G3 Genes | Genomes | Genetics

INVESTIGATION

Comparison of ChIP-Seq Data and a Reference Motif Set for Human KRAB C2H2 Zinc Finger Proteins

Marjan Barazandeh,* Samuel A. Lambert,† Mihai Albu,* and Timothy R. Hughes*,†,‡,1
*Terrence Donnelly Centre for Cellular and Biomolecular Research and †Department of Molecular Genetics, University of Toronto, Ontario M5S 1A8, Canada and ‡Canadian Institutes for Advanced Research, Toronto, Ontario M5G 1M1, Canada
ORCID IDs: 0000-0001-8222-008X (S.A.L.); 0000-0002-8721-4719 (T.R.H.)

ABSTRACT KRAB C2H2 zinc finger proteins (KZNFs) are the largest and most diverse family of human transcription factors, likely due to diversifying selection driven by novel endogenous retroelements (EREs), but the vast majority lack binding motifs or functional data. Two recent studies analyzed a majority of the human KZNFs using either ChIP-seq (60 proteins) or ChIP-exo (221 proteins) in the same cell type (HEK293). The ChIP-exo paper did not describe binding motifs, however. Thirty-nine proteins are represented in both studies, enabling the systematic comparison of the data sets presented here. Typically, only a minority of peaks overlap, but the two studies nonetheless display significant similarity in ERE binding for 32/39, and yield highly similar DNA binding motifs for 23 and related motifs for 34 (MoSBAT similarity score >0.5 and >0.2, respectively). Thus, there is overall (albeit imperfect) agreement between the two studies. For the 242 proteins represented in at least one study, we selected a highest-confidence motif for each protein, utilizing several motif-derivation approaches, and evaluating motifs within and across data sets. Peaks for the majority (158) are enriched (96% with AUC >0.6 predicting peak *vs.* nonpeak) for a motif that is supported by the C2H2 "recognition code," consistent with intrinsic sequence specificity driving DNA binding in cells. An additional 63 yield motifs enriched in peaks, but not supported by the recognition code, which could reflect indirect binding. Altogether, these analyses validate both data sets, and provide a reference motif set with associated quality metrics.

KEYWORDS
KRAB C2H2 zinc finger proteins
endogenous retroelements
DNA-binding motif
ChIP-seq
C2H2 recognition code

The human genome encodes ~350 KRAB C2H2 zinc finger proteins (KZNFs), which encode a Kruppel-Associated Box (KRAB) domain, which is best known for its repressor activity (Schultz *et al.* 2002), followed by a tandem array of C2H2 zinc finger (ZNF) domains (up to 40), which mediate sequence-specific DNA binding. The ZNFs each contact three or more bases, and typically bind in tandem with an offset of three bases. The DNA sequence motifs recognized by the array of ZNFs therefore often resemble concatenation of the base preferences for the individual ZNFs (Wolfe *et al.* 2000), which can be predicted to some degree on the basis of "specificity residues" at positions –1, 2, 3, and 6 of the DNA-contacting α helix (Najafabadi *et al.* 2015a). The modular fashion of DNA recognition by C2H2-ZF proteins apparently facilitates adaptation, with evidence for positive selection on the specificity residues of many KZNFs (Emerson and Thomas 2009), such that many encode a unique sequence specificity (Najafabadi *et al.* 2015a).

The best-characterized function of the KRAB domain is to recruit TRIM28 (aka KAP1), which represses transcription by subsequent recruitment of SETDB1, a histone H3 lysine 9 (H3K9) trimethylase (Schultz *et al.* 2002). TRIM28 is involved in silencing endogenous retroelements (ERE); this observation led to the now widely accepted theory that KZNFs evolve rapidly to silence EREs (Matsui *et al.* 2010; Rowe and Trono 2011). A KZNFs *vs.* EREs "arms race" model (Jacobs *et al.* 2014) provides a readily understood mechanism for the evolution of new KZNFs, but does not explain the retention of so many of them; presumably, they take on other host functions. The KRAB domains vary in primary sequence, with an average sequence identity of ~40%, and KZNFs vary in their protein–protein interactions (Schmitges *et al.* 2016), suggesting that they may also vary in effector function.

doi: https://doi.org/10.1534/g3.117.300296
Manuscript received September 22, 2017; accepted for publication November 8, 2017; published Early Online November 16, 2017.

Supplemental material is available online at www.g3journal.org/lookup/suppl/doi:10.1534/g3.117.300296/-/DC1.
[1] Corresponding author: Room 1302, 160 College St., Toronto, ON M5S 3E1, Canada. E-mail: t.hughes@utoronto.ca

Detailed study of the KZNFs requires knowledge of their intrinsic DNA binding preferences (*i.e.*, motifs), and their genomic binding sites. General characteristics of their motifs can be gleaned from their protein sequences, *e.g.*, that two proteins should bind distinct motifs, but current "recognition codes" are too error-prone to obtain high-confidence motifs (Najafabadi *et al.* 2015b). Furthermore, it remains difficult to determine which C2H2 domains in a large array are likely to bind DNA (Brayer and Segal 2008); KZNFs typically carry many more tandem ZNFs than needed to specify individual loci in the human genome (2–40; median 12), and known motifs for human ZNF proteins often correspond to only part of the ZNF array (Najafabadi *et al.* 2015a). An additional challenge with ZNF proteins is that they often fail to yield motifs from *in vitro* analyses such as Protein Binding Microarrays (Badis *et al.* 2009; Wang 2014) and HT-SELEX (Jolma *et al.* 2013), possibly due to misfolding, lack of obligate cofactors, or the fact that many KZNFs have long binding sites, which would be poorly represented among the sequences in these experiments. Systematic application of these approaches has consequently yielded motifs for very few of the KZNFs (Badis *et al.* 2009; Jolma *et al.* 2013, 2015; Weirauch *et al.* 2013). Other approaches to determining sequence specificity of C2H2 proteins have been described, including SMILE-seq (Isakova *et al.* 2017), B1H (Noyes *et al.* 2008), and analysis of selected sites by context-dependent models (Zhao *et al.* 2009), but, to our knowledge, none has been tested on a large number of human proteins.

ChIP-seq and related methods [*e.g.*, ChIP-exo (Imbeault *et al.* 2017)] represent an alternative to rapidly obtain motifs for C2H2 zinc finger proteins. ChIP-seq is a relatively challenging means to obtain high-confidence motifs for TFs, due to the fact that the number of binding sites is relatively small (in comparison to, *e.g.*, HT-SELEX), that the genome sequence is highly nonrandom (with many sequence biases and repeated sequences), and that proteins can associate with DNA indirectly, such that cofactor motifs are often obtained (Encode-Project-Consortium 2012; Mathelier and Wasserman 2013). For KZNF proteins, the fact that EREs are often bound represents an additional confounding variable, because motif-finding tools [*e.g.*, MEME (Bailey *et al.* 2009)] typically assume that different binding sites are drawn from independent random sequences, while EREs are related by common origin. To address these issues, we have previously described Recognition Code-Assisted Discovery of regulatory Elements (RCADE), a computational method that employs the zinc finger recognition code to predict primary binding motifs from ChIP-seq data specifically for C2H2-ZF proteins (Najafabadi *et al.* 2015a). We utilized RCADE in two large-scale analyses of human C2H2-ZFs proteins (using tagged, inducible heterologous expression constructs in HEK293 cells) to produce motifs for dozens of human C2H2-ZFs (Najafabadi *et al.* 2015b; Schmitges *et al.* 2016).

A recent study described a similar ChIP-exo analysis of 221 human KZNFs (Imbeault *et al.* 2017), also in HEK293 cells (but using different constructs). This study did not describe DNA binding motifs, however. Here, we applied our existing data analysis pipeline to the data from the Imbeault *et al.* (2017) study, and compared the results to our own, and also to HT-SELEX, SMILE-seq, and other motifs from the literature, where available. Overall, the comparison reveals that the two data sets usually produce comparable motifs, and also largely agree on the sets of EREs bound by the individual KZNFs. This agreement is obtained even though most of the individual peaks do not overlap, suggesting that both studies sample from a larger set of genomic target sites. We introduce a web portal that summarizes the results of our analyses (http://kznfmotifs.ccbr.utoronto.ca/), and produce a set of high-confidence motifs, each together with a confidence statistic that represents its ability to predict binding sites in living cells.

MATERIALS AND METHODS

Reprocessing the ChIP-exo data

Primary and processed ChIP-seq and ChIP-exo data were obtained from GEO accession numbers GSE76496 (Schmitges *et al.* 2016), and GSE78099 (Imbeault *et al.* 2017). The Trono ChIP-exo data were reprocessed as previously described (Najafabadi *et al.* 2015b; Schmitges *et al.* 2016). Briefly, the ChIP-exo reads were trimmed to 50 nucleotides and mapped to the human genome (hg19; build GRCh37) using Bowtie 2 (Langmead and Salzberg 2012). We used the "–very-sensitive" preset option, which allows the retention of one alignment for the multi-mapped reads; otherwise many ERE instances cannot be detected. The multi-mapping may affect the reads containing EREs due to their similar sequences by aligning them to the genomic repetitive regions other than their origin, and could result in erroneous peak calling. However, as we and others have described before (Najafabadi *et al.* 2015b), this is unlikely to be a problem regarding the enrichment of EREs, since these reads typically map to another instance of the same ERE, and vice versa (Day *et al.* 2010).

