Published online 4 May 2018

Nucleic Acids Research, 2018, Vol. 46, No. 11 ***5395–5409***
doi: 10.1093/nar/gky338

Towards a map of *cis*-regulatory sequences in the human genome

Meng Niu, Ehsan Tabari, Pengyu Ni and Zhengchang Su*

Department of Bioinformatics and Genomics, College of Computing and Informatics, The University of North Carolina at Charlotte, 9201 University City Blvd., Charlotte, NC 28223, USA

Received January 17, 2018; Revised April 14, 2018; Editorial Decision April 16, 2018; Accepted April 19, 2018

ABSTRACT

Accumulating evidence indicates that transcription factor (TF) binding sites, or *cis*-regulatory elements (CREs), and their clusters termed *cis*-regulatory modules (CRMs) play a more important role than do gene-coding sequences in specifying complex traits in humans, including the susceptibility to common complex diseases. To fully characterize their roles in deriving the complex traits/diseases, it is necessary to annotate all CREs and CRMs encoded in the human genome. However, the current annotations of CREs and CRMs in the human genome are still very limited and mostly coarse-grained, as they often lack the detailed information of CREs in CRMs. Here, we integrated 620 TF ChIP-seq datasets produced by the ENCODE project for 168 TFs in 79 different cell/tissue types and predicted an unprecedentedly completely map of CREs in CRMs in the human genome at single nucleotide resolution. The map includes 305 912 CRMs containing a total of 1 178 913 CREs belonging to 736 unique TF binding motifs. The predicted CREs and CRMs tend to be subject to either purifying selection or positive selection, thus are likely to be functional. Based on the results, we also examined the status of available ChIP-seq datasets for predicting the entire regulatory genome of humans.

INTRODUCTION

Since the Human Genome Project was completed 16 years ago (1,2), researchers have used powerful computational and experimental methods (3) to gain a good understanding of coding sequences in the human reference genome. In contrast, although *cis*-regulatory sequences are as important as coding sequences in specifying various phenotypes of organisms (4–6), only recently have we begun to decipher them systematically due to difficulties in characterizing them using traditional method (7). These *cis*-regulatory sequences (i.e. promoters, enhancers, silencers and insulators) are also called *cis*-regulatory modules (CRMs) (8), because they are made of clusters of short *cis*-regulatory elements (CREs) recognized by specific transcription factors (TFs) (9). Thus, CREs are also called TF binding sites. A TF can bind tens of thousands of similar, yet degenerate, CREs in the genome to regulate many genes (thereafter, we refer to a set of CREs recognized by the same TF as a motif). A growing body of evidence indicates that it is mainly CRMs, rather than coding sequences, that account for interspecies divergence and intra-species diversity (10–30). Further, recent genome-wide association studies (GWAS) have found that most complex trait-associated single nucleotide variations (SNVs) do not reside in coding sequences, but rather lie in non-coding regions (NCRs, including introns and intergenic regions) (31,32) and often overlap with or are in linkage disequilibrium (LD) with CREs (33). Complex trait-associated variants have also been shown to systematically disrupt CREs of TFs related to the traits (33), and variation in CREs affects DNA binding, chromatin modification, transcription of genes (34–38), and complex traits/diseases (22,39–43). More recently, it was reported that CREs determine chromatin modification and gene expression patterns (34–36,44–46). Therefore, a better understanding of CREs and CRMs encoded in the human reference genome is necessary for personalized medicine to prevent and treat complex diseases (18,41–43,45–52).

Although it is time consuming to identify CREs and CRMs using traditional methods, the development of a plethora of next-generation sequencing (NGS)-based techniques has allowed genome-wide characterization of CREs and CRMs. These methods include 1) ChIP-seq for locating the CREs of a TF (53–55) and for various histone markers (56); 2) DNase-seq (57–59), ATAC-seq (60) and FAIRE-seq (58) for locating free nucleosome regions; 3) Hi-C for measuring the physical proximity of linearly distal DNA segments (61,62); and 4) RNA-seq for profiling the transcriptomes in cells or tissues (63). As a result, enormous data including TF ChIP-seq data are being produced

*To whom correspondence should be addressed. Tel: +1 704 687 7996; Fax: +1 704 687 8867; Email: zcsu@uncc.edu
Present address: Meng Niu, Department of Genetics, Cell Biology and Anatomy, College of Medicine, University of Nebraska Medical Center. Email: meng.niu@unmc.edu.

in human cells or tissues by individual labs world-wide and large consortia such as the ENCODE (64–66), NIH Roadmap Epigenomics (67,68), Genotype-Tissue Expression (GTEx) (69,70), FANTOM (71–74), all are aimed to identify all functional elements in the human genome (64–66). Although highly challenging, it is now possible to predict at least most of CREs and CRMs in the human genome by integrating these enormous data, particularly TF ChIP-seq data collected for different TFs in various cells/tissues and developmental stages.

Ren's group was the first to use hidden Markov models (HMMs) to predict CRMs based on multiple histone modification markers (75,76). Ernst *et al.* (77,78) extended the idea and developed ChromHMM to segment the genome into different functional types according to histone modification profiles in a cell/tissue type. Hoffman *et al.* (79,80) developed Segway using dynamic Bayesian networks for the same purpose. Other machine-learning methods have been proposed to predict enhancers in a cell/tissue type based on histone modification profiles. For example, Firpi *et al.* (81) developed CSI-ANN using time-delay neural networks (NNs), Rajagonal *et al.* (82) developed RFECS using a random forest, and Villanrroel *et al.* (83) developed ChromaGenSVM and Kleftogiannis *et al.* (84) developed DEEP using support vector machines (SVMs). Primary sequence features have also been used to predict tissue-specific enhancers using SVM (84,85). Although these methods can predict CRMs with cell/tissue type specificity, their accuracy is quite low (86). Even the best-performing tools (DEEP and CSI-ANN) have only 49.8% and 45.2%, respectively, of their predicted CRMs overlapping with the DNase I hypersensitivity sites (DHSs) in Hela cells (84). Moreover, none of these methods can pin down the exact locations of CRMs and their constituent CREs, thus the predictions are only coarse-grained, lacking specifics about CRE in CRMs, such as their numbers and locations, and importance of each position for TF binding. Surprisingly, although TF ChIP-seq data can provide the most accurate information of CREs of ChIP-ed TFs (87) and their possible combinatory patterns in CRMs, to the best of our knowledge, no existing algorithm is able to mine a large number of TF ChIP-seq datasets to more accurately predict CREs and CRMs in the human genome.

To fill these gaps, we have recently developed an algorithm called DePCRM (88) for predicting CREs and CRMs in eukaryotic genomes by integrating a large number of TF ChIP datasets, and have successfully used it to predict an unprecedentedly complete map of CREs and CRMs in the *Drosophila melanogaster* genome. However, compared with the *D. melanogaster* genome (139.5 Mb), the human genome (3.2 Gb), is 22.9 times larger, encoding more genes (21 000 versus 13 600), more TFs (2886 versus 1030), and more complex gene regulatory networks for more complex phenotypes. ChIP-seq datasets obtained from human tissues or cells can be 10 times larger than those from *D. melanogaster* cells/tissues, making their analysis and integration more challenging. Moreover, given the great efforts that have been made world-wide to generate a large number of ChIP-seq datasets from various human cell/tissue types, it is interesting to see how the way that these data were generated is effective, and how much additional data we may need to predict a complete map of CREs and CRMs in the genome.

To address the questions, we predicted a map of CREs and CRMs in the human genome at single-nucleotide resolution using our algorithm by integrating a total of 620 ChIP-seq datasets for 168 TFs in 79 different cell/tissue types. The map includes 305 912 CRMs containing 736 unique CRE motifs. The predicted CRMs recovered 51.3% of known enhances in the datasets, and 14.8% of our predicted CRMs overlaps with DNase I hypersensitive sites (DHSs). Moreover, both the predicted CRMs and CREs tend to be more conserved than corresponding randomly selected sequences, thus, they are likely to be functional. Using these datasets, we also analyzed the saturation trend of TF binding motif predictions in three different scenarios to address questions such as what the most effective ways are to generate TF ChIP-seq data, and how many datasets we may need to predict a complete map of CREs and CRMs in the human genome.

MATERIALS AND METHODS

Datasets and processing

A total of 620 ChIP-seq binding peaks datasets for 168 TFs in 79 different cell/tissue types were downloaded from the UCSC Genome Browser database (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/). The binding peaks were identified by the peak-calling and refinery procedure designed by Kundaje and colleagues (89). A total of 897 experimentally verified the sequences containing enhancers in the human genome (version hg19) were downloaded from the VISTA Enhancer Browser database (90). These human enhancer fragments have an average length of 1,950 bp. Coordinates of a total of 1 281 988 non-overlapping DHSs in 125 tissue/cell types produced by ENCODE were downloaded from the UCSC Genome Browser database (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegDnaseClustered/). To predict CRMs around the summits of binding peaks, we extended the binding peaks shorter than 3 kb to up to 3 kb by padding equal length of flanking genomic sequences to the two ends, as most of the known human enhancer segments from VISTA are shorter than 3 kb.

Measurement of the overlap of binding peaks in two datasets

We define the overlapping level of extended binding peaks in two datasets d_i and d_j as,

$$S_o\ (d_i, d_j) = o(d_i, d_j)/|d_i| + o(d_i, d_j)/|d_j| \qquad (1)$$

where $|d_i|$ and $|d_j|$ are the number of binding peaks in d_i and d_j, respectively, and $o(d_i, d_j)$ the number of overlapping sequences between d_i and d_j.

