Supplementary Information for:

"C2H2 zinc finger proteins greatly expand the human regulatory lexicon"

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

[1] Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto M5S 3E1, Canada

[2] Department of Molecular Genetics, University of Toronto, Toronto M5S 1A8, 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] Department of Computer Science, University of Toronto, Toronto M5S 2E4, Canada

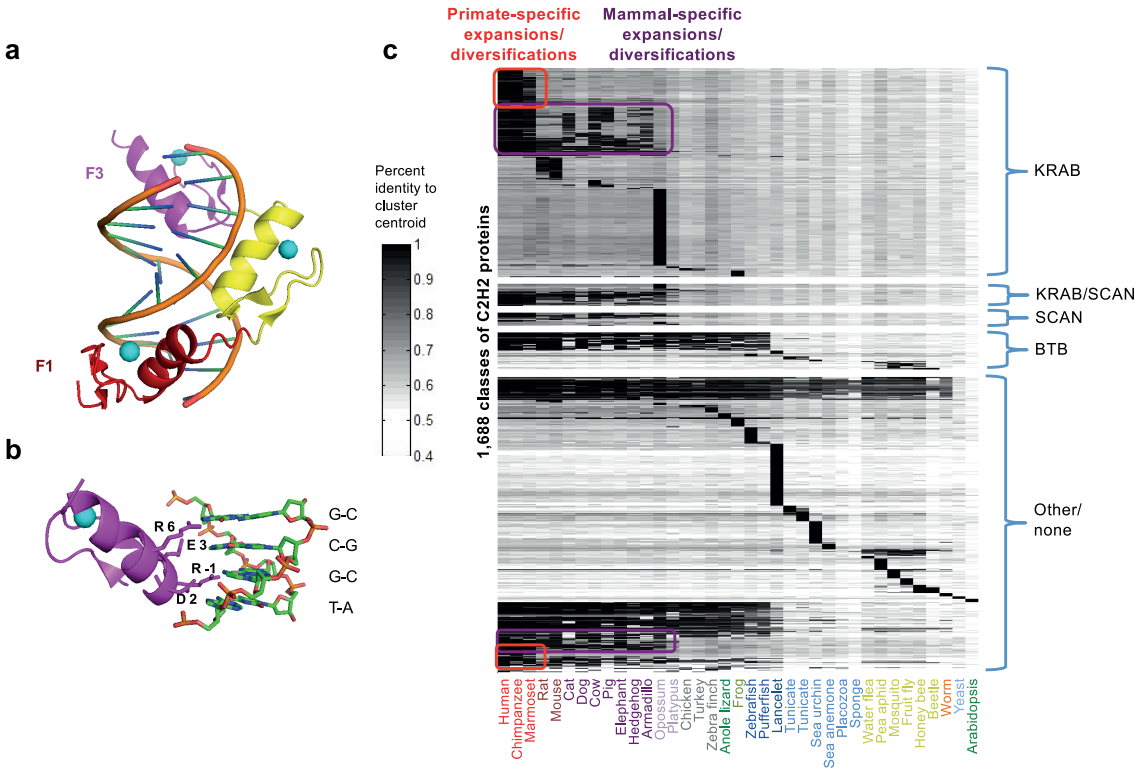
[5] Department of Electrical and Computer Engineering, University of Toronto, Toronto, M5S 3G4, Canada

[6] Canadian Institutes For Advanced Research

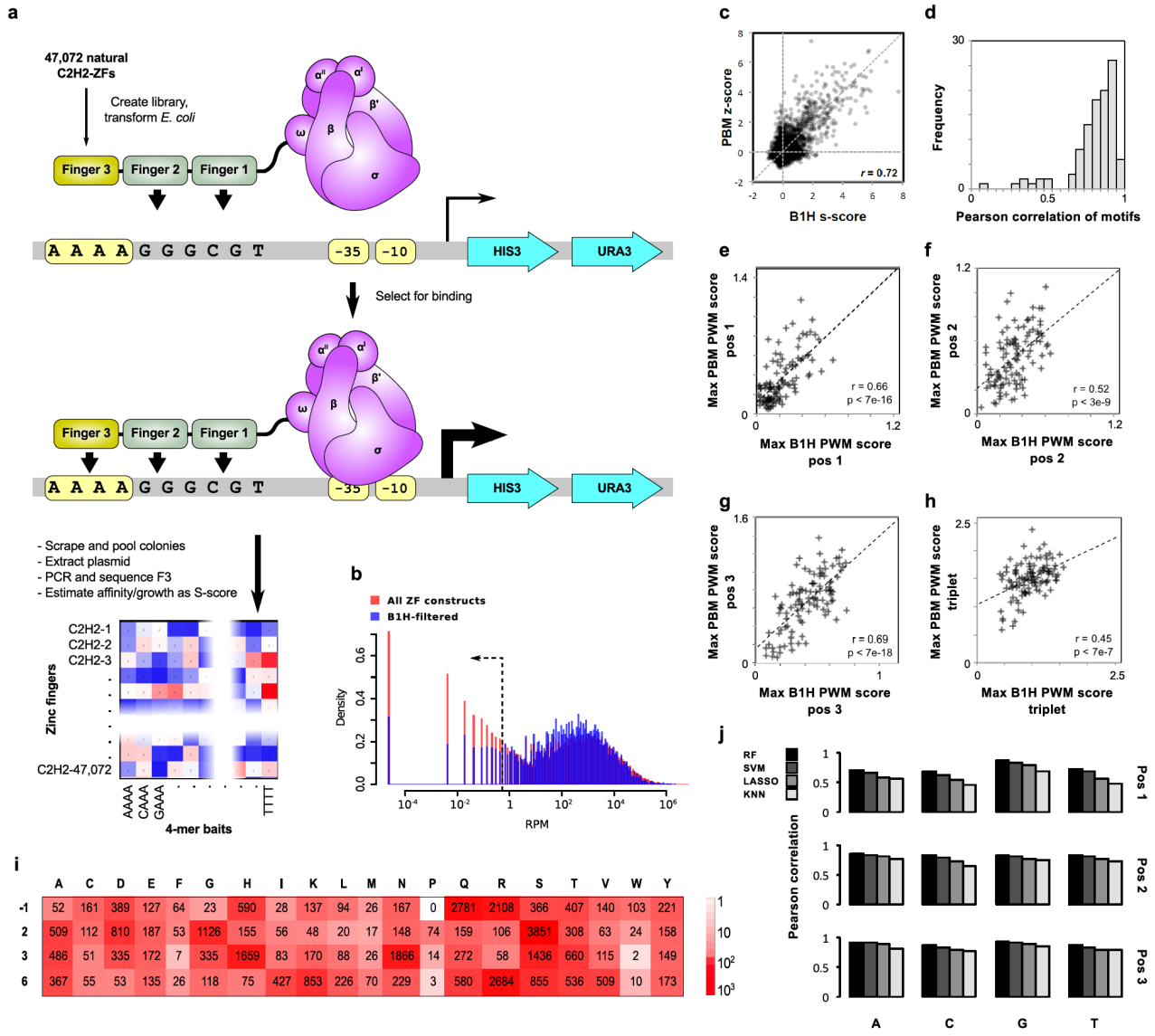
* These authors contributed equally to this work.