MACS 1.4.2 was used for peak calling (Zhang *et al.* 2008), where the provided input DNA data set (a combination of randomly selected peaks from all ChIP-exo) was used as the control. The biological replicates were merged into single samples, retaining all the peaks from all replicates, and merging the peaks in a maximum of 50 bp distance from each other into a single peak. For further analyses, the top 500 peaks with the highest enrichment scores were picked.

Motif analysis

Non-ERE peaks were filtered using RepeatMasker track of the UCSC Table Browser (Karolchik *et al.* 2004). We used RCADE (Najafabadi *et al.* 2015a) using default settings and MEME (Bailey *et al.* 2009) with CentriMo, on both all and non-ERE peaks, giving priority to those obtained by RCADE and in non-ERE peaks, as will be described in the text. For further detail, see Schmitges *et al.* (2016). Enrichment of motifs from ChIP data and other sources were tested by calculating AUROC for differentiating the top 500 ChIP peaks in a range of ±250 bp from the summit from dinucleotide shuffled sequences, using the single maximum PWM scoring match for each 501-base sequence as its score.

Motif similarity analysis

Motif similarity was measured using the Motif Similarity Based on Affinity of Targets (MoSBAT) method (Lambert *et al.* 2016) using the energy scores option. To make similarity measurements consistent across datasets we set the sequence background to $n = 100{,}000$ and $l = 100$.

ERE enrichment

The ERE enrichment was calculated as the proportion of the top 500 peaks overlapping any transposon and retroelement instances present in the RepeatMasker. The Pearson correlations (r) between the overlapped EREs of all possible Hughes and Trono pairs (matching and unmatching within the same set of 39) were calculated using R (v. 3.2.3).

Data availability

The detailed results of our analyses for the 242 KZNFs are available at the web portal of the paper (http://kznfmotifs.ccbr.utoronto.ca/). For each KZNF, it contains motifs from different sources, a heatmap of the pairwise motif similarity scores between different motifs, the overlap between all peaks and top 500 peaks of the Hughes and Trono ChIP data, where available, and the ERE enrichment of the peaks.

RESULTS

Figure 1 presents an overview of the data analyzed in this study and the analysis steps, the results of which are detailed below. Briefly, we compiled ChIP-seq ("Hughes") or ChIP-exo ("Trono") data for KZNF proteins from two different studies. Hughes data (60 KZNFs) is from Schmitges *et al.* 2016, while Trono data (221 KZNFs) is from Imbeault *et al.* (2017). We also identified motif data from other sources that correspond to each of the proteins in the combined list of 242 KZNFs (Matys *et al.* 2006; Encode-Project-Consortium 2012; Jolma *et al.* 2013; Kulakovskiy *et al.* 2013; Najafabadi *et al.* 2015b; Mathelier *et al.* 2016; Isakova *et al.* 2017; Yin *et al.* 2017). The Venn diagram in Supplemental Material, Figure S1 shows the overlaps among the data types.

For all analyses, we used two versions of the peaks from the Trono data. The first is the set of peaks reported in the original paper (Imbeault *et al.* 2017), and the second is the set of peaks obtained by rerunning the raw Trono reads using the same peak-calling system employed for the Hughes data. These "Trono reprocessed" peaks are, on average, 10-fold more numerous than the "Trono original" peaks, but they are only 1.8-fold less compared to those from the Hughes data, for the 39 KZNFs that overlap (average 40,660 for Hughes and 22,483 for Trono reprocessed).

For each KZNF, we computed the overlap in peaks (for the 39 present in both Hughes and Trono data), and the proportion of the Top 500 peaks (or maximum number of peaks where the number of peaks was <500) that overlap with each of 934 types of transposons and retroelements cataloged by the RepeatMasker track of the UCSC Table Browser (Karolchik *et al.* 2004). We also generated new motifs (using both MEME and RCADE) for both the Trono original and Trono reprocessed peak sets. As in our previous studies, we favored RCADE motifs, which are supported by the recognition code and therefore more likely to represent primary binding sites, but retained MEME motifs if RCADE produced no results [(a result of the algorithm failing to converge, or nonenrichment of any of the predicted motifs in the ChIP-seq peaks; see Najafabadi *et al.* (2015a)]. We also favored motifs derived from non-ERE peaks. We then scored the similarity among the motifs for each protein (choosing one per data set), and also scored the area under the receiver operating characteristic curve (AUROC) for all available motifs on the Hughes, the Trono original, and the Trono reprocessed data sets. The AUROC scores reflect the ability of the motif to discriminate peak sequences from the background (dinucleotide shuffled peaks). Finally, we selected a representative motif for each protein, favoring those with evidence for direct binding, and those with the highest AUROC on any ChIP data type. The web site accompanying this paper (http://kznfmotifs.ccbr.utoronto.ca/) contains files used in all the analyses herein, including the numerical data underlying the figures, as well as a visualization of the motifs and all the analyses for each protein considered. It also contains all the motifs as PWMs, and lists the representative motifs, together with confidence metrics for each: AUROC on ChIP data, similarity to most similar independently generated motif (if available), and method generated (RCADE, MEME, or external source).

Overlap in peaks and EREs bound between Hughes and Trono data

We first considered the overlap in peaks and the overlap in EREs bound between the Hughes and Trono data, for the 39 proteins that overlapped between the two studies. Figure 2 shows the percent peak overlap for each protein, calculated as the percentage of the top 500 Hughes peaks that overlap Trono peaks and vice versa in a range of ±250 bp from the peak summits, with the four bar graphs representing comparisons between Hughes data and Trono original and Trono reprocessed data. On average, 35% of the Top 500 Hughes and Trono reprocessed peaks overlapped in both comparisons, albeit with a considerable spread. At random, the overlap should be zero in all these comparisons, because the 500 peaks encompass a miniscule fraction of the genome. We observed no correlation between the degree of peak overlap and quality scores of the ChIP experiments (Kharchenko *et al.* 2008; Landt *et al.* 2012) (data not shown).

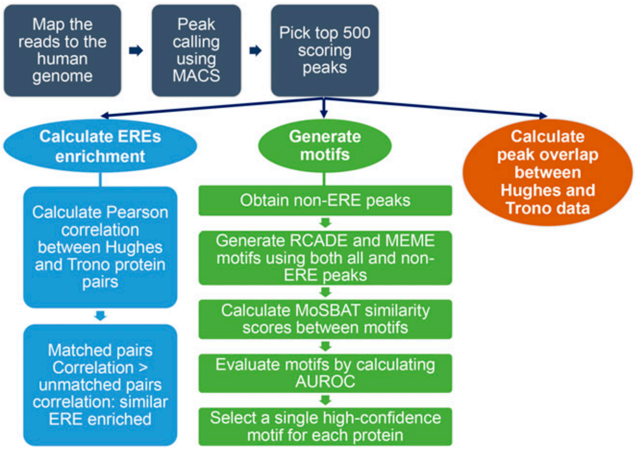

Map the reads to the human genome
Peak calling using MACS
Pick top 500 scoring peaks
Calculate EREs enrichment
Generate motifs
Calculate peak overlap between Hughes and Trono data
Calculate Pearson correlation between Hughes and Trono protein pairs
Matched pairs Correlation > unmatched pairs correlation: similar ERE enriched
Obtain non-ERE peaks
Generate RCADE and MEME motifs using both all and non-ERE peaks
Calculate MoSBAT similarity scores between motifs
Evaluate motifs by calculating AUROC
Select a single high-confidence motif for each protein

Figure 1 Overview of the data analysis steps and methods utilized in this study.

Our use of 500 peaks throughout is for convenience and uniformity; different numbers and proportions of peaks will yield slightly different numbers, but with generally similar conclusions (data not shown). For example, we note that the Trono original peaks often have a generally higher proportion of overlap with Hughes, relative to all other comparisons made. This phenomenon is explained by the data processing; there are a substantially lower number of peaks in the Trono original data (<500 in 17/39 KZNFs compared) and they are presumably the highest enriched based on the description in Imbeault *et al.* (2017). This trend is not evident with the Trono reprocessed data, and similar (albeit again not identical) outcomes were obtained with Trono original and Trono reprocessed data in the analyses below.