Finding motifs in binding peak datasets

We used DREME (91) to identify all possible motifs in each of the extended binding peak dataset for its computational efficiency and ability to return enough number of

over-represented motifs in a dataset. As DREME requires a negative dataset for more accurate prediction, we generated a random sequence set for each input dataset using a third order Markov chain model based on the transition probabilities of the sequences in the dataset. As the size of a dataset becomes large, even a fast algorithm such as DREME cannot finish in a practical time, thus we split a dataset with a size over 10,000 peaks into multiple sub-datasets with similar number of peaks smaller than 10 000, i.e. the size of sub-datasets is equal to $N/$ (mod(N/10 000) +1), where $N >$ 10 000 is the number of binding peaks in the original dataset.

Prediction of CREs and CRMs

We used our DePCRM program developed earlier (88) to predict CRE and CRMs in the genome based on the motifs found in all datasets or sub-datasets with minor modifications. Briefly, for each pair of motifs $M_d(i)$ *and* $M_d(j)$ found in the same dataset d, we compute a motif co-occurring score S_c defined as,

$$S_c\ (M_d\,(i)\,,\ M_d\,(j)) = \frac{o\,(M_d\,(i)\,,\ M_d\,(j))}{\max\,(|\,M_d\,(i)\,|,\,|\,M_d\,(j)|)} \qquad (2)$$

where $|M_d(i)|$ and $|M_d\,(j)|$ are the number of binding peaks containing CREs of motifs $M_d(i)$ and $M_d\,(j)$, respectively; and $o(M_d(i),\ M_d(j))$ the number of binding peaks containing CREs of both the motifs in the dataset. We select motif pairs with an $S_c \geq \alpha$ as co-occurring motif pairs (CPs) for further analysis. The cutoff α is chosen such that the predicted motifs are maximally excluded, and at the same time, those in known CRMs are maximally included. Then for each pair of datasets a and b, we compute a similarity score S_s between each pair of CPs $P[M_a(i),\ M_a(j)]$ from a and $P[M_b\,(m),\ \mathrm{M_b}\,(n)]$ from b, defined as,

$$\begin{aligned} &S_s\{P[M_a(i),\ M_a(j)],\ \ P[M_b(m),\ M_b(n)]\} \\ &= max_{k\in\{i,j\},\ l\in\{m,n\}}\{Sim[M_a(k),\ M_b(l)]\} \\ &+Sim[M_a(r),\ M_b(s)], r \in \{i,\ j\},\ r \neq k; s \in \{m, n\}, s \neq l \end{aligned} \qquad (3)$$

where $Sim(M,\ N)$ is the similarity score between motifs M and N using a metric called SPIC (92–94). Note that the maximization operation is only on the first $Sim(M,\ N)$ score, because we want to reward the scenario that two motifs each from $P[M_a(i),\ M_a(j)]$ and $P[M_b\ (m),\ \mathrm{M_b}\ (n)]$, respectively, are highly similar to each other, but the two other remaining motifs may not be so. Next, we construct a CP similarity graph using the CPs as the nodes, connecting two CPs with an edge with their score S_s being the weight if $S_s \geq \beta$, and removing isolated nodes. We choose a β value such that the density (defined as the number of edges divided by the number of nodes) of resulting graph is as low as possible, meanwhile the graph contains as many as possible connected nodes/CPs. We use the Markov Chain Clustering algorithm (MCL) (95) to cut the graph into dense subgraphs, each corresponding to a cluster of repetitively occurring CPs across multiple datasets. We discard the clusters containing fewer than τ CPs ($\tau = 2$ in this study, i.e. we only discard singleton CPs). Presumably, the remaining clusters contain highly similar CPs for two certain TFs. We call these clusters CP clusters (CPCs). To identify co-occurring patterns containing more than two motifs, we calculate a co-occurring score S_{cpc} for each pair of CPCs, C_i and C_j over all the datasets defined as

$$\begin{aligned} &S_{CPC}(C_i,\,C_j) \\ &= \tfrac{1}{D}\sum_{k\,=\,1}^{D}\tfrac{1}{\mathrm{N}[\Omega_{d_k}(C_i,\,C_j)]}\sum_{(P_s\in C_i,\,P_t\in C_j)\subset\Omega_{d_k}(C_i,\,C_j)}\left[\tfrac{o(P_s,\,P_t)}{|P_s|}+\tfrac{o(P_s,\,P_t)}{|P_t|}\right] \end{aligned} \qquad (4)$$

where D is the number of datasets containing CPs of both C_i and C_j; $\Omega_{d_k}(C_i,\ C_j)$ the set of the CPs in C_i and C_j from the same dataset d_k; $\mathrm{N}[\Omega_{d_k}(C_i,\ C_j)]$ the number of unique comparisons among the CPs in $\Omega_{d_k}(C_i,\ C_j)$; P_s and P_t two CPs from C_i and C_j, respectively; $\mathrm{o}(P_s,\ P_t)$ the number of binding peaks in which P_s and P_t co-occur; and $|P|$ the size of P. We construct a CPC co-occurring graph using each CPC as a node, and connecting two CPCs C_i and C_j by an edge with the S_{cpc} being the weight if $S_{CPC}\ (C_i,\ C_j) \geq \gamma$.The cutoff γ is chosen based on the bimodal distribution of the S_{cpc} sores. We apply MCL to cut the CPC co-occurring graph into dense sub-graphs. Each of these sub-graphs is assumed to correspond to a possible combination of their motifs to form a CRM based on the datasets used. For this reason, we refer to these CPC clusters as CRM components (CRMCs). Some motifs in the CRMCs may have overlapping CREs and can be highly similar to one another. They are likely the same or similar CREs of the same TF or closely related ones of the same family. Thus, we combine such highly similar and possibly redundant motifs into unique ones. We call each of these combined motifs a unique motif or U-motif. We then represent each motif in the identified CRMCs by the U-motif that it belongs. We project the predicted CREs of all the CRMCs back to their locations in the genome, and if the projected CREs overlap with one another, we merge them in one. We then connect two adjacent CREs if their distance is shorter than a preset value δ ($\delta = 150$ bp in this study), but the connection cannot span over an exon unless it contains at least a CRE. We predict each segment of the sequences connected by the CREs as a CRM.

Prediction saturation analysis

We analyzed the saturation trends of predicted U-motifs in the following three scenarios: (i) changes in the number of predicted U-motifs with increasing number of datasets for different TFs from the same cell/tissue type; (ii) changes in the number of predicted U-motifs with increasing number of datasets in different cell/tissue types for the same TF and (iii) changes in the number of predicted U-motifs with increasing number of randomly selected datasets. Specifically, for the first two scenarios, we used the U-motifs predicted using the 620 datasets as the standard set and count the number of the U-motifs whose biding sites are located in the randomly selected datasets. For the third scenario, we randomly selected different numbers ($n = 100$, 200, 250, 300 and 350) of datasets from the 620 datasets and applied the algorithm to each of the randomly selected datasets with the same parameter settings. For randomly selected datasets, we repeated the process five times for difference choice of datasets and present the averaged results to minimize the effect caused by different combinations of datasets used. We fitted the results to a sigmoid function,

$$f\ (n) = \alpha + \delta \times \left(1 - \frac{1}{1 + \frac{n}{\beta}}\right) \qquad (5)$$

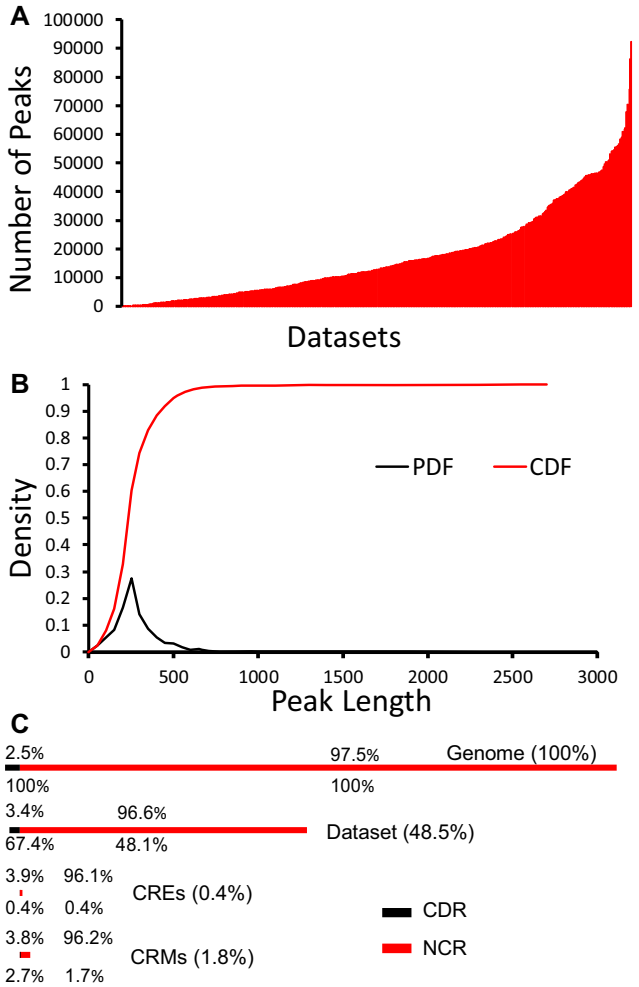

Figure 1. Features of the ChIP-seq datasets. (**A**). Number of binding peaks in the 620 ChIP-seq datasets sorted by their sizes in the ascending order. (**B**) Distribution of binding peak lengths in the 620 datasets. Vast majority (99.9%) of them are shorter than 800 bp. (**C**) Coverage of CDRs and NCRs in the genome by the extended peaks (datasets), predicted CREs and CRMs. The numbers above the lines are the proportions of CDRs and NCRs in the corresponding sequence categories; the numbers below the lines are the proportions of CDRs and NCRs with respect to the entire CDRs and NCRs in the genome, respectively.

where α, β and δ are constants, and n the number of datasets used for the predictions.