Correspondence should be addressed to T.R.H. (t.hughes@utoronto.ca).

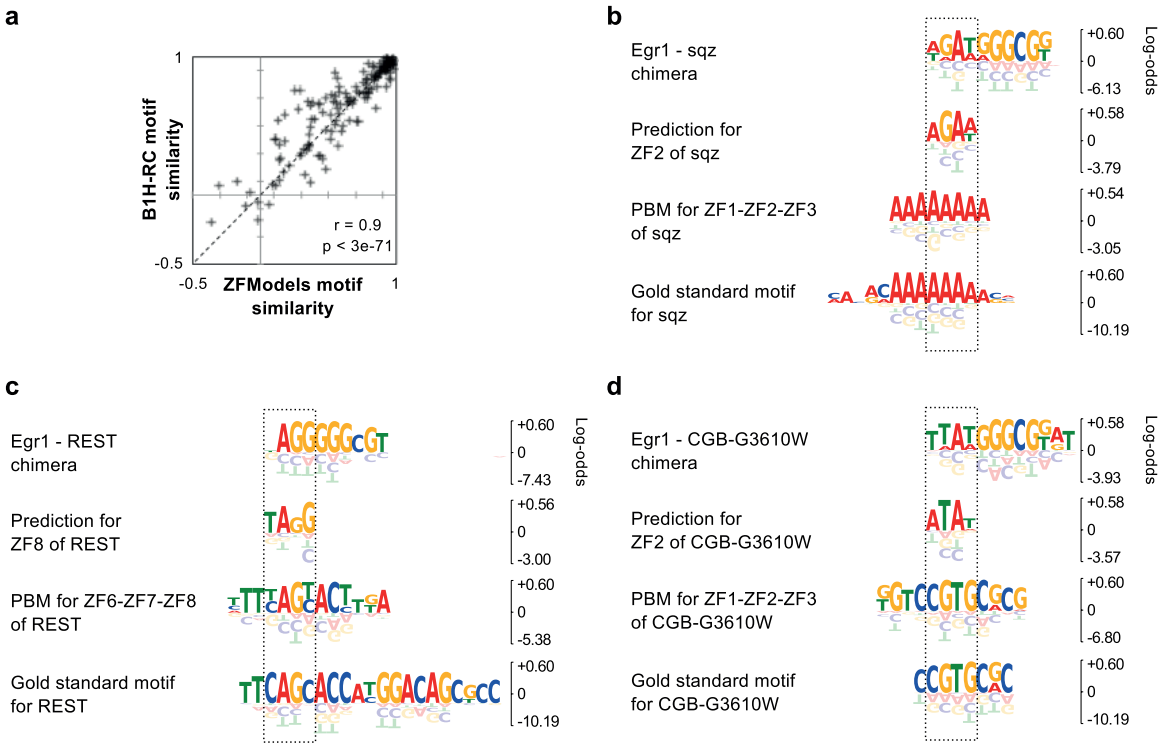
Nature Biotechnology: doi:10.1038/nbt.3128



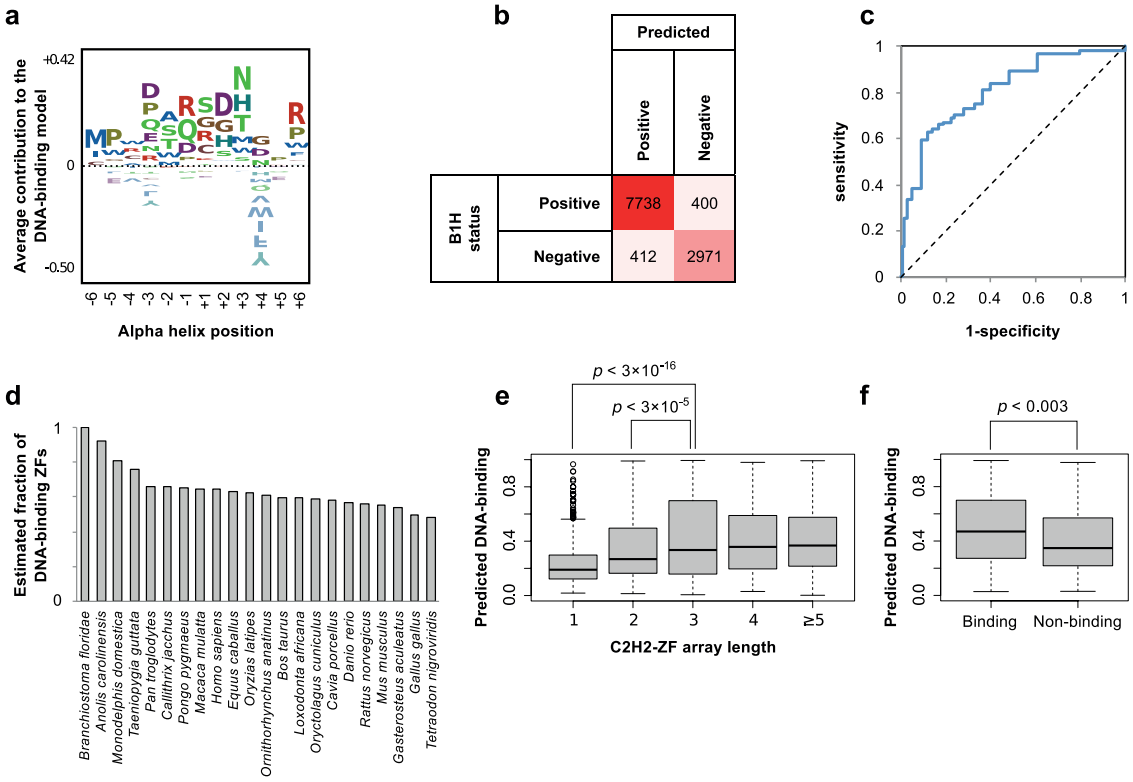
Supplementary Figure 1. DNA-binding structure and lineage-specific expansion of C2H2-ZF proteins. **(a)** Structure of Egr1 (Zif268) bound to its consensus recognition site (PDB# 1AAY) [1]. The three C2H2-ZFs are indicated in different colors. By convention, the finger numbers follow their order in the protein, and are shown bottom to top, but the recognition site is given from top to bottom. **(b)** Close-up of the DNA-contacting residues; F3 is rotated about the vertical axis relative to **(a)**. **(c)** Lineage-specific expansion of C2H2-ZFs. All pairwise identities among all 15,838 C2H2-ZF proteins sequences in the genomes shown were calculated using ClustalW. Affinity Propagation grouped them into 1,688 clusters, which we take to represent classes of orthologs or paralogs. Each row represents one of these 1,688 groups. Identity of the cluster centroid to the closest protein in each genome is shown. The vertical axis reflects approximate divergence time from human; the horizontal axis was sorted by the effector domain (if any) present in the cluster centroid, and then clustered using hierarchical agglomerative clustering.



