A DNA recognition code for probing the *in vivo* functions of zinc finger transcription factors at domain resolution

Berat Dogan[1,2,3,*], Senthilkumar Kailasam[1,2,*], Aldo Hernández Corchado[1,2], Hamed S. Najafabadi[1,2,¶]

[1] McGill University and Genome Quebec Innovation Centre, Montreal, QC, Canada
[2] Department of Human Genetics, McGill University, Montreal, QC, Canada
[3] Current address: Department of Biomedical Engineering, Inonu University, Malatya, Turkey

* These authors contributed equally to this work.
¶ Correspondence to: Hamed S. Najafabadi, hamed.najafabadi@mcgill.ca

Abstract

Multi-zinc finger proteins are the largest class of human transcription factors with integral roles in genome regulation and function. Often using a subset of their zinc finger domains, these proteins bind to diverse DNA sequences that make up the majority of the human regulatory lexicon. However, the molecular code that underlies the interaction of the zinc finger with DNA is incompletely understood, and for most multi-zinc finger proteins the zinc finger subset that is responsible for *in vivo* DNA binding is unknown. Here, we present a computational model, derived from a compendium of *in vivo* and *in vitro* binding preferences, that captures context-specific binding of zinc fingers to DNA. This model can predict the binding specificity of each zinc finger in its natural context more accurately than existing models, and together with molecular dynamics analyses reveals new structural aspects of DNA recognition by zinc finger proteins, including novel amino acid residues that contribute to sequence specificity. Furthermore, by combining this computational model with *in vivo* binding data, we identify the sequence preference of each protein and the zinc finger subset that is responsible for *in vivo* DNA binding, providing DNA binding maps and the associated domains for ~30% of all human multi-zinc finger proteins. We show that compared to non-DNA-binding zinc finger domains, DNA-binding zinc fingers are under stronger selective pressure across species and are depleted of genetic variants in the human population. Finally, we propose that a combination of context as well as intrinsic zinc finger features determines the ability to bind to DNA *in vivo*.

Introduction

Cys2His2 zinc finger proteins (C2H2-ZFPs) make up the largest class of human transcription factors (TFs): the human genome encodes ~750 C2H2-ZFPs, which constitutes ~45% of all human TFs. Most C2H2-ZFPs recognize distinct DNA sequences, which form the most diverse regulatory lexicon of all human TFs[1]. These proteins are characterized by the presence of multiple DNA-binding domains known as Cys2His2 zinc fingers (ZFs). Each ZF typically interacts with three to four nucleotides[2], and the amino acid-base interactions of consecutive ZFs determine the overall DNA sequence specificity of each C2H2-ZFP.

The molecular principles that dictate the relationship between the amino acid sequence of ZFs and their preference for specific DNA sequences have been studied for decades. Earlier studies mostly focused on the 3D structure of a few C2H2-ZFPs[3] and a limited number of mutation analyses[4] to derive simple models of DNA recognition by ZFs, highlighting the role of four "specificity residues" in determining the binding preference (**Fig. 1a**). Extensive *in vitro* binding data from thousands of ZFs has enabled recent studies to generate more complex "recognition codes" by correlating the amino acid sequences of the ZFs with their binding preferences[1,5-7]. These machine-learning-based recognition models have enabled the discovery of integral and unexpected roles for C2H2-ZFPs[8] and new insights into their evolution[9,10].

However, these models have substantial limitations; most importantly, they are derived from *in vitro* experiments in which individual ZFs are tested in an unnatural context (e.g. as fusion to ZFs of other proteins[1,6]), disregarding the influence of adjacent ZFs on DNA binding specificity[11]. Furthermore, most recognition models are based on the identity of the four specificity residues, ignoring the potential contribution of other ZF positions. As a result of these limitations, the predictions made by existing recognition codes only partially match the observed *in vivo* preferences of C2H2-ZFPs[1,11].

Understanding the relationship between ZFs and *in vivo* DNA binding has been further complicated by the fact that not all ZFs of a C2H2-ZFP engage with the DNA: The human C2H2-ZFPs on average contain ~10 ZFs[1], which would correspond to a binding footprint of ~30 nucleotides. However, the binding sites of C2H2-ZFPs are often much shorter[12], suggesting that only a fraction of the ZFs interact with the DNA while other ZFs might be involved in other functions such as mediating protein-protein[13] or protein-RNA[14] interactions. Currently, the DNA-engaging ZFs of only a small subset of C2H2-ZFPs have been characterized[15-17], preventing a comprehensive functional stratification of ZFs.

Here, we take advantage of a large set of recently published *in vivo* C2H2-ZFP binding preferences[18,19] and combine them with *in vitro* data[1] to derive a recognition code of ZF-DNA interaction that is significantly more accurate than existing models. We show that this code captures the contribution of non-canonical ZF residues as well as adjacent ZFs to DNA binding preference, and provides an amino acid-resolution map of ZF-DNA interaction. Furthermore, by combining this recognition code with ChIP-seq and ChIP-exo binding data, we identify the ZF domains that engage with DNA *in vivo*, and examine the evolutionary pressures acting on these DNA-engaging ZFs across species and across individuals in the human population.

Results

Base-specificity of C2H2-ZF proteins is not limited to canonical contact residues

The canonical model of C2H2-ZF interaction with DNA primarily includes four "specificity residues", each of which interacts with four specific sites on the DNA. However, this model is obtained from a limited number of protein-DNA complex structures, mutation experiments, and *in vitro* data. In order to examine whether other ZF residues might contribute to sequence-specificity, we correlated the protein sequence of 836 ZF domains from 157 human C2H2-ZF proteins with their *in vivo* binding preferences. Specifically, we used ChIP-seq and ChIP-exo data from two previously published large-scale datasets[18,19] in order to identify the *in vivo* binding preference of each protein. We then divided the *de novo*-identified motifs into three-nucleotide partitions (triplets) each representing the binding preference of one ZF domain, and examined whether the amino acid identity at each ZF position is informative about the base preference at different positions of the triplet. We observed highly significant dependencies between amino acid identity and DNA base identity for all expected canonical interactions (**Fig. 1b**). However, the significant correlations were not limited to these positions, and included other interactions, e.g. between residue +2 and DNA triplet position +3, and residue +6 and all three DNA triplet positions as well as DNA position +3 of the upstream triplet (**Fig. 1b**).