RESULTS

Extended binding peaks of different datasets for cooperative TFs have large overlaps

Although the 620 datasets were collected from 79 cell/tissue types for 168 TFs, they were sampled highly unevenly, as few cell/tissue types such K526 and Gm12878 (Supplementary Figures S1A and B) and TFs such as CTCF and POL2 (Supplementary Figures S1C and D) have a large number of datasets, while the vast majority of other cell/tissue types and TFs have only one or two datasets. Each dataset also contains a highly varying number (1–92 358) of binding peaks (Figure 1A) with 366 of them containing >10 000 peaks. The 620 datasets contain a total of 10,894,581 peaks. The vast majority (99.9%) of them have a length shorter than 800 bp (Figure 1B). After the length extension, the datasets have a total length of 32 677 939 091 bp, which are 10.42 times the human genome (3 137 161 264 bp). However, they only cover 48.5% (1 522 974 911) of the genome (Figure 1C), indicating that these extended sequences have significant overlaps. Of the 1 522 974 911 bp genome sequence covered by the datasets, 1 470 721 860 bp (96.6%) are NCRs, consisting of 48.1% of NCRs (3 059 588 382 bp) in the genome (Figure 1C). The remaining 52 253 051 pb (3.4%) extended sequences are in the coding regions (CDRs), consisting of 67.4% of CDRs (77 572 882 bp) in the genome (Figure 1C). In addition, 789 (88%) of the 897 enhancers from VISTA are located in our extended peaks.

To see the overlapping patterns in the datasets, we computed a pair-wise overlapping score S_o among the 620 datasets using formula (1), and clustered the datasets based on the scores. As shown in Figure 2A, there are clearly numerous clusters formed by some datasets, indicating that their sequences highly overlap with one another. Interestingly, datasets of TFs that are known to work cooperatively in regulating genes form a cluster. For example, the small cluster highlighted in Figure 2B is formed by the datasets of TFs ZNF274, KAP1 and SETDB1. It has been shown that knockdown of ZNF274 with siRNAs reduced the levels of KAP1 and SETDB1 binding to the ZNF274 binding regions, suggesting that ZNF274 is involved in the recruitment of KAP1 and SETDB1 to specific regions of the human genome (96). Another cluster highlighted in Figure 2C for is formed by datasets of TFs RAD21, CTCF and SMC3. It is well known that RAD21 and SMC3 are the members of the cohesin complex, and that cohesin co-localizes with CTCF at more than 80% of CTCF binding locations (97). Therefore, these results indicate that the 620 datasets contain sufficient information to predict at least a portion of CREs and CRMs in the human genome by exploring repeated co-occurring motif patterns.

Identification of motifs

Our goal is to find all possible binding motifs of the ChIP-ed TF and of its cooperators in each dataset. To facilitate motif finding in the 366 large datasets containing >10 000 binding peaks, we split them into a total 1,150 datasets, ending up with a total of 1433 datasets, each contains <10 000 binding peaks. We found a varying number (0–121) of motifs in each dataset, depending on the quality and size of the dataset (Figure 3A). There are 14 datasets containing 1∼355 peaks, in which DREME was not able to find any motif, so they were filtered out at this step. To see the effects of splitting a large dataset in smaller ones on the motif finding results, we randomly split three datasets with 22 314 (S_1), 30 924 (S_2) and 40 670 (S_3) peaks in three, four and five sub-datasets, respectively, so each sub-dataset contained <10 000 peaks, and found motifs in each of the resulting sub-datasets. We repeated this process by 10 times. As shown in Figure 3B, in all the three cases, the number of motifs identified in each subset are quite similar and are similar to the number of motifs identified by the way of splitting used in the algorithm. The identified motifs in each subset for the same dataset are

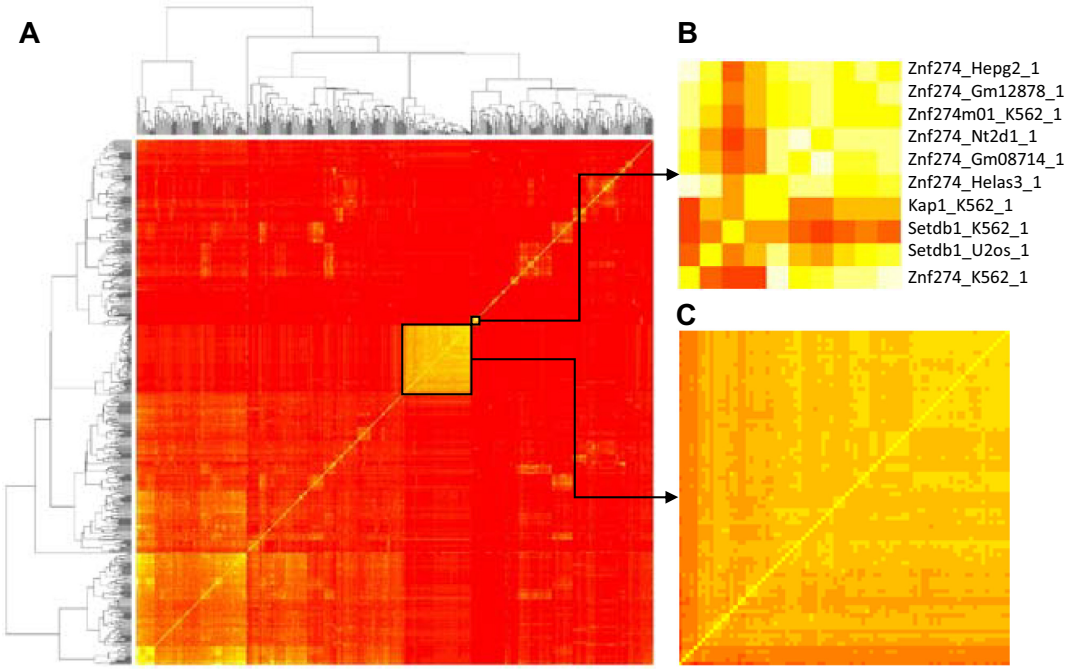

Figure 2. Overlaps of extended binding peaks. (**A**) Hierarchical clustering of the 620 datasets for 168 TFs based on the pair-wise overlapping scores S_o of the extended peaks. (**B**) Blow-up of a cluster of datasets for functionally related TFs ZNF274, KAP1 and SETDB1. **C.** Blow-up of a cluster of datasets for functionally related TFs RAD21, CTCF, SMC3 and ZNF143, etc.

also very similar (data not shown). Therefore, the way to splitting a large dataset does not significantly affect motifs identified in the dataset. The returned motifs generally have high information content (Figure 3C). Overall, we identified 50 856 putative motifs in the datasets, containing a total of 78 456 541 putative CREs. Interestingly, putative CREs were found in all the 789 VISTA enhancers that are located in the extended peaks.

Prediction of CREs and CRMs in the human genome

We applied DePCRM to these 50 856 putative motifs to predict CRMs and CREs in the genome. Since these input motifs may contain a large number spurious ones due to the high false positive rate of motif-finders including DREME, DePCRM identifies overrepresented co-occurring motif patterns as possible CRMs by gradually filtering out spurious motifs based on the assumption that truth motifs in CRMs are more likely than do spurious ones to co-occur in the same sequence. DePCRM first identifies highly co-occurring motif pairs (CPs) in each dataset using the motif co-occurring scores S_c (formula 2) for pairs of putative motifs found in each dataset. As shown in Figure 3D, the distribution of S_c is strongly skewed toward right, indicating that the low-scoring Gaussian-like component is likely due to spurious motif pairs that occur by chance. To find a proper cutoff α for S_c such that most spurious motif-pairs are filtered out, while at least most the true motif pairs are kept, we plotted the percentage of total putative motif pairs with $S_c > \alpha$ and percentage of VISTA enhancers containing putative CREs of motif pairs with $S_c > \alpha$ as functions α. As shown in Figure 3E, when $\alpha = 0.34$, 28 581 (56.2%) of the 50 856 input motifs were filtered out, while only 74 (9.38%) of the 789 VISTA enhancers were lost. Thus, we chose $S_c > \alpha = 0.34$ as the cutoff, resulting in 22 265 (43.8%) motifs forming 76 764 CPs.

