nature biotechnology

C2H2 zinc finger proteins greatly expand the human regulatory lexicon

Hamed S Najafabadi[1,7], Sanie Mnaimneh[1,7], Frank W Schmitges[1,7], Michael Garton[1], Kathy N Lam[2], Ally Yang[1], Mihai Albu[1], Matthew T Weirauch[3,4], Ernest Radovani[2], Philip M Kim[1,2,5], Jack Greenblatt[1,2], Brendan J Frey[1,4–6] & Timothy R Hughes[1,2,4]

Cys2-His2 zinc finger (C2H2-ZF) proteins represent the largest class of putative human transcription factors. However, for most C2H2-ZF proteins it is unknown whether they even bind DNA or, if they do, to which sequences. Here, by combining data from a modified bacterial one-hybrid system with protein-binding microarray and chromatin immunoprecipitation analyses, we show that natural C2H2-ZFs encoded in the human genome bind DNA both *in vitro* and *in vivo*, and we infer the DNA recognition code using DNA-binding data for thousands of natural C2H2-ZF domains. *In vivo* binding data are generally consistent with our recognition code and indicate that C2H2-ZF proteins recognize more motifs than all other human transcription factors combined. We provide direct evidence that most KRAB-containing C2H2-ZF proteins bind specific endogenous retroelements (EREs), ranging from currently active to ancient families. The majority of C2H2-ZF proteins, including KRAB proteins, also show widespread binding to regulatory regions, indicating that the human genome contains an extensive and largely unstudied adaptive C2H2-ZF regulatory network that targets a diverse range of genes and pathways.

C2H2-ZF proteins are best known as transcription factors in which tandem, modular C2H2-ZF domains each contact three or more bases, with the sequence preferences controlled to a large degree by four 'specificity residues' at amino acid positions −1, 2, 3 and 6 of the DNA-contacting alpha helix[1,2] (**Supplementary Fig. 1a**,**b**). Independent expansions of C2H2-ZF family members in many metazoans, including primates (**Supplementary Fig. 1c**), involve rearrangement of C2H2-ZF domains as well as diversification of the specificity residues and suggest a prominent role in adaptive evolution[3–7]. The ~700 human C2H2-ZF proteins represent the largest class of putative human transcription factors by far[8,9]. The biological functions of the large majority of human and mouse C2H2-ZF proteins are unknown, and those that are characterized have a wide variety of molecular and genetic roles[7], suggesting that the remainder will also be functionally diverse. Consistent with this notion, dozens of C2H2-ZF loci have been implicated in association studies with diseases including neonatal diabetes mellitus[10], mental retardation[11] and polydactyly[12].

The fact that ~50% of human C2H2-ZF proteins contain a KRAB domain, whose cofactor TRIM28 is involved in silencing both exogenous retroviruses and EREs[13–15], has led to the proposal that their diversity stems from a function in silencing EREs[14]. This hypothesis is supported by the correspondence between the number of retroviral LTRs and the number of C2H2-ZF domains across metazoan genomes[16]. This role is almost entirely hypothetical, however, as there are few concrete examples of C2H2-ZF proteins with this function, and most bind exogenous retroviruses[14,17–19].

C2H2-ZF proteins are also the only major class of human transcription factors in which a large majority (~80%) has no known DNA-binding motif[3,20]. A long-standing goal has been to predict the sequence preferences of C2H2-ZF proteins directly from their amino acid sequence[1,2], which could in turn reveal genomic binding sites and potential biological functions. However, current C2H2-ZF 'recognition codes' are based largely on mutation of specificity residues within well-defined templates and remain incomplete and error-prone[21,22], possibly because of incomplete mapping between the identity of specificity residues and their base preferences, the influence of amino acid positions outside of the four specificity residues[23] and the impact of neighboring C2H2-ZF domains[24]. In addition, human C2H2-ZF proteins contain an average of ~10 C2H2-ZF domains, leading to predicted binding sites of ~30 bases—many more than necessary to specify individual sites in the human genome. It is possible not all of the domains engage DNA simultaneously, further complicating prediction of genomic binding sites. Moreover, for the majority of C2H2-ZF proteins it is unknown whether they even bind DNA. There are many examples in which C2H2-ZF domains bind RNA, protein and other ligands[25–27]. To our knowledge, there are no known sequence rules that indicate what kind of ligand(s) each C2H2-ZF domain binds.

[1] Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada. [2] Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [3] Center for Autoimmune Genomics and Etiology (CAGE) and Divisions of Biomedical Informatics and Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA. [4] Canadian Institutes for Advanced Research, Toronto, Ontario, Canada. [5] Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [6] Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada. [7] These authors contributed equally to this work. Correspondence should be addressed to T.R.H. (t.hughes@utoronto.ca).

Received 12 July 2014; accepted 15 December 2014; published online 18 February 2015; doi:10.1038/nbt.3128

RESULTS

A recognition code derived from natural C2H2-ZF domains

To establish a more complete C2H2-ZF recognition code that encompasses features of natural C2H2-ZF proteins, we determined the DNA sequence preferences of 8,138 distinct natural C2H2-ZF domains, sampled from all eukaryotes. To do this, we modified the bacterial one-hybrid (B1H) system[28,29] by replacing the third C2H2-ZF domain ("F3") of the three-fingered protein Egr1 with a library of 47,072 natural C2H2-ZF domains. We then screened this library against 200 of the 256 possible NNN NGG GCG variants of the optimal Egr1 binding site, GCG TGG GCG[28,30,31], each in duplicate, providing a minimum of 40 independent measurements for each nucleotide at each of the four variable positions (additional assays beyond ~100 variants did not improve results in any of the analyses described below; thus, we did not assay all 256 variants owing to cost considerations). In each B1H assay, the bacteria that harbor this library were selected for expression of a positive reporter that is activated only when the C2H2-ZF construct binds the NNN NGG GCG sequence. This allowed us to identify specific protein-DNA by high-throughput sequencing of the library after selection, obtaining an 'S-score' for each C2H2-ZF and each DNA 4-mer that, in theory, represents relative binding energy (**Supplementary Fig. 2a**). We chose F3 as the finger to replace because the primary triplet (i.e., the first three bases) in the nine-base recognition site is the sole triplet contacted by only one finger; we also varied the fourth base which is contacted by both F3 and F2. We filtered the data to retain the 8,138 C2H2-ZF domains with the most clear and reproducible sequence preferences (**Fig. 1a** and **Supplementary Fig. 2b**). Motifs generated for each C2H2-ZF domain were very similar to those obtained from Protein Binding Microarray (PBM) data using selected individual constructs (**Fig. 1b**, **Supplementary Fig. 2c**,**d** and **Supplementary Table 1**). The motifs were highly diverse, and most contained considerable degeneracy. We observed an apparent continuum of sequence selectivity, which we also confirmed by PBM (**Supplementary Table 1** and **Supplementary Fig. 2e**–**h**). The 8,138 C2H2-ZF domains encompassed all 20 amino acids at three of the specificity residue positions and 19 at the fourth (**Supplementary Fig. 2i**). All of the data, including the motifs for each of the C2H2-ZF domains and the PBM validations below, are available on our project website at http://hugheslab.ccbr.utoronto.ca/supplementary-data/C2H2_B1H/ and in **Supplementary Data**.

These data both confirm and extend prior knowledge about the base preferences conferred by individual residues within the C2H2-ZF domain[1], motivating us to develop an improved recognition code model. The established DNA-contacting 'specificity residues' display the strongest relationship to base preferences, but all residues in the DNA-contacting alpha helix have a significant association with at least one base position, and most correlate with base preferences at multiple positions (**Fig. 1c**). Changes in the nonspecificity residues alone could cause a decrease in motif correlation (**Fig. 1d**), indicating that these associations are not accounted for by covariance with the known specificity residues. However, inclusion of these residues in a recognition code based on random forests (following ref. 22 and hereafter referred to as the B1H-RC, available online at http://zifrc.ccbr.utoronto.ca/) slightly reduced the ability of the code to correctly predict gold-standard motifs of C2H2-ZF proteins (the "GS set"; see Online Methods), suggesting that the contributions of nonspecificity residues may be complex and difficult to model. Alternative machine learning strategies, including regression, SVM and nearest-neighbor, performed even less well (**Supplementary Fig. 2j**).

