.
45. Choo,Y., Sanchez-Garcia,I. and Klug,A. (1994) In vivo repression by a site-specific DNA-binding protein designed against an oncogenic sequence. *Nature*, **372**, 642–645.
46. Dreier,B., Beerli,R.R., Segal,D.J., Flippin,J.D. and Barbas,C.F. III (2001) Development of zinc finger domains for recognition of the 5′-ANN-3′ family of DNA sequences and their use in the construction of artificial transcription factors. *J. Biol. Chem.*, **276**, 29466–29478.
47. Noyes,M.B. (2012) Analysis of specific protein-DNA interactions by bacterial one-hybrid assay. *Methods Mol. Biol.*, **786**, 79–95.
48. Christensen,R.G., Gupta,A., Zuo,Z., Schriefer,L.A., Wolfe,S.A. and Stormo,G.D. (2011) A modified bacterial one-hybrid system yields improved quantitative models of transcription factor specificity. *Nucleic Acids Res.*, **39**, e83.
49. Meng,X., Brodsky,M.H. and Wolfe,S.A. (2005) A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nat. Biotechnol.*, **23**, 988–994.
50. Hurt,J.A., Thibodeau,S.A., Hirsh,A.S., Pabo,C.O. and Joung,J.K. (2003) Highly specific zinc finger proteins obtained by directed domain shuffling and cell-based selection. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 12271–12276.
51. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 10915–10919.
52. Jiang,P. and Singh,M. (2010) SPICi: a fast clustering algorithm for large biological networks. *Bioinformatics*, **26**, 1105–1111.
53. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
54. Persikov,A.V. and Singh,M. (2011) An expanded binding model for Cys2His2 zinc finger protein-DNA interfaces. *Phys. Biol.*, **8**, 035010.
55. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C.Y., Chou,A., Ienasescu,H. *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–D147.
56. Newburger,D.E. and Bulyk,M.L. (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **37**, D77–D82.
57. Schaefer,U., Schmeier,S. and Bajic,V.B. (2011) TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins. *Nucleic Acids Res.*, **39**, D106–D110.
58. Zhu,L.J., Christensen,R.G., Kazemian,M., Hull,C.J., Enuameh,M.S., Basciotta,M.D., Brasefield,J.A., Zhu,C., Asriyan,Y., Lapointe,D.S. *et al.* (2011) FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res.*, **39**, D111–D117.

59. Wang,J., Zhuang,J., Iyer,S., Lin,X., Whitfield,T.W., Greven,M.C., Pierce,B.G., Dong,X., Kundaje,A., Cheng,Y. *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.*, **22**, 1798–1812.
60. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
61. Gabriel,R., Lombardo,A., Arens,A., Miller,J.C., Genovese,P., Kaeppel,C., Nowrouzi,A., Bartholomae,C.C., Wang,J., Friedman,G. *et al.* (2011) An unbiased genome-wide analysis of zinc-finger nuclease specificity. *Nat. Biotechnol.*, **29**, 816–823.
62. Pattanayak,V., Ramirez,C.L., Joung,J.K. and Liu,D.R. (2011) Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nat. Methods*, **8**, 765–770.
63. Wolfe,S.A., Greisman,H.A., Ramm,E.I. and Pabo,C.O. (1999) Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. *J. Mol. Biol.*, **285**, 1917–1934.
64. Choo,Y. and Klug,A. (1994) 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.*, **91**, 11163–11167.
65. Isalan,M., Klug,A. and Choo,Y. (1998) Comprehensive DNA recognition through concerted interactions from adjacent zinc fingers. *Biochemistry*, **37**, 12026–12033.
66. Zhu,C., Gupta,A., Hall,V.L., Rayla,A.L., Christensen,R.G., Dake,B., Lakshmanan,A., Kuperwasser,C., Stormo,G.D. and Wolfe,S.A. (2013) Using defined finger-finger interfaces as units of assembly for constructing zinc-finger nucleases. *Nucleic Acids Res.*, **41**, 2455–2465.
67. Dreier,B., Fuller,R.P., Segal,D.J., Lund,C.V., Blancafort,P., Huber,A., Koksch,B. and Barbas,C.F. III (2005) Development of zinc finger domains for recognition of the 5′-CNN-3′ family DNA sequences and their use in the construction of artificial transcription factors. *J. Biol. Chem.*, **280**, 35588–35597.
68. Emerson,R.O. and Thomas,J.H. (2009) Adaptive evolution in zinc finger transcription factors. *PLoS Genet.*, **5**, e1000325.
69. Lu,D., Searles,M.A. and Klug,A. (2003) Crystal structure of a zinc-finger-RNA complex reveals two modes of molecular recognition. *Nature*, **426**, 96–100.
70. Brayer,K.J. and Segal,D.J. (2008) Keep your fingers off my DNA: protein-protein interactions mediated by C2H2 zinc finger domains. *Cell Biochem. Biophys.*, **50**, 111–131.
71. Giesecke,A.V., Fang,R. and Joung,J.K. (2006) Synthetic protein-protein interaction domains created by shuffling Cys2His2 zinc-fingers. *Mol. Syst. Biol.*, **2**, 2006.0011.