Figure 3 provides a detailed view of the agreement in EREs bound in the two studies. Our comparison statistic is the Pearson correlation across all ERE classes, where the value for each is the proportion represented among the Top 500 peaks. Figure 3A shows that the distribution of values for the 39 matched KZNFs between the Hughes and Trono reprocessed data sets is much different from that for mismatched KZNFs; 82% (32/39) of the matched pairs exceed a correlation achieved by only 8% of the mismatched pairs. The distribution is quite similar when comparing Hughes and Trono original data, with 82% of the matched pairs exceed a correlation achieved by 5% of the mismatched pairs (data not shown). Figure 3B provides a visual confirmation that the individual transposons and EREs types represented in the three peak sets for each of the 39 proteins are largely in agreement.

Overall, we take this outcome to indicate that a large majority of data in both data sets correctly identifies the spectrum of EREs recognized, assuming that the overlapping KZNFs are an unbiased sample from each of the two studies. There can be good agreement on the EREs bound even when the peak overlap is relatively low; however, the higher peak overlap is usually associated with higher EREs correlation (Figure 3A). One interpretation of these observations is that both data sets are drawing from a substantially larger set of *bona fide* genomic binding sites, but both are subject to noise, and neither has been sequenced to saturation.

Similarity of motifs from Hughes, Trono, and external data

We next compared the motifs obtained from the two ChIP data sets for the 39 overlapping proteins, as well as against the motifs obtained from

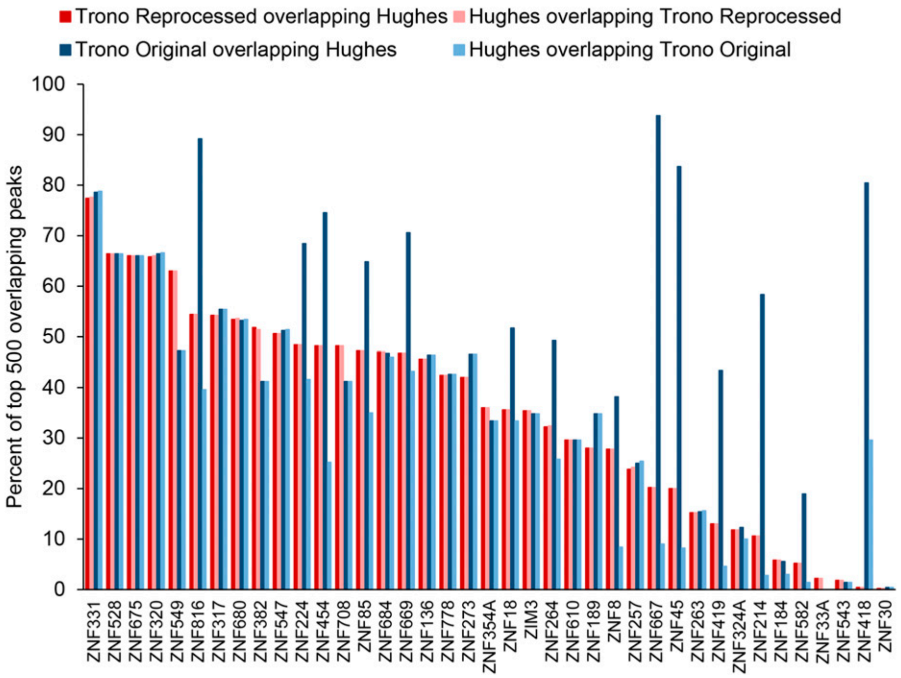


Figure 2 Peak overlaps for the 39 shared KZNFs between the Trono original and Hughes data (dark blue bars), Hughes data and Trono original data (light blue bars), Trono reprocessed and Hughes data (dark red bars), and Hughes and Trono reprocessed data (light red bars).

other sources. For Hughes data, we used the motifs directly from the publication Schmitges *et al.* (2016). For the Trono data, we generated motifs for both the Trono original and Trono reprocessed peaks, using the same procedure employed in Schmitges *et al.* (2016). Briefly, we used either RCADE (Najafabadi *et al.* 2015a) or MEME (Bailey *et al.* 2009) to identify motifs based on the ±250 bp sequences around the summits of the top 500 peaks of each protein. We used motifs obtained by RCADE, using non-ERE peaks, if the algorithm was successful, since the fact that EREs within a given class are related by common descent can confound motif-finding algorithms. If non-ERE peaks did not generate significant motifs, we included the ERE peaks in the analysis. We used MEME to generate motifs if no RCADE motifs were obtained, selecting the top-scoring motif using CentriMo; similar to the RCADE motifs, priority was given to the motifs obtained by non-ERE peaks. Overall, 80% (48/60) of the motifs from the Hughes data, 74% (164/221) from the Trono original data, and 77% (170/221) from the Trono reprocessed were obtained from RCADE. RCADE was successful in at least one of the three data sets in the 39 shared proteins, and in many cases, on the same proteins from the Hughes and Trono laboratories (30/39 in common between Hughes, and at least one of the Trono original or Trono reprocessed data sets, 1/39 from Hughes only, 1/39 from Trono original only, and 7/39 from both Trono original and Trono reprocessed, but not Hughes).

To score similarity of motifs, we used MoSBAT (Lambert *et al.* 2016), which computes the Pearson correlation of the predicted affinities for two different motifs to thousands of randomly generated short sequences. MoSBAT is appropriate for this analysis because it is nonparametric, and requires no adjustments for differences in motif length. Figure 4A shows an example confirming similarity among multiple published motifs for ZIM3 (similar figures are shown for all KZNFs at http://kznfmotifs.ccbr.utoronto.ca/). In this example, only 35% of the top 500 Hughes peaks overlap Trono original and Trono reprocessed peaks, but RCADE produces very similar motifs with both data sets ($r = 0.90$ and $r = 0.89$, respectively). Figure 4B shows the motif similarity scores between the Hughes data and the Trono original and reprocessed data, and the corresponding motifs for the 39 overlapping KZNFs, illustrating that virtually all bear a clear visual resemblance, with only 3/39 (8%) and 7/39 (18%) of the KZNFs yielding similarity scores <0.10 between Hughes motifs and Trono original and Trono reprocessed motifs, respectively. Overall, roughly half of the overlapping KZNFs between the Hughes and Trono data (19/39; 49% for Trono Original and 18/39; 46% for Trono Reprocessed) yielded motif similarity scores >0.50, a value typically obtained from different experiments using the same transcription factor (TF) (Lambert *et al.* 2016). Furthermore, the majority of the overlapping KZNFs (33/39; 85% for Trono Original and 27/39; 69% for Trono Reprocessed) are clearly comparable by visual inspection (Figure 4B), and bear a similarity score of >0.20, which exceeds the score of the 98% of the mismatched KZNFs. As with the ERE overlap, in most cases the motifs tend to be similar, even when peak overlap between the two data sets is low. Higher peak overlap usually results in higher motif similarity, however (Figure 4C).

We also compared the ChIP-derived motifs with those from other sources, taken from the initial publications or TF databases. If there was more than one motif from the same source available for a KZNF, the one with the highest AUROC was selected for further analysis (41 in total; Matys *et al.* 2006; Encode-Project-Consortium 2012; Jolma *et al.* 2013; Kulakovskiy *et al.* 2013; Mathelier *et al.* 2016; Isakova *et al.* 2017; Yin *et al.* 2017), and 5/22 of Hughes KZNFs, 14/22 of Trono KZNFs, and 3/22 of both data sets overlapped with at least one external motif (Figure S1). For 23/41, the external in-vitro motif was similar (MoSBAT >0.2) to at least one ChIP-derived motif. Surprisingly, 17 of the external motifs outperformed the ChIP-derived motifs in predicting ChIP peaks (see the section *A reference motif set for human KZNFs* for details), among which six belonged to the SMiLE-seq motif sets (seven in total; Isakova *et al.* 2017), six to the HT-SELEX motif sets (12 in total; Jolma *et al.* 2013; Yin *et al.* 2017), and five to one of the other sources (10 in total; Matys *et al.* 2006; Encode-Project-Consortium 2012; Kulakovskiy *et al.* 2013). This outcome not only confirms that motifs derived *in vitro* are often consistent with *in vivo* genomic binding sites, but illustrates that the greater depth of the assays can produce more accurate motifs. Some cases, however, are discrepant (11/41 cases score <0.1 in MoSBAT compared to all corresponding ChIP-derived motifs), and will require resolution (see *Discussion*).