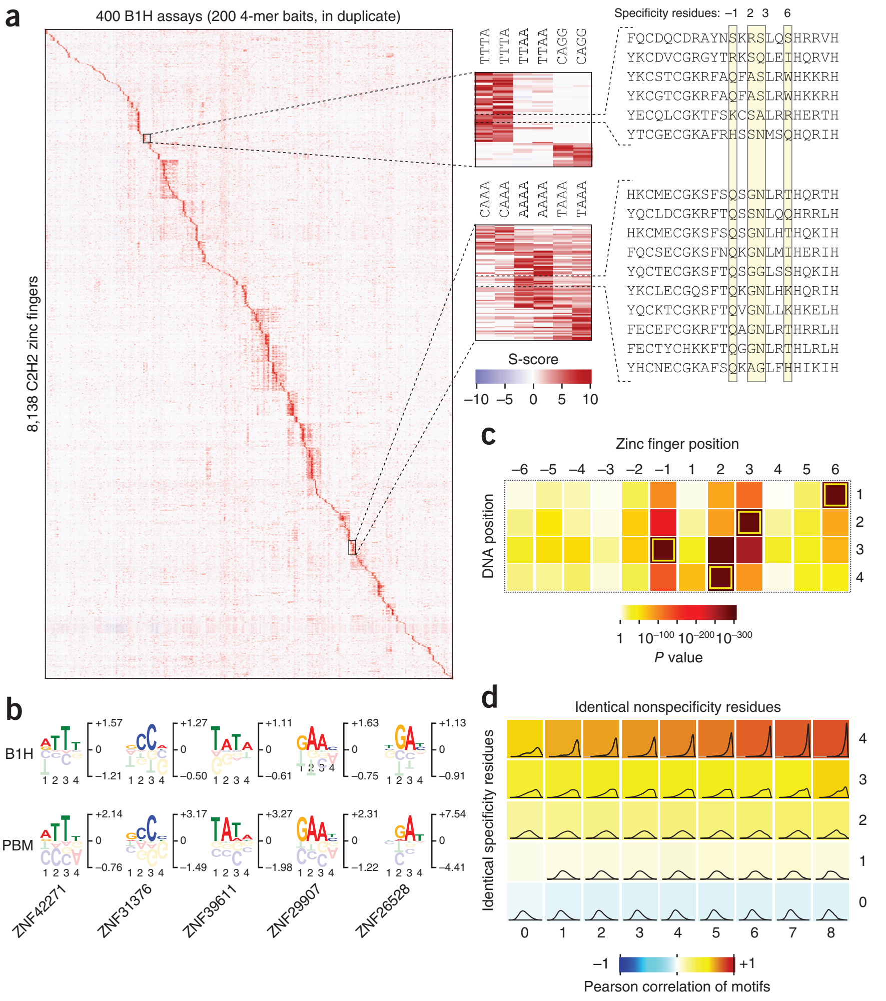


Figure 1 B1H data and PBM confirmations. (**a**) Matrix of inferred binding energy to target DNA (s-scores) for 8,138 individual C2H2-ZF domains. Following two-dimensional hierarchical clustering, large groups were rearranged to achieve a diagonal appearance. Pullouts show groups of C2H2-ZF domains with similar data profiles. (**b**) Comparison of PBM- and B1H-derived motifs for select C2H2-ZF domains. See **Supplementary Figure 2** and **Supplementary Table 1** for all PBM-B1H comparisons. (**c**) Relationship between different positions of C2H2-ZF alpha helix and the target DNA. *P* values are obtained using Pearson's chi-squared test, comparing the amino acid present at each zinc finger position to the most preferred nucleotide at each DNA position. Boxes highlight zinc finger-DNA contacts from the canonical C2H2-ZF-DNA recognition model[1]. (**d**) Distribution of motif similarities for pairs of C2H2-ZF domains as a function of the number of identical specificity residues and nonspecificity residues of the alpha helix. The color of each box represents the average of the distribution.

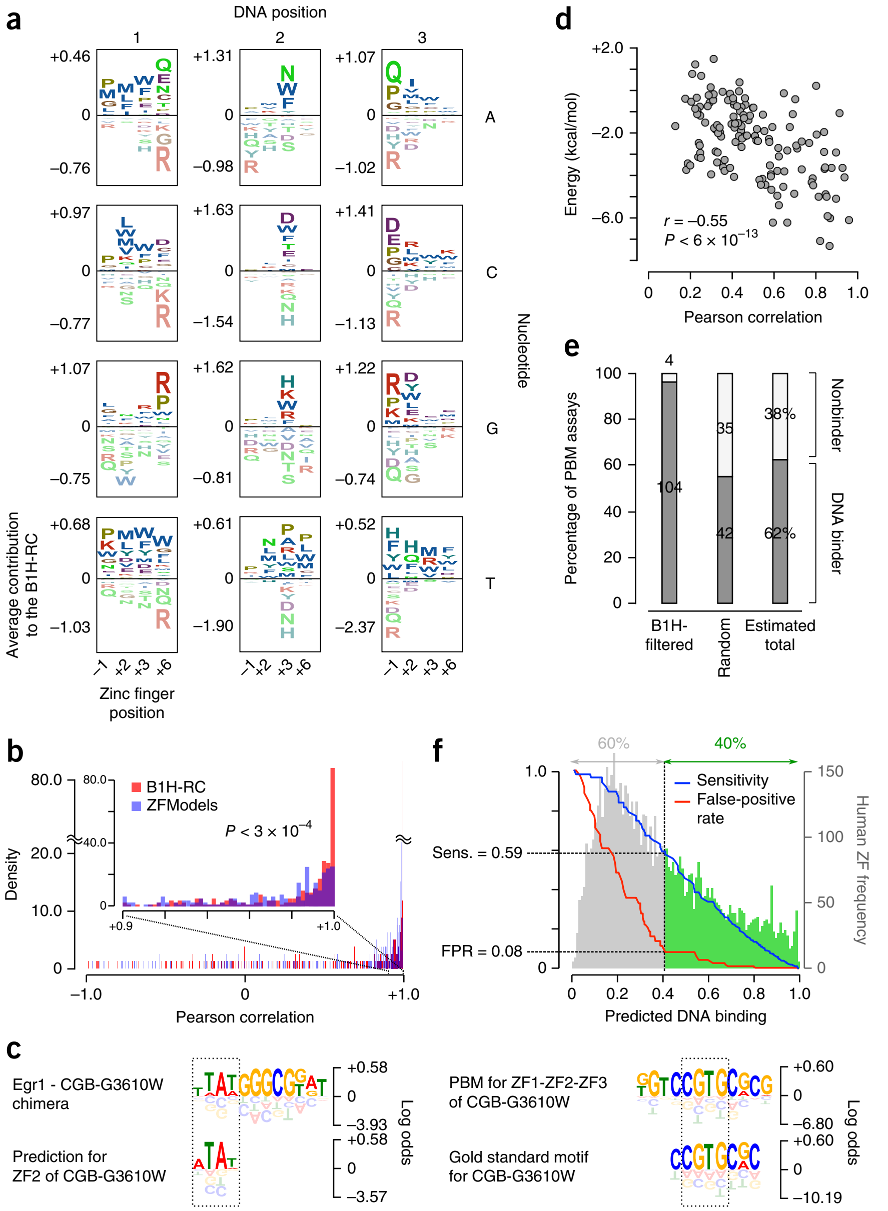


Figure 2 The B1H recognition code. (**a**) Average contribution of each amino acid at each C2H2-ZF specificity residue to the B1H-RC. (**b**) Comparison of B1H-RC to ZFModels[22]. Predicted motifs from the two models were aligned to the gold standard (GS) motifs and similarity of predictions across the four bases at each alignment column was measured. B1H-RC predictions are significantly more accurate than ZFModels predictions (two-sided paired *t*-test, $P < 3 \times 10^{-4}$). (**c**) An example showing the effect of neighboring zinc fingers on sequence preference. The PBM motif for ZF2 of *Cryptococcus gattii* protein CGB-G3610W fused to ZF1-ZF2 of Egr1 (top, left) is highly similar to the motif predicted by B1H-RC (bottom, left), whereas the PBM motif of the same zinc finger in its natural context is dramatically different (top, right). Previously published motif for CGB-G3610W is also shown (bottom, right), with the boxes representing the part of the motifs that corresponds to ZF2. (**d**) Binding energy calculated from structural modeling correlates with similarity of B1H-RC motifs to GS motifs. The *y* axis shows the energy calculated for binding of each protein to its most preferred sequence based on the B1H-RC motif, and the *x* axis represents the similarity of the predicted B1H-RC motifs to the GS motifs. (**e**) Summary of PBM results for 108 Egr1-variants that were included in the B1H training set and 77 randomly selected Egr1-variants that were not included. Given that the B1H training set constitutes 17% of the initial B1H pool, we estimate that 62% of all C2H2-ZFs would bind to DNA in the Egr1 context (last column). (**f**) Distribution of scores for human C2H2-ZF domains based on a model trained to identify DNA-binding activity. Sensitivity and false-positive rate (FPR) are determined from PBM experiments.

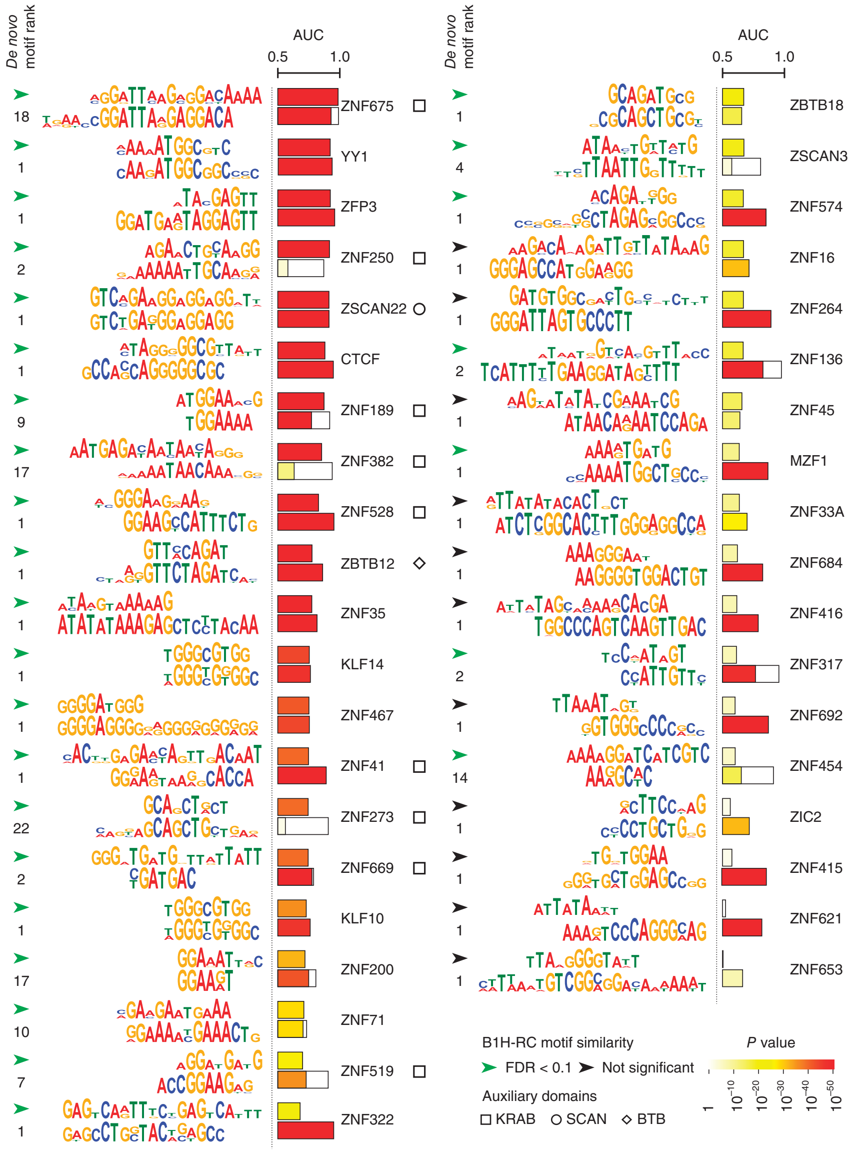
The B1H-RC model uses an ensemble of decision trees to predict the relative free energy of binding to each nucleotide at each of the three DNA contact positions. These decision trees can potentially encode complex relationships among the specificity residues. **Figure 2a** shows a simplified view of the B1H-RC by depicting the relative contribution of specificity residue identities at each position of the DNA motif, indicating that although the B1H-RC supports current models of a one-to-one zinc finger residue to DNA residue interaction, additional zinc finger positions likely affect nucleotide preference at each binding site position, such as the effect of position 2 of the alpha helix on nucleotide preference at position 3 of the DNA triplet.