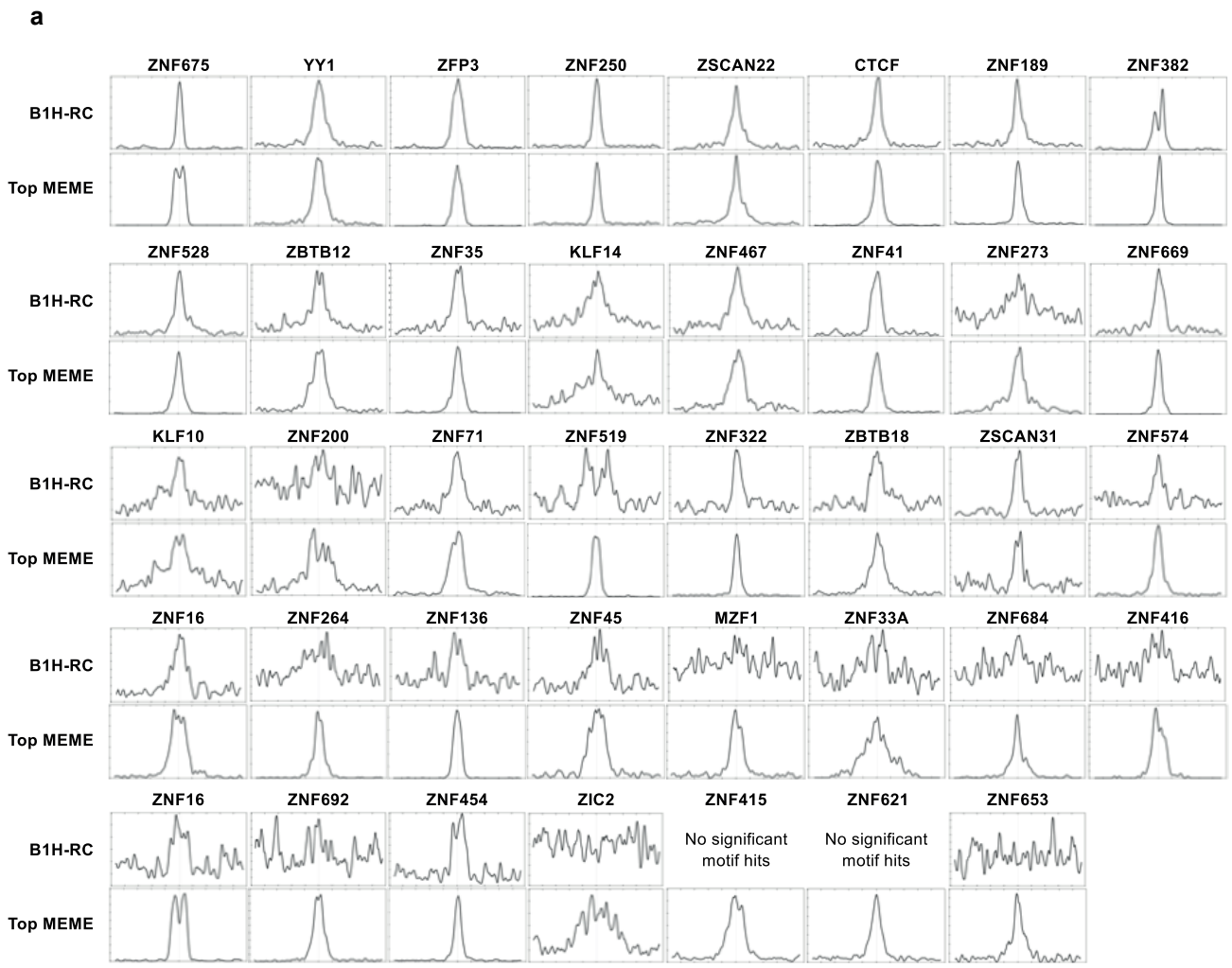
Supplementary Figure 2. The B1H system, and comparison to PBM. **(a)** Schematic representation of B1H system used in this study. **(b)** The B1H data were filtered to retain the 8,138 C2H2-ZFs with the most clear and reproducible sequence preferences (see Methods). Constructs that were present in the transformed library at abundances <0.7 RPM (dashed arrow) are under-represented in this B1H-filtered set (blue), indicating a higher rate of false negatives for these constructs. Less than 15% of all ZF constructs fall below the 0.7 RPM threshold, and are thus negatively affected. **(c)** B1H data correlate strongly with PBM data for 104 C2H2-ZFs that show sequence-specificity in both assays. Each point represents a DNA triplet-protein pair; x-axis shows the average s-score of the protein in B1H assays that use the triplet as the bait, and the y-axis represents the average PBM z-score of 8-mers that contain the triplet in the expected binding context (NNNNGGGC, triplet underlined). **(d)** Histogram of PBM vs. B1H motif similarities. **(e-g)** Scatter plots of selectivity of motifs derived from B1H and PBM data. The x- and y-axes show the maximum PWM score at indicated positions of the B1H and PBM motifs, respectively, corresponding to the selectivity of the PWMs at each position. **(h)** Similarly, the maximum triplet score based on the PWMs corresponds to overall selectivity of the motif. **(i)** Amino acid representations at the specificity residues among the B1H set of 7,984. Number of occurrences of each amino acid at each position is indicated. 154 C2H2-ZFs that had alpha helices of non-canonical lengths were not considered for this table. **(j)** Comparison of different machine learning approaches for prediction of B1H data. The y-axis corresponds to Pearson correlation of observed B1H s-scores vs. predicted values in 10-fold cross-validations. RF: Random Forest, SVM: Support Vector Machine, LASSO: Least Absolute Shrinkage and Selection Operator, KNN: K-Nearest Neighbor (k=1).



Supplementary Figure 3. B1H-RC predictions are affected by context-dependent factors. **(a)** Per-ZF comparison of B1H-RC and ZFModels predictions with respect to GSTD motifs. Each dot represents a single C2H2-ZF, with y-axis corresponding to the Pearson correlation of the predicted B1H-RC motif vs. the corresponding part of the GSTD motif, and the x-axis corresponding to the Pearson correlation of predicted ZFModels motifs. B1H-RC shows a significant improvement over ZFModels (paired t-test, $p < 4\times10^{-4}$). However, both methods generally have difficulty with the same C2H2-ZF domains. **(b-d)** Context-specificity of C2H2-ZF sequence recognition. The PBM motifs for ZF1-ZF2 of Egr1 fused to a single ZF from each of the three proteins sqz (*Drosophila melanogaster*), REST (*Homo sapiens*), and CGB-G3610W (*Cryptococcus gattii*) is shown in the first row of each panel. The second row shows the motif predicted by B1H-RC, and the PBM motif of the same zinc finger in its natural context is shown in the third row. Previously published motif for each protein is shown in the last row.