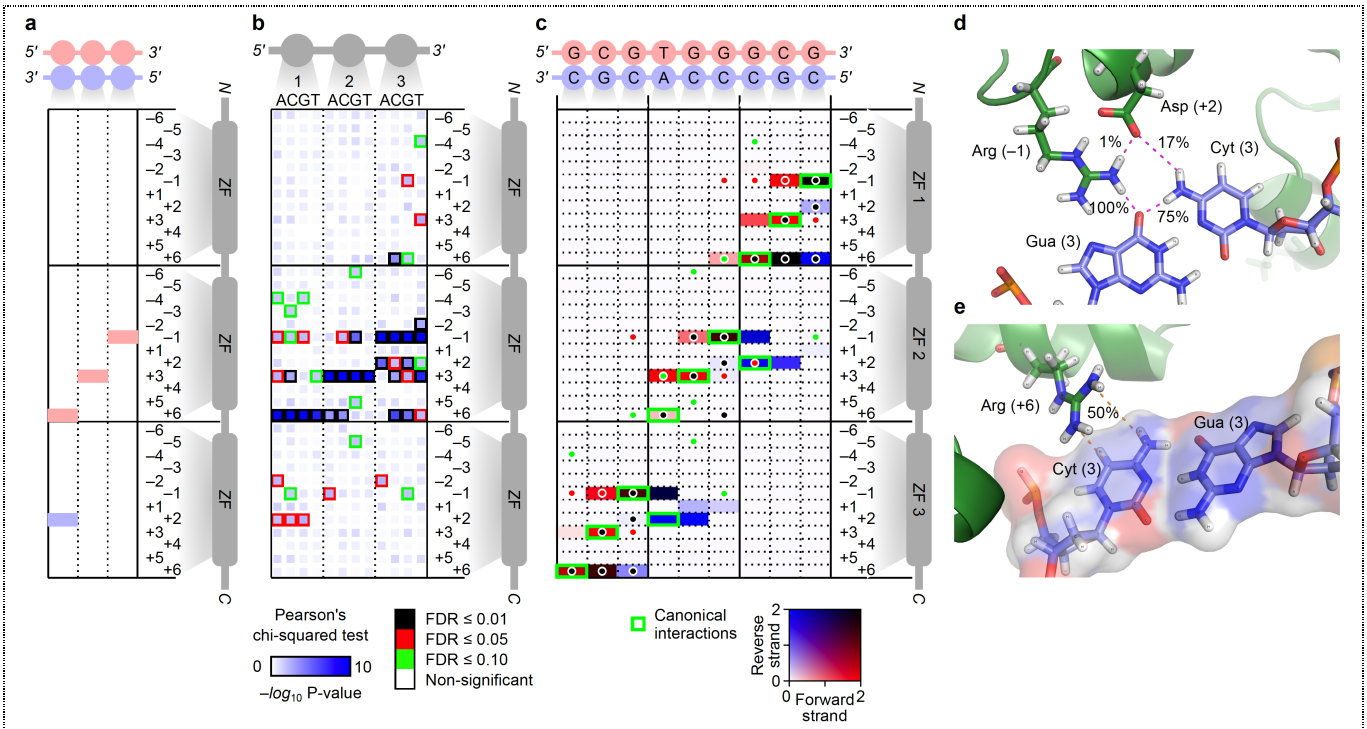

Figure 1. The canonical model of C2H2-DNA interaction only explains a subset of possible contacts. (**a**) Schematic representation of the canonical interactions between ZF positions –1, +2, +3, and +6 with a DNA triplet. Red and blue depict interaction with forward and reverse strand, respectively. Residue numbers are relative to the start of ZF alpha helix. (**b**) Associations between the identity of different ZF residues and each of four bases at different DNA positions, based on *in vivo* motifs. After splitting *in vivo* motifs into triplets, the probability of each base at each triplet position was discretized into two bins, and a chi-square test was used to examine the association of discretized probabilities with amino acid identity at each ZF position, including the N-terminal and C-terminal neighbors of the directly contacting ZF. The color gradient represents the P-value of the test. The border color of the squares represents the level of significance after Benjamini-Hochber correction for multiple hypothesis testing. Data underlying this figure are included in **Table S1**. (**c**) A 1μs-long MD simulation was used to measure the average all-atom contacts between each residue of Egr1 and the nucleotide bases of the forward and reverse strand of the target DNA, based on PDB structure 4X9J. The color gradient stands for the average number of contacts observed per frame of MD simulations, with red and blue representing contact with forward and reverse DNA strands, respectively. The dots represent significant associations based on correlation analysis of *in vivo* binding preferences, with the dot colour mirroring the border colour in panel a. Underlying data are provided in **Table S2**. (**d**) Structural illustration of a non-canonical hydrogen bond between position +2 of Egr1 ZNF1 and the reverse strand base at position 9 of the DNA sequence (i.e. position 3 of the third triplet), which can be observed at ~17% occupancy. (**e**) The planar guanidino group of Arginine in position +6 of ZF1 forms a π-π stacking interaction with reverse strand base of the DNA position 9 (~50% occupancy).

In order to understand whether these non-canonical associations may actually correspond to interactions between ZF residues and target DNA bases, we used molecular dynamics (MD) to track the amino acid-base contacts in the Egr1-DNA complex. Interestingly, many of the correlations observed from *in vivo* binding preferences can be captured as hydrogen bonds and other non-covalent interactions (e.g. van der Waals, ionic, or stacking interactions) in the Egr1-DNA complex (**Fig. 1c**). Example structures highlighting some of these novel contacts are shown in **Fig. 1d,e**, including a non-canonical hydrogen bond and a π-π stacking interaction between DNA and positions +2 and +6 of ZF1, respectively. Together, these analyses suggest that DNA base-specificity might not be limited to canonical contact residues of the C2H2-ZF domain. Given that these associations can be captured using available *in vivo* binding preferences, we set to develop a computational recognition code that takes into account these non-canonical interactions, including the interactions between a DNA triplet and neighboring C2H2-ZFs.

Combining *in vitro* and *in vivo* data results in an improved context-aware recognition code

The *in vivo* binding preferences of C2H2-ZFPs provide the opportunity to capture non-canonical ZF-DNA contacts as well as the effect of neighboring ZFs on DNA sequence recognition. The available *in vitro* data[1], on the other hand, provide a higher coverage and larger training dataset for modeling ZF-DNA interactions.

Therefore, we sought to obtain a "compound recognition code" (hereafter referred to as C-RC), which combines *in vitro* and *in vivo* data in order to obtain optimal prediction accuracy. We performed a systematic feature selection to identify the ZF residues that maximize prediction accuracy when they are used as input to random forest regression models[20] that predict specificity for each of four bases at each DNA triplet position (**Fig. 2a**). We identified different features that optimize prediction accuracy in different contexts – the contexts that we considered included whether the ZF was located at the N-terminus, C-terminus, or the middle of a DNA-engaging ZF-array (**Fig. S1**).

In addition, we also examined whether different encodings of the amino acids can improve prediction accuracy by reducing the number of parameters of the recognition code. Previous attempts at developing a C2H2-ZF recognition code have considered the 20 amino acids as unrelated identities, e.g. through one-hot encoding[1]. However, amino acids with similar biochemical and/or structural properties often have similar propensities for interaction with different bases. Therefore, we specifically examined the possibility that encoding the amino acids by their biochemical properties (**Fig. 2b**) may reduce the complexity of the parameters that our computational model needs to learn, and therefore increase the model accuracy and generalizability. This encoding indeed allowed the random forest to learn simpler rules that were shared among amino acids with similar properties (**Fig. 2c** and **Fig. S2**), and resulted in significantly better predictions in different training-validation scenarios (**Fig. 2d**).

The final recognition code consists of 36 random forests: 12 random forests for each of the N-terminal, C-terminal, or middle ZF contexts, with each random forest predicting the preference for binding to one of the four possible bases at one of the three DNA triplet positions (**Fig. S1**). We benchmarked the performance of our C-RC using different evaluation schemes. First, we used 50 randomly selected ZF-target pairs from ChIP-seq data, which were excluded from all stages of the random forest training, including feature selection. As shown in **Fig. 2e**, C-RC outperformed a previous B1H-based recognition code (B1H-RC)[1] in predicting the base preferences for 11 out of 12 base-position combinations. Next, we compared the accuracy of C-RC with two available approaches, the B1H-RC[1] and ZFModels[5], using protein-binding microarray (PBM) data from a set of non-animal C2H2-ZFPs[10] and a set of synthetic C2H2-ZFPs created from concatenating different ZFs from different proteins[21]. C-RC produced motifs that were significantly more similar to the PBM motifs than both B1H-RC and ZFModels predictions in both datasets (**Fig. 2f,g**), suggesting that the new code outperforms the state-of-the-art irrespective of the nature of the reference data (*in vivo* or *in vitro*) or the source of the protein (human, non-animal, or synthetic).

The new recognition code reflects known and novel structural features of ZF-DNA interaction

Random forests are ensembles of large number of trees, which are not directly interpretable. To understand how C-RC works, we examined the output of the code for 100,000 randomly generated ZFs, each with one adjacent ZF on each side. By correlating the output of C-RC with the amino acid identities at each of the central or adjacent ZF positions, we were able to obtain a simplified picture of the average effect of each amino acid at each position on the recognition of different bases. This simplified representation ignores more complicated features that might be encoded in the random forest model, such as the non-linear dependencies of different positions; nonetheless, it reveals key features, such as the associations between different positions of the ZF and target the DNA. Many of these associations are consistent with the canonical model of ZF-DNA interaction as shown in **Fig. 3a**. However, the model also encodes strong associations that are not part of the canonical model, most prominently between position 3 of the DNA triplet and position +6 of the N-terminal neighboring ZF (**Fig. 3b**).