The B1H-RC outperformed a recently described predictor, ZFModels[22] on the same GS set (**Fig. 2b**); however, both methods had difficulty with the same C2H2-ZF domains (**Supplementary Fig. 3a**). Because the GS set was derived from proteins containing C2H2-ZF arrays, we asked whether these failures were due to a "neighbor effect," in which the specificity of a finger may be influenced by adjacent fingers. Indeed, in three different cases, we found that when a C2H2-ZF domain was placed in the Egr1-F3 position, it bound an entirely different motif than it did when in its native protein (**Fig. 2c** and **Supplementary Fig. 3b**–**d**). We then sought to determine whether the neighbor effect could be rationalized structurally, using molecular dynamics simulations (see Methods). These simulations suggested that some zinc fingers, when placed next to their natural adjacent fingers, do not bind with high affinity to the B1H-RC predicted targets. This low binding affinity in the natural context explains the low similarity between the corresponding known and predicted motifs (**Fig. 2d**), suggesting that changes in sequence preference are likely due to structural changes induced by neighboring C2H2-ZF domains.

Most human C2H2-ZF proteins bind DNA

To examine whether sequence properties of the C2H2-ZF domains predict whether they bind DNA in a sequence-specific manner, we trained a random forest model to discriminate between C2H2-ZFs that do and do not have sequence specificity (see Methods, **Supplementary Fig. 4a**,**b**). To validate this model, we used PBMs to examine the sequence preferences of 77 randomly selected Egr1-F3 variants that were in the initial B1H pool, but not among the B1H-filtered set, in addition to 108 PBMs that overlapped the B1H-filtered set (**Fig. 2e** and **Supplementary Table 1**). The PBM results were largely consistent with the model predictions (**Supplementary Fig. 4c**),

Figure 3 Comparison of B1H-RC to ChIP-seq results for 39 human C2H2-ZF proteins. For each protein, the predicted B1H-RC motif is shown on the top, and the *de novo* ChIP-seq motif with the smallest similarity *P* value is shown at the bottom. The numbers represent the rank of the shown motif in the list of *de novo* discovered motifs. All *de novo* motifs that are shown are centrally enriched in the ChIP-seq peaks. Green arrows indicate B1H-RC motifs that are significantly similar to the *de novo* ChIP-seq motif (false-discovery rate (FDR) < 0.1). The AUC of each motif for distinguishing the top 500 peaks from dinucleotide-shuffled sequences is shown on the right (color gradient represents the AUC *P* value, Mann-Whitney *U* test). In cases where the best-matching ChIP-seq motif is not the top-ranking *de novo* motif, the AUC of the top-ranking motif is also indicated.



with 42 of the 77 randomly selected variants yielding motifs by PBM (55%). In agreement with this fraction, the model indicates that 62% of all human C2H2-ZF domains are likely DNA binding (**Fig. 2f**), as 40% of all human C2H2-ZF domains are above a threshold with 59% recall and only an 8% false positive rate. This number should be considered as a lower estimate of the actual fraction of DNA-binding C2H2-ZF domains in human, as many C2H2-ZF domains may only bind in their natural context and thus produce false-negative results in the Egr1-F3 context that was examined here. These data indicate that the majority of human C2H2-ZF domains are proficient for sequence-specific DNA binding; similar figures are obtained for most other genomes (**Supplementary Fig. 4d**). The model also gives low scores to the human C2H2-ZF domains with no adjacent C2H2-ZFs (**Supplementary Fig. 4e**), consistent with the observation that DNA binding usually requires two or more tandem C2H2-ZF domains[1].

To ask whether long human C2H2-ZF proteins generally bind specific DNA sequences *in vivo*, and whether they recognize sequences resembling those predicted by the B1H-RC, we analyzed 71 randomly selected GFP-tagged human C2H2-ZF proteins, containing between 3 and 28 C2H2-ZF domains, using ChIP-seq in HEK293 cells. The peaks from 39 proteins (54%) yielded a strongly enriched motif located near the peak center on average (**Supplementary Fig. 5a**). Among these 39, 35 (90%) also displayed significant enrichment of motifs predicted by the B1H-RC for a contiguous subset of the C2H2-ZF domains (Bonferroni corrected $P < 0.001$), often with an area under the curve (AUC) value similar to that of motifs derived directly from the ChIP-seq data (**Fig. 3**). In almost all cases, the B1H-RC predicted motif is preferentially located in the center of the peaks, consistent with direct DNA binding (**Supplementary Fig. 5a**). Moreover, for 28 of these 35 proteins (80%), there is a substantial similarity between a B1H-RC predicted motif, and at least one motif derived directly from the ChIP-seq data (**Fig. 3**). Thus, although some of the motifs may be erroneous (e.g., ChIP-seq motifs corresponding to other transcription factors expressed in HEK293 cells), there is evidence that a large majority represent the bona fide DNA binding preference of the C2H2-ZF proteins (**Fig. 3**).

The motifs obtained from the *in vivo* data can be assigned to individual C2H2-ZF domains in the assayed protein; on average, these comprise 45% of the C2H2-ZF domains in each protein, often with clear similarity between the *in vivo* motifs and those obtained directly from B1H assays (**Supplementary Fig. 5b**). These domains tend to score significantly higher in the random forest model trained to identify DNA-binding activity ($P < 0.003$, **Supplementary Fig. 4f**). These results strongly indicate that most human C2H2-ZF proteins utilize at least some of their C2H2-ZF domains for direct DNA binding *in vivo*, and that there is no fundamental discrepancy between the B1H-RC and *in vivo* binding sites. Given that 38 of the 71 proteins were assayed only once, and that HEK293 cells may be lacking cofactors or conditions required for particular proteins or individual C2H2-ZF domains to bind DNA, our figures likely underestimate the degree of DNA binding for C2H2-ZF proteins in general.

C2H2-ZF proteins have diverse functions

To probe the biological functions of C2H2-ZF proteins, we compared their motif-containing genomic binding sites to gene annotations,

Table 1 Functional characteristics of 39 human C2H2-ZF proteins and their binding sites

Name	Auxiliary domains	Enriched repeats	DHS fraction[a]	H3K9me3 enrichment[b]	Conserved binding site[c]	Top enriched function/phenotype[d]
ZNF189	KRAB	L2a/b/c	**0.62**	**Yes**	**Yes**	Muscle cell differentiation (4.2)
ZNF317	KRAB	L1M4/4b/4c/Ec/Ef/Eg	**0.52**	**Yes**	No	–
ZNF41	KRAB	L1MEd/g, MER89	0.34	**Yes**	No	–
ZNF675	KRAB	THE1-int, MST-int	0.12	**Yes**	No	–
ZNF136	KRAB	L1M1/2/4b/4c/Ef	0.11	**Yes**	No	–
ZNF250	KRAB	L1M2/3/4/A3/B2/B3	0.07	**Yes**	No	–
ZNF382	KRAB	L1HS, L1PA2/3	0.06	**Yes**	No	–
ZNF45	KRAB	LTR18A, L1MEb/c/f	0.29	**Yes**	**Yes**	–
ZNF264	KRAB	MLT1A/A0/B	0.49	**Yes**	No	Hypoperistalsis (3.1)
ZNF669	KRAB	HERV9/17-int	0.43	**Yes**	No	–
ZNF273	KRAB	MER52A/C/D	0.29	**Yes**	No	–
ZNF528	KRAB	L1PB1/2/a	0.24	**Yes**	**Yes**	–
ZNF519	KRAB	MLT1A/B	0.38	No	No	–
ZNF416	KRAB	MLT2B1/2/4	0.48	No	No	Cell diff. in kidney development (2.7)
ZNF454	KRAB	L1M5/C	0.35	No	No	Limb morphogenesis (2.8)
ZNF33A	KRAB	SVA-E/D/F	0.02	No	No	–
ZNF621	KRAB	–	**0.88**	**Yes**	**Yes**	Regulation of synapse structure (2.5)
ZNF684	KRAB	–	0.31	**Yes**	No	–
ZSCAN22	SCAN	–	**0.64**	**Yes**	No	Hyaluronan metabolic process (5.5)
MZF1	SCAN	–	**0.71**	No	No	–
ZSCAN31	SCAN	–	**0.58**	No	No	–
ZBTB18	BTB	–	**0.89**	No	No	–
ZBTB12	BTB	–	**0.83**	**Yes**	***Yes***	Protein kinase C activity (4.5)
ZNF35	–	L1M5/B3/C4a/E4a	0.27	**Yes**	**Yes**	Midbrain development (3.2)
ZFP3	–	L2b/c	0.44	**Yes**	**Yes**	Osteoblast development (3.9)
ZNF200	–	–	**0.96**	No	***Yes***	Small nuclear RNP complex (17.3)
ZNF574	–	–	**0.96**	No	***Yes***	Structural constituent of ribosome (3.5)
CTCF	–	–	**0.92**	No	***Yes***	–
ZNF467	–	–	**0.92**	No	**Yes**	–
KLF10	–	–	**0.91**	No	***Yes***	Aciduria (3.5)
YY1	–	–	**0.89**	No	***Yes***	–
KLF14	–	–	**0.70**	No	**Yes**	–
ZIC2	–	–	**0.60**	No	***Yes***	–
ZNF71	–	–	**0.52**	No	**Yes**	Regulation of synaptic transmission (2.9)
ZNF322	–	–	**0.74**	**Yes**	No	–
ZNF415	–	–	**0.73**	**Yes**	No	–
ZNF692	–	–	**0.88**	No	No	–
ZNF653	–	–	**0.56**	No	No	–
ZNF16	–	–	0.43	No	**Yes**	Abnormal dendritic cell number (4.7)