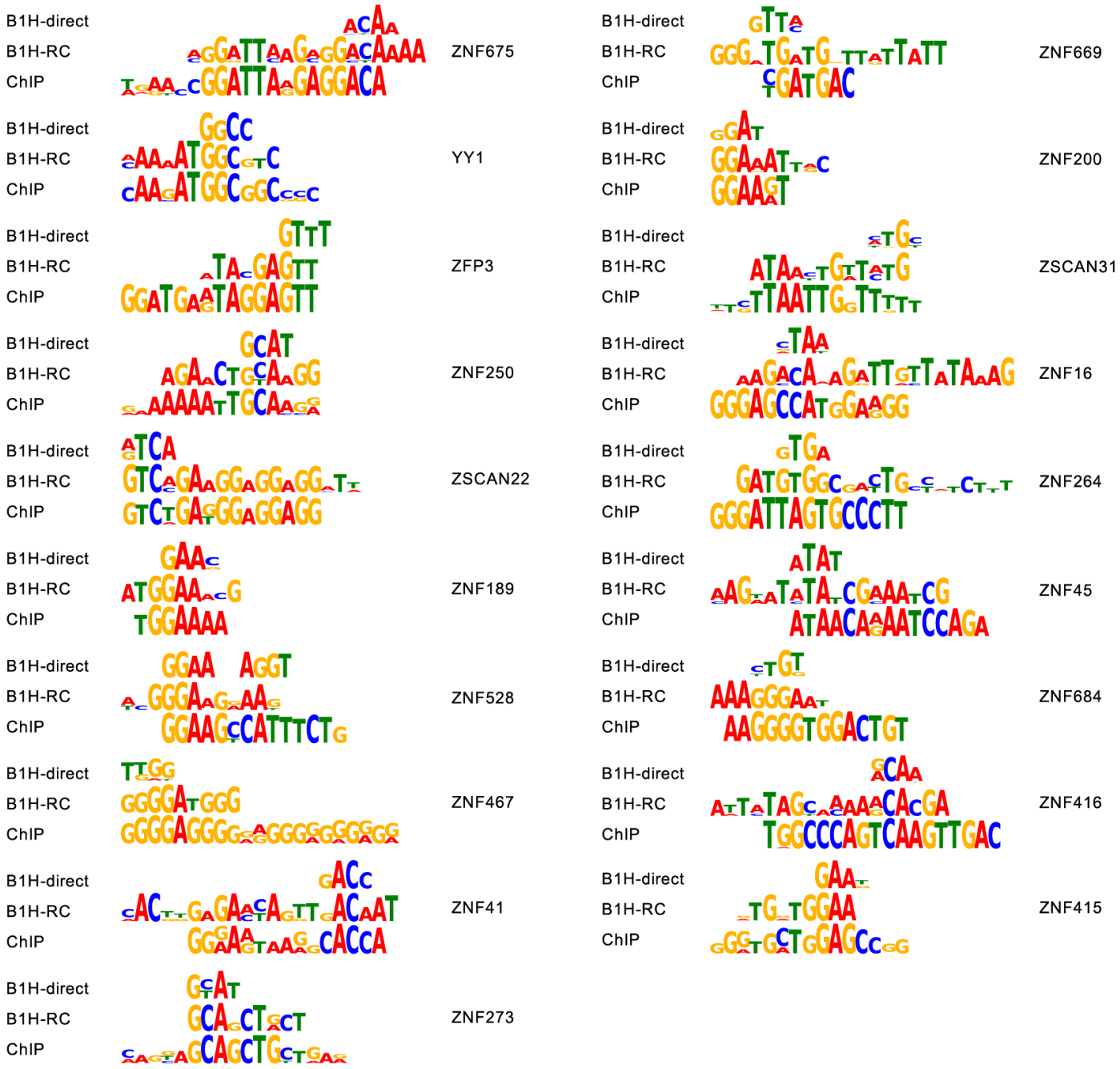


Supplementary Figure 4. Predicting DNA-binding activity from C2H2-ZF sequence. A random forest model was trained to distinguish DNA-binding C2H2-ZFs from non-binding ZFs. **(a)** The average contribution of each amino acid at each position to the random forest model. **(b)** Results of 10-fold cross-validation of the model on the B1H set. **(c)** The ROC curve for validation of the model using PBM of 185 C2H2-ZFs that were excluded from the training set. **(d)** Estimated number of C2H2-ZFs of different organisms that would work in the Egr1-F3 context based on the DNA-binding model. The shown organisms are selected to represent different vertebrate taxa, including the majority of organisms studied in ref [2]. **(e)** Human ZFs that are in C2H2-ZF arrays of length 1 or 2 have significantly lower predicted DNA-binding scores than ZFs in longer arrays. **(f)** ZFs that are inferred to be involved in DNA-binding in 35 ChIP-seq experiments have significantly higher predicted DNA-binding scores than other ZFs of the same proteins (Student's t-test).