To understand whether these novel associations are structurally relevant, we performed MD simulations of the CCCTC-binding factor (CTCF) in complex with its cognate DNA, and indeed identified a non-canonical interaction between position +3 of the DNA triplet that is adjacent to ZF5 and residue +6 of ZF4 (**Fig. 3c,d**). Consistent with this observation, C-RC predicts that mutations in residue +6 of ZF4 on average negatively affect this interaction (**Fig. 3e**). In fact, systematic application of this *in silico* mutation strategy revealed a quantitative relationship between the associations predicted by C-RC and the strength of H-bonds observed in MD simulations, (**Fig. 3f**), suggesting that C-RC not only enables the identification of non-canonical interactions, but also allows us to predict the strength of canonical interactions given the sequence of the ZFP and its target DNA.

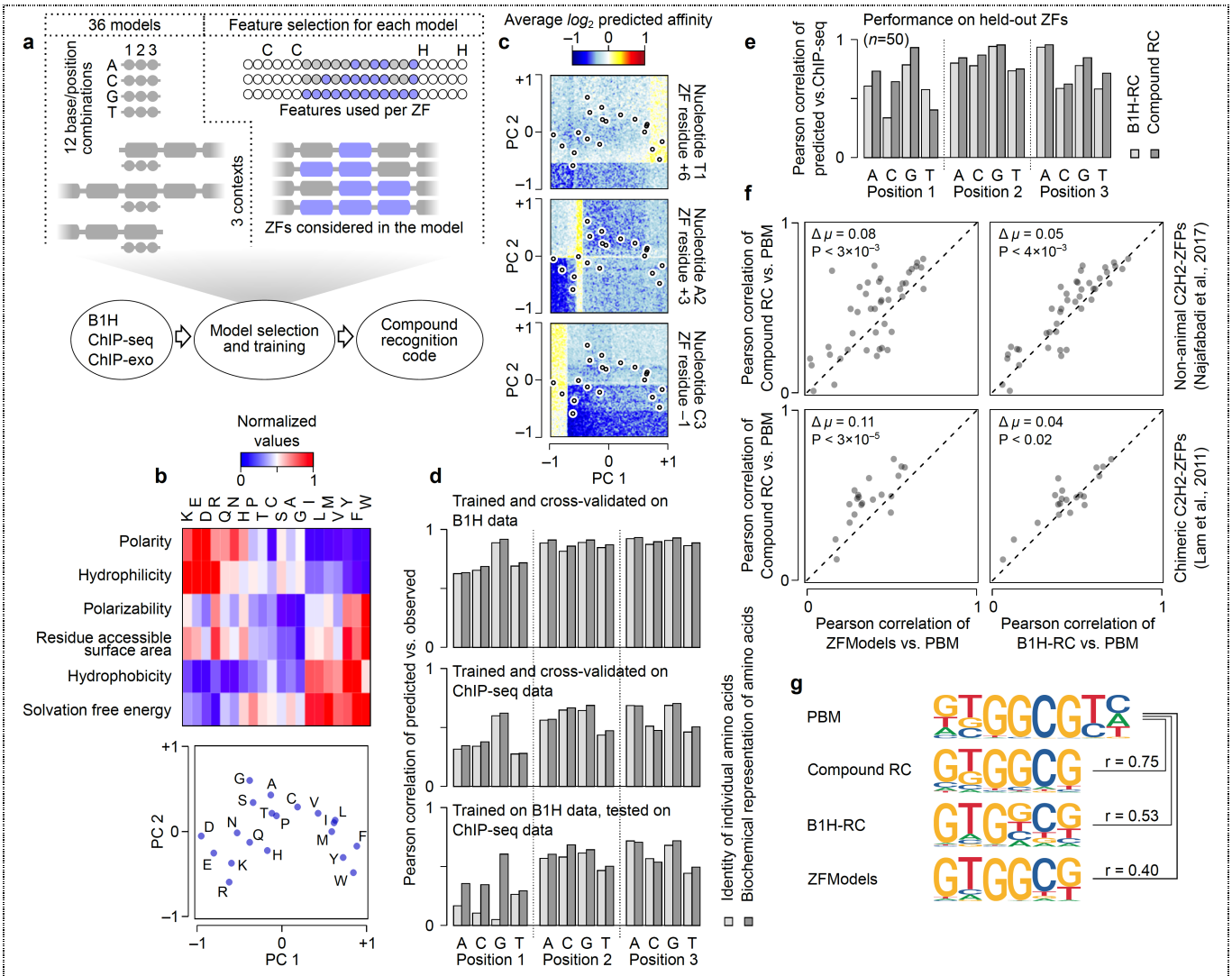

Figure 2. Training and evaluation of C-RC. (**a**) Schematic representation of the procedure for training C-RC. Thirty-six models were trained separately (top left), corresponding to recognition of four bases at three DNA positions in three different contexts. Feature selection for each model included selection of residues to be considered for each ZF, as well as the choice of ZFs (top right). See **Fig. S1** for detailed representation of the selected features for each model. (**b**) Encoding the amino acids by their biochemical properties. The heatmap (top) represents the biochemical properties we considered. The PCA-transformed values are shown in the scatterplot at the bottom. Underlying data are provided in **Table S3**. (**c**) Visual representation of example rules learned by random forest after encoding the amino acids by their biochemical properties. We generated 40,000 "pseudo-amino acids" by dividing the PCA plot in panel (b) into a 200×200 grid, and then generated 40,000 random ZFs by sampling (without replacement) these pseudo-amino-acids for each of the 12 ZF positions. The color gradient in the graphs shows the predicted affinity of these random ZFs for recognition of base T at position 1 of DNA triplet (T1), recognition of A2, or recognition of C3, as indicated beside each graph (see **Fig. S2** for additional base-position combinations). The ZFs are projected on the scatterplot based on the PCA coordinates of the pseudo-amino-acid at position +6, +3, or –1, as indicated beside each graph. (**d**) The performance of the recognition code for predicting the probability of each base at each triplet position, when the amino acids are encoded as categorical variables (20 individual identities) or using the PCA-transformed biochemical properties. We used 5-fold cross-validation on B1H motifs[1] (top) or ChIP-seq motifs[18] (middle), or trained the recognition code on B1H motifs and tested on ChIP-seq motifs (bottom), ensuring that any ZF present in the ChIP-seq data was removed from the B1H training set. (**e**) Comparison of the performance of C-RC and a B1H-based recognition code[1] (B1H-RC) for predicting the base probabilities at different triplet positions for 50 ChIP-seq ZFs that were excluded from the feature selection procedure and model training. (**f**) Systematic comparison of C-RC, B1H-RC, and ZFModels predictions vs. PBM motifs from non-animal C2H2-ZFPs[10] and chimeric C2H2-ZFP constructs[21]. Pearson correlation similarity of motif pairs are calculated by MoSBAT-e[22]. P-values are based on two-tailed paired t-test. (**g**) As an example, MoSBAT-e comparison of the motifs predicted by C-RC, B1H recognition code, and ZFModels[5] vs. the motif obtained by PBM[10] for a C2H2-ZFP from *Rhizopus delemar* (UniProt ID I1BM86) is shown.