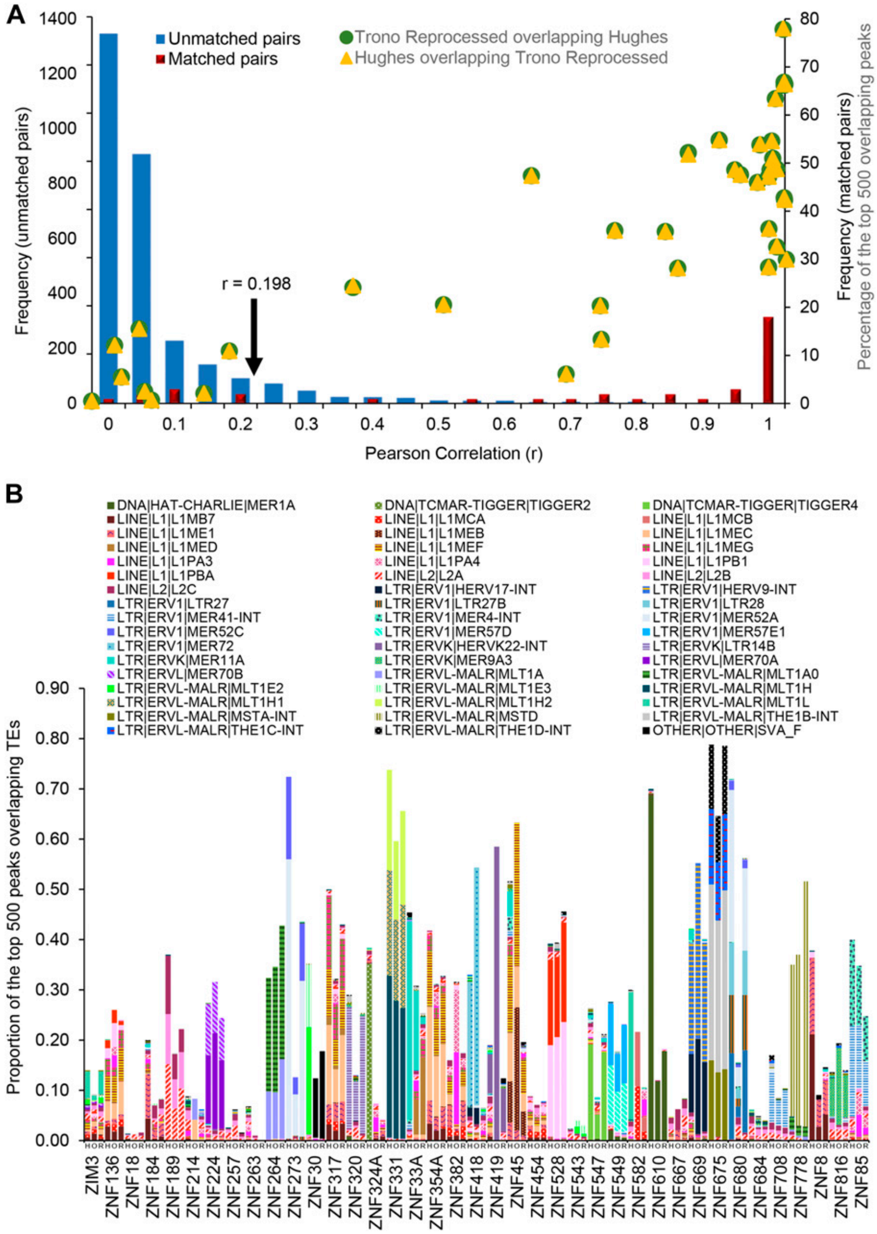


Figure 3 Overview of the ERE enrichment in Hughes and Trono ChIP data. (A) Pearson Correlation between the 39 Hughes and Trono reprocessed overlapping KZNFs (matched pairs; red bars) and nonoverlapping KZNFs (unmatched pairs: 2964 comparisons; blue bars) and the frequency of the KZNF pairs at each given correlation. The arrow indicates the correlation beyond which 82% of the matched pairs and 8% of the unmatched pairs lie. The percentage of the peak overlap between the Hughes and Trono reprocessed (yellow dots) and Trono reprocessed and Hughes (green dots) at corresponding correlations are also presented. (B) Fraction of the top 500 overlapping KZNFs enriched in TEs (ERE instances and transposons). In total, 51 single TE instances were enriched with a fraction of >0.1. H, Hughes; O, Trono Original; R, Trono Reprocessed.

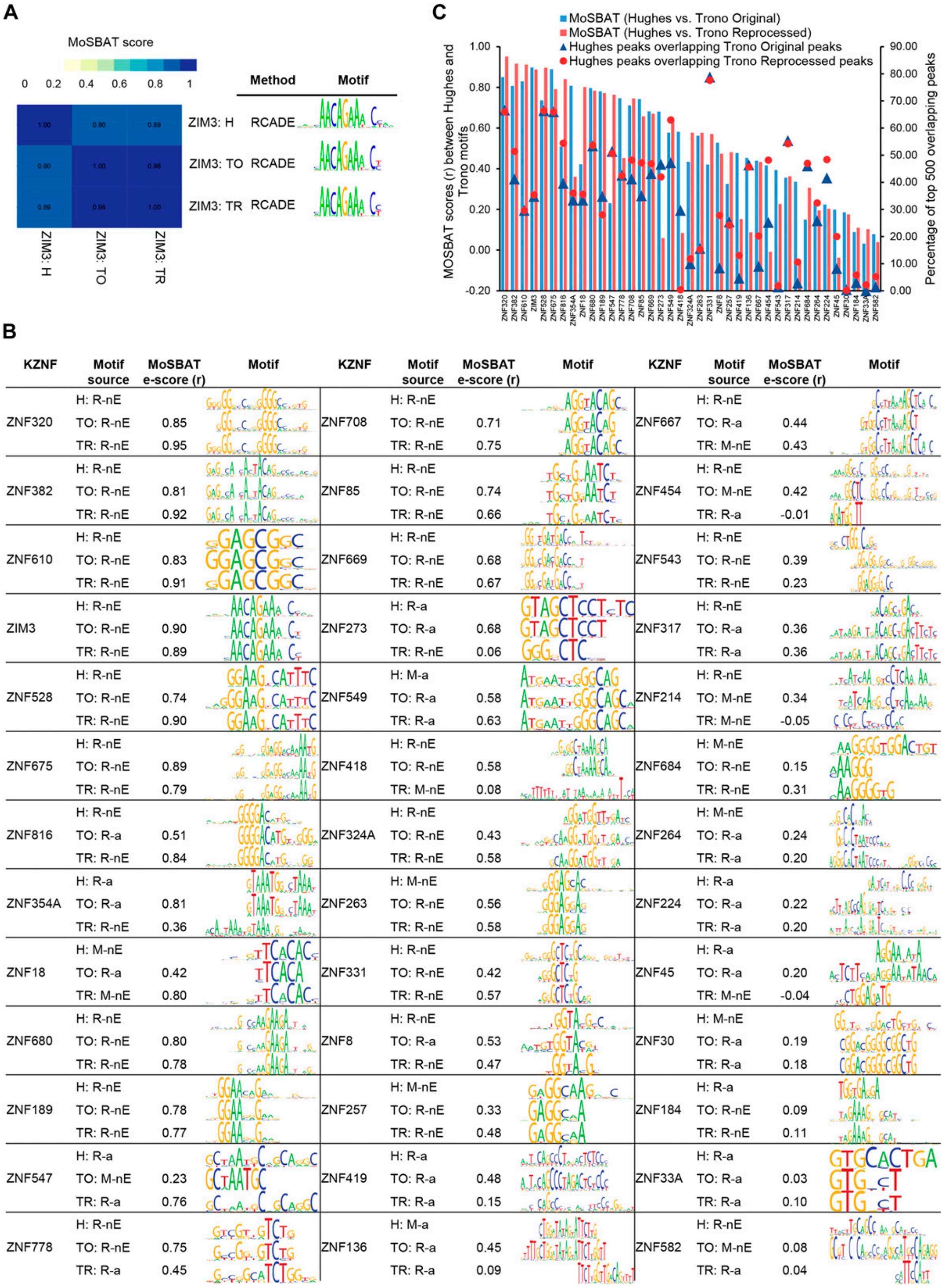
We also found that the vast majority of motifs obtained from one of the ChIP data sets were significantly enriched in peaks from the other. Figure 5 depicts AUROC values (Top 500 *vs.* dinucleotide shuffled sequences) for ChIP data and other sources tested on the three data sets: Hughes, Trono original, and Trono reprocessed. The left third of Figure 5 shows that, except for a few cases, the ChIP-derived motifs are comparable at predicting the ChIP peaks of the other data set, as well as the one from which they are derived. The right two-thirds of Figure 5 show that motifs from other sources are also generally good predictors of the ChIP peaks.

These analyses show that the two data sets typically identify similar binding motifs, most of which are also supported by the recognition code and/or by independent data.

A reference motif set for human KZNFs

The comparisons above provide confidence metrics for the motifs obtained for each of the 242 KZNFs examined, representing reproducibility and predictive power both within and among data types, consistency with the zinc finger recognition code, and quality statistics for the data used to derive the motifs. We used these metrics to choose a single best current motif for each protein, as such a motif set is useful for many types of analyses (Matys *et al.* 2006; Encode-Project-Consortium 2012; Kulakovskiy *et al.* 2013; Mathelier *et al.* 2016).