[a]Significant enrichments are boldface and underscored (FDR < 0.001). [b]Asymmetric enrichments are underscored. [c]Underscore denotes higher conservation in EREs; italic represents higher conservation in nonrepeat genome. [d]FDR < 0.01. Only terms with >2-fold enrichment of binding sites are shown. GO and REACTOME terms are given priority over phenotypes. The fold enrichment of each term is shown in parentheses.

chromatin state and genomic conservation (**Table 1**). GREAT analysis[32], which identifies enrichment of peaks near genes sharing functional properties, indicates that C2H2-ZF proteins play diverse roles in human physiology and suggests that many are multifunctional. ZSCAN22 peaks, for example, are more than fivefold enriched in hyaluronan metabolism genes, >4-fold enriched in genes involved in Charcot-Marie-Tooth disease and >3-fold enriched for semaphorin interactions as well as in spleen development (GREAT binomial and hypergeometric FDR < 0.01). Genes in these categories have no overlap with each other, yet their expression shows strong anti-correlation with ZSCAN22 across different tissues/cell lines (**Supplementary Fig. 6a**). In fact, expression of the majority of C2H2-ZF proteins examined is significantly correlated or anti-correlated with that of their ChIP-seq targets (FDR < 0.01, **Supplementary Fig. 6b**), indicating a regulatory link between them. C2H2-ZF protein binding sites are also enriched in many cell type-specific DNaseI hypersensitive regions (**Supplementary Fig. 6c**), suggesting a role in regulation of cell type–specific transcriptional programs.

The ChIP-seq results also confirm the hypothesis that many KRAB-C2H2-ZF proteins bind endogenous repeat elements (EREs)[14,16]; 16 of the 18 KRAB-C2H2-ZF proteins bound a specific ERE or class of related EREs to a large extent (**Fig. 4a**), but only two of the 21 non-KRAB proteins (ZFP3 and ZNF35) bound EREs by the same criteria. The bound ERE families encompassed endogenous retroviruses (ERVs), the hominid-specific SVA elements[33] and long interspersed nuclear element (LINE)s of diverse ages (**Fig. 4a**). These EREs often show a distinct asymmetric H3K9me3 signal relative to the C2H2-ZF protein binding site, compatible with the predicted orientation of the KRAB domain (**Fig. 4b** and **Table 1**), consistent with local trimethylation of H3K9 through interaction of the KRAB domain with TRIM28 and, subsequently, the histone methyltransferase SETDB1 (ref. 34). The estimated ages of C2H2-ZF arrays are often similar to those of the EREs to which they bind (**Fig. 4c**), suggesting that the proteins adapted in response to and/or were fixed by the arrival of the EREs. The only major exceptions are the two proteins that bind the newest EREs: ZNF33A bound primarily to hominoid-specific SVA elements, and ZNF382 bound to the currently active LINE L1HS-like elements. The fact that these proteins are largely conserved across all mammals suggests that they may have other functions and were recently co-opted to bind these new EREs. This notion is supported by the previously described function of ZNF382 as a tumor suppressor[35] and by a strong signature of recent positive selection at the ZNF33A locus in the human lineage (**Supplementary Fig. 7**).

Most of the EREs bound, and the proteins that bind them, date from early in the mammalian radiation ~100 Mya (**Fig. 4c**). As these EREs have presumably not been active as retrotransposons for tens

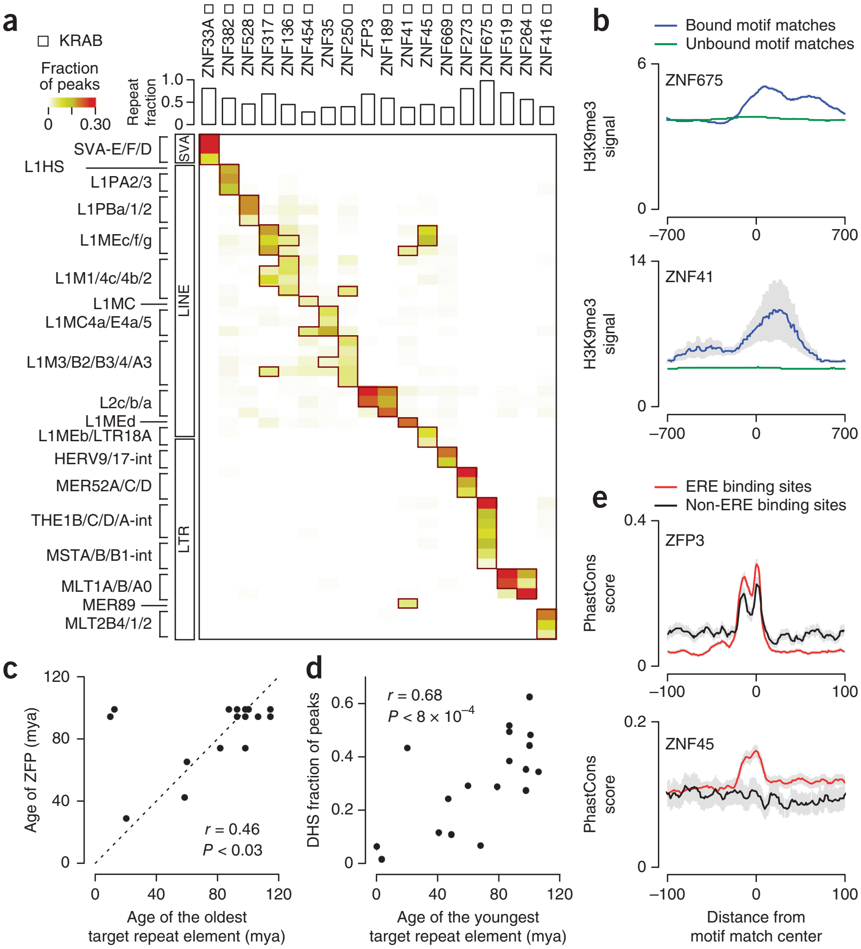


Figure 4 The majority of KRAB family proteins bind to EREs. (**a**) Fraction of the top 500 ChIP-seq peaks of each protein that overlap different EREs. Significant enrichments are highlighted with black boxes (Fisher's exact test, FDR < 0.001). The total fraction of repeat-overlapping peaks is shown at the top. (**b**) Histone H3-K9 trimethylation signal around the binding sites of KRAB C2H2-ZF proteins. Blue represents the average H3K9me3 signal around motif matches within ChIP-seq peaks, and green represents the signal around motif matches that do not overlap peaks. The gray area corresponds to the s.d. of the estimated average. KRAB domain is located at the N terminus, which is oriented downstream of the binding site (i.e., positive values on the *x* axis; orientation determined based on the B1H-RC motifs). (**c**) The estimated age of each protein corresponds to the age of the repeats that it binds to. The age of each protein was estimated based on the time of divergence of the organisms that contain a homolog (>70% identity over the domains inferred to be involved in DNA binding). The age of the EREs was estimated based on their average divergence from the consensus sequence. ZFP, zinc finger protein. (**d**) Fraction of top 500 ChIP-seq peaks of each protein that overlap DHS sites correlates with the age of the youngest repeat element that the protein binds to. (**e**) Conservation of C2H2-ZF binding sites within and outside EREs. Red represents motif matches within ChIP-seq peaks that overlap EREs, and black represents motif matches within ChIP-seq peaks outside EREs. The gray area shows the s.d. of the estimated average.

of millions of years, the binding of KRAB proteins may serve a role other than silencing. One possibility is suppression of nonallelic homologous recombination (a function previously proposed for the H3K9me3 marks found over tandem C2H2-ZF arrays themselves[36]), thereby mediating genomic stability in regions rich with EREs, which are known to otherwise be prone to segmental duplications[37]. It is also conceivable that the C2H2-ZF proteins binding EREs have acquired additional regulatory functions. Consistent with this notion, the proportion of peaks that lie in DNaseI hypersensitive sites (DHSs), a proxy for enhancers, increases with the age of the youngest EREs bound by each protein (the youngest ERE sets a minimum age for the protein functioning primarily to repress the ERE) (**Fig. 4d**). One example is the KRAB C2H2-ZF protein ZNF189, which binds the ancient LINE L2 elements, as well as DHS. ZNF189 binding sites are also enriched for several neural and muscle-related functions. In particular, ZNF189 sites are >11-fold enriched in genes whose absence causes the thin myocardium phenotype in mammals (GREAT binomial $P < 10^{-200}$, hypergeometric $P < 2 \times 10^{-4}$). Intriguingly, ZNF189 has highest expression in heart, and the major ZNF189 isoform in heart lacks the KRAB domain (**Supplementary Fig. 8a**). Thus, ZNF189 may maintain a nonrepressive function in heart cells. Alternative splicing of the KRAB domain appears to be widespread and more prevalent than other exons of C2H2-ZF protein-coding transcripts (**Supplementary Fig. 8b**), suggesting a mechanism for diversification of C2H2-ZF functions and incorporation in tissue-specific regulatory programs.

The EREs themselves may also carry regulatory functions, as they are known to seed the genome with sites for transcription factors to bind[38,39]. For example, L2b repeats are enriched in several neural and muscle-related terms, such as genes involved in nervous system development (GREAT binomial $P < 2 \times 10^{-53}$, hypergeometric $P < 2 \times 10^{-14}$) and genes involved in cardiovascular system development (binomial $P < 3 \times 10^{-33}$, hypergeometric $P < 3 \times 10^{-8}$), suggesting that LINE L2 elements might have played a role in seeding the ZNF189 binding sites near these genes. Furthermore, we found several EREs that harbor conserved binding sites for C2H2-ZF proteins, including examples where the binding sites are more conserved within EREs compared to non-ERE binding sites (**Fig. 4e** and **Table 1**).