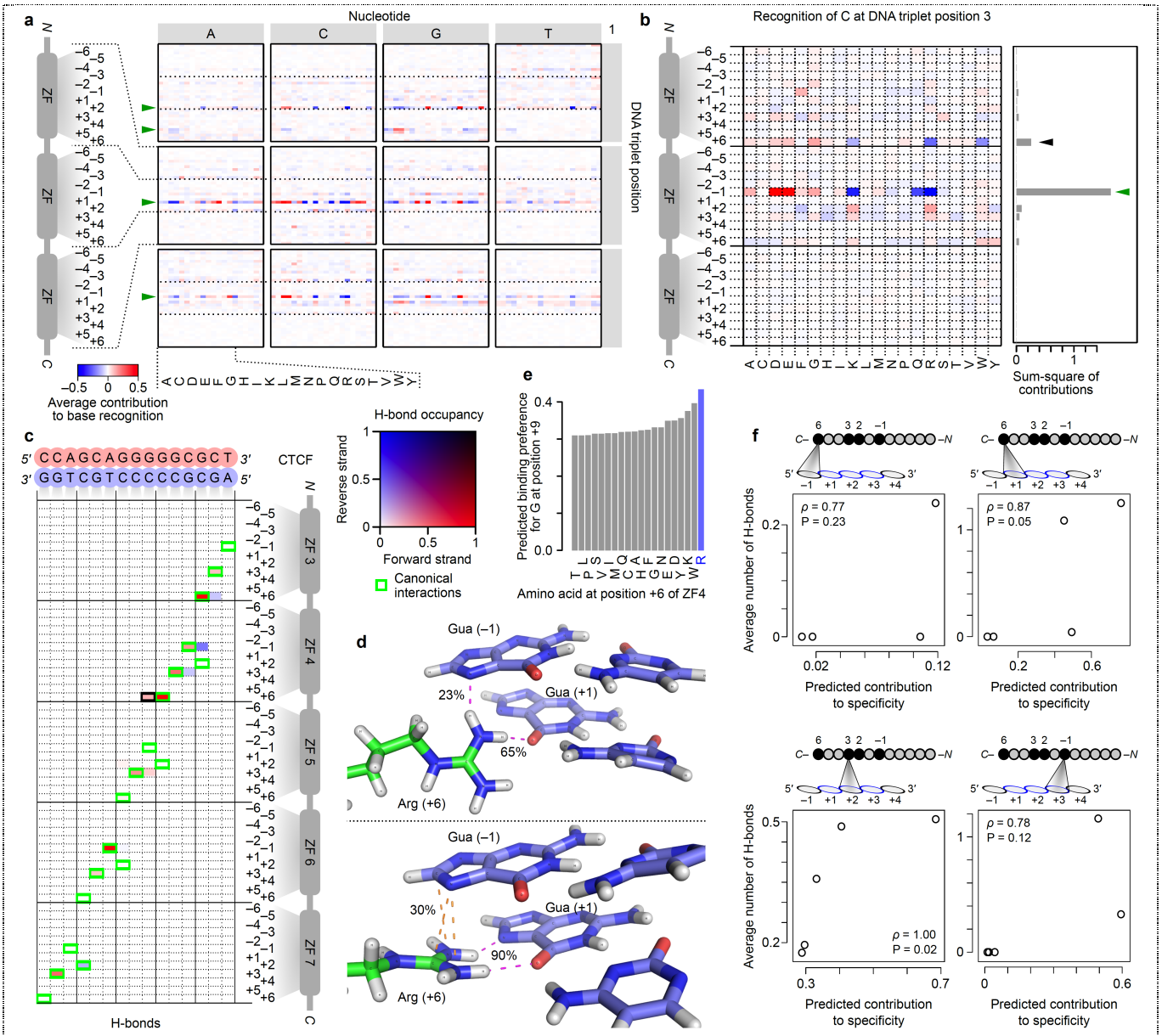

Figure 3. C-RC quantitatively predicts amino acid-base interactions. (**a**) The average contribution of each amino acid at different ZF positions for recognition of each base at different DNA triplet positions. Each panel represents recognition of one base, indicated above the figure, in a specific DNA triplet position, indicated on the right; each column represents one amino acid, each row represents one ZF position, and the color gradient denotes the contribution toward specificity (red: increased preference for the specified base; blue: decreased preference). The specificity residues that, according to the canonical model, contribute to the recognition of each DNA position are shown with green arrows. (**b**) Recognition of base C at position 1 of the DNA triplet. The bar graph on the right shows the squared sum of values at each ZF position. The black arrow represents a potential new interaction with position +6 of the N-terminal neighboring ZF. (**c**) H-bonds identified from MD simulations of CTCF in complex with its target DNA. The interactions in the canonical C2H2-ZF model are shown with green border. The non-canonical interaction highlighted in panel b is shown here with black border. Underlying data are provided in **Table S4**. (**d**) The Arginine at position +6 of ZNF4 forms a non-canonical interaction with position –1 of its cognate DNA triplet. The interaction exists in two dominant conformations: a hydrogen bond (23% of MD simulation frames) between the Arginine and the base (top) and a stacking overlap (30%) between Arginine and the guanidino group and the DNA base (bottom). (**e**) The predicted preference of variants at position +6 of ZF4 for binding to G at DNA position 9 (i.e. position –1 relative to the ZF4-associated triplet). The wild-type amino acid is highlighted in blue. (**f**) Scatterplots for the recognition code-predicted associations vs. MD simulation-based H-bonds. In each plot, each dot represents one ZF. The graphs represent the amino acid-base pairs for which at least one H-bond and at least one non-zero association based on the recognition code was found. The spearman correlations are shown, along with their associated P-values (two-tailed).

A domain-resolution map of the *in vivo* interactions between human C2H2-ZFPs and their genomic targets

Human C2H2-ZFPs on average contain ~10 ZFs per protein[1], which is substantially longer than what would be expected from the binding preference of most transcription factors, suggesting that only a fraction of the ZFs of most proteins engage with DNA. Combining *in vivo* binding data with the predicted DNA preferences of the ZFs enables the identification of the ZF domains that engage with DNA *in vivo*[23]. We therefore sought to combine our new recognition code with the genomic binding sites of human C2H2-ZFPs to characterize their *in vivo* binding preferences and the ZF domains that engage with the DNA. We used a previously described framework[23], which starts by predicting the binding preferences of all possible ZF arrays contained in a C2H2-ZFP, and then identifies the ZF array whose binding preference maximally explains the observed *in vivo* binding sequences, along with its optimized binding motif. We applied this framework to previously published ChIP-seq and ChIP-exo data[18,19], limiting the results to cases where the C-RC predictions and *in vivo*-optimized motifs had significant similarity at $P<0.001$. We identified significant motifs for 109 ChIP-seq and 126 ChIP-exo datasets (**Fig. 4a** and **Fig. S3**, respectively), encompassing a total of 209 unique proteins.

Several lines of evidence from the analysis of the *in vivo* data point to substantially improved performance of the C-RC compared to the state-of-the-art, adding to the benchmarking results presented in the previous sections. First, in both ChIP-seq and ChIP-exo data, the C-RC was able to predict *in vivo*-enriched motifs for more C2H2-ZFPs than B1H-RC (**Fig. 4b**). Particularly, in the ChIP-exo data, the new recognition code was able to increase the number of proteins with significant motifs by ~25%. Secondly, for datasets that produced significant motifs using both the B1H-RC and the new recognition code, C-RC was able to identify a larger number of ZFs that engage with DNA (**Fig. 4b**). Thirdly, for these common proteins, the motifs identified by C-RC had overall higher quality, as measured by the area under the receiver operating characteristic curve (AUROC) for distinguishing real binding sites from dinucleotide-randomized sequences (**Fig. 4b**).