The ranking system, intended to capture motifs that correspond to both ChIP data and external information, if available, is shown in Figure S2. We gave highest confidence to any motif (*in vitro* or ChIP-derived) that predicts "test" ChIP data (the data that they were not trained on: one of the Hughes, Trono original, or Trono reprocessed data sets) with AUROC at least 0.1 greater than all other motifs (Motif is uniquely predictive of test peaks = Class A). In the absence of a Class A motif, we then favored *in vitro* derived motifs that predict the test ChIP data better than, or almost as well as, the ChIP motifs (AUROC difference <0.1; *In vitro* motif predictive of test peaks = Class B), since *in vitro* data are not impacted by extraneous factors such as indirect binding. If



KZNF	Motif source	MoSBAT e-score (r)	Motif
ZNF320	H: R-nE		
	TO: R-nE	0.85	
	TR: R-nE	0.95	
ZNF382	H: R-nE		
	TO: R-nE	0.81	
	TR: R-nE	0.92	
ZNF610	H: R-nE		
	TO: R-nE	0.83	
	TR: R-nE	0.91	
ZIM3	H: R-nE		
	TO: R-nE	0.90	
	TR: R-nE	0.89	
ZNF528	H: R-nE		
	TO: R-nE	0.74	
	TR: R-nE	0.90	
ZNF675	H: R-nE		
	TO: R-nE	0.89	
	TR: R-nE	0.79	
ZNF816	H: R-nE		
	TO: R-a	0.51	
	TR: R-nE	0.84	
ZNF354A	H: R-a		
	TO: R-a	0.81	
	TR: R-nE	0.36	
ZNF18	H: M-nE		
	TO: R-a	0.42	
	TR: M-nE	0.80	
ZNF680	H: R-nE		
	TO: R-nE	0.80	
	TR: R-nE	0.78	
ZNF189	H: R-nE		
	TO: R-nE	0.78	
	TR: R-nE	0.77	
ZNF547	H: R-a		
	TO: M-nE	0.23	
	TR: R-a	0.76	
ZNF778	H: R-nE		
	TO: R-nE	0.75	
	TR: R-a	0.45	
ZNF708	H: R-nE		
	TO: R-nE	0.71	
	TR: R-nE	0.75	
ZNF85	H: R-nE		
	TO: R-nE	0.74	
	TR: R-nE	0.66	
ZNF669	H: R-nE		
	TO: R-nE	0.68	
	TR: R-nE	0.67	
ZNF273	H: R-a		
	TO: R-a	0.68	
	TR: R-nE	0.06	
ZNF549	H: M-a		
	TO: R-a	0.58	
	TR: R-a	0.63	
ZNF418	H: R-nE		
	TO: R-nE	0.58	
	TR: M-nE	0.08	
ZNF324A	H: R-nE		
	TO: R-nE	0.43	
	TR: R-nE	0.58	
ZNF263	H: M-nE		
	TO: R-nE	0.56	
	TR: R-nE	0.58	
ZNF331	H: R-nE		
	TO: R-nE	0.42	
	TR: R-nE	0.57	
ZNF8	H: R-nE		
	TO: R-a	0.53	
	TR: R-nE	0.47	
ZNF257	H: M-nE		
	TO: R-nE	0.33	
	TR: R-nE	0.48	
ZNF419	H: R-a		
	TO: R-a	0.48	
	TR: R-a	0.15	
ZNF136	H: M-a		
	TO: R-a	0.45	
	TR: R-a	0.09	
ZNF667	H: R-nE		
	TO: R-a	0.44	
	TR: M-nE	0.43	
ZNF454	H: R-nE		
	TO: M-nE	0.42	
	TR: R-a	-0.01	
ZNF543	H: R-nE		
	TO: R-nE	0.39	
	TR: R-nE	0.23	
ZNF317	H: R-nE		
	TO: R-a	0.36	
	TR: R-a	0.36	
ZNF214	H: R-nE		
	TO: M-nE	0.34	
	TR: M-nE	-0.05	
ZNF684	H: M-nE		
	TO: R-nE	0.15	
	TR: R-nE	0.31	
ZNF264	H: M-nE		
	TO: R-a	0.24	
	TR: R-a	0.20	
ZNF224	H: R-a		
	TO: R-a	0.22	
	TR: R-a	0.20	
ZNF45	H: R-a		
	TO: R-a	0.20	
	TR: M-nE	-0.04	
ZNF30	H: M-nE		
	TO: R-a	0.19	
	TR: R-a	0.18	
ZNF184	H: R-a		
	TO: R-nE	0.09	
	TR: R-nE	0.11	
ZNF33A	H: R-a		
	TO: R-a	0.03	
	TR: R-a	0.10	
ZNF582	H: R-nE		
	TO: M-nE	0.08	
	TR: R-a	0.04	

Figure 4 Similarity between ChIP-derived motifs. (A) Similarity between the Hughes and Trono motifs for ZIM3. The heat map on the left indicates the MoSBAT similarity e-scores between each pair compared. The motifs and the motif-finding methods are represented on the right. H, Hughes; TO, Trono Original; TR, Trono Reprocessed. (B) The MoSBAT e-scores between Hughes motifs and Trono original and Trono reprocessed motifs and the corresponding aligned motifs for the 39 overlapping KZNFs. H, Hughes; TO, Trono Original; TR, Trono Reprocessed; R, RCADE; M, MEME; a, all peaks; nE, nonERE peaks. (C) MoSBAT similarity e-scores for the 39 overlapping KZNFs between the Hughes data and Trono original (blue) and Trono reprocessed (red). The dots indicate the percentage of the overlap between the Hughes and Trono original peaks (blue) and Hughes and Trono reprocessed peaks (red).