DISCUSSION

Our analyses indicate that C2H2-ZF proteins are primarily DNA-binding proteins. The B1H assay, despite employing only a single context (the F3 position of Egr1) and a high-throughput assay, yielded sequence specificity for nearly 20% of all eukaryotic C2H2-ZF domains tested and motifs encompassing all possible base preferences. PBM assays confirm the motifs obtained by B1H and also indicate that in fact a majority of C2H2-ZF domains display sequence specificity in the Egr1-F3 position, indicating that our B1H analysis has a considerable false-negative rate. ChIP-seq also indicates that the majority of human C2H2-ZF proteins bind DNA *in vivo*. Due to the fact that neighboring C2H2-ZF domains can influence the DNA binding characteristics of a C2H2-ZF domain, the B1H-RC predicted motifs for natural, multi-C2H2-domain proteins often contain inaccuracies. Nonetheless, B1H-RC motifs typically resemble those obtained directly from ChIP-seq data, providing a means to pinpoint motifs that reflect direct binding, as well as residues of the protein that interact with DNA.

The prevalence of C2H2-ZFs, frequent conservation of their binding sites and association of their binding sites with genes of diverse functions suggest that they are as integral to metazoan genome function and evolution as the more classical "evo-devo" and other conserved pathways that currently dominate the literature on gene regulation[40]. C2H2-ZF proteins represent a largely unstudied regulatory system[41], and, to our knowledge, the extent to which they are involved in adaptive gene regulatory programs has not been systematically explored before. We propose that this system operates in parallel with well-characterized developmental processes and has different operating principles. Most developmental transcription factors are highly conserved, have low-information-content motifs and often target loci containing multiple binding sites, enabling *cis*-regulatory shuffling over evolutionary time without dramatically impacting regulatory programs[42–45]. In contrast, in the C2H2-ZF-based system, the motifs are often large and information-rich, and the *trans*-regulators also frequently shuffle by rearranging and modifying their domains[4,5,7]. We note that the auxiliary domains found in tetrapod C2H2-ZF proteins (primarily KRAB, SCAN and BTB) are very rarely found in other proteins, which could minimize interference with conserved developmental mechanisms.

We estimate that the full collection of sequence preferences of full-length human C2H2-ZF proteins is likely to encompass 450 distinct motifs (average 0.6 motif per protein) (**Supplementary Fig. 9a–d**). In contrast, the combination of all other transcription factor classes is expected to contain just over 350 independent motifs (average 0.4 motif per protein) (**Supplementary Fig. 9e**,**f**). *In vitro* selection data also support the diversity of C2H2-ZF motifs[20]. Virtually any DNA fragment the size of an ERE would therefore contain potential binding sites for multiple human KRAB-C2H2-ZF proteins, given their large numbers; in fact, it is surprising that non-KRAB C2H2-ZF proteins generally do not bind to specific EREs, as the motifs for KRAB proteins are not enriched in EREs relative to non-KRAB motifs (data not shown). One possible explanation is that recruitment of TRIM28 by the KRAB domain could facilitate early access of the KRAB C2H2-ZF proteins to newly replicated DNA, via the association of TRIM28 with SMARCAD1, which localizes to DNA replication sites via binding PCNA[36]. In addition, although motifs for KRAB proteins are enriched in the EREs that they bind, they do also bind to other sites throughout the genome, suggesting that cells must have mechanisms to overcome KRAB-based silencing at native genes, such as alternative splicing that removes the KRAB domain. We anticipate that our results will facilitate the future discovery of new regulatory mechanisms and the architecture of the poorly studied adaptive C2H2-ZF regulatory network.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes and data availability. GEO: GSE52523. Additional data tables including library sequences, plasmid and insert sequences, and results of analysis of B1H, PBM and ChIP-seq data are available at http://hugheslab.ccbr.utoronto.ca/supplementary-data/C2H2_B1H/ and in **Supplementary Data**.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to S. Wolfe for providing B1H reagents and protocols, F. Aidoo, H. Zheng, H. Tang, P. Young, T. Kanagalingam, D. Torti and the Donnelly Sequencing Centre for technical support, and E. Chan, H. van Bakel and X. Chen for computational support and analyses. This work was supported by grants from the Canadian Institutes of Health Research (MOP-77721 and MOP-111007 to T.R.H., MOP-272138 to T.R.H., J.G. and Andrew Emili), and funding from the Canadian Institutes for Advanced Research to T.R.H., B.J.F. and M.T.W. H.S.N. was supported by a Canadian Institutes of Health Research Banting Fellowship, F.W.S. by a European Molecular Biology Organization postdoctoral fellowship, and K.N.L. by a Natural Science and Engineering Research Council CGS-M.

AUTHOR CONTRIBUTIONS

H.S.N., S.M., F.W.S. and T.R.H. conceived and designed the experiments. S.M. performed the B1H experiments, with contributions from K.N.L. F.W.S. performed the ChIP-seq experiments, with contributions from E.R. S.M. and A.Y. performed the PBM experiments. H.S.N. analyzed the data and developed the computational models. M.G. performed the structural modeling. M.A., M.T.W. and T.R.H. contributed to data analysis. J.G. contributed reagents and materials. P.M.K., J.G. and B.J.F. provided critical advice and commentary on data analysis. H.S.N. prepared the figures. T.R.H. conceived the study and supervised the project, and H.S.N. and T.R.H. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Wolfe, S.A., Nekludova, L. & Pabo, C.O. DNA recognition by Cys2His2 zinc finger proteins. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 183–212 (2000).
2. Klug, A. The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annu. Rev. Biochem.* **79**, 213–231 (2010).
3. Emerson, R.O. & Thomas, J.H. Adaptive evolution in zinc finger transcription factors. *PLoS Genet.* **5**, e1000325 (2009).
4. Nowick, K. *et al.* Gain, loss and divergence in primate zinc-finger genes: a rich resource for evolution of gene regulatory differences between species. *PLoS ONE* **6**, e21553 (2011).
5. Hamilton, A.T. *et al.* Evolutionary expansion and divergence in the ZNF91 subfamily of primate-specific zinc finger genes. *Genome Res.* **16**, 584–594 (2006).
6. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
7. Stubbs, L., Sun, Y. & Caetano-Anolles, D. Function and evolution of C2H2 zinc finger arrays. *Subcell. Biochem.* **52**, 75–94 (2011).
8. Weirauch, M.T. & Hughes, T.R. A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. *Subcell. Biochem.* **52**, 25–73 (2011).
9. Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. & Luscombe, N.M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10**, 252–263 (2009).
10. Mackay, D.J. *et al.* Hypomethylation of multiple imprinted loci in individuals with transient neonatal diabetes is associated with mutations in ZFP57. *Nat. Genet.* **40**, 949–951 (2008).
11. Kleefstra, T. *et al.* Zinc finger 81 (ZNF81) mutations associated with X-linked mental retardation. *J. Med. Genet.* **41**, 394–399 (2004).
12. Kalsoom, U.E. *et al.* Whole exome sequencing identified a novel zinc-finger gene ZNF141 associated with autosomal recessive postaxial polydactyly type A. *J. Med. Genet.* **50**, 47–53 (2013).
13. Rowe, H.M. *et al.* KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature* **463**, 237–240 (2010).
14. Rowe, H.M. & Trono, D. Dynamic control of endogenous retroviruses during development. *Virology* **411**, 273–287 (2011).
15. Matsui, T. *et al.* Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. *Nature* **464**, 927–931 (2010).
16. Thomas, J.H. & Schneider, S. Coevolution of retroelements and tandem zinc finger genes. *Genome Res.* **21**, 1800–1812 (2011).
17. Carlson, K.A. *et al.* Molecular characterization of a putative antiretroviral transcriptional factor, OTK18. *J. Immunol.* **172**, 381–391 (2004).
18. Wolf, D. & Goff, S.P. Embryonic stem cells use ZFP809 to silence retroviral DNAs. *Nature* **458**, 1201–1204 (2009).
19. Jacobs, F.M. *et al.* An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* **516**, 242–245 (2014).
20. Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).
21. Persikov, A.V. & Singh, M. *De novo* prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. *Nucleic Acids Res.* **42**, 97–108 (2014).
22. Gupta, A. *et al.* An improved predictive recognition model for Cys2-His2 zinc finger proteins. *Nucleic Acids Res.* **42**, 4800–4812 (2014).
23. Wolfe, S.A., Grant, R.A., Elrod-Erickson, M. & Pabo, C.O. Beyond the "recognition code": structures of two Cys2His2 zinc finger/TATA box complexes. *Structure* **9**, 717–723 (2001).
24. Isalan, M., Choo, Y. & Klug, A. Synergy between adjacent zinc fingers in sequence-specific DNA recognition. *Proc. Natl. Acad. Sci. USA* **94**, 5617–5621 (1997).

25. 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**, 111–131 (2008).
26. Brayer, K.J., Kulshreshtha, S. & Segal, D.J. The protein-binding potential of C2H2 zinc finger domains. *Cell Biochem. Biophys.* **51**, 9–19 (2008).
27. Iuchi, S. Three classes of C2H2 zinc finger proteins. *Cell. Mol. Life Sci.* **58**, 625–635 (2001).
28. Meng, X., Brodsky, M.H. & Wolfe, S.A. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nat. Biotechnol.* **23**, 988–994 (2005).
29. Noyes, M.B. *et al.* A systematic characterization of factors that regulate *Drosophila* segmentation via a bacterial one-hybrid system. *Nucleic Acids Res.* **36**, 2547–2560 (2008).
30. Swirnoff, A.H. & Milbrandt, J. DNA-binding specificity of NGFI-A and related zinc finger transcription factors. *Mol. Cell. Biol.* **15**, 2275–2287 (1995).
31. Berger, M.F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* **24**, 1429–1435 (2006).
32. McLean, C.Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
33. Wang, H. *et al.* SVA elements: a hominid-specific retroposon family. *J. Mol. Biol.* **354**, 994–1007 (2005).
34. Ayyanathan, K. *et al.* Regulated recruitment of HP1 to a euchromatic gene induces mitotically heritable, epigenetic gene silencing: a mammalian cell culture model of gene variegation. *Genes Dev.* **17**, 1855–1869 (2003).
35. Cheng, Y. *et al.* KRAB zinc finger protein ZNF382 is a proapoptotic tumor suppressor that represses multiple oncogenes and is commonly silenced in multiple carcinomas. *Cancer Res.* **70**, 6516–6526 (2010).
36. *Drosophia* 12 Genes Consortium. *et al.* Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**, 203–218 (2007).
37. She, X., Cheng, Z., Zollner, S., Church, D.M. & Eichler, E.E. Mouse segmental duplication and copy number variation. *Nat. Genet.* **40**, 909–914 (2008).
38. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **22**, 1798–1812 (2012).
39. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42**, 631–634 (2010).
40. Carroll, S.B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25–36 (2008).
41. Weirauch, M. *et al.* Determination and inference of Eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
42. Dermitzakis, E.T. & Clark, A.G. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* **19**, 1114–1121 (2002).
43. Sanges, R. *et al.* Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage. *Genome Biol.* **7**, R56 (2006).
44. Odom, D.T. *et al.* Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat. Genet.* **39**, 730–732 (2007).
45. Wunderlich, Z. & Mirny, L.A. Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet.* **25**, 434–440 (2009).