Based on the motifs identified by combining C-RC and ChIP-seq data, on average about 56% of the ZFs of each protein appear to engage with DNA *in vivo* (45% for ChIP-exo data), although these ZFs vary substantially in terms of their sequence specificity, as measured by the information content of the 3-nucleotide motif that corresponds to each ZF (**Fig. 4a** and **Fig. S3**). Interestingly, we observed that DNA-engaging ZFs are overall more conserved across vertebrates than non-binding ZFs or ZFs that bind to DNA with low sequence specificity (measured by information content of their associated motifs. This trend can be seen when we aggregate all ZFs from all C2H2-ZFPs that have ChIP-seq data (**Fig. 5a**), as well as when we analyze the ZF domains of each C2H2-ZFP separately (**Fig. 5b**). We also separately analyzed the conservation pattern of ZF domains in C2H2-ZFPs that contain a KRAB domain – these C2H2-ZFPs are relatively recent and often involved in repression of transposable elements[1,9,18,19]. Surprisingly, despite their overall lower conservation across vertebrates due to their more recent origin, the correlation between conservation and DNA-binding can be clearly seen in KRAB-containing C2H2-ZFPs (**Fig. S4a**), suggesting that DNA-binding ZFs are more conserved than non-binding ZFs both in KRAB-containing and non-KRAB C2H2-ZFPs.

The higher conservation of DNA-engaging ZFs suggests that mutations in these regions are overall more deleterious than mutations in ZFs that do not bind DNA (which may potentially have other functions). We set out to confirm this hypothesis by analysis of population-level genetic variations in human ZFs. We analyzed genetic variation data from gnomAD, which encompasses variants identified from >140,000 individuals, and observed that missense variations in ZFs that engage with DNA with high specificity are significantly less frequent than missense variations in ZFs that do not engage with DNA or have low sequence specificity (**Fig. 5c**). We also specifically examined the frequency of rare missense variants (minor allele frequency $< 10^{-5}$) and more common variants (minor allele frequency > 0.001), reasoning that extremely rare variants are likely to be recent and, therefore, have not been filtered by negative selection yet, as opposed to common variants[24]. We observed that common variants are significantly depleted from DNA-engaging ZFs compared to non-binding or low-specificity ZFs (**Fig. 5d**), suggesting that these ZFs are under stronger negative selection. This trend holds for both KRAB-containing C2H2-ZFPs and non-KRAB C2H2-ZFPs (**Fig. 5d**), confirming the pattern that we observed by analysis of ZF conservation across vertebrates. Nonetheless, we observed overall stronger depletion of common variants in non-KRAB C2H2-ZFPs, to the extent that non-DNA-binding ZFs of these proteins appear to harbor the same ratio of common and rare variants as DNA-binding ZFs of KRAB proteins (**Fig. 5d**). This trend suggests a stratified model of selective pressure on the ZFs based on their function, in which the mutations at the DNA-binding ZFs of non-KRAB proteins are overall the most deleterious ones, whereas mutations at the non-DNA-binding ZFs of KRAB proteins are least affected by negative selection (**Fig. 5e**).

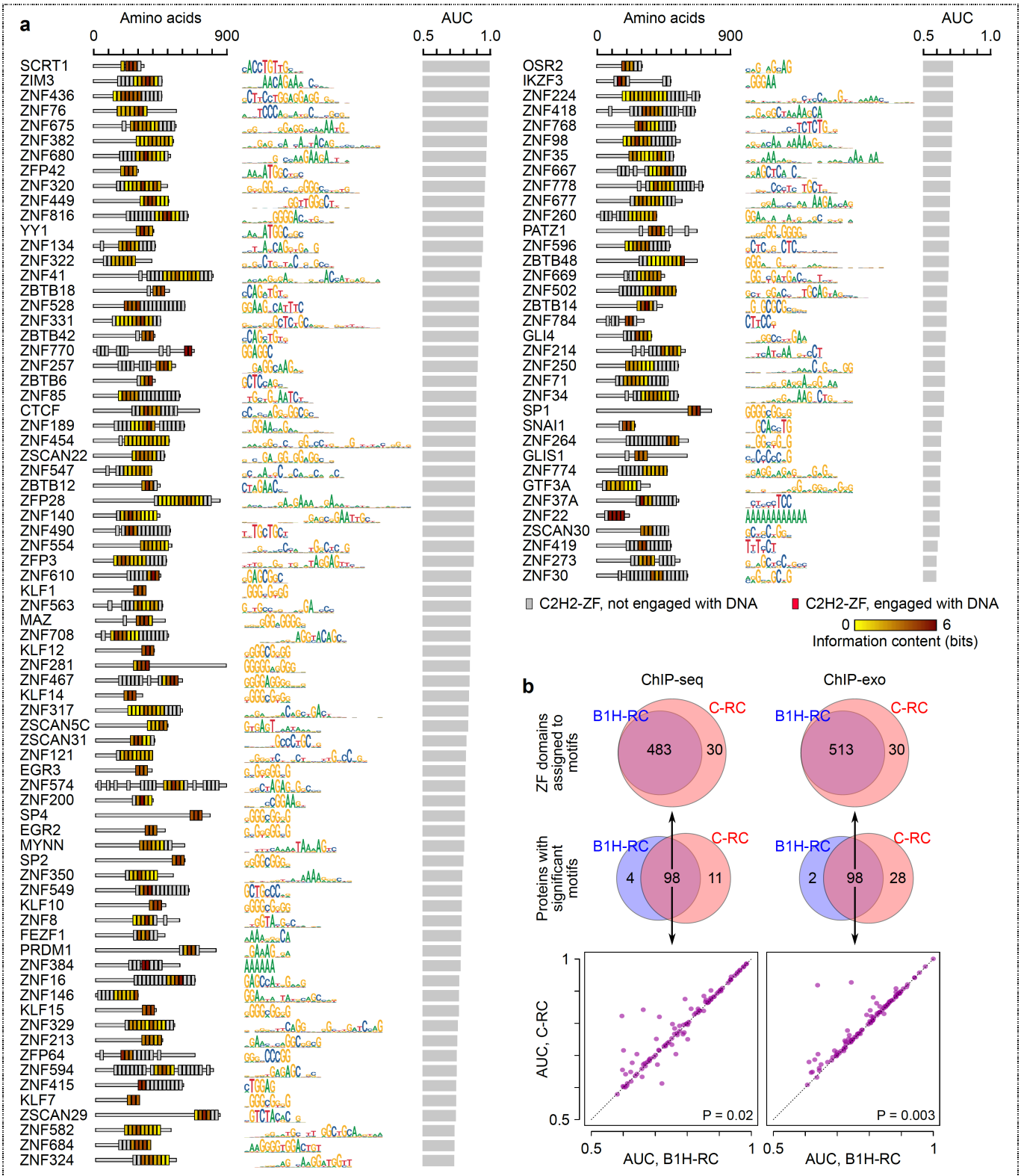

Figure 4. C-RC improves identification of motifs and the associated ZFs from *in vivo* data. (**a**) Motifs identified by recognition code-assisted analysis of ChIP-seq data for C2H2-ZFPs. For each protein, the domain structure is shown on the left (only the ZF domains). The bar graph shows the AUROC for distinguishing the binding site sequences from dinucleotide-shuffled sequences. The ZFs that correspond to the identified motifs are highlighted, with the colour gradient representing the sequence-specificity of the ZF (as measured by information content). See **Fig. S3** for ChIP-exo motifs. Position-specific frequency matrices are provided in **Data Files S1 and S2**. (**b**) Comparison of the number of motifs (middle) and the number of ZFs associated with the motifs (top), as identified by the B1H-RC or C-RC using ChIP-seq data[18] (left) or ChIP-exo data[19] (right). For common motifs, the AUROC values of the optimized versions are compared in the scatterplots at the bottom.