To further enrich true motif pairs, DePCRM next identifies repeatedly occurring CPs in multiple datasets by clustering highly similar CPs in different datasets. To this end, DePCRM constructs a CP similarity graph using the CPs as the nodes, and S_s scores $> \beta$ (formula 3) as the weights on the edges (see Materials and Methods). To find the optimal vale of β, we plotted the density of the graph as a function of β. As shown in Figure 3F, with the increase in β, the density of the graph drops rapidly, but the dropping stops around $\beta = 1.8$, while the number of nodes (CPs) in the graph starts decreasing rapidly (Figure 3G). Thus, we set $\beta = 1.8$ to construct the CP similarity graph. Applying the Markov chain clustering (MCL) algorithm (95) to the graph resulted in 13 364 CP clusters (CPCs) containing 53 278 CPs involving 20 640 motifs.

To identify larger combinatorial motif patterns, DePCRM then identifies CPCs whose CPs tend to co-occur in the same sequence by constructing a CPC co-occurring graph based on a S_{cpc} (formula 4) cutoff scores γ (see Materials and Methods). Interestingly, the distribution of S_{cpc} displays a well-separated bimodal distribution (Figure 3H); the low-scoring peak is likely mainly due to motif patterns occurring by chance, while the high-scoring one is likely attributable to truly cooperative motifs in CRMs. Thus, we set $\gamma = 0.37$ (the value at the valley between the two peaks). Applying the MCL algorithm to the resulting CPC co-occurring graph resulted in a total of 846 CRMCs involving 12 022 putative motifs, each containing 2∼184 CPCs (Supplementary Figure S2). Therefore, the algorithm eventually

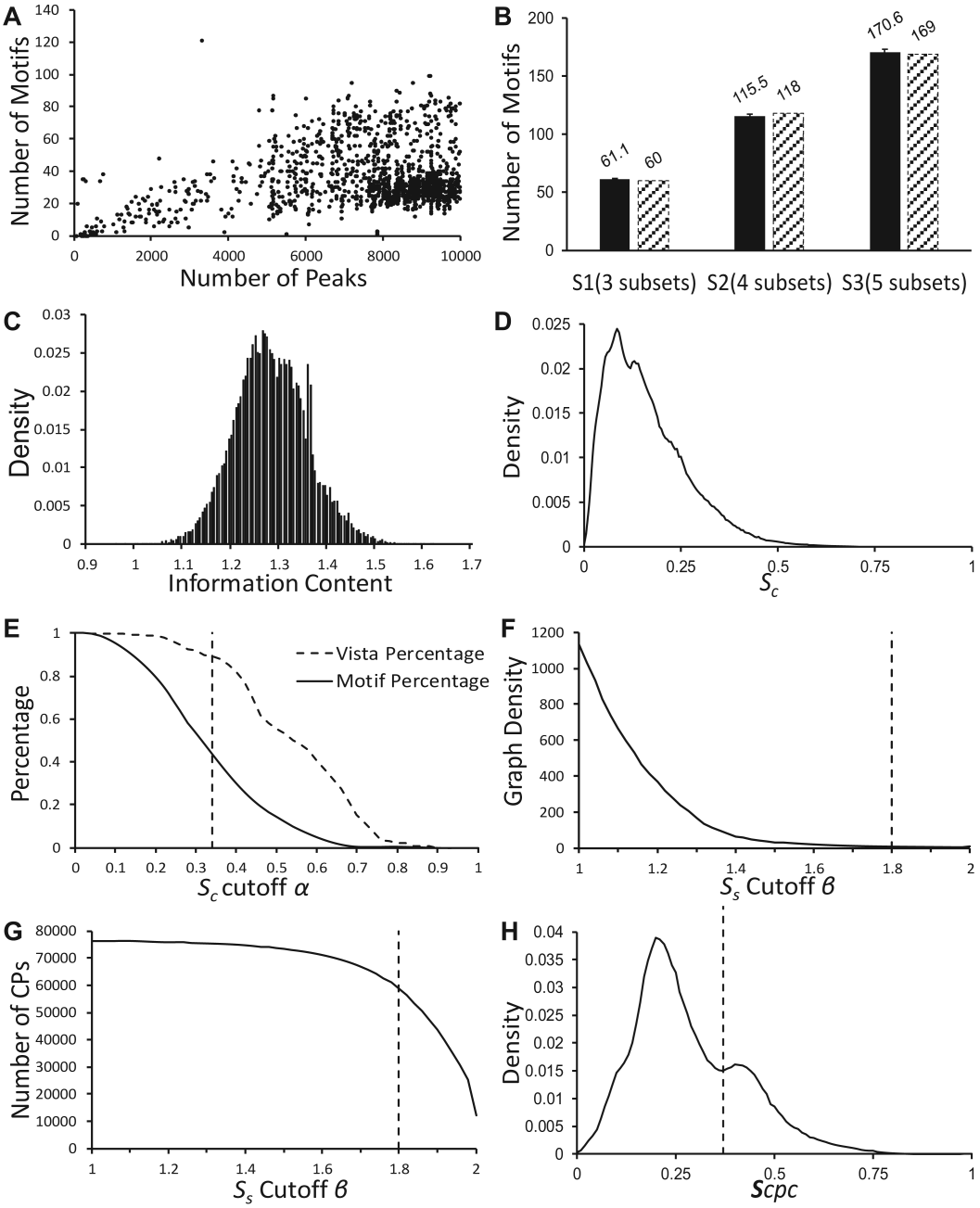

Figure 3. Results from the key steps of the prediction pipeline. (**A**) Number of motifs found as a function of the size of the 1433 sub-/datasets. (**B**) Average ($n = 10$) numbers of motifs found in the sub-datasets for datasets S1, S2 and S3, randomly split in three, four and five sub-datasets, respectively (black bars), and the number of motifs found by the way of splitting used in this study (hatched bars). (**C**) Distribution of the information contents of the identified motifs. (**D**) Distribution of motif co-occurring scores S_c. (**E**) Proportions of the motifs (solid line) and the VISTA enhancers (dashed line) that are remained as a function of S_c score cutoff α. We choose $\alpha = 0.34$ (the vertical line) to exclude as many as possible predicted motifs, and at the same time, to include as many as possible predicted motifs in known CRMs. (**F**, **G**) Density of the CP similarity graph and number of CRMs in the graph as a function of the S_s cutoff β. We choose $\beta = 1.8$ (the vertical lines) so that the density of the graph is largely minimized, and at the same time, the number of nodes/CPs in the graph is largely maximized. (**H**) Distribution of CPC co-occurring scores S_{CPC}. The vertical line indicates the S_{CPC} cutoff $\gamma = 0.37$ at the deepest valley between the two peaks, for constructing the CPC co-occurring graph.

filtered a total of 38 826 (76.4%) of the 50,848 input motifs, which are likely spurious predictions.

However, some of the 12,022 identified motifs are highly similar to one another, they may share the same CREs given the fact that sequences of some datasets may have significant overlaps (Figure 2). By combing highly similar and/or overlapping motifs via clustering (see Material and Methods), we identified 736 U-motifs (Supplementary Figure S2B), each containing 1–608 highly similar motifs, and 46–204 896 unique CREs. As examples, Figures 4A and 5B shows U-motif 367 and U-motif 389 and their respective four and three motif members. When compared with the known motifs from Jolma *et al.* (98) and JASPAR CORE vertebrate (99), 370 (50.3%) of the U-motifs are highly similar to known motifs in Human at $P < 0.001$ using TOMTOM (100), suggesting that they are likely to be true motifs. For example, U-motif 367 and U-motif 389 are very similar to the JASPER motifs HSFY2 (Figure 4A) and ZIC3 (Figure 4B), respectively. We replaced the motifs in the CRMCs with the U-motifs that they belong to and represented each CRMC by their constituent U-motifs.

Projecting the CREs in these 846 CRMCs back to the human genome (Materials and Methods) resulted in a total of 1 178 913 non-overlapping CREs with 1 140 005 (96.7%) being in NCR and the remaining 38 908 (3.3%) being in CDRs. These 1 178 913 CREs cover 11 772 217 bp (0.4%) genome sequence, of which 11 313 594 bp (96.1%) are in NCRs consisting 0.4% of NCRs, and the remaining 458 623 bp (3.9%) are in CDRs, consisting of 0.4% of CDRs (Figure 1C). By connecting these putative CREs, we predicted a total of 305 912 non-overlapping CRMs, 264 035 (86.3%) of which are entirely located in NCRs, and the remaining 41 877 (13.7%) are at least partially located in CDRs. These 305 912 CRMs cover 55 605 307 bp (1.8%) of genome sequence, of which 534 919 752 bp (96.2%) are in NCRs, consisting 1.7% of NCRs, and the remaining 2 113 332 bp (3.8%) are in CDRs, consisting of 2.7% of CDRs (Figure 1C). These putative CRMs tend to have shorter lengths than those of the known CRMs (Figure 5A). Furthermore, the putative CRMs harbor 2–89 CREs with a median of 2, and only a small portion of the putative CRMs tends to have a short distance between adjacent two putative CREs (Figure 5B). These results suggest that we might have missed certain CREs in the predicted CRMs, particularly at the two ends, presumably due to insufficient information in the limited number of ChIP datasets used in this study. In other words, some of our predicted CRMs might consist of only a part of otherwise longer real CRMs, possibly missing CREs at the two ends of CRMs. Clearly, in order to make more accurate and complete predictions, more and highly diverse ChIP datasets are needed.