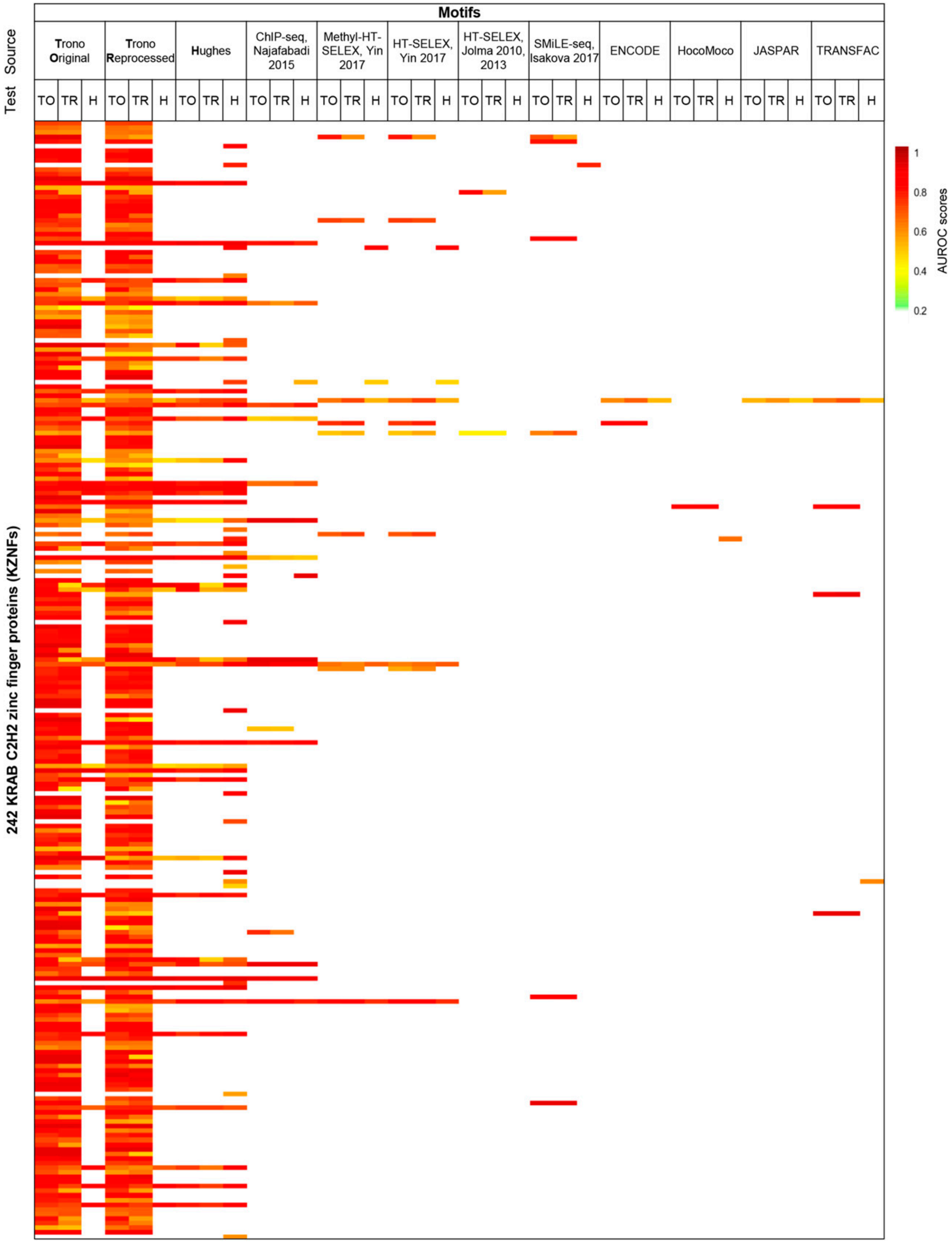
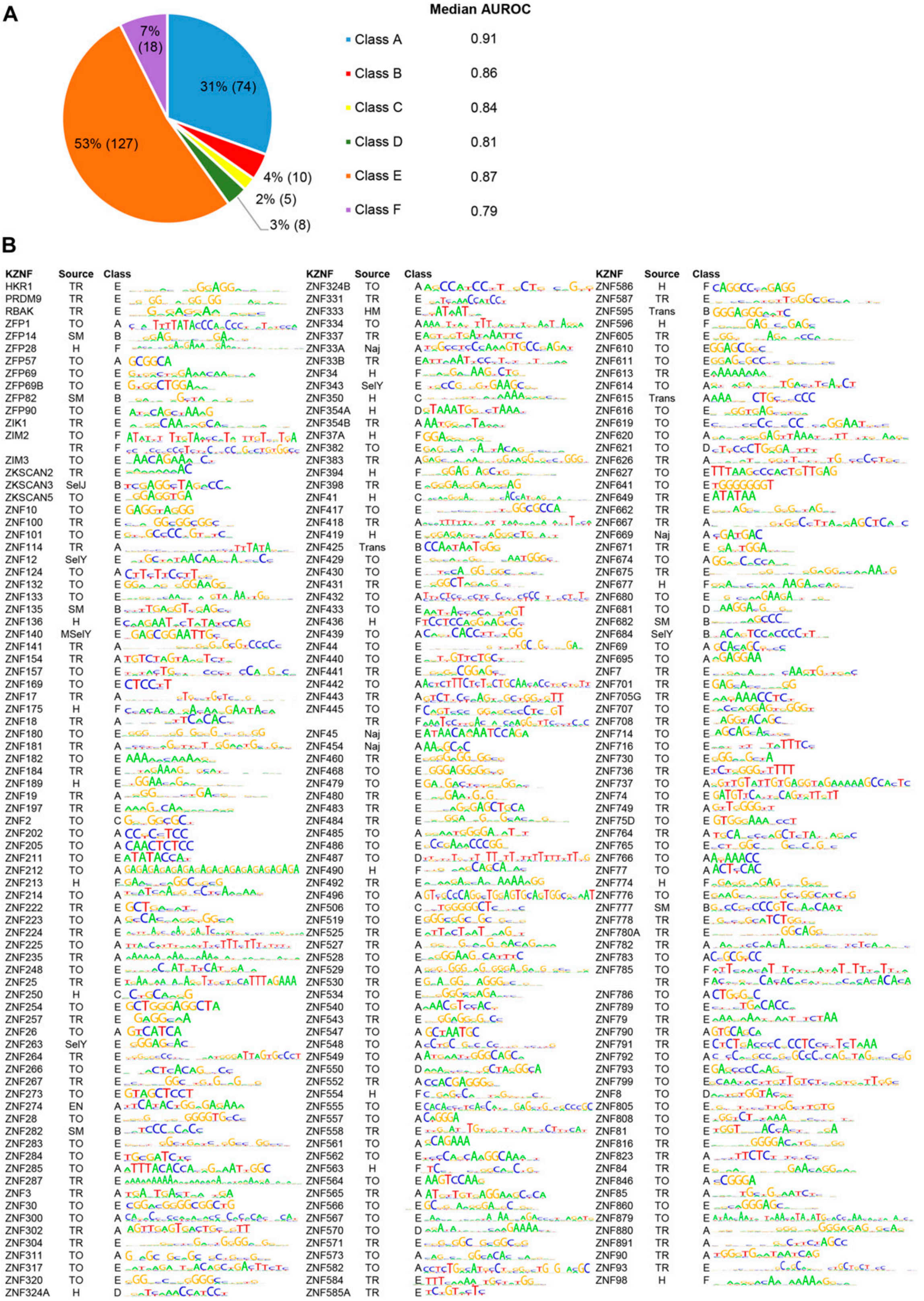
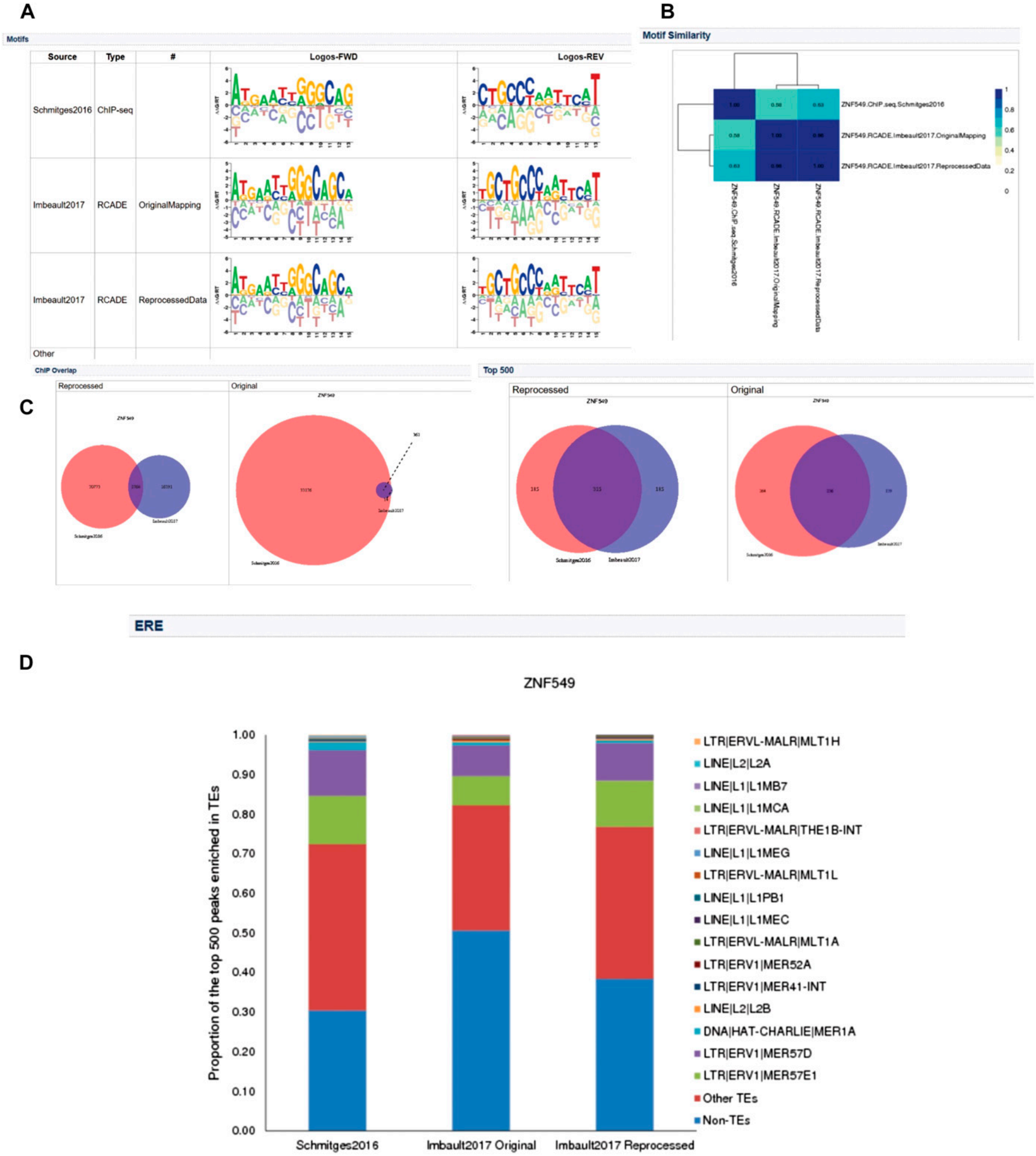


Figure 5 AUROC of the Hughes and Trono motifs and external motifs overlapping any of the two data sets. Heat map represents the AUROC value of each motif tested on Hughes, Trono original, or Trono reprocessed peaks. The first row at the top indicates the source of the motif, and the second row indicates the test data set. TO, Trono Original; TR, Trono Reprocessed; H, Hughes. White indicates no data is available. A full version of the figure that includes the KZNFs IDs is available at the web portal of the paper (http://kznfmotifs.ccbr.utoronto.ca/figures.html).