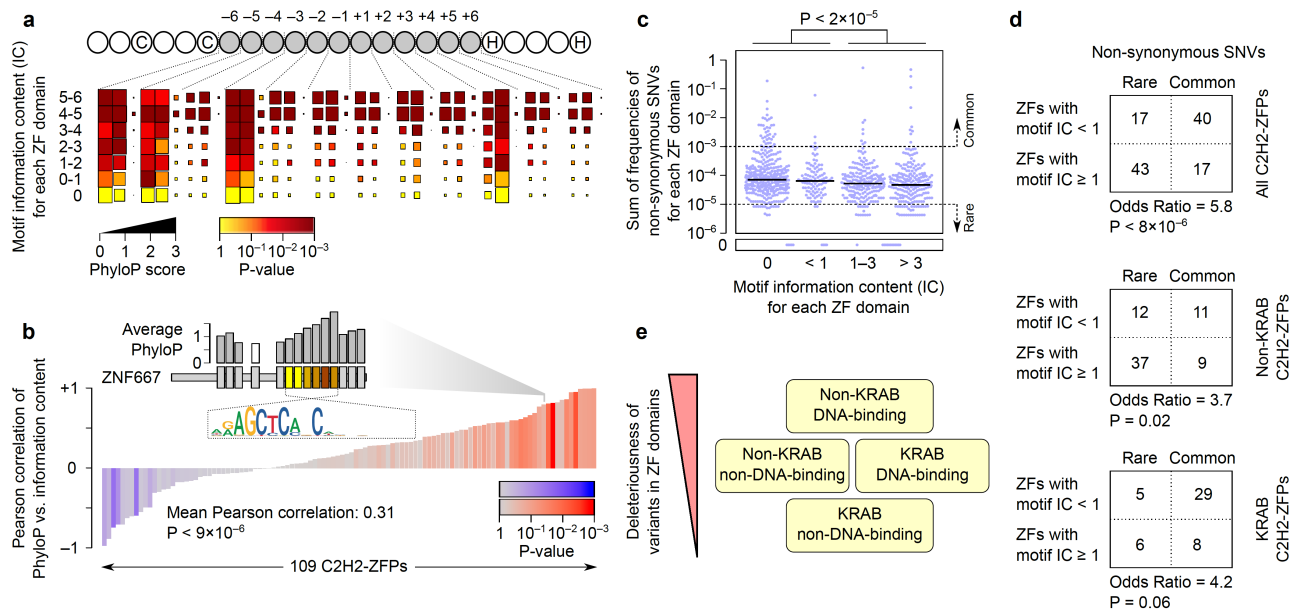

Figure 5. DNA-binding ZFs are under stronger purifying selection for both KRAB and non-KRAB C2H2-ZFPs. (**a**) PhyloP[25] conservation score profile across ZFs that do not engage with DNA (i.e. IC=0) and those that engage with varying sequence-specificity. Each set of three columns represents one codon in the ZF-encoding sequence, and each row represents one set of ZFs, binned by the information content (IC) of their associated triplet in the *in vivo* motif. The box size represents the average phyloP score, whereas the color gradient represents the difference of conservation compared to ZFs that do not engage with DNA (P-value based on one-tailed t-test). Data provided in **Table S5a**. (**b**) Correlation of the phyloP score profile of each C2H2-ZFP with the IC of the ZFs. The inset illustrates an example in which the phyloP score of the ZFs (bar plot above the ZNF667 domain structure) correlates with the IC of the ZF-associated triplets (color of the ZF domains, similar to **Fig. 4a**). For each C2H2-ZFP, the P-value associated with the Pearson correlation is shown using the color of the associated bar (two-tailed correlation t-test). The overall P-value for deviation of the mean of correlations from zero is also shown (two-tailed t-test). Also see **Fig. S4**. (**c**) Minor allele frequencies of non-synonymous variants for ZFs that do not engage with DNA (IC=0), ZFs with low sequence-specificity (IC<1), and ZFs with high sequence-specificity (IC≥1). Each dot represents one ZF, and the y-axis shows the sum of frequencies of all non-synonymous minor alleles that overlap with the ZF. P-value is based on two-tailed t-test. Data provided in **Table S5b**. (**d**) Overlap of rare and common non-synonymous variants with ZFs that engage with DNA with high specificity (IC≥1) and ZFs that do not engage or have low specificity (IC<1). The top table represents all C2H2-ZF proteins with ChIP-seq-based motifs, whereas the next two tables are stratified based on absence or presence of a KRAB domain in the protein. (**e**) A schematic model showing the inferred deleteriousness of variants in different ZFs, as suggested by the frequencies of common and rare variants.

Discussion

Our analyses indicate that the compound recognition code (C-RC) provides a more accurate model of DNA binding by C2H2-ZFs than existing models, as a result of incorporating *in vivo* binding data during model construction. Interestingly, this improvement was achieved by inclusion of *in vivo* binding preferences of less than 840 human zinc fingers, leaving open the possibility of further improvements as additional *in vivo* data become available for C2H2-ZFPs. Furthermore, our molecular dynamics analyses suggest that C-RC can quantitatively infer the contribution of different ZF residues to DNA specificity, including new potential interactions beyond the canonical model such as π-π stacking of planar amino acids with DNA bases. These interactions appear to occur as a result of alternate conformations of the amino acid side chains at the DNA-protein interface, explaining why they were not reported earlier in the X-ray crystal structures of C2H2-ZPFs, which often represent only the most stable conformation. We also note that the ability of C-RC to model the base-specific interactions of different ZF residues raises the possibility that it can be used to predict the effect of missense mutations, providing an approach for understanding the functional consequences of genetic variation and mutations in these proteins.

By combining the C-RC with ChIP-seq and ChIP-exo data, we were able to obtain a domain-resolution map of ZFP-DNA interactions for a total of 209 C2H2-ZFPs, which suggested that ~50% of ZF domains engage with DNA *in vivo*. We observed that DNA-engaging ZFs are more conserved across species than ZF domains that do not engage with DNA *in vivo*. This trend in conservation, however, is not limited to the base-contacting residues and can be observed in all ZF residues, suggesting that non-base-contacting residues are also under stronger selective pressure in ZFs that engage with DNA compared to ZFs that do not engage with DNA *in vivo*. This observation is consistent with a role of non-base-contact residues in DNA-binding (e.g. through interaction with DNA backbone, as suggested by recent studies[10]). On the other hand, the relative lower sequence conservation in ZFs that do not bind DNA may reflect alternative functions that allow higher tolerance to mutations in the protein sequence.

Interestingly, this pattern can be observed not only in non-KRAB C2H2-ZFPs, but also in KRAB proteins, whose functions in regulating gene expression are less understood. This observation suggests that mutations in the DNA-binding domains of KRAB proteins are deleterious, which is also supported by depletion of common genetic variants in DNA-binding ZFs at the population level. This is a surprising finding given that there are few known disease-linked KRAB proteins, and highlights the need for better characterization of the functions of these proteins in genome regulation and cell function. A question that needs to be addressed is whether the KRAB proteins that we analyzed form a representative sample from this subclass of C2H2-ZFPs. We limited our analyses to C2H2-ZFP binding sites that do not overlap endogenous repeat elements (EREs) due to their confounding effect on motif finding[23]; however, a large number of the KRAB protein binding sites are in EREs[1,18,19], which may result in the identification of high-quality motifs for fewer KRAB C2H2-ZFPs. Indeed, we were able to obtain high-quality motifs for only 25% of the KRAB proteins that almost exclusively bind to ERE regions (i.e. >90% of top 500 peaks in EREs). In contrast, for the rest of KRAB proteins, we had >97% success rate in finding high-quality motifs that match the recognition code predictions, suggesting that our results may be biased against exclusive ERE binders. We note, however, that this group likely forms a relative minority of KRAB proteins (12 out of 55 KRAB proteins with ChIP-seq data match our definition of exclusive ERE binder).