Predicted CREs in NCRs are more likely to be evolutionarily constrained than randomly selected sequences

It is widely recognized that functional elements such as protein-coding exons and regulatory sequences are usually under negative (purifying) or positive selection while non-functional sequences are often selectively neutral or nearly so. Therefore, we compared the conservation levels of nucleotides in putative CREs in the 264 035 predicted CRMs located in NCRs with those of the same number and length of sequences randomly selected from NCRs. We quantified the conservation level of each nucleotide using GERP++ (101), which estimates the substitution rate at each position in the human genome based on multiple alignments of 34 mammalian genomes. GERP++ computes a rejected substitution (RS) score for each position relative to selectively neutral sequences. Thus, a positive RS score indicates purifying selection at the position, thus it is conserved; a negative RS score might indicate positive selection at the position; and a RS score around 0 suggest that the position is selectively neutral or nearly so. As shown in Figure 5C, the average RS scores of a putative CRE and of a randomly selected NCR sequence (50 repeats) have remarkably different distributions ($P < 2.2 \times 10^{-302}$, Kolmogorov–Smirnov test). More specifically, the distribution for randomly selected NCR sequences is narrower with a high peak at score $= 0$ (areas in the window (–0.6, 0.2) is 44.35%); by contrast, that of putative CREs in NCRs is broader with only a small peak at 0 (areas in the window (–0.6, 0.2) is 32.40%), indicating that randomly selected NCR sequences are more likely to be selectively neutral or nearly so as expected. Moreover, compared with the randomly selected sequences, the predicted CREs in NCRs are either more likely to be negatively selected with a RS score ≥ 0.2 (30.86% versus 20.90%), or more likely to be moderately positively selected with a RS score within [–2, –0.6] (32.37% versus 28.92%). Similar results were obtained for the average score of all nucleotides in a CRM (Figure 5D). Therefore, our predicted CREs and CRMs in NCRs are likely to be functional and thus authentic. The evaluation of the predicted CREs and CRMs in CDRs will be addressed in a separate manuscript.

Functional elements revealed by independent studies are highly enriched in our predicted CRMs

DHSs are the regions in the genome that have less condense structure in certain cells or tissues, and thus are highly sensitive to cleavage by DNase I enzyme. They are also likely bound by TFs in these cells or tissues, working as CRMs. A large number of DHSs in 125 human cell or tissue types have been recently determined by the ENCODE consortium, hence we used them as additional line of independent evidence to further validate our predicted CRMs. We consider a DHS is recovered by a predicted CRM if the DHS overlaps with predicted CRM by at least a single nucleotide. Of the 1 281 988 non-overlapping DHSs (total length: 388 420 483 bp, 12.38% of the genome) from the ENCODE consortium, 1 059 387 (82.64%, total length: 330 454 362 bp, 10.53% of the genome) are located in the extended peaks, indicating that they are also highly enriched in the datasets. As shown in Figure 5E, 156 153 (14.76%) of the DHSs in the extended peaks are recovered by our predicted CRMs; by contrast, the same number and length of sequences randomly selected from the genome covered by the extended peaks can only recover 77 711 (7.35%) of the DHSs. This number is close to the expected recovery rate by chance (7.67%). Thus, our predicted CRMs recovered twice more DHSs than did randomly selected sequences, although the DHSs were derived from far more (125) cell/tissue types than the 79 cell/tissue types from which the datasets we used

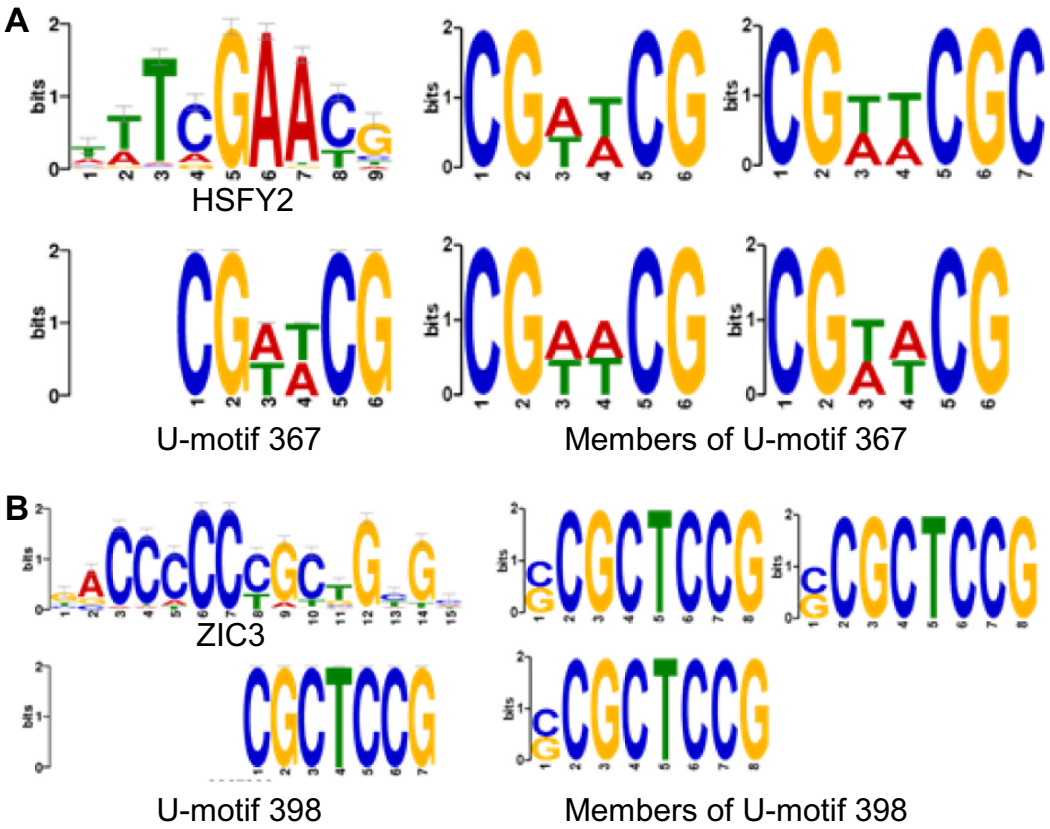

Figure 4. Examples of U-motifs and their member motifs. (**A**) U-motif 367, its matched known motif for TF HSFY2 and its four motif members. (**B**) U-motif 389, its matched known motif for TF ZIC3 and its three motif members.

were derived, and the DHSs may include all active CRMs in these 125 cell/tissue types, while our predicted CRMs are largely limited by the used datasets for a small number of TFs.

Additionally, we validated our predicted CRMs with the 789 experimentally verified enhancer segments from VISTA (90), which are located in the extended peaks. Although we used them to help set the S_c cutoff in the very early step of the algorithm, if our algorithm does not work, they can still be lost as vast majority of the input motifs are dropped out by the algorithm. However, our predicted CRMs recover 405 (51.3%) of the 789 enhancers. By contrast, the same number and length of sequence randomly selected from the extended peaks only recover an average of 202 (25.5%) of the 789 enhancers (50 repeats, Figure 5F). This number is close to the expected recovery rate by chance (25.75%). Thus, our predicted CRMs recovered twice more known enhancers than did randomly selected sequences. Taken together, all these three lines of independent evidence indicate that our predicted CREs and CRMs are likely to be authentic.

Comparison with existing methods for predicting TF binding motifs

Since no similar method for predicting CREs and CRMs at a genome scale by integrating a large number of TF ChIP-seq datasets has been seen in the literature to our best knowledge, we compared our predicted motifs with those reported by the ENCODE consortium who attempted to identify all possible TF motifs using the ENCODE TF ChIP-seq datasets (102,103). These studies performed independent motif finding in each dataset using only the top 500 representative binding peaks with a length of 100bp using MEME-ChIP (104). As a result, they were only able to identify 79 unique motifs (clustered from 1092 motifs) in 457 ChIP-seq datasets for 119 TFs (102,103). Thus, we identified 8.3 times more U-motifs (736 clustered from 12,022 motifs) using only 163 more datasets (620 versus 457). In addition, Kheradpour and Kellis (105) developed a pipeline that combined five motif-finders (AlignACE (106), MDscan, MEME (107), Weeder (108,109) and Trawler (110,111)) to systematically predict TF motifs in 427 ENCODE ChIP-seq datasets for 123 TFs belonging to 84 protein families. Again, they used only the top 250 representative binding peaks in one of two partitions of each dataset for motif finding, identifying a total of 468 motifs matching known motifs of the target TFs and 293 new motifs that did not match any known motifs. As the former 468 motifs are very similar to each other, presumably recognized by different TFs of the same family (105), we clustered them to avoid redundancy, resulting in 56 unique motifs. Therefore this study identified a total of 349 (56 + 293) unique motifs, which is less than half we identified. More importantly, as we stated earlier, more than 50% of our 736 predicted U-motifs match