KZNF	Source	Class	KZNF	Source	Class	KZNF	Source	Class
HKR1	TR	E	ZNF324B	TO	A	ZNF586	H	F
PRDM9	TR	E	ZNF331	TR	E	ZNF587	TR	E
RBAK	TR	E	ZNF333	HM	E	ZNF595	Trans	B
ZFP1	TO	A	ZNF334	TO	A	ZNF596	H	F
ZFP14	SM	B	ZNF337	TR	E	ZNF605	TR	E
ZFP28	H	F	ZNF33A	Naj	A	ZNF610	TO	E
ZFP57	TO	A	ZNF33B	TR	E	ZNF611	TO	E
ZFP69	TO	E	ZNF34	H	F	ZNF613	TR	E
ZFP69B	TO	E	ZNF343	SelY	E	ZNF614	TO	A
ZFP82	SM	B	ZNF350	H	C	ZNF615	Trans	A
ZFP90	TO	E	ZNF354A	H	D	ZNF616	TO	E
ZIK1	TR	E	ZNF354B	TR	A	ZNF619	TO	E
ZIM2	TO	F	ZNF37A	H	F	ZNF620	TO	A
	TR	F	ZNF382	TO	E	ZNF621	TO	D
ZIM3	TO	E	ZNF383	TR	A	ZNF626	TR	A
ZKSCAN2	TR	E	ZNF394	H	F	ZNF627	TO	E
ZKSCAN3	SelJ	B	ZNF398	TR	E	ZNF641	TO	E
ZKSCAN5	TO	E	ZNF41	H	C	ZNF649	TR	E
ZNF10	TO	E	ZNF417	TO	E	ZNF662	TR	E
ZNF100	TR	E	ZNF418	TR	A	ZNF667	TR	A
ZNF101	TO	E	ZNF419	H	E	ZNF669	Naj	A
ZNF114	TR	A	ZNF425	Trans	B	ZNF671	TR	E
ZNF12	SelY	E	ZNF429	TO	E	ZNF674	TO	A
ZNF124	TO	A	ZNF430	TO	E	ZNF675	TR	E
ZNF132	TO	E	ZNF431	TR	E	ZNF677	H	F
ZNF133	TO	E	ZNF432	TO	A	ZNF680	TO	E
ZNF135	SM	B	ZNF433	TO	E	ZNF681	TO	D
ZNF136	H	E	ZNF436	H	F	ZNF682	SM	B
ZNF140	MSelY	E	ZNF439	TO	A	ZNF684	SelY	B
ZNF141	TR	A	ZNF44	TO	E	ZNF69	TO	A
ZNF154	TR	A	ZNF440	TO	E	ZNF695	TO	A
ZNF157	TO	E	ZNF441	TR	E	ZNF7	TR	E
ZNF169	TO	E	ZNF442	TO	A	ZNF701	TR	E
ZNF17	TR	A	ZNF443	TR	A	ZNF705G	TR	E
ZNF175	H	F	ZNF445	TO	F	ZNF707	TO	E
ZNF18	TR	A		TR	F	ZNF708	TR	E
ZNF180	TO	E	ZNF45	Naj	E	ZNF714	TO	E
ZNF181	TR	A	ZNF454	Naj	A	ZNF716	TO	E
ZNF182	TO	E	ZNF460	TR	E	ZNF730	TO	E
ZNF184	TR	E	ZNF468	TO	E	ZNF736	TR	E
ZNF189	H	E	ZNF479	TO	E	ZNF737	TO	A
ZNF19	TR	A	ZNF480	TR	E	ZNF74	TO	E
ZNF197	TR	E	ZNF483	TR	E	ZNF749	TR	A
ZNF2	TO	C	ZNF484	TR	E	ZNF75D	TO	E
ZNF202	TO	A	ZNF485	TO	A	ZNF764	TR	A
ZNF205	TO	A	ZNF486	TO	E	ZNF765	TO	E
ZNF211	TO	E	ZNF487	TO	D	ZNF766	TO	A
ZNF212	TO	A	ZNF490	H	F	ZNF77	TO	A
ZNF213	H	F	ZNF492	TR	E	ZNF774	H	F
ZNF214	TO	A	ZNF496	TO	A	ZNF776	TO	E
ZNF222	TR	E	ZNF506	TO	C	ZNF777	SM	B
ZNF223	TO	A	ZNF519	TO	E	ZNF778	TR	E
ZNF224	TR	E	ZNF525	TR	E	ZNF780A	TR	E
ZNF225	TO	A	ZNF527	TR	A	ZNF782	TR	A
ZNF235	TR	A	ZNF528	TO	E	ZNF783	TO	A
ZNF248	TO	E	ZNF529	TO	A	ZNF785	TO	F
ZNF25	TR	E	ZNF530	TR	E		TR	F
ZNF250	H	C	ZNF534	TO	E	ZNF786	TO	A
ZNF254	TO	E	ZNF540	TO	A	ZNF789	TO	E
ZNF257	TR	E	ZNF543	TR	E	ZNF79	TR	E
ZNF26	TO	A	ZNF547	TO	A	ZNF790	TR	A
ZNF263	SelY	E	ZNF548	TO	A	ZNF791	TR	E
ZNF264	TR	E	ZNF549	TO	A	ZNF792	TO	A
ZNF266	TO	E	ZNF550	TO	D	ZNF793	TO	E
ZNF267	TR	E	ZNF552	TR	A	ZNF799	TO	E
ZNF273	TO	E	ZNF554	H	F	ZNF8	TO	D
ZNF274	EN	A	ZNF555	TO	E	ZNF805	TO	E
ZNF28	TO	E	ZNF557	TO	A	ZNF808	TO	E
ZNF282	SM	B	ZNF558	TR	E	ZNF81	TO	E
ZNF283	TO	E	ZNF561	TO	A	ZNF816	TR	E
ZNF284	TO	E	ZNF562	TO	E	ZNF823	TR	A
ZNF285	TO	A	ZNF563	H	F	ZNF84	TR	E
ZNF287	TR	E	ZNF564	TO	E	ZNF846	TO	A
ZNF3	TR	A	ZNF565	TR	A	ZNF85	TR	A
ZNF30	TO	E	ZNF566	TO	E	ZNF860	TO	E
ZNF300	TO	A	ZNF567	TO	E	ZNF879	TO	E
ZNF302	TR	A	ZNF570	TO	D	ZNF880	TR	A
ZNF304	TR	E	ZNF571	TR	E	ZNF891	TR	A
ZNF311	TO	A	ZNF573	TO	A	ZNF90	TR	A
ZNF317	TO	E	ZNF582	TO	A	ZNF93	TR	E
ZNF320	TO	E	ZNF584	TR	E	ZNF98	H	F
ZNF324A	H	D	ZNF585A	TR	E			

Figure 6 The reference motif set for the 242 KZNFs. (A) Percentage and number of the motifs (in parentheses) fit into classes A–F and the median AUROC values of each group. (B) The reference motif for each of the 242 KZNFs. Source refers to motif origin (TO, Trono Original; TR, Trono Reprocessed; H, Hughes; Naj, ChIP-seq (Najafabadi *et al.* 2015b); SM, SMiLE-seq (Isakova *et al.* 2017); SelY, HT-SELEX (Yin *et al.* 2017); MSelY, Methyl-HT-SELEX (Yin *et al.* 2017); SelJ, HT-SELEX (Jolma *et al.* 2013); EN, ENCODE; Trans, TRANSFAC; HM, HocoMoco). The class is the selection class that each motif falls into. For ZIM2, ZNF445 and ZNF785, both motifs from class F are represented.



Source	Type	#	Logos-FWD	Logos-REV
Schmitges2016	ChIP-seq			
Imbeault2017	RCADE	OriginalMapping		
Imbeault2017	RCADE	ReprocessedData		
Other				

Figure 7 Web portal of ZNF549 containing all the analyses described (http://kznfmotifs.ccbr.utoronto.ca/report.php?name=ZNF549). (A) Motifs for the same KZNF derived from different sources. (B) MoSBAT similarity heat maps between all motifs. (C) Overlap between Hughes peaks and Trono reprocessed (left) and Trono original (right) peaks for all peaks and top 500 peaks. (D) ERE enrichment for the Hughes and Trono ChIP peaks.

no motifs satisfied these criteria, we selected ChIP motifs that are supported by the recognition code (*i.e.*, RCADE) (RCADE motifs = Class C), followed by the ChIP motifs that predict their training data better than all other motifs, where the AUROC was at least 0.1 greater than the others (ChIP motif is uniquely predictive of training peaks = Class D), followed by those that predict the test ChIP data sets even slightly better than the others (AUROC = 0.01–0.09 greater than the others) (Motif is most predictive of test peaks = Class E), and, finally, all remaining motifs including single motifs and motifs for three KZNFs (ZIM2, ZNF445 and ZNF785) that cannot be discriminated by any of the above criteria (Others = Class F).

Figure 6A shows the number and proportion of KZNFs currently falling into each class, and Figure 6B shows the motifs. In this scheme, the largest classes of motifs are Class A (in which one motif clearly outperforms all others on test data), and Class E (where all motifs are roughly equivalent). The median AUROC for each class does not vary greatly (Figure 6A). Most of the motifs are supported by the recognition code: 158/242 (65%) are RCADE motifs, 96% of which have AUC >0.6 predicting peak *vs.* nonpeak. As previously reported for KZNFs, these reference motifs are highly diverse, and tend to be longer and more information-rich than motifs for other TF types, which tend to be only 6–12 bases long (Badis *et al.* 2009; Jolma *et al.* 2013).