Previous studies have shown that, at least *in vitro* and when fused to ZF1-2 of the Egr1 protein, not all ZF domains are able to bind to DNA, suggesting that there are intrinsic differences between DNA-binding and non-binding ZFs. Do these intrinsic differences explain the observation that only a fraction of human ZF domains engage with DNA *in vivo*? We have found that ZFs that bind to DNA *in vivo* have, on average, higher sequence specificity than ZFs that do not engage with DNA (**Fig. S4c**), as predicted by our recognition code. However, sequence specificity alone is a modest predictor of *in vivo* DNA binding. Furthermore, ZFs that do not engage with DNA *in vivo* have, on average, significantly higher sequence specificity than ZFs in pseudogenes (which are presumably non-functional), suggesting that at least some of the non-engaging ZFs are intrinsically capable of binding to DNA (**Fig. S4c**). This notion is also supported by direct comparison of our *in vivo* binding map with *in vitro* bacterial one-hybrid data[1], which shows that while there is a statistically significant overlap between ZFs that bind to DNA *in vitro* and *in vivo* (odds ratio=1.7, $P < 0.003$, Fisher's exact test), at least 15% of ZFs that do not engage with DNA *in vivo* are still capable of binding to DNA *in vitro* (**Fig. S4d**). This raises the possibility that DNA-binding capability is, at least to some extent, determined by the context, which may include the ZF position in the array as well as the identity of adjacent ZFs. More strikingly, ~55% of ZFs that do not bind DNA *in vitro* interact with DNA *in vivo*, often with high specificity, enforcing the notion that *in vitro* binding assays at best only partially reflect the *in vivo* function of ZFs.

Finally, we note the possibility that some ZFs may engage with DNA in alternative binding modes that we have not yet characterized: A few C2H2-ZFPs are known to employ alternative sets of ZFs to bind to DNA[15-17], suggesting that some ZFs might be engaged only in a subset of *in vivo* binding sites – such ZFs would not be part of the core motifs that we have presented in this paper, which can potentially explain their higher-than-random sequence specificity. Therefore, the factors that determine the ability of ZFs to engage with DNA may depend not only on the C2H2-ZFP itself, but also on its binding sites.

Methods

Obtaining the data set of C2H2-ZF preferences for training the recognition code

We obtained the *in vivo* binding sites of 313 proteins from two previous studies[18,19], representing 131 proteins with ChIP-seq and 221 proteins with ChIP-exo data in HEK293 cells. The binding sites of each protein were directly downloaded from GEO datasets associated with these publications (GEO accessions GSE76494 and GSE78099). For each datasets, we ran RCADE[23] on the top 500 peaks that did not overlap any endogenous repeat elements (EREs), as described previously[26] (ERE coordinates were obtained from the RepeatMaser track of the UCSC Genome Browser[27]). For successful runs, we included the top-ranking motif of each protein in our training dataset. For the proteins that had both ChIP-seq and ChIP-exo data and resulted in a significant motif in both cases, we used the ChIP-exo motif, given the higher resolution of ChIP-exo compared to ChIP-seq. We split each motif into its constituent triplets (total of 836 triplets), and also extracted the ZF associated with each triplet, as well as the two ZFs on the two ends of each direct ZF. For the triplet at the 5' end of each motif, no ZF from the C-terminus of the associated ZF was used. Similarly, for the triplet at the 3' end of each motif, no ZF from the N-terminus of the associated ZF was used (see **Fig. S1** for a schematic representation). We set aside 50 randomly selected ZFs from this training dataset, so that they could be used at the end for testing the obtained model. We augmented the training dataset with *in vitro* B1H-based motifs from 8138 ZFs[1]. Since in these B1H experiments only the binding preference of a single ZF was queried at a time, the associated ZF of each motif does not have any N-terminal or C-terminal adjacent ZFs.

Biochemical encoding of amino acids

For each amino acid, we extracted six different biochemical properties[28] (**Table S3a**). Given the strong correlations among these properties, we used principal component analysis (PCA) and used the transformed coordinates of the amino acids on the first two components (**Table S3b**) for downstream analyses.

Feature selection and training of the recognition code

We considered four possible input configurations for predicting the preference of the ZF for a DNA triplet: (a) only the ZF that is directly in contact with the triplet, (b) the direct ZF plus its N-terminal neighbor, (c) the direct ZF plus its C-terminal neighbor, and (d) the direct ZF plus its two neighbors.

For each of these configurations, we considered different set of input features (**Figure S1b**): (i) only the four canonical residues of the ZFs (i.e. residues –1, +2, +3, and +6), (ii) the seven residues that showed the highest correlations with the DNA preference according to Chi-square test of *in vivo* data (i.e. residues –4, –2, –1, +1, +2, +3, and +6), and (iii) all 12 residues between the second Cys and the first His in the ZF (i.e. the X_{12} in the Pfam pattern $X_2CX_{2,4}CX_{12}HX_{3,4,5}H$). For each configuration a-d, we tried each feature set i-iii for predicting each of the four bases at each position of the triplet, and chose the feature set that maximized the Pearson correlation of predicted vs. target values during 5-fold cross-validation.

Finally, for each of the three possible ZF contexts, we compared the performances of the compatible configurations and selected the best-performing configuration by 5-fold cross-validation. The ZF contexts were: (1) ZFs that are at the N-terminus of the DNA-binding ZF array (compatible with configurations a and c), (2) ZFs that are at the C-terminus of the DNA-binding array (compatible with configurations a and b), and (3) ZFs that in the middle of the DNA-binding array (compatible with configurations a, b, c, and d). The configuration that was selected for each context and the corresponding feature set are shown in **Fig. S1c**.

Evaluating the performance of the final mode

After configuration- and feature-selection based on 5-fold cross-validation results, the final recognition code was tested on 50 randomly selected held-out ZFs that were not included at any stage of training. We used the Pearson correlation of predicted vs. observed base probabilities as the measure of performance. Also, we used MoSBAT-e[22] to compare the predictions of the code with motifs obtained from protein binding microarray analysis of a set of non-animal C2H2-ZFPs[10] or chimeric C2H2-ZFPs constructed from fusion of different ZFs[21].

Molecular dynamics simulations of Egr1 and CTCF

CTCF-DNA complex was obtained by combining two crystal structures (PDB accessions 5T0U, 5UND)[29], and EGR1-DNA complex was obtained from PDB accession 4X9J[30]. The simulations were carried out using AMBER16 package[31] with the *ff14SB* force field. The system was immersed in a rectangular box of TIP3P water

model[32] with neutralizing concentration of ions. Additionally, Na+/Cl–ions matching the ionic strength of 0.1 M were included. Each system was energy-minimized using 2500 steps of steepest descent followed by 2500 steps of conjugated gradient using a harmonic restriction on the solute with a value of 40 kcal/mol·Å. The heating process was carried from 0 to 300 K using Langevin dynamics and subsequently followed by equilibration for 500 ps (NPT). Finally, we carried out unrestrained MD for 1 μs under NPT conditions. The particle mesh Ewald (PME) method[33,34] was used to handle long-range electrostatic interactions and a cutoff of 12.0 Å was used for the simulations. SHAKE algorithm[35] was used to contain hydrogen bonds and to allow a longer integration step. All simulations were carried out with a time step of 2 fs using the *pmemd* module in AMBER16[36]. The *cpptraj* module was used for trajectory analyses[37]. The criterion for hydrogen bonding was set at ≤3.0 Å distance between electron donor atom and hydrogen of electron acceptor atom with 120-degree angle cutoff. Heavy atoms involved in stacking interaction were identified as described in MolBridge[38]. A perpendicular distance cutoff of 3.2 Å between the planar atoms was used. The simulated structure and contacts were visualized using PyMOL[39].

Identification of the *in vivo* binding preferences and DNA-engaging domains of human C2H2-ZFPs

We created a modified version of RCADE[23], called RCADE2, which uses our new compound recognition code (C-RC) to predict the binding preferences of different ZF arrays, evaluate their enrichment in *in vivo* binding sequences (compared to dinucleotide-shuffled sequences), and then optimize the motifs to maximize AUROC. RCADE2 is available at https://github.com/csglab/RCADE2, and includes features such as HTML output and integrated positional analysis of the identified motifs. Using RCADE2, we analyzed ChIP-seq and ChIP-exo data from two previous publications[18,19] as described above, including removal of peaks that overlapped EREs to ensure that our analyses are not confounded by sequence homology among repeat elements, and then using the summit position of the top 500 peaks with the largest scores (i.e. smallest p-values) for motif finding. For successful runs, we extracted the top-scoring optimized motif that had significant similarity with the seed motif (i.e. the motif predicted by C-RC) at $P < 0.001$, and marked the ZFs associated with that motif as the DNA-engaging ZFs. For each DNA-engaging ZF, we also calculated the information content of the 3-nucleotide motif that it encodes as $IC = \sum_{i,j} p_{i,j} log_2(4 \times p_{i,j})$, where $p_{i,j}$ denotes the probability of observing base j at position i of the triplet motif ($1 \leq i \leq 3$).