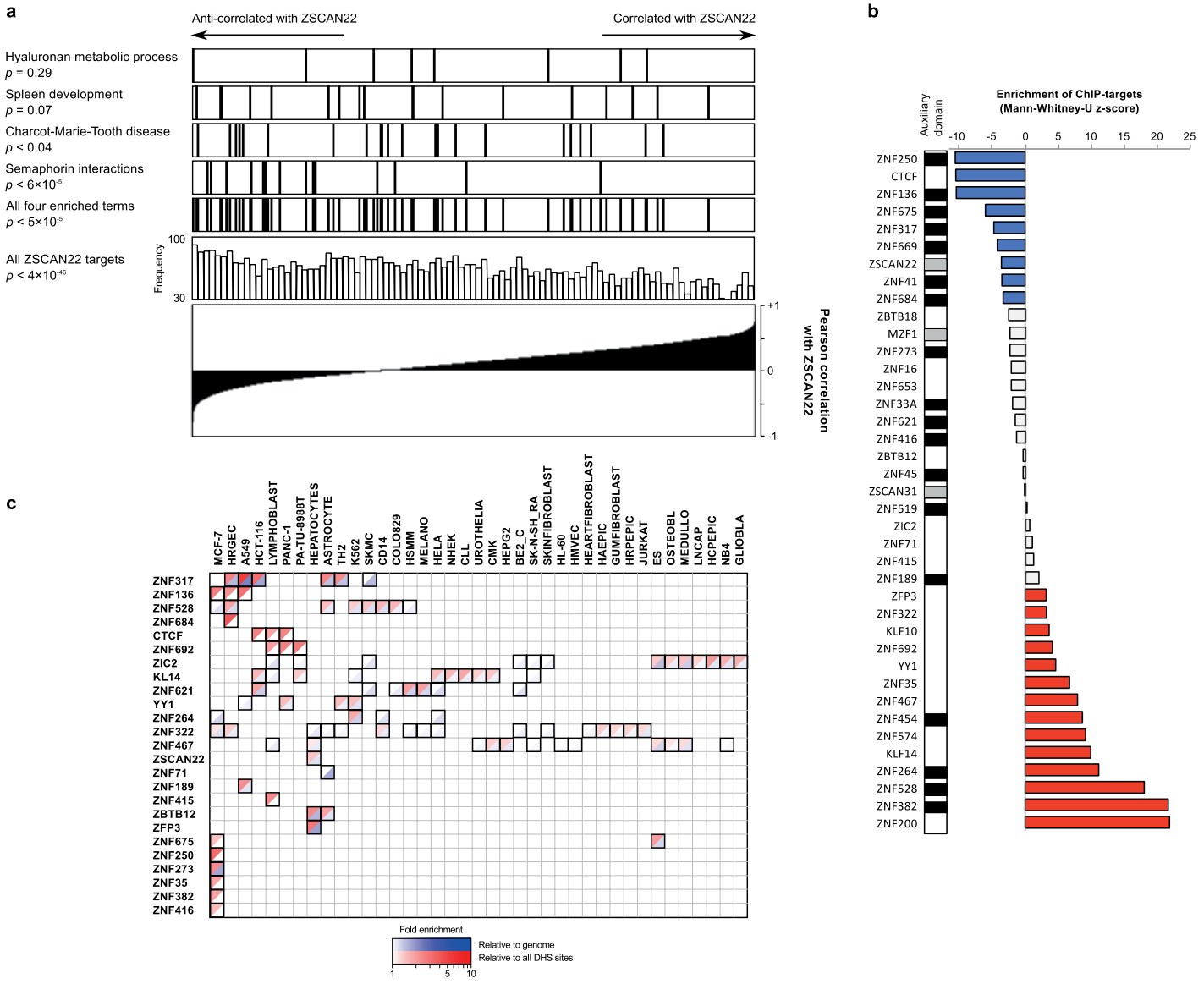


Supplementary Figure 5. C2H2-ZF proteins directly bind to DNA. (a) Centrality of B1H-RC and top *de novo* motifs in ChIP-seq peaks for 39 human C2H2-ZF proteins. For each motif, centrality was calculated using CentriMo for the top 500 peaks, including the region [-250,250] around the peak summits. Central enrichment of motifs in ChIP-seq peaks is often an indication of direct DNA binding[3]. **(b, on the next page)** Motifs obtained for individual C2H2-ZF domains from B1H and ChIP-seq are often similarity. Only proteins are included that are assayed by ChIP-seq in this study, and also have at least one C2H2-ZF domain in the B1H-filtered set.