Web portal for KZNFs

Finally, we generated a web interface that assembles all the data described herein, providing all the relevant data files, as well as a graphical interface for each KZNF. Figure 7 provides an illustration for ZNF549, a KZNF with a Class A motif. Motifs from different sources are shown, if available (Figure 7A). A heatmap of the pairwise motif similarity scores between different motifs is then shown (Figure 7B). The peak overlap between the Hughes and Trono ChIP data are illustrated, where available, for all peaks and top 500 peaks (Figure 7C). Finally, the ERE enrichment of the peaks is given (Figure 7D).

DISCUSSION

The overall conclusion of these analyses is that there is reasonable agreement between the Hughes and Trono lab ChIP data, and also external data, on the motifs, genomic binding sites in HEK293 cells, and EREs enriched among those sites. This finding enables us to conclude that ChIP-seq—despite being a notoriously difficult method from which to derive motifs and to predict primary regulatory "targets"—is a viable strategy for both obtaining motifs and for identifying a potential physiological role for these proteins, or at least an initial driving evolutionary force (repression of EREs). These analyses also indicate that most human KZNFs are *bona fide* sequence-specific DNA-binding proteins.

Despite general overall agreement, there are discrepancies between the data sets. When considering both Trono original and Trono reprocessed data together, 23% of the experiments disagree on at least one of these parameters (EREs bound, $r < 0.3$ and motifs, $r < 0.2$); 3/39 (8%) of the experiments disagree on both. Because disagreement can stem from noise or error in only one of the two data sets, it is possible that these issues are less prevalent in the individual data sets (*e.g.*, 11.5% error in each of them would lead to an overall error of 23%). KZNFs for which the data sets disagree, and for which there are no motifs that relate to the recognition code (44/242), might be considered prime candidates for rerunning the ChIP assays, possibly in other cell types, and/or analysis by alternative assays such as SMILE-seq. Ideally, an *in vitro*-derived motif for each protein will eventually also be available, enabling confirmation that the motifs enriched in ChIP data represent *bona fide* direct sequence recognition events. In the meantime, the recognition code appears to provide a reasonable substitute.

The absence of a complete set of motifs for human TFs (Weirauch *et al.* 2014) remains a glaring shortcoming in the study of human gene regulation. Our collection of reference motifs includes 166 proteins for which there was previously no known motif. We anticipate that the reference set will be of general utility in the study of regulatory mechanisms and networks. It will also facilitate exploration of the targeting mechanisms of these proteins to EREs—the motifs alone are largely insufficient (Najafabadi *et al.* 2015b; Schmitges *et al.* 2016)—as well as their roles in human genetics and disease, because SNPs in noncoding regions often impact motif scores for TFs (Deplancke *et al.* 2016).

ACKNOWLEDGMENTS

We are grateful to Hamed Najafabadi, Laura Campitelli, Arttu Jolma, and Matt Weirauch for input on the methodology and web site, and grants from the Canadian Institutes of Health Research (FDN-148403) and the National Science and Engineering Research Council (NSERC) (RGPIN-2016-05643) for financial support. M.B. was supported in part by a C.H. Best Fellowship. S.L. is supported by a doctoral scholarship from NSERC.

LITERATURE CITED

Badis, G., M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke *et al.*, 2009 Diversity and complexity in DNA recognition by transcription factors. Science 324: 1720–1723.

Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant *et al.*, 2009 MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37: W202–W208.

Brayer, K. J., and D. J. Segal, 2008 Keep your fingers off my DNA: protein-protein interactions mediated by C2H2 zinc finger domains. Cell Biochem Biophys 50: 111–131.

Day, D. S., L. J. Luquette, P. J. Park, and P. V. Kharchenko, 2010 Estimating enrichment of repetitive elements from high-throughput sequence data. Genome Biol. 11: R69.

Deplancke, B., D. Alpern, and V. Gardeux, 2016 The genetics of transcription factor DNA binding variation. Cell 166: 538–554.

Emerson, R. O., and J. H. Thomas, 2009 Adaptive evolution in zinc finger transcription factors. PLoS Genet. 5: e1000325.

Encode-Project-Consortium, 2012 An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74.

Imbeault, M., P. Y. Helleboid, and D. Trono, 2017 KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. Nature 543: 550–554.

Isakova, A., R. Groux, M. Imbeault, P. Rainer, D. Alpern *et al.*, 2017 SMiLE-seq identifies binding motifs of single and dimeric transcription factors. Nat. Methods 14: 316–322.

Jacobs, F. M. J., D. Greenberg, N. Nguyen, M. Haeussler, A. D. Ewing *et al.*, 2014 An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature 516: 242–245.

Jolma, A., J. Yan, T. Whitington, J. Toivonen, K. R. Nitta *et al.*, 2013 DNA-binding specificities of human transcription factors. Cell 152: 327–339.

Jolma, A., Y. Yin, K. R. Nitta, K. Dave, A. Popov *et al.*, 2015 DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature 527: 384–388.

Karolchik, D., A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet *et al.*, 2004 The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32: D493–D496.

Kharchenko, P. V., M. Y. Tolstorukov, and P. J. Park, 2008 Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat. Biotechnol. 26: 1351–1359.

Kulakovskiy, I. V., Y. A. Medvedeva, U. Schaefer, A. S. Kasianov, I. E. Vorontsov *et al.*, 2013 HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. Nucleic Acids Res. 41: D195–D202.

Lambert, S. A., M. Albu, T. R. Hughes, and H. S. Najafabadi, 2016 Motif comparison based on similarity of binding affinity profiles. Bioinformatics 32: 3504–3506.

Landt, S. G., G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli *et al.*, 2012 ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 22: 1813–1831.

Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods 9: 357–359.

Mathelier, A., and W. W. Wasserman, 2013 The next generation of transcription factor binding site prediction. PLOS Comput. Biol. 9: e1003214.

Mathelier, A., O. Fornes, D. J. Arenillas, C. Y. Chen, G. Denay *et al.*, 2016 JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 44: D110–D115.

Matsui, T., D. Leung, H. Miyashita, I. A. Maksakova, H. Miyachi *et al.*, 2010 Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. Nature 464: 927–931.

Matys, V., O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land *et al.*, 2006 TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 34: D108–D110.

Najafabadi, H. S., M. Albu, and T. R. Hughes, 2015a Identification of C2H2-ZF binding preferences from ChIP-seq data using RCADE. Bioinformatics 31: 2879–2881.

Najafabadi, H. S., S. Mnaimneh, F. W. Schmitges, M. Garton, K. N. Lam *et al.*, 2015b C2H2 zinc finger proteins greatly expand the human regulatory lexicon. Nat. Biotechnol. 33: 555–562.

Noyes, M. B., X. Meng, A. Wakabayashi, S. Sinha, M. H. Brodsky *et al.*, 2008 A systematic characterization of factors that regulate *Drosophila* segmentation via a bacterial one-hybrid system. Nucleic Acids Res. 36: 2547–2560.

Rowe, H. M., and D. Trono, 2011 Dynamic control of endogenous retroviruses during development. Virology 411: 273–287.

Schmitges, F. W., E. Radovani, H. S. Najafabadi, M. Barazandeh, L. F. Campitelli *et al.*, 2016 Multiparameter functional diversity of human C2H2 zinc finger proteins. Genome Res. 26: 1742–1752.

Schultz, D. C., K. Ayyanathan, D. Negorev, G. G. Maul, and F. J. Rauscher, III, 2002 SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. Genes Dev. 16: 919–932.

Wang, J., 2014 Quality *vs.* accuracy: result of a reanalysis of protein-binding microarrays from the DREAM5 challenge by using BayesPI2 including dinucleotide interdependence. BMC Bioinformatics 15: 289.

Weirauch, M. T., A. Cote, R. Norel, M. Annala, Y. Zhao *et al.*, 2013 Evaluation of methods for modeling transcription factor sequence specificity. Nat. Biotechnol. 31: 126–134.

Weirauch, M. T., A. Yang, M. Albu, A. G. Cote, A. Montenegro-Montero *et al.*, 2014 Determination and inference of eukaryotic transcription factor sequence specificity. Cell 158: 1431–1443.

Wolfe, S. A., L. Nekludova, and C. O. Pabo, 2000 DNA recognition by Cys2His2 zinc finger proteins. Annu. Rev. Biophys. Biomol. Struct. 29: 183–212.

Yin, Y., E. Morgunova, A. Jolma, E. Kaasinen, B. Sahu *et al.*, 2017 Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science 356: pii: eaaj2239.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson *et al.*, 2008 Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9: R137.

Zhao, Y., D. Granas, and G. D. Stormo, 2009 Inferring binding energies from selected binding sites. PLOS Comput. Biol. 5: e1000590.

Communicating editor: D. Gresham