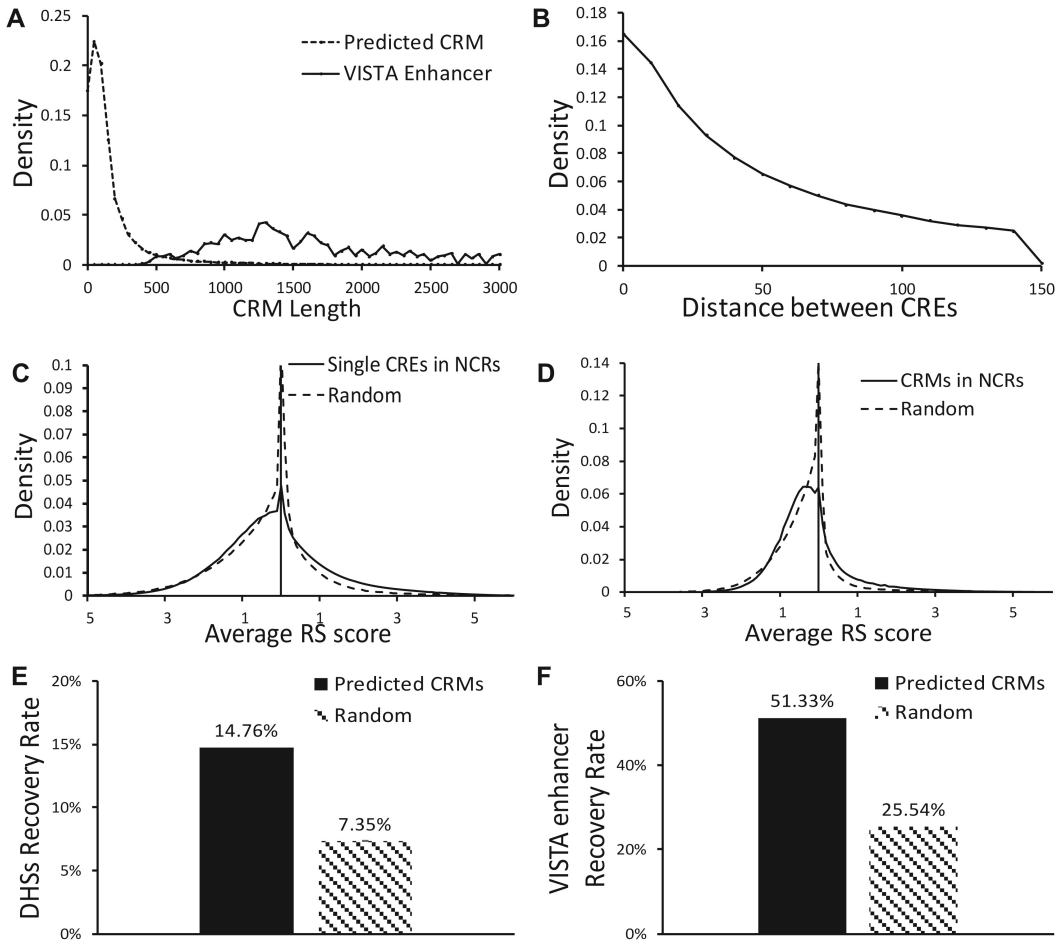

Figure 5. Validation of the predicted CREs and CRMs. (**A**) Distributions of the lengths of the known Vista enhancers and predicted CRMs. (**B**) Distribution of the distances (bp) between two adjacent CREs in a CRM. (**C**) Distributions of average RS scores of single predicted CREs in NCRs and the same number and length of sequences randomly selected from NCRs. (**D**) Distributions of average RS scores of a predicted CRM in NCRs and the same number and length of sequences randomly selected from NCRs. (**E**) Enrichment of DNase I hypersensitive sites in the predicted CRMs. (**F**) Enrichment of VISTA enhancers in the predicted CRMs.

known motifs, hence at least the vast majority of them are likely to be authentic.

Saturation analysis of predicted U-motifs

Next, we analyzed the trends of changes in the numbers of U-motifs predicted using an increasing number of datasets in three scenarios (Materials and Methods). In the first scenario, we used as examples cell lines in which enough number of datasets for different TFs are available, including K562 (the first human immortalized myelogenous leukemia cell line, 144 datasets), GM12878 (a lymphoblastoid cell line, 88 datasets) and Helas3 (a sub-clone of the HeLa cell line, 59 datasets). We plotted the number of recovered U-motifs as a function of the number of selected datasets. However, as some of these datasets did not contribute to the prediction in the cell lines (often for lightly sampled TFs, see below), we excluded them from this analysis. The results in the K562 cells based on 113 datasets (31 were excluded) show a saturation trend for the predicted U-motifs, as indicated by the decreasing slopes of the trend lines from the first quarter to the fourth quarter of the plot (Figure 6A). The trend of saturation is notable when as few as ∼10 datasets were selected for the prediction (Figure 6A). The saturation trend can be well fitted to a sigmoid function (formula 5, Figure 6B). Extrapolation of the fitting function suggests that up to ∼580 U-motifs (Figure 6B) could be predicted if ∼1200 such datasets in the K562 cells would be used. Therefore, the 113 datasets in the K562 cells predict 62.07% (360) of the saturation prediction. However, as indicated by the fitting curve (Figure 6B), the number of predicted U-motifs increases slowly when more than 200 datasets would be used, suggesting that generation of more than 200 such datasets in this cell line by the same strategy as currently used would not be cost-effective for predicting U-motifs. Interestingly, very similar results were obtained using the datasets from the GM12878 (Supplementary Figures S3A and B) and HeLaS3 (Supplementary Figures S3C and D) cells for different TFs. In both cases, the number of predicted U-motifs also decreases rapidly when >200 datasets would be used.

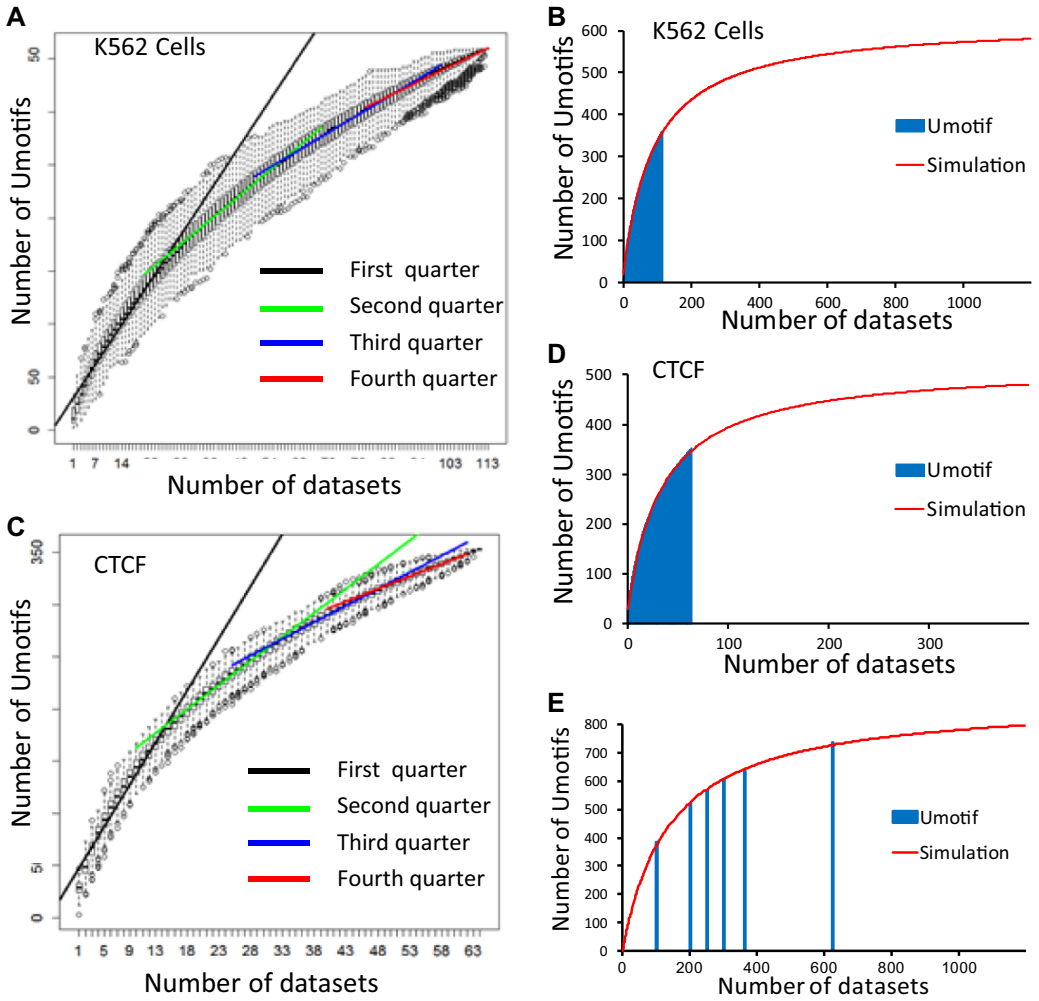

Figure 6. Trends of predicted U-motifs using an increasing number of datasets in three scenarios. (**A**) Number of predicted U-motifs as a function of the number of datasets used, collected from the K562 cells for different TFs. Trend lines are plotted for the first, second, third and fourth quarters of the datasets used. The data points are presented using box-plot based on 50 repeats. (**B**) Fitting the numbers to a sigmoid function of predicted U-motifs using varying numbers of datasets collected from the K562 cells, and extrapolation of the saturation trend. (**C**) Number of predicted U-motifs as a function of the number of datasets used, generated for TF CTCF in different cell/tissue types. Trend lines are plotted for the first, second, third and fourth quarters of the datasets used. The data points are presented using box-plot based on 50 repeats. (**D**) Fitting the numbers to a sigmoid function of predicted U-motifs using varying numbers of datasets generated for CTCF, and extrapolation of the saturation trend. (**E**) Number of predicted U-motifs as a function of the number of randomly selected datasets used. The data points are average of five repeats, and the curve is the fitting of the results to a sigmoid function and its extrapolation.