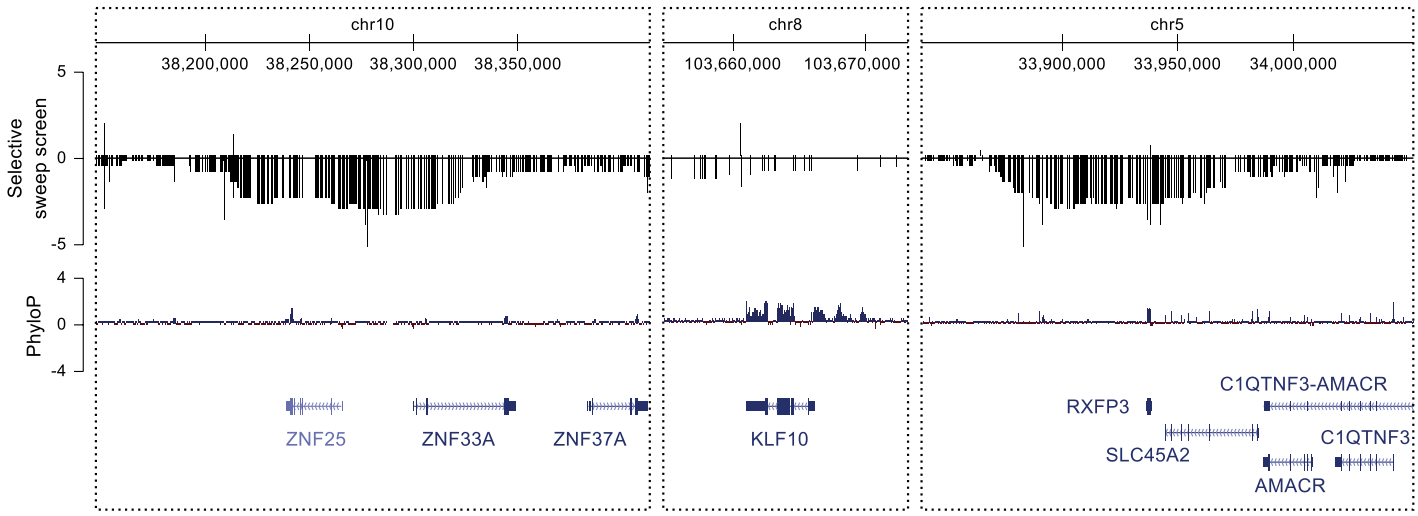
b



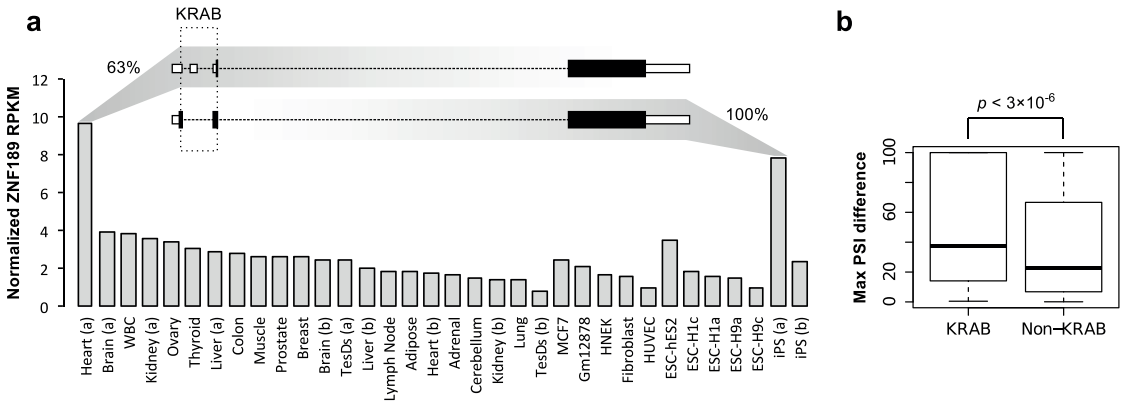
Supplementary Figure 5 – continued



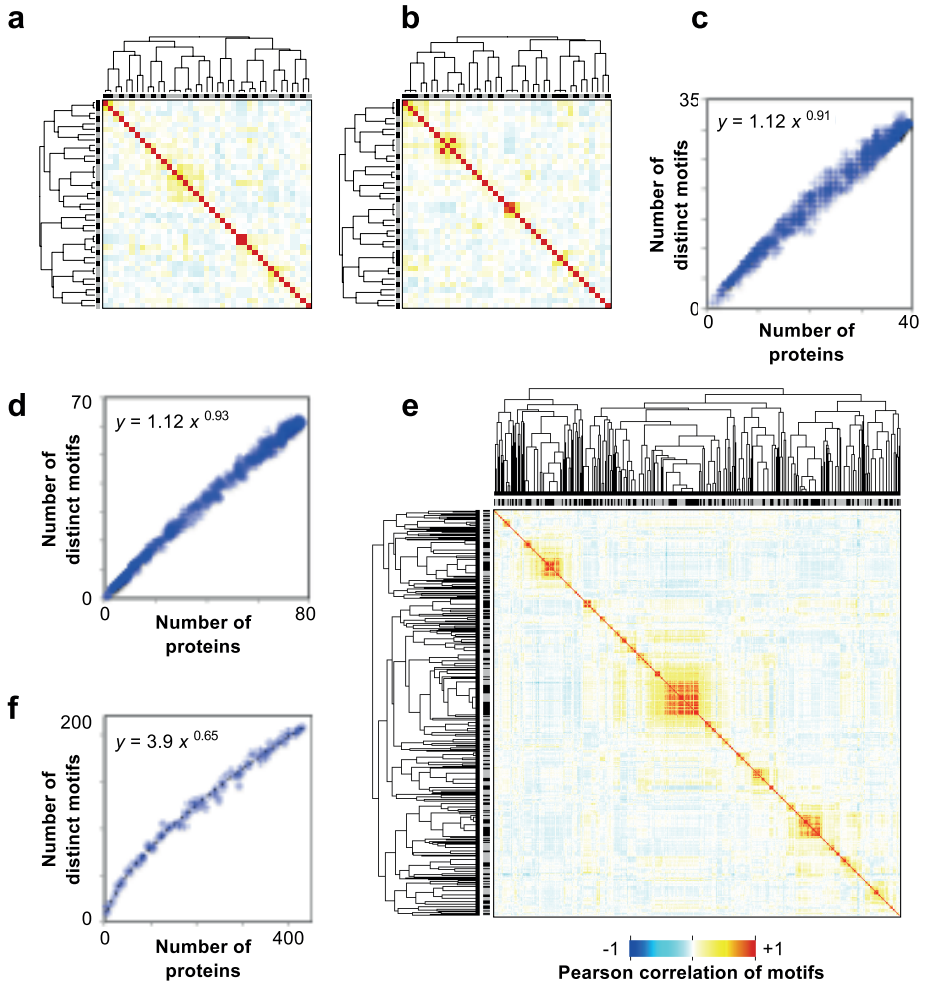
Supplementary Figure 6. C2H2-ZF proteins are involved in cell type-specific regulatory programs. (a) Expression of genes whose regulatory regions contain at least one ZSCAN22 binding site (i.e. overlap with a ChIP-seq peak containing the ZSCAN22 motif) is anti-correlated with expression of ZSCAN22 across 37 tissues/cell lines based on previously described RNA-seq data [4]. This figure shows genes that are within each of four different enriched functional terms, as well as all ZSCAN22 targets. *P*-values are calculated based on Mann-Whitney U test of ranks. **(b)** Expression of C2H2-ZF proteins is significantly correlated or anti-correlated with their targets across 112 cell lines [5]. Red and blue represent significant correlations and anti-correlations, respectively. Of the 9 proteins that are significantly anti-correlated with their targets, 8 proteins have a KRAB domain (black box) or a SCAN domain (grey box), representing a 1.7-fold enrichment of repressive domains among anti-correlated proteins compared to positively correlated proteins (Fisher's exact test, p-value < 0.007). **(c)** Cell type-specific DHS sites are enriched for binding sites of different C2H2-ZF proteins. Black boxes mark enrichments that are significant either relative to genome or relative to all DHS sites (Fisher's exact test, FDR < 0.01), with significant enrichments relative to each of these two backgrounds shown in blue and red, respectively.



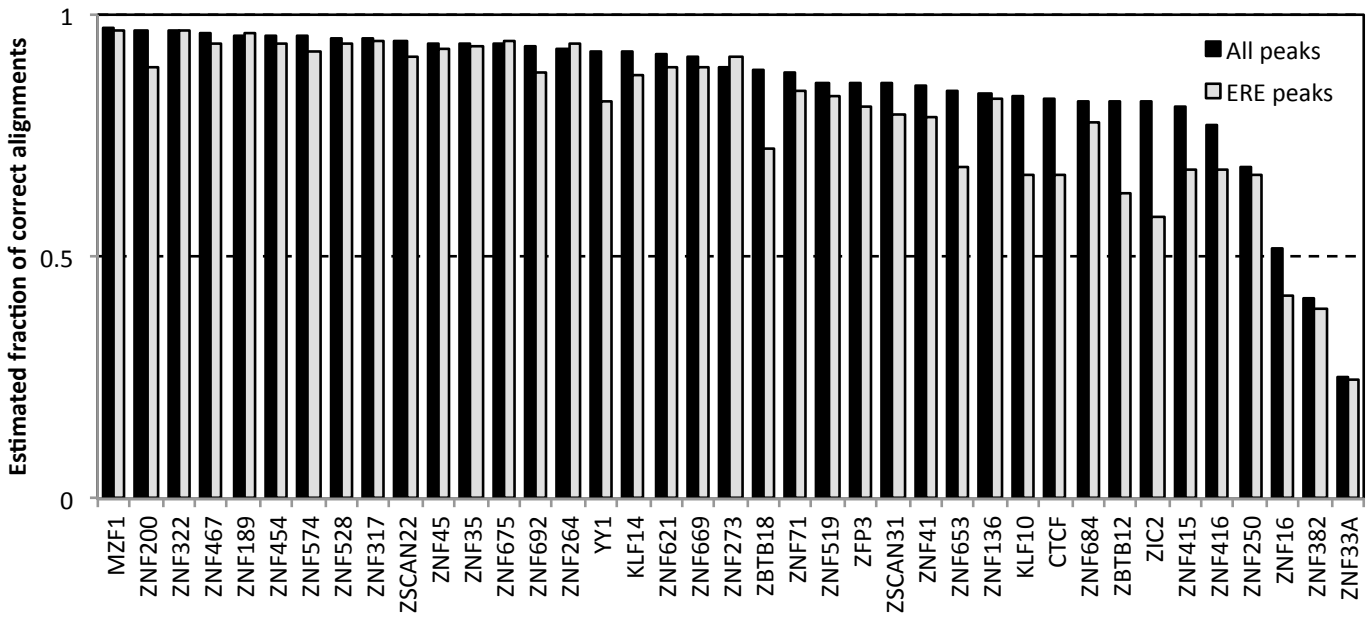
Supplementary Figure 7. Selective sweep at ZNF33A locus. The selective sweep score [6] measures the frequency of derived (non-ancestral) alleles in the human genome to the frequency of derived alleles in the corresponding Neanderthal genomic section, with negative numbers reflecting lower frequencies in the Neanderthal genome, suggesting selective sweep by early positive selection in the human lineage. The left panel shows the ZNF33A locus. The two other panels are for comparison: the middle panel shows the locus for KLF10, a C2H2 zinc finger protein that is not expected to have undergone positive selection, and the right panel illustrates the locus for SLC45A2, a skin color gene under strong positive selection [7].