ONLINE METHODS

B1H prey library and bait plasmids. We obtained individual natural C2H2-ZF domain sequences from Pfam (v. 22.0) and from predicted protein sequences in Ensembl (v. 55) scanned with PF00096, obtaining a total of 47,072 unique C2H2-ZF domains. We back-translated the sequences for expression in *E. coli*, added flanking vector homology regions to each sequence and ordered the DNA as an "OLS" pool from Agilent Technologies. We amplified the pool by PCR and used oligo-directed mutagenesis to create the library of Zif268 F3 variants in plasmid pTH5311, following[46]. In this library, each F3 variant is linked to F1-F2 of Zif268 (Egr1) using a linker with constant length and sequence. To obtain the prey library, we subcloned inserts (F1-F2-F3) from the phage library into pTH4792, a version of the B1H vector 1352-omega (UV2)-Paired[47] modified for subcloning. We constructed bait plasmids by inserting double-stranded oligonucleotides containing variants of the Zif268/Egr1 binding site 5′-GCGGCCGCTNNNNGGGCGGCACTCCGGAGGCGCGCCGAATTC-3′ into the NotI and EcoRI sites of pH3U3 (ref. 28).

B1H screens. We performed B1H screens essentially as described[48], with the modification that we first transformed the bacterial selection strain (US0*ΔpyrFΔhisB*) with the bait plasmid, then made electrocompetent cells, then transformed these with the prey library of C2H2-ZF variants. We plated ~10^7 transformants on three medium plates (150 mm) containing 1 mM 3-Amino-1,2,4-triazole (3-AT), 100 μg/ml ampicillin, 25 μg/ml kanamycin, 10 μM IPTG, no histidine and 0.2 mM uracil. We incubated the plates at 37 °C for 2 days, or more if needed to obtain colonies (typically hundreds to thousands of visible colonies of varying size). We scraped the colonies from the plates and prepared plasmid DNA which we PCR amplified using Illumina barcoded primers.

Sequencing and data processing. Sequencing employed Illumina HiSeq 2000 Version 3 chemistry, typically with single-end 130 base reads, at the Donnelly Sequencing Centre at the University of Toronto. Before mapping, reads were trimmed at the 3′ end, so that for each read the minimum Phred quality score of any base of the remaining part was 12, there were no more than ten bases with Phred quality scores <20 and the reads, after removing the constant vector region and amplification adaptors, were not longer than 80 bases. The trimmed reads were further filtered to remove those shorter than 50 bases. Reads were mapped to the library using Bowtie[49], allowing a maximum of 2 mismatches at the first 28 bases and a maximum of 70 for the sum of quality values at all mismatched read positions, retaining only reads that were uniquely mapped.

Removing batch effects and normalizing s-scores. We initially obtained normalized read counts by dividing the read counts within each sample by the average read count for that sample, adding a pseudo-count and converting to log-scale, followed by row-normalization so that the average of each row (i.e., each zinc finger across different samples) would become zero. We then identified clusters of experiments with similar experimental/read mapping biases. In order to do so, we first selected a subset of zinc fingers that showed minimal sign of sequence-specific enrichment, reasoning that the measurements for such zinc fingers would best reflect experimental biases rather than DNA recognition by zinc fingers. We performed ANOVA *F*-test on each zinc finger across experiments, selecting zinc fingers that had between-bait variability smaller than within-bait variability (i.e., $F < 1$; regardless of the degree of freedom, this threshold represents cases where the association between bait and normalized read counts is insignificant). The initial normalized measurements for these zinc fingers were then used to identify clusters of similar experiments using Affinity Propagation[50] and manual cluster refinement. Then, the raw read counts for experiments within each cluster were renormalized as above, except that row-normalization was performed for each experiment cluster separately, so that the average normalized scores for each zinc finger across each experiment cluster would become zero. These normalized scores are called B1H s-score throughout the paper. The s-score reflects the logarithm of the growth rate of bacterial clones that harbor the corresponding zinc finger construct, which is assumed to be proportional to the occupancy of the binding site by the protein up to a certain saturation level[51]. Thus, the s-score represents a measure of binding free energy.

Motif derivation from B1H data and filtering C2H2-ZF domains. We used a stepwise linear regression to derive motifs (i.e., position weight matrices, PWMs) from s-scores, in a procedure modified from (ref. 52), as follows: starting from the s-score vector for one C2H2-ZF variant across the experiments and the 4-nucleotide bait sequences, at each step of the regression, one base at one of the four positions (i.e., one "feature") is selected randomly, based on a probability that scales linearly with the squared Pearson correlation of that feature with the s-score vector. The value of that feature (which corresponds to the selected base at the selected position of the PWM) is then increased by 0.01 if the Pearson correlation is positive, or decreased by 0.01 if the Pearson correlation is negative. The residual of the s-score vector is then calculated, replacing the s-score vector for the next iteration of updating the PWM. This process is repeated until the sum of squared Pearson correlations of the residual s-score vector with the vectors of all 16 features is <0.0001. C2H2-ZF domains whose motifs were not predictive of the s-scores in leave-one-out cross-validation experiments were removed (i.e., if the predicted s-score profile was not correlated with the experimental s-score profile at Benjamini-corrected FDR < 10%).

The B1H recognition code (B1H-RC). For each nucleotide at each of the positions 1-3 of the zinc finger binding site, we used random forest regression[53] to train a model that could predict the B1H-derived motifs (above) based on the sequence of the alpha helix of the zinc finger. We obtained the best results for prediction of a set of gold standard motifs (described in the next section) when we included only the canonical residues −1, +2, +3 and +6 of the helix in the model. This is most likely because the inclusion of noncanonical residues leads the model to learn properties that are specific to the Egr1-context used in the B1H assays, which are not necessarily generalizable to the natural context of the protein.

The random forest model is composed of 500 decision trees, whose outputs are averaged to produce the final "majority" prediction. Due to dependencies among nodes in each decision tree, random forest can potentially encode combinatorial effects of several zinc finger residues at the same time. However, for visualization purposes, **Figure 2a** depicts a linear representation of our random forest model, essentially indicating that, if all other residues were kept constant, what the average change in binding energy would be if a particular amino acid of the zinc finger was modified. This was obtained by predicting the binding preferences of 10^6 random amino acid sequences with uniform amino acid probabilities at the four contact residues, followed by regression to identify the average contribution of each amino acid at each position. For predicting the binding motif of an individual C2H2-ZF domain, we used the output of each of the 12 trained random forest models in order to obtain the corresponding value in the PWM.

To obtain a full-length motif of a protein from individual PWMs of C2H2-ZF domains, PWMs were concatenated, with the PWM of the C-terminal C2H2-ZF domain placed at the 5′ side of the full-length motif. Multi-zinc finger proteins were split into two or more C2H2-ZF "arrays" if at least two adjacent C2H2-ZF domains were connected with a linker shorter than 4 amino acids or longer than 6 amino acids (the canonical linker length for modular C2H2-ZF proteins is 4-6 amino acid), and motifs were predicted for each C2H2-ZF array separately. For analysis of motif diversity (**Supplementary Fig. 9d**) as well as comparison to gold standard motifs (next section), only proteins that did not have unusual linker lengths and, thus, did not need to be split were considered.

Full-length predicted PWMs were converted to position frequency matrices (PFMs), assuming that the values of the predicted PWMs were linearly related to the logarithm of likelihoods of the bases at different positions. We further assume that at the position with the maximum base selectivity, the most preferred base is 50 times more likely to be recognized than the least preferred base. Although this 50-fold preference is rather arbitrary, it worked reasonably well for prediction of motifs based on comparison with gold standard motifs. This leads to the formula $p_{ij} = c_i \times \exp(w_{ij}/b)$, where p_{ij} is the PFM value of base *j* at position *i*, c_i is a normalization factor so that the sum of PFM values at each position *i* is 1.0, w_{ij} is the PWM value of base *j* at position *i* and *b* is a normalization factor enforcing the maximum 50-fold preference mentioned above.

For proteins with ChIP-seq experiments, all possible zinc finger sub-arrays of lengths 3-7 were used for motif prediction, and the array with maximum AUC score for the top 500 ChIP-seq peaks was selected. AUC scores were obtained by first calculating the affinity score of each motif for each peak sequence centered around the peak summit (explained later in this document), as well as for shuffled peak sequences with the same dinucleotide frequency as the original sequences (generated using the 'fasta-dinucleotide-shuffle' tool of MEME suite[54]). The AUC scores were then calculated as the area under the ROC curve for distinguishing original sequences from shuffled sequences based on the affinity scores. Motif affinity scores were calculated using the PFMs as described before[55].

Comparison to Gold Standard motifs. We compiled a set of 64 "gold standard" (GS) motifs from the literature and available databases for natural C2H2-ZF proteins from different organisms. We used the collection of motifs reported for C2H2-ZF proteins in the CisBP database[41], including only C2H2-ZF proteins with canonical linker lengths (4–6 amino acids). In order to remove redundant motifs, we first selected, for each protein with multiple motifs, a single representative motif, as follows: if the protein had more than two motifs, we selected the motif that had the largest "sum of similarities" to other motifs. The similarity of a pair of motifs was defined as the Pearson correlation of their affinity scores across 50,000 random sequences of length 100 bp, with affinity scores calculated as described previously[55]. If the protein had only two motifs but also had a characterized homolog, we selected the motif that was most similar to the homolog motif (reasoning that this motif was supported by an independent experiment from a similar protein). If the protein had only two motifs and no homologs, we selected one motif randomly. We further removed similar motifs from different proteins by performing Affinity Propagation clustering of the motifs[50], selecting only the "exemplar" motif from each cluster.