In the second scenario, we used well studied TFs for which a relatively large number of datasets from different cell/tissue types are available, including CTCF (64 datasets) that is involved in insulator activity (112), V(D)J recombination (113), and regulation of chromatin architecture (114); NRSF (12 datasets) that is involved in the repression of neural genes in non-neuronal cells (115); and NF-kB (10 datasets) that is involved in the immune and inflammatory responses, developmental processes, cellular growth, and apoptosis. The results for the 64 datasets for CTCF display a trend of saturation when as few as ∼10 datasets were selected for the prediction (Figure 6C). The result can be well fitted to a sigmoid function, and extrapolation of the fitting suggests that up to 480 U-motifs (Figure 6D) could be predicted if ∼380 such datasets for CTCF would be used. Therefore, the 64 datasets available for CTCF predict 72% (346) of the saturation prediction of U-motifs (480) for datasets for CTCF. However, as indicated by the fitting curve (Figure 6D), the number of predicted U-motifs increases slowly when >100 datasets are used, suggesting that generation of >100 datasets for this TF by the same strategy as currently used would not be cost-effective for predicting U-motifs. Similar results were obtained for NRSF (Supplementary Figures S4A and B) and NFKB (Supplementary Figures S4C and D).

In the third scenario, we calculated the number of predicted U-motifs using a varying number (100, 200, 250 300, 350) of randomly selected datasets as well as the entire 620 datasets. As shown in Figure 6E, the number of predicted U-motifs increased with the increase in the number of

datasets used, but it rapidly enters a saturation phase when ∼350 datasets were used. Extrapolation of the fitting function suggests that we could predict 796 U-motifs using a sufficiently large number of datasets produced by the current strategy. Thus, we have predicted vast majority (736, 92.46%) of the 796 U-motifs using the 620 datasets. Leaving out all the 199 datasets from cell/tissue types and for TFs that were lightly sampled (with at most two datasets) had little effect on the of saturation pattern (data not shown). Therefore, the rapid saturation of the number of predicted U-motifs when only 350 datasets were used might be due to the fact that the datasets used in the prediction are heavily biased to a few cell/tissue types and TFs, while most cell/tissue types and TFs were under sampled (Supplementary Figure S1). This result suggests that it would not be very cost effective to generate even more than ∼350 such biased datasets for predicting U-motifs.

Applications of the predicted CREs and CRMs

To facilitate the utilization of these predicted CREs and CRMs by the research community, we have developed a webserver (http://bioinfo.uncc.edu/pcrms) for queries and visualization of the predictions and their genomic contexts. For example, a user can search for predicted CREs and CRMs surrounding a gene of interest. Figure 7A shows the results of querying the server with gene name *CASZ*1, which is involved in blood vessel assembly and morphogenesis (116). The result shows that a predicted CRM c31586754 (chr1:10781329–10781587) with a length of 259bp is located in the first intron of *CASZ*1. This putative CRM completely overlaps with the VISTA enhancer element 389 (chr1:10781239–10781744) with a length of 506 bp (90), thus c31586754 is authentic. We predicted that this CRM contains six putative CREs chr1:10781329–10781337, chr1:10781401–10781409, chr1:10781425–10781433, chr1:10781443–10781451, chr1:10781493–10781501 and chr1:10781579–10781587 for U-motifs U-motif 0, U-motif 1, U-motif 8, U-motif 31, U-motif 92 and U-motif 254, which match known motifs of TFs KLF1, SOX6, FOXP1, RARA, ETV6 and ZFX, respectively. Thus, it is interesting to experimentally test whether *CASZ1* is co-regulated by these six TFs. Figure 7B shows another example, where a predicted CRM c31828239 (chr6: 41570097–41570164) with a length of 73bp is located in the second intron of *FOXP4* that plays a crucial role in brain development and autism (117,118). This putative CRM contains three CREs chr6:41570097–41570105, chr6:41570130–41570139 and chr6:41570156–41570164 for U-motifs U-motif 24, U-motif 132 and U-motif 476, which match known motifs of TFs KLF16, HINFP and GLIS2, respectively. Hence, it is interesting to investigate whether these three TFs co-regulate the expression of FOXP4 via this putative CRM.

DISCUSSION

The DePCRM algorithm predicts CREs and CRMs largely based on the fact that similar TF combinatorial patterns are often repeatedly used to regulate multiple the same or different regulons in different cell/tissue types, developmental stages or physiologically conditions. As the predicted motifs by motif-finders in large datasets may contain a large portion of spurious ones, and number of possible combinations of TFs is extremely large, DePCRM predicts CREs and CRMs by rapidly filtering out spurious motifs and combinations using a branch and bound approach. More specifically, it identifies possible real motif combinatorial patterns in a large number of ChIP datasets through iteratively filtering out randomly occurring spurious motifs, thereby effectively reducing the searching space in each step. Having successfully demonstrated that DePCRM works for the *D. Melanogaster* genome (88), we applied the algorithm to the much larger human genome with more and bigger ChIP datasets. In order to make it work more efficiently on large human datasets, we modified the algorithm by splitting the large datasets into smaller ones. Such splitting has little effect on the motif-finding results, due probably to the information redundancy in large ChIP datasets (Figure 3B). Use all the 620 TF ChIP-seq datasets from ENCODE available to us, we have predicted an unprecedentedly complete map of 305,912 CRMs containing 1 178 913 CREs in the human genome at single nucleotide resolution.

Three lines of independent evidence indicate that our predicted CREs and CRMs are likely to be authentic. First, our predicted CREs and CRMs in NCRs are more likely to have gone either strongly negative selection, or moderately positive selection (Figure 6C and D), indicating that they are highly likely to be functional. This observation is in excellent agreement with the consensus that regulatory sequences tend to be more conserved due to negative selection, or to undergo rapid turnover by degrading existing CREs (death), or gaining new CREs (birth) due to positive selection, a process called CRE turnover (119). CRE turnover plays a more pivotal role in evolutionary divergence of organisms than previously thought (13,120), including the evolution for human-specific functions including intelligence. Second, our predicted CRMs recovered twice more DHSs in the extended peaks than expected by chance. Finally, our predicted CRMs recovered twice more Vista enhancers in the extended peaks than expected by chance.

In principle, to predict all CREs and CRMs in the human genome, we need a sufficiently large number of diverse and less biasedly sampled ChIP-seq datasets from various cell/tissue types and for various TFs, so that information about all possible combinatory regulations among all TFs would be included. Therefore, it is interesting to evaluate the status of the available datasets and the strategy that have been used to generate them to reach the goal. To this end, we analyzed the saturation trends of the numbers of predicted U-motifs under three scenarios based on the 620 ENCODE datasets. When a large number of datasets for different TFs are available in a cell/tissue type (Figure 6A and B), and a larger number of datasets in different cell/tissue types are available for a TF (Figure 6C and D), the trends of saturation develop rapidly with the increasing number of datasets used, presumably due to the facts that these datasets are biased to well-studied cooperative TFs in relevant cell types. Thus, this strategy is highly effective for revealing functional CREs for cooperative TFs in relevant cell lines. On the other hand, when all the datasets were considered (Figure 6E),

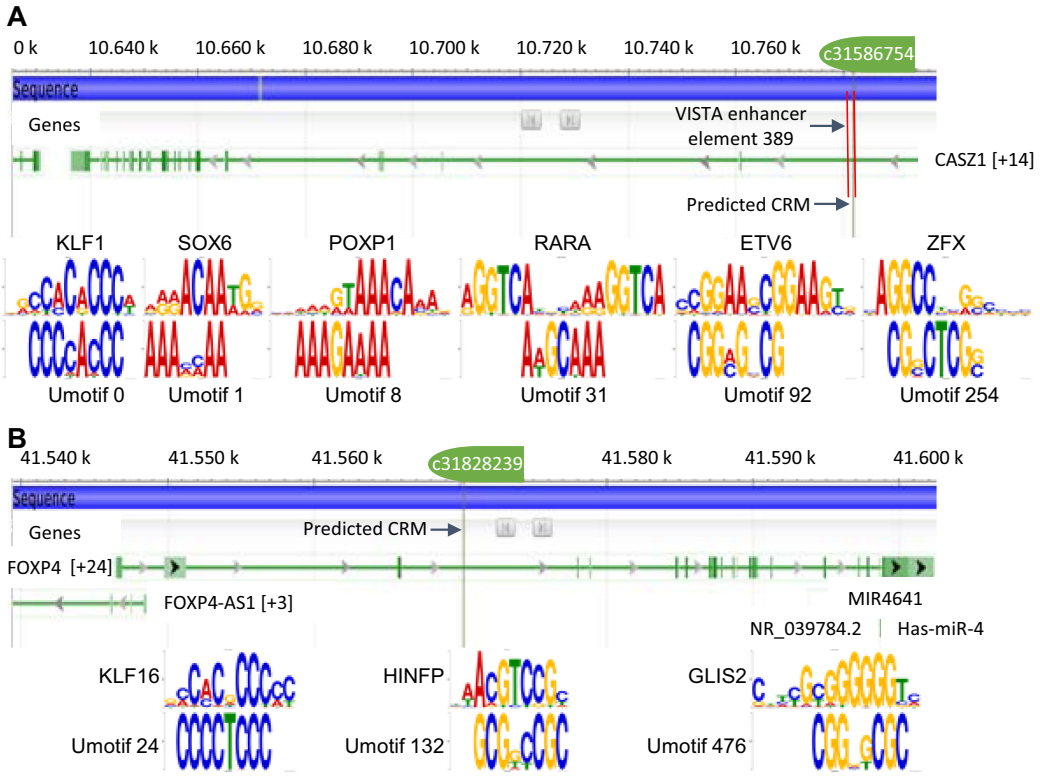

Figure 7. Examples of predicted CRMs and constituent CREs in genomic contexts. (**A**) A predicted CRMs c31586754 located in the first intron of the *CASZ1* gene overlaps entirely with the VISTA enhancer 389. c31586754 contains six CREs of U-motifs U-motif 0, U-motif 1, U-motif 8, U-motif 31, U-motif 92 and U-motif 254, which match known motifs of TFs KLF1, SOX6, FOXP1, RARA, ETV6, and ZFX, respectively. (**B**) A predicted CRM c31828239 is located in the second intron of the *FOXP4* gene. c31828239 contains three CREs of U-motifs U-motif 24, U-motif 132 and U-motif 476, which match known motifs of TFs KLF16, HINFP and GLIS2, respectively.