Supplementary Figure 8. Alternative splicing of KRAB proteins. (a) ZNF189 is most highly expressed in heart and in iPS cells. The major isoform that is expressed in heart, however, contains a cassette exon with an in-frame stop codon, resulting in translation of the ZNF189 mRNA from a downstream start codon and, thus, skipping the KRAB domain. Data are from ref [4]. Thick black boxes represent coding sequence regions. **(b)** Exons of KRAB protein-coding genes generally show a higher degree of tissue-specific alternative splicing compared to exons of other C2H2-ZF protein-coding genes. Tissue-specific splicing for each exon was calculated as the maximum "percentage spliced-in" (PSI) minus the minimum PSI across 37 tissues/cell lines [4]. Only exons that affect the coding sequence are included. For KRAB proteins, 42% of alternatively spliced exons affect the KRAB domain, and an additional 54% affect the non-zinc finger part of the protein, often at the N-terminal side of the zinc fingers. *P*-value was calculated using Mann-Whitney U test.



Supplementary Figure 9. Uniqueness of sequences recognized by C2H2-ZF proteins and other TFs. **(a,b)** Affinity propagation (AP) clustering [8] of motifs for 39 human C2H2-ZF proteins. Panel **(a)** shows the clustering of motifs predicted by B1H-RC, and panel **(b)** shows the clustering of *de novo* ChIP-seq motifs that are most similar to B1H-RC motifs. **(c)** The standard curve of motif diversity vs. number of proteins, constructed by sub-sampling of different numbers of motifs from panel **(b)**, followed by AP clustering. Extrapolation of the fitted curve suggests that ~450 distinct motifs are encoded by ~720 human C2H2-ZF proteins. **(d)** Roughly the same estimate is obtained from clustering B1H-RC motifs predicted for single-array ZF proteins with 4-7 C2H2-ZFs. **(e)** Affinity propagation of all available motifs for non-C2H2 human TFs, either obtained directly from experiment or by homology mapping of motifs (Weirauch et al., Cell, *In press*) (1072 motifs for 430 TFs). **(f)** Standard curve of motif diversity against protein number for non-C2H2 TFs. Extrapolation suggests that a total of ~350 distinct motifs are encoded by ~950 non-C2H2 human TFs.



Supplementary Figure 10. Fraction of correctly mapped reads in peak regions. MAPQ scores were used to estimate the fraction of correctly mapped reads that overlap peaks in each experiment. Probability of correct mapping for each read was calculated from its mapping quality (MAPQ) score as p=1-exp($10^{-MAPQ/10}$), and the average of p of all peak-overlapping reads was calculated for each experiment in order to obtain the expected fraction of correctly mapped reads.

Supplementary Table 1. Summary of PBM results and comparison to B1H. C2H2-ZFs are identified by their OLS IDs. For C2H2-ZFs with both B1H and PBM motifs, the base selectivity, represented by maximum PWM score, at different positions is compared between the two motifs. The overall Pearson correlation of B1H vs. PBM motifs is also indicated.

OLS ID	Organism	OLS category	PBM status	Max PWM score								Correlation of PBM vs. B1H
				Pos. 1		Pos. 2		Pos. 3		Triplet		
				B1H	PBM	B1H	PBM	B1H	PBM	B1H	PBM	
				0.7	1.2	0.7	1.1	0.7	1.4	1.6	2.4	1.0
15445	Mus musculus	B1H-set	Success									
25707	Anole lizard	B1H-set	Success									
15438	Cnidaria	B1H-set	Success									
17172	Aedes aegypti	B1H-set	Success									
43353	Sea squirt	B1H-set	Success									
45719	Zebrafish	B1H-set	Success									
30467	Anole lizard	B1H-set	Success									
36768	Anopheles mosquito	B1H-set	Success									
17239	C. neoformans	B1H-set	Success									
16521	P. nodorum	B1H-set	Success									
26686	Anole lizard	B1H-set	Success									
21597	Anopheles gambiae	B1H-set	Success									
17171	Anopheles gambiae	B1H-set	Success									
23734	Equus caballus	B1H-set	Success									
23735	Mus musculus	B1H-set	Success									
40570	Dog	B1H-set	Success									
26093	Anole lizard	B1H-set	Success									
43867	Sea squirt	B1H-set	Success									
4108	Macaca fascicularis	B1H-set	Success									
3264	Mus musculus	B1H-set	Success									
39058	Cow	B1H-set	Success									
42271	Sloth	B1H-set	Success									
37919	Cow	B1H-set	Success									
25025	Anole lizard	B1H-set	Success									
15886	Aedes aegypti	B1H-set	Success									
26528	Anole lizard	B1H-set	Success									
5967	Mus musculus	B1H-set	Success									
3489	Homo sapiens	B1H-set	Success									
31376	Anole lizard	B1H-set	Success									
12402	Anopheles gambiae	B1H-set	Success									
13399	Homo sapiens	B1H-set	Success									
13400	T. nigroviridis	B1H-set	Success									
29907	Anole lizard	B1H-set	Success									
7212	Mus musculus	B1H-set	Success									
37961	Cow	B1H-set	Success									
15707	Homo sapiens	B1H-set	Success									
15685	Brachydanio rerio	B1H-set	Success									
37169	Cow	B1H-set	Success									
43744	Sea squirt	B1H-set	Success									
37542	Cow	B1H-set	Success									
6114	Mus musculus	B1H-set	Success									
42425	Sloth	B1H-set	Success									
3912	Homo sapiens	B1H-set	Success									
46758	Zebrafish	B1H-set	Success									
45173	Zebrafish	B1H-set	Success									
19114	Anopheles gambiae	B1H-set	Success									
44516	Sea squirt	B1H-set	Success									
12109	Aspergillus oryzae	B1H-set	Success									
13124	Dugesia japonica	B1H-set	Success									
15673	Primates	B1H-set	Success									
34341	Anole lizard	B1H-set	Success									
13741	Homo sapiens	B1H-set	Success									
11781	Brachydanio rerio	B1H-set	Success									
11124	Xenopus laevis	B1H-set	Success									
12682	Ciona intestinalis	B1H-set	Success									
29680	Anole lizard	B1H-set	Success									
5926	Aedes aegypti	B1H-set	Success									
9308	Homo sapiens	B1H-set	Success									
12681	Mus musculus	B1H-set	Success									
39611	Cow	B1H-set	Success									
27985	Anole lizard	B1H-set	Success									
9753	Ciona intestinalis	B1H-set	Success									
38776	Cow	B1H-set	Success									
15605	Mus musculus	B1H-set	Success									
24970	Anole lizard	B1H-set	Success									
42469	Sloth	B1H-set	Success									
27384	Anole lizard	B1H-set	Success									
14526	T. nigroviridis	B1H-set	Success									
26550	Anole lizard	B1H-set	Success									
4626	Macaca fascicularis	B1H-set	Success									
7460	Mus musculus	B1H-set	Success									
10605	Aedes aegypti	B1H-set	Success									
37855	Cow	B1H-set	Success									
27479	Anole lizard	B1H-set	Success									
24033	C. briggsae	B1H-set	Success									
15531	D. melanogaster	B1H-set	Success									
5322	Homo sapiens	B1H-set	Success									
44831	Sea squirt	B1H-set	Success									
16355	Sus scrofa	B1H-set	Success									
6120	Homo sapiens	B1H-set	Success									
31820	Anole lizard	B1H-set	Success									
37146	Cow	B1H-set	Success									
14627	D. melanogaster	B1H-set	Success									
26892	Anole lizard	B1H-set	Success									
6212	Mus musculus	B1H-set	Success									
17480	D. melanogaster	B1H-set	Success									
22787	S. cerevisiae	B1H-set	Success									
10732	Anopheles gambiae	B1H-set	Success									
37799	Cow	B1H-set	Success									
32786	Anole lizard	B1H-set	Success									
2820	Mus musculus	B1H-set	Success									
38732	Cow	B1H-set	Success									
28598	Anole lizard	B1H-set	Success									
40844	Dog	B1H-set	Success									
17269	Neurospora crassa	B1H-set	Success									
25474	Anole lizard	B1H-set	Success									
16467	Brachydanio rerio	B1H-set	Success									
44223	Sea squirt	B1H-set	Success									
31276	Anole lizard	B1H-set	Success									
30111	Anole lizard	B1H-set	Success									
23321	Gibberella zeae	B1H-set	Success									
34199	Anole lizard	B1H-set	Success									
45658	Zebrafish	B1H-set	Success									
25245	Anole lizard	B1H-set	Success									