For comparison of the predicted motifs with GS motifs, we aligned the PWMs predicted by the B1H-RC approach for each protein with the PWM of the experimental GS motif, so that the Pearson correlation of the vectors of scores of the two PWMs over all possible L-mers were maximized, where L corresponds to the length of the alignment. This maximum Pearson correlation was compared to a distribution of maximum Pearson correlations from randomized versions of the two PWMs in order to calculate the alignment P value. Randomized PWMs were constructed by shuffling the values within each column (position) of each PWM. In practice, an analytical approach was used for calculating the Pearson correlation of the scores of two PWMs, in a given alignment of length L, over all possible L-mers, as well as an analytical approach for estimating the alignment P values, which will be described elsewhere (H.S.N. and T.R.H., unpublished data).

Pearson correlations were also calculated for each individual zinc finger of the GS proteins (**Supplementary Fig. 3a**). We used the alignments to determine the region in each experimental motif that corresponded to each zinc finger and then calculated the Pearson correlation of the predicted motif of each zinc finger with the corresponding region within the experimental motif.

Structural modeling of C2H2-ZF proteins. Structural models were generated with the Jackal[56] program using PDB file 1AAY as the template. The DNA was elongated by four base pairs at each end using X3DNA[57] such that any end effects (termini melting) would not affect the protein bound nucleotides. Zinc 2+ ions were approximated using four mass-less dummy atoms in addition to the core zinc, each with a 1/2 positive charge, in a tetrahedral arrangement, for correctly orientated coordination of the four C2H2-ZF side-chains[58]. Using a 36-member subset of the gold standard set, one model was made for every possible three-finger array. The DNA component was mutated in each case using Chimera[59] to produce the sequence determined experimentally for that C2H2-ZF array. All models were prepared for Amber MD simulation using the WHATIF web interface[60] to build in any missing atoms and identify protonation states. They were then explicitly solvated in a 10 nm3 box of TIP3P water using TLEAP in AMBER 10 (ref. 61). Sodium counter-ions were added for overall charge neutrality and periodic boundary conditions were applied. Bonds to hydrogen were constrained using SHAKE[62] to permit a 2 fs time step, and the particle mesh Ewald[63] algorithm was used to treat long-range electrostatic interactions. The nonbonded cut-off was set at 12.0 Å. Systems were energy minimized using a combination of steepest descent and conjugate gradient methods. MD calculations were carried out with the PMEMD module of AMBER 10 in conjunction with the FF99 Barcelona forcefield[64], which is specifically customized for nucleic acids. The FF99 Stony Brook forcefield[65] was used for the protein. Each system was equilibrated and heated over 100 ps to 300K and positional restraints were gradually removed. A Berendsen thermostat and barostat was used throughout for both temperature and pressure regulation[66]. 20 ns of conformational space exploration was obtained for each array. During calculations a snapshot was saved every 2 ps. Root mean square deviation (r.m.s. deviation) was evaluated to assess the equilibration of each run. RMS clustering of the trajectory frames was carried out using the MMTSB toolset[67] kclust, with the radius set to 2.5 Å and maxerr to 1. This produced a set of 45 representative conformations. Further Jackal modeling was then carried out using these conformations as the templates. Every possible three-fingered array in the full gold standard set, together with their cognate B1H-RC predicted motifs, was committed to each template producing 138 × 45 models. Binding energy calculations were performed on the resulting complexes using FoldX[68]. The lowest energy model for each B1H-RC prediction was selected (as the closest to native conformation) and the binding energy for these was plotted (**Fig. 2d**). This approach makes C2H2-ZF binding energy calculations on much larger data sets computationally feasible.

The specificity code. We also trained a random forest model to discriminate 8,030 filtered C2H2-ZF domains in the B1H set from 3,383 C2H2-ZF domains that did not show sequence-specificity in the B1H experiments, despite being highly abundant in the original B1H library. We excluded zinc fingers that were used for PBM validations, as described below. The model obtained 0.95 sensitivity and 0.95 precision in tenfold cross-validation (**Supplementary Fig. 4b**). To assess the performance of this model on independent observations, we used the outcome of 185 PBM experiments for zinc fingers not included in the specificity model training, consisting of 108 zinc fingers that passed the sequence-specificity threshold in the B1H assays and 77 that did not. We used the random forest model probability output as the predictor and identified the minimum probability threshold below which the rate of gain in sensitivity was smaller than the rate of loss of specificity, corresponding to 59% sensitivity and 92% specificity (**Supplementary Fig. 4c**). We used this threshold to identify potential DNA-binding zinc fingers in various organisms, including human, and to estimate the number of DNA-binding zinc fingers in each organism (**Supplementary Fig. 4d**). Briefly, given the 59% sensitivity and 8% false positive rate, the fraction of DNA-binding zinc fingers was estimated as $r = (p\text{-}0.08)/0.51$, where p is the fraction of zinc fingers predicted by the model as positive.

Protein binding microarrays (PBMs). PBM methods, including quality control and data analysis, followed the procedure described previously[69]. We PCR amplified, subcloned and sequence-verified inserts from the B1H prey library into pTH5325 or pTH6838. We examined a total of 185 constructs by PBMs, corresponding to two categories of proteins: 108 constructs were examined to confirm the triplet obtained by B1H and to test a range of motifs with various degrees of sequence selectivity, and 77 constructs represented a random selection of C2H2-zinc fingers from among those that did not pass the sequence-specificity cutoff in the B1H assays. A summary of information about these experiments is given in **Supplementary Table 1**.

PBM motifs were derived from 8-mer Z-scores as follows: for each PBM experiment, 8-mers that represented the expected binding context were selected (i.e., those matching NNNNGGGC; bait sequence is underlined). Then, a PWM was learned for the variable part, by performing linear regression on the Z-scores, similar to the procedure described in "Motif derivation from B1H data." Comparisons of Z-scores vs. s-scores and motifs for individual proteins are shown in **Supplementary Figure 2** and **Supplementary Table 1**.

***In vivo* analysis of binding sites of C2H2-ZF proteins.** *Tissue culture and cell lines:* HEK293 cells (Invitrogen) were cultured in Dulbecco's modified Eagle's medium with 10% FBS and antibiotics as described previously[70,71]. Gateway-compatible entry clones were cloned into the pDEST pcDNA5/FRT/TO-eGFP vector[72] according to the manufacturer's instructions and co-transfected into Flp-In T-REx 293 cells together with the pOG44 Flp recombinase expression plasmid. Cells were selected for FRT site-specific recombination into the genome, following instructions by the manufacturer. Expression of the gene of interest was induced by addition of doxycycline to the culture medium 24 h before harvesting.

doi:10.1038/nbt.3128

Chromatin immunoprecipitation and sequencing: Chromatin immunoprecipitation was performed as described in ref. 73. In brief, 10^7–10^8 HEK293 cells were cross-linked for 10 min in 1% formaldehyde. Lysates were sonicated to a DNA fragment length range of 200–300 bp using a Bioruptor (Diagenode). GFP-tagged transcription factors were immunoprecipitated with a polyclonal anti-GFP antibody (ab290, Abcam) and Dynabeads Protein G (Invitrogen). Subsequently, crosslinks were reversed at 65 °C overnight and bound DNA fragments were purified (EZ-10 Spin Column PCR Product Purification kit, Bio Basic). Sequencing libraries were constructed using the NEBNext ChIP-Seq Library Prep kit (NEB) according to the manufacturer's instructions. Libraries were sequenced (single end reads) on the Illumina HiSeq 2500 to a minimum depth of 20 million 51-nucleotide reads.

ChIP-seq data analysis. *Read mapping:* The 3′ end of reads were trimmed so that the final reads were 50 nucleotides long, and the reads were mapped to the human genome build GRCh37 using Bowtie 2 (ref. 74), with "–very-sensitive" preset of parameters. This parameter set allows one alignment to be reported for each read that is mapped to multiple positions on the genome, in addition to alignments of uniquely mappable reads. Duplicate reads, which often represent PCR amplification artifacts, were removed using SAMtools[75].

This procedure may result in alignment of a fraction of the reads to repetitive regions other than the one they originated from. However, this should not affect our conclusions about enrichment of EREs among the binding targets of the proteins because the potentially misaligned reads will almost certainly still map to the same ERE type as that of the read origin[76]. Furthermore, as **Supplementary Figure 10** shows, on average 85% of the reads that fall within ChIP-seq peaks are mapped correctly (median 90% correct read mapping per experiment). The main exceptions are ZNF33A and ZNF382, which bind to the youngest and, thus, least diverged EREs in the genome, i.e., SVA and L1HS repeats. Because of their minimal sequence divergence, most reads that originate from these repeats cannot be aligned uniquely to an exact ERE instance.

Peak-calling: We performed two rounds of peak-calling: an initial round was performed without an explicit background model of read distributions, in order to identify ChIP experiments with similar experimental/read-mapping biases. We then constructed a background model for each sample and repeated the peak-calling in order to identify the final set of protein-specific peaks.

For the initial round, peaks were identified using MACS v1.4 (ref. 77) at P value < 10^{-5}, with fragment length specified based on the results of cross-correlation analysis[78] using SPP package[79] (average 150 bp). Then, the overlap of the identified peaks for each pair of experiments was calculated as the total length of the intersection of peaks divided by the total length of the union of peaks from two experiments. Then, for each experiment x, the top ten experiments with the highest overlap were identified, so that none of these ten experiments would represent a biological replicate of experiment x. The reads from these ten experiments were then randomly sampled, with probabilities proportional to the extent of overlap of each experiment with experiment x, so that a total of 5×10^7 reads would be obtained. These reads were then used as the background model in order to perform a second, final round of peak-calling for experiment x, using MACS as above.