Analysis of cross-species conservation and population-level genetic variation at ZFs

We obtained missense single-nucleotide variants (SNVs) and their frequencies from the Genome Aggregation Database (gnomAD[40]) for genomic build GRCh37.24. Only single nucleotide variants that passed the high-quality filter in gnomAD (PASS filter) were included for further analysis. Multiple variants for the same genomic location were split using BCFTools. ZF domains were annotated based on the presence of Pfam pattern $X_2CX_{2,4}CX_{12}HX_{3,4,5}H$ in the protein sequence, and the SNVs from the canonical transcript were mapped to their corresponding amino acid residue in the ZF. To obtain per-residue genetic variation, allele frequencies of the three codon positions were summed. We used a similar approach to analyze per-base conservation at the ZF coding sequences, using phyloP[25] conservation scores that were obtained from the UCSC Genome Browser (phyloP46way track, hg19).

Acknowledgements

This work was supported by funds from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2018-05962 to H.S.N) and resource allocations from Compute Canada. H.S.N holds a Canada Research Chair funded by the Canadian Institutes of Health Research. B.D. is supported by a postdoctoral fellowship from the Scientific and Technological Research Council of Turkey (TUBITAK, grant no. 1059B191600423). A.H.C. is supported by a Globalink Graduate Fellowship from Mitacs.

Author contributions

B.D. and H.S.N. developed the computational methods and analyzed the data. S.K. performed the structural analyses and contributed to analysis of genetic variations. A.H.C. contributed to data analysis and algorithm implementation. H.S.N. conceived and directed the study. H.S.N wrote the manuscript with contribution from B.D. and S.K.

References

1. Najafabadi, H.S. *et al.* C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat Biotechnol* **33**, 555-62 (2015).
2. Wolfe, S.A., Nekludova, L. & Pabo, C.O. DNA recognition by Cys2His2 zinc finger proteins. *Annu Rev Biophys Biomol Struct* **29**, 183-212 (2000).
3. Pavletich, N.P. & Pabo, C.O. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science* **252**, 809-17 (1991).
4. Nardelli, J., Gibson, T.J., Vesque, C. & Charnay, P. Base sequence discrimination by zinc-finger DNA-binding domains. *Nature* **349**, 175-8 (1991).
5. Gupta, A. *et al.* An improved predictive recognition model for Cys(2)-His(2) zinc finger proteins. *Nucleic Acids Res* **42**, 4800-12 (2014).
6. Persikov, A.V. *et al.* A systematic survey of the Cys2His2 zinc finger DNA-binding landscape. *Nucleic Acids Res* **43**, 1965-84 (2015).
7. Persikov, A.V., Osada, R. & Singh, M. Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics* **25**, 22-9 (2009).
8. Myers, S. *et al.* Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* **327**, 876-9 (2010).
9. Jacobs, F.M. *et al.* An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* **516**, 242-5 (2014).
10. Najafabadi, H.S. *et al.* Non-base-contacting residues enable kaleidoscopic evolution of metazoan C2H2 zinc finger DNA binding. *Genome Biol* **18**, 167 (2017).
11. Garton, M. *et al.* A structural approach reveals how neighbouring C2H2 zinc fingers influence DNA binding specificity. *Nucleic Acids Res* **43**, 9147-57 (2015).
12. Barazandeh, M., Lambert, S.A., Albu, M. & Hughes, T.R. Comparison of ChIP-Seq Data and a Reference Motif Set for Human KRAB C2H2 Zinc Finger Proteins. *G3 (Bethesda)* **8**, 219-229 (2018).
13. Brayer, K.J. & Segal, D.J. Keep your fingers off my DNA: protein-protein interactions mediated by C2H2 zinc finger domains. *Cell Biochem Biophys* **50**, 111-31 (2008).
14. Brown, R.S. Zinc finger proteins: getting a grip on RNA. *Curr Opin Struct Biol* **15**, 94-8 (2005).
15. Baxley, R.M. *et al.* Deciphering the DNA code for the function of the Drosophila polydactyl zinc finger protein Suppressor of Hairy-wing. *Nucleic Acids Res* **45**, 4463-4478 (2017).
16. Han, B.Y., Foo, C.S., Wu, S. & Cyster, J.G. The C2H2-ZF transcription factor Zfp335 recognizes two consensus motifs using separate zinc finger arrays. *Genes Dev* **30**, 1509-14 (2016).
17. Nakahashi, H. *et al.* A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep* **3**, 1678-1689 (2013).
18. Schmitges, F.W. *et al.* Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Res* **26**, 1742-1752 (2016).
19. Imbeault, M., Helleboid, P.Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550-554 (2017).
20. Breiman, L. Random Forests. *Machine Learning* **45**, 5-32 (2001).
21. Lam, K.N., van Bakel, H., Cote, A.G., van der Ven, A. & Hughes, T.R. Sequence specificity is obtained from the majority of modular C2H2 zinc-finger arrays. *Nucleic Acids Res* **39**, 4680-90 (2011).

22. Lambert, S.A., Albu, M., Hughes, T.R. & Najafabadi, H.S. Motif comparison based on similarity of binding affinity profiles. *Bioinformatics* **32**, 3504-3506 (2016).
23. Najafabadi, H.S., Albu, M. & Hughes, T.R. Identification of C2H2-ZF binding preferences from ChIP-seq data using RCADE. *Bioinformatics* **31**, 2879-81 (2015).
24. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548 e24 (2019).
25. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110-21 (2010).
26. Dogan, B. & Najafabadi, H.S. Computational Methods for Analysis of the DNA-Binding Preferences of Cys2His2 Zinc-Finger Proteins. *Methods Mol Biol* **1867**, 15-28 (2018).
27. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**, D493-6 (2004).
28. Yousef, A. & Charkari, N.M. A novel method based on physicochemical properties of amino acids and one class classification algorithm for disease gene identification. *J Biomed Inform* **56**, 300-6 (2015).
29. Hashimoto, H. *et al.* Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA. *Mol Cell* **66**, 711-720 e3 (2017).
30. Zandarashvili, L., White, M.A., Esadze, A. & Iwahara, J. Structural impact of complete CpG methylation within target DNA on specific complex formation of the inducible transcription factor Egr-1. *FEBS Lett* **589**, 1748-53 (2015).
31. Case, D.A. *et al.* The Amber biomolecular simulation programs. *J Comput Chem* **26**, 1668-88 (2005).
32. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. & Klein, M.L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79**, 926-935 (1983).
33. Galindo-Murillo, R., Bergonzo, C. & Cheatham III, T.E. Molecular Modeling of Nucleic Acid Structure. *Current Protocols in Nucleic Acid Chemistry* **54**, 7.5.1-7.5.13 (2013).
34. Essmann, U. *et al.* A smooth particle mesh Ewald method. *The Journal of Chemical Physics* **103**, 8577-8593 (1995).
35. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H.J.C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **23**, 327-341 (1977).
36. Salomon-Ferrer, R., Götz, A.W., Poole, D., Le Grand, S. & Walker, R.C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *Journal of Chemical Theory and Computation* **9**, 3878-3888 (2013).
37. Roe, D.R. & Cheatham, T.E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation* **9**, 3084-3095 (2013).
38. Kumar, P., Kailasam, S., Chakraborty, S. & Bansal, M. MolBridge: a program for identifying nonbonded interactions in small molecules and biomolecular structures. *Journal of Applied Crystallography* **47**, 1772-1776 (2014).
39. Schrodinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. (2015).
40. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-91 (2016).