OLS ID	Organism	OLS category	PBM status
10036	Mus musculus	B1H-set	Fail
13015	D. melanogaster	B1H-set	Fail
15419	Bos taurus	B1H-set	Fail
30200	Anole lizard	B1H-set	Fail
692	Ashbya gossypii	Random	Success
1483	Mus musculus	Random	Success
3475	Homo sapiens	Random	Success
4251	Mus musculus	Random	Success
5992	D. pseudoobscura	Random	Success
6002	Homo sapiens	Random	Success
7675	Mus musculus	Random	Success
7754	Homo sapiens	Random	Success
9599	Homo sapiens	Random	Success
10510	Xenopus laevis	Random	Success
13379	D. melanogaster	Random	Success
13699	Mus musculus	Random	Success
15000	Anopheles gambiae	Random	Success
15660	Ciona intestinalis	Random	Success
15873	Homo sapiens	Random	Success
16252	S. cerevisiae	Random	Success
16285	Mus musculus	Random	Success
16320	Mus musculus	Random	Success
16400	D. melanogaster	Random	Success
16727	S. cerevisiae	Random	Success
17492	D. melanogaster	Random	Success
17821	Oryza sativa	Random	Success
19279	Aedes aegypti	Random	Success
20426	T. nigroviridis	Random	Success
20496	Aedes aegypti	Random	Success
23312	C. globosum	Random	Success
23322	C. globosum	Random	Success
24152	Ciona intestinalis	Random	Success
26471	Anole lizard	Random	Success
28130	Anole lizard	Random	Success
32622	Anole lizard	Random	Success
32985	Anole lizard	Random	Success
33254	Anole lizard	Random	Success
35742	Anole lizard	Random	Success
38884	Cow	Random	Success
39218	Cow	Random	Success
39596	Cow	Random	Success
40940	Dog	Random	Success
41340	Guinea pig	Random	Success
43502	Sea squirt	Random	Success
46164	Zebrafish	Random	Success
46986	Zebrafish	Random	Success
6521	Brachydanio rerio	Random	Fail
7645	Mus musculus	Random	Fail
11096	Mus musculus	Random	Fail
11332	Mus musculus	Random	Fail
13847	S. cerevisiae	Random	Fail
14324	D. pseudoobscura	Random	Fail
14741	Aedes aegypti	Random	Fail
14973	Gallus gallus	Random	Fail
15213	Aedes aegypti	Random	Fail
17701	Aedes aegypti	Random	Fail
17739	T. nigroviridis	Random	Fail
17763	Aedes aegypti	Random	Fail
17810	Xenopus laevis	Random	Fail
19145	S. cerevisiae	Random	Fail
20539	D. melanogaster	Random	Fail
20649	D. pseudoobscura	Random	Fail
21173	D. melanogaster	Random	Fail
21435	T. nigroviridis	Random	Fail
21685	Brachydanio rerio	Random	Fail
22766	T. nigroviridis	Random	Fail
23596	Mus musculus	Random	Fail
29654	Anole lizard	Random	Fail
31325	Anole lizard	Random	Fail
31996	Anole lizard	Random	Fail
33867	Anole lizard	Random	Fail
35605	Anole lizard	Random	Fail
37095	A. mosquito	Random	Fail
38336	Cow	Random	Fail
39956	Cow	Random	Fail
41045	Dog	Random	Fail
42925	Sloth	Random	Fail
43127	Sloth	Random	Fail
44906	Sea squirt	Random	Fail
46118	Zebrafish	Random	Fail
46932	Zebrafish	Random	Fail

REFERENCES

1 Elrod-Erickson, M., Rould, M. A., Nekludova, L. & Pabo, C. O. Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. *Structure* **4**, 1171-1180 (1996).

2 Thomas, J. H. & Schneider, S. Coevolution of retroelements and tandem zinc finger genes. *Genome Res* **21**, 1800-1812, doi:10.1101/gr.121749.111 (2011).

3 Bailey, T. L. & Machanick, P. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res* **40**, e128, doi:10.1093/nar/gks433 (2012).

4 Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172-177, doi:10.1038/nature12311 (2013).

5 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, doi:10.1101/gr.152140.112 (2013).

6 Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710-722, doi:10.1126/science.1188021 (2010).

7 Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913-918, doi:10.1038/nature06250 (2007).

8 Frey, B. J. & Dueck, D. Clustering by passing messages between data points. *Science* **315**, 972-976, doi:10.1126/science.1136800 (2007).