De novo motif discovery: For each ChIP experiment, the top 500 peaks with the highest enrichment were used for *de novo* motif discovery. Peaks were centered around their "summits" as identified by MACS, and their sequences were submitted to MEME-ChIP[80] for *de novo* motif discovery, with the following parameters: ZOOPS mode, minimum MEME width 6, maximum MEME width 30, maximum 5 MEME motifs and DREME E-value cutoff 0.05. The *de novo* motifs were then analyzed using CentriMo[81] and experiments with at least one centrally enriched *de novo* motif were selected for further analysis (including 39 C2H2-ZF proteins).

Analysis of enrichment in endogenous retroelements (EREs). EREs of the human genome were obtained from the RepeatMasker track of the UCSC Table Browser[82] and were used to examine binding of the C2H2-ZF proteins to EREs. We calculated the number of top 500 peaks from each ChIP experiment that overlapped at least one ERE element, and then identified significant enrichments using Fisher's exact test. Specifically, for each experiment x, we asked whether the top 500 peaks from x showed significantly higher overlap with EREs compared to the pool of top 500 peaks from each of the other experiments. P values were adjusted for multiple hypotheses testing, and significant enrichments at Benjamini-corrected FDR < 0.01 were identified. We identified a total of 18 proteins with peaks that were enriched in EREs. For each of these proteins, we repeated the same procedure as above, using each repeat type separately, in order to identify the specific ERE types that were bound by each protein.

Functional enrichment analysis. For each of the 39 C2H2-ZF proteins, we first selected a single PWM as the most reliable motif for that protein. For each protein, we gave priority to the B1H-RC motif if it showed similar or better AUC value than the *de novo* motifs (**Fig. 3**). The second priority was given to the *de novo* motif with the highest similarity to the B1H-RC if it had a similar AUC value to the top-ranking MEME motif. Otherwise, the top-ranking MEME motif was selected.

Then, for each motif, we identified all hits on the human genome using FIMO[83] at P value < 0.0001 and selected hits that overlapped at least one peak in the corresponding ChIP experiment. The selected motif hits were then used to identify enriched functions using GREAT[84]. For each protein x, we used three different backgrounds for GREAT analysis: (i) the entire human genome, (ii) the set of all motif hits for protein x in the human genome regardless of overlap with the ChIP peaks (in order to control for sequence biases introduced by motif scanning) and (iii) the set of all peaks from the 39 proteins (in order to control for biases introduced by the ChIP experimental procedure). For backgrounds (ii) and (iii), we subsampled a random set of maximum 100,000 entries due to computational limitations. In order to identify significant functional enrichments, we required a term to have significant hypergeometric P-values at FDR < 0.01 for all three backgrounds, significant binomial P value at FDR < 0.01 for the genomic background, ≥2-fold enrichment of peaks in the associated regions relative to all the three backgrounds, as well as ≥5 associated genes.

Analysis of conservation and H3K9 methylation. We used the PhastCons track from UCSC[85] in order to identify conserved motif hits, as well as H3K9 methylation data from ENCODE K562 cell line[86] for identification of methylation patterns around C2H2-ZF protein binding sites. Motif hits within ChIP-seq peaks were identified, and in case of PhastCons, were divided into ERE-overlapping and nonoverlapping groups. Then, the average PhastCons score or H3K9me3 signal at different positions relative to the center of motif hits were calculated, with positive positions facing the 3′ side of the motif hit and negative positions facing the 5′ side of the motif. For PhastCons, we calculated the conservation for all three categories of motifs, i.e., the B1H-RC motif, the top-ranking MEME motif and the MEME motif with the highest similarity to the B1H-RC motif (if not the same as the top-ranking MEME motif). For H3K9me3, we used the same motif as that used for GREAT analysis (see above). Wherever possible, the *de novo* motifs were re-oriented to match the corresponding B1H-RC motifs, ensuring that the KRAB domain is oriented toward the 3′ side of the motif.

Correlation of expression between C2H2-ZF proteins and their targets. For each C2H2-ZF protein, we used the associated genes identified by GREAT (see above), in order to identify C2H2-ZF proteins whose targets are significantly correlated or anti-correlated in terms of expression across 112 different ENCODE cell lines (data from ref. 87). Following the procedure described previously[55], for each C2H2-ZF protein x, we sorted all human genes in the descending order of the Pearson correlation of their expression with that of x. Then, we identified significant enrichment of targets of x at the top or bottom of this sorted list using Mann-Whitney U test, correcting for multiple hypotheses testing (i.e., 39 proteins) at Benjamini-corrected FDR < 0.01.

Overlap with cell type-specific DHS sites. We used previously described DHS-cell type assignments[87] in order to identify DHS sites that are specific to a single ENCODE cell type. We then calculated the overlap of motif-containing ChIP-seq peaks with DHS sites specific to each cell type and identified significant enrichments using Fisher's exact test, using a procedure similar to that described in the section "Analysis of enrichment in endogenous retroelements (EREs)". We performed two different tests: (i) we asked whether among all binding sites of protein x, there was an enrichment of DHS sites

doi:10.1038/nbt.3128

specific to each cell type *y*, with pool of binding sites from all other proteins as the background; (ii) we asked whether among binding sites of protein *x* that overlapped any DHS, there was an enrichment of DHS sites specific to each cell type *y*, with a pool of DHS-overlapping binding sites from all other proteins as the background (**Supplementary Fig. 6c**).

Estimating the number of independent motifs. We used Affinity Propagation (AP)[50] to measure the diversity of motifs among C2H2-ZF proteins as well as other transcription factors. We used B1H-RC motifs of all human C2H2-ZF proteins with 4–7 zinc fingers in which all the linkers had the canonical length (4–6 amino acids, 77 proteins), B1H-RC motifs for the 39 human C2H2-ZF proteins with ChIP-seq or the *de novo* ChIP-seq motifs that were most similar to the B1H-RC motifs for the 39 human C2H2-ZF proteins with ChIP-seq experiments (the latter two groups are shown in **Fig. 3**). For the first group, the full-length motif was obtained by concatenating the individual zinc finger motifs for all zinc fingers of each protein. For the second group, see the section "the B1H recognition code (B1H-RC)."

To cluster the motifs, for pairs of PFMs within each category, we calculated the Pearson correlation of logarithm of affinities across 50,000 random 100 bp sequences with uniform base distributions. Affinities were calculated using PSAMs as described before[55]. The pairwise similarity matrix that was obtained with this approach for each category of proteins was then used as input to AP to cluster the PFMs, with the "preference" parameter of the AP algorithm set to 0.3. Also, within each group, the similarity matrix was sampled at various depths to contain only a subset of the proteins, in order to obtain a standard curve for predicting motif diversity as a function of number of proteins. The results are shown in **Supplementary Figure 9**.

46. Tonikian, R., Zhang, Y., Boone, C. & Sidhu, S.S. Identifying specificity profiles for peptide recognition modules from phage-displayed peptide libraries. *Nat. Protoc.* **2**, 1368–1386 (2007).
47. Gupta, A. *et al.* An optimized two-finger archive for ZFN-mediated gene targeting. *Nat. Methods* **9**, 588–590 (2012).
48. Meng, X. & Wolfe, S.A. Identifying DNA sequences recognized by a transcription factor using a bacterial one-hybrid system. *Nat. Protoc.* **1**, 30–45 (2006).
49. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
50. Frey, B.J. & Dueck, D. Clustering by passing messages between data points. *Science* **315**, 972–976 (2007).
51. Stormo, G.D. & Zhao, Y. Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.* **11**, 751–760 (2010).
52. Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. Least angle regression. *Ann. Stat.* **32**, 407–499 (2004).
53. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
54. Bailey, T.L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
55. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177 (2013).
56. Petrey, D. *et al.* Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins* **53** (suppl. 6), 430–435 (2003).
57. Lu, X.J. & Olson, W.K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **31**, 5108–5121 (2003).
58. Pang, Y.P. Successful molecular dynamics simulation of two zinc complexes bridged by a hydroxide in phosphotriesterase using the cationic dummy atom method. *Proteins* **45**, 183–189 (2001).
59. Pettersen, E.F. *et al.* UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
60. Vriend, G. WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **8**, 52–56 (1990).
61. Case, D.A. *et al.* The Amber biomolecular simulation programs. *J. Comput. Chem.* **26**, 1668–1688 (2005).
62. 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. *J. Comput. Phys.* **23**, 327–341 (1977).
63. Toukmaji, A., Sagui, C., Board, J. & Darden, T. Efficient particle-mesh Ewald based approach to fixed and induced dipolar interactions. *J. Chem. Phys.* **113**, 10913 (2000).
64. Pérez, A. *et al.* Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.* **92**, 3817–3829 (2007).
65. Hornak, V. *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **65**, 712–725 (2006).
66. Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., DiNola, A. & Haak, J.R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684 (1984).
67. Feig, M., Karanicolas, J. & Brooks, C.L. III. MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J. Mol. Graph. Model.* **22**, 377–395 (2004).
68. Guerois, R., Nielsen, J.E. & Serrano, L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387 (2002).
69. 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**, 4680–4690 (2011).
70. Chen, G.I. *et al.* PP4R4/KIAA1622 forms a novel stable cytosolic complex with phosphoprotein phosphatase 4. *J. Biol. Chem.* **283**, 29273–29284 (2008).
71. Moffat, J. *et al.* A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell* **124**, 1283–1298 (2006).
72. Skarra, D.V. *et al.* Label-free quantitative proteomics and SAINT analysis enable interactome mapping for the human Ser/Thr protein phosphatase 5. *Proteomics* **11**, 1508–1516 (2011).
73. Schmidt, D. *et al.* ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. *Methods* **48**, 240–248 (2009).
74. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
75. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
76. Day, D.S., Luquette, L.J., Park, P.J. & Kharchenko, P.V. Estimating enrichment of repetitive elements from high-throughput sequence data. *Genome Biol.* **11**, R69 (2010).
77. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
78. Landt, S.G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
79. Kharchenko, P.V., Tolstorukov, M.Y. & Park, P.J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* **26**, 1351–1359 (2008).
80. Machanick, P. & Bailey, T.L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011).
81. Bailey, T.L. & Machanick, P. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.* **40**, e128 (2012).
82. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 (2004).
83. Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
84. McLean, C.Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
85. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
86. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
87. Sheffield, N.C. *et al.* Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res.* **23**, 777–788 (2013).

doi:10.1038/nbt.3128