our analysis suggests that using only 620 datasets we were able to predict 92.46% (736) of the saturation number (796) of U-motifs that would be predicted if >1000 such datasets were used. In fact, using more than 350 such datasets are no longer cost-effective for predicting as many as possible U-motifs. This might be due to the fact that the highly biased sampling of the datasets led to biased predictions of CREs and CRMs working in few cell/tissue types or for few TFs that were heavily sampled (Supplementary Figure S1), while missing those working in the majority of cell/tissue types or for the majority of TFs that were lightly sampled. Therefore, more diverse and less biased datasets for various TFs and from various cell/tissue types are urgently needed for more cost-effective prediction of U-motifs and CRMs in the human genome.

In addition, our results allow us to estimate the lower bound of the size or proportion of the human genome that are involved in transcriptional regulation. With the 620 datasets covering about half of (48.5%) the genome, we predicted 736 U-motifs and 305 912 CRMs. Extrapolating these results, we estimate that there are at least 1518 U-motifs and 630 746 CRMs in the human genome. This estimate of U-motif number is consistent with the number of TFs encoded in the human genome which is 2,000∼3,000 as estimated by early studies (121,122), considering the possibility that a U-motif may include several highly similar motifs of multiple TFs in the same protein family. Additionally, using these 620 datasets, our predicted CREs and CRMs covers 0.4% and 1.8% of the human genome, respectively (Figure 1C). Assuming that these results are extendable to the other part of the genome that are not covered by the datasets, we estimate that at least 0.83% and 3.71% of the genome might code for CREs and CRMs, respectively. We anticipate that when more less biased datasets are available in the future, more accurate predictions and estimates can be made. As 2.5% of the genome are CDRs (Figure 1C), we estimate that at least 6.21% (2.5 + 3.71) of human genome are functional, which is in agreement with the estimate that that ∼7% of the human genome are conserved and thus are functional (101).

DATA AVAILABILITY

The predicted CREs and CRMs are available at http://bioinfo.uncc.edu/pcrms for various searches.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Science Foundation [DBI1661332 to Z.S.]; National Institutes of Health [R01GM106013 to Z.S.]. Funding for open access charge: National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
2. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
3. Temple,G., Gerhard,D.S., Rasooly,R., Feingold,E.A., Good,P.J., Robinson,C., Mandich,A., Derge,J.G., Lewis,J., Shoaf,D. *et al.* (2009) The completion of the Mammalian Gene Collection (MGC). *Genome Res.*, **19**, 2324–2333.
4. Maston,G.A., Evans,S.K. and Green,M.R. (2006) Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, **7**, 29–59.
5. Narlikar,L. and Ovcharenko,I. (2009) Identifying regulatory elements in eukaryotic genomes. *Brief. Funct. Genomic Proteomic*, **8**, 215–230.
6. Alexander,R.P., Fang,G., Rozowsky,J., Snyder,M. and Gerstein,M.B. (2010) Annotating non-coding regions of the genome. *Nat. Rev. Genet.*, **11**, 559–571.
7. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
8. Hardison,R.C. and Taylor,J. (2012) Genomic approaches towards finding cis-regulatory modules in animals. *Nat. Rev. Genet.*, **13**, 469–483.
9. Davidson,E.H. (2006) *The Regulatory Genome: Gene Regulatory Networks in Development And Evolution*. Academic Press.
10. Rubinstein,M. and de Souza,F.S. (2013) Evolution of transcriptional enhancers and animal diversity. *Philos. Trans. R Soc. Lond. B Biol. Sci.*, **368**, 20130017.
11. Douglas,A.T. and Hill,R.D. (2014) Variation in vertebrate cis-regulatory elements in evolution and disease. *Transcription*, **5**, e28848.
12. Evans,N.C., Swanson,C.I. and Barolo,S. (2012) Sparkling insights into enhancer structure, function, and evolution. *Curr. Top. Dev. Biol.*, **98**, 97–120.
13. Wittkopp,P.J. and Kalay,G. (2012) Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.*, **13**, 59–69.
14. Fraser,H.B. (2013) Gene expression drives local adaptation in humans. *Genome Res.*, **23**, 1089–1096.
15. Ye,K., Lu,J., Raj,S.M. and Gu,Z. (2013) Human expression QTLs are enriched in signals of environmental adaptation. *Genome Biol. Evol.*, **5**, 1689–1701.
16. Babak,T., Garrett-Engele,P., Armour,C.D., Raymond,C.K., Keller,M.P., Chen,R., Rohl,C.A., Johnson,J.M., Attie,A.D., Fraser,H.B. *et al.* (2010) Genetic validation of whole-transcriptome sequencing for mapping expression affected by cis-regulatory variation. *BMC Genomics*, **11**, 473.
17. Stranger,B.E., Montgomery,S.B., Dimas,A.S., Parts,L., Stegle,O., Ingle,C.E., Sekowska,M., Smith,G.D., Evans,D., Gutierrez-Arcelus,M. *et al.* (2012) Patterns of cis regulatory variation in diverse human populations. *PLoS Genet.*, **8**, e1002639.
18. Vernot,B., Stergachis,A.B., Maurano,M.T., Vierstra,J., Neph,S., Thurman,R.E., Stamatoyannopoulos,J.A. and Akey,J.M. (2012) Personal and population genomics of human regulatory variation. *Genome Res.*, **22**, 1689–1697.
19. Zheng,W., Zhao,H., Mancera,E., Steinmetz,L.M. and Snyder,M. (2010) Genetic analysis of variation in transcription factor binding in yeast. *Nature*, **464**, 1187–1191.
20. Kasowski,M., Grubert,F., Heffelfinger,C., Hariharan,M., Asabere,A., Waszak,S.M., Habegger,L., Rozowsky,J., Shi,M., Urban,A.E. *et al.* (2010) Variation in transcription factor binding among humans. *Science*, **328**, 232–235.
21. Haraksingh,R.R. and Snyder,M.P. (2013) Impacts of variation in the human genome on gene regulation. *J. Mol. Biol.*, **425**, 3970–3977.
22. Fu,W., O'Connor,T.D. and Akey,J.M. (2013) Genetic architecture of quantitative traits and complex diseases. *Curr. Opin. Genet. Dev.*, **23**, 678–683.
23. Siepel,A. and Arbiza,L. (2014) Cis-regulatory elements and human evolution. *Curr. Opin. Genet. Dev.*, **29**, 81–89.
24. King,M. and Wilson,A. (1975) Evolution at two levels in humans and chimpanzees. *Science*, **188**, 107–116.
25. Gazave,E., Marques-Bonet,T., Fernando,O., Charlesworth,B. and Navarro,A. (2007) Patterns and rates of intron divergence between humans and chimpanzees. *Genome Biol.*, **8**, R21.
26. Pai,A.A., Bell,J.T., Marioni,J.C., Pritchard,J.K. and Gilad,Y. (2011) A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS Genet.*, **7**, e1001316.
27. Arbiza,L., Gronau,I., Aksoy,B.A., Hubisz,M.J., Gulko,B., Keinan,A. and Siepel,A. (2013) Genome-wide inference of natural selection on human transcription factor binding sites. *Nat. Genet.*, **45**, 723–729.
28. Domene,S., Bumaschny,V.F., de Souza,F.S., Franchini,L.F., Nasif,S., Low,M.J. and Rubinstein,M. (2013) Enhancer turnover and conserved regulatory function in vertebrate evolution. *Philos. Trans. R Soc. Lond. B Biol. Sci.*, **368**, 20130027.
29. Lappalainen,T. and Dermitzakis,E.T. (2010) Evolutionary history of regulatory variation in human populations. *Hum. Mol. Genet.*, **19**, R197–R203.
30. Reilly,S.K., Yin,J., Ayoub,A.E., Emera,D., Leng,J., Cotney,J., Sarro,R., Rakic,P. and Noonan,J.P. (2015) Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science*, **347**, 1155–1159.
31. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 9362–9367.
32. Ramos,E.M., Hoffman,D., Junkins,H.A., Maglott,D., Phan,L., Sherry,S.T., Feolo,M. and Hindorff,L.A. (2014) Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur. J. Hum. Genet.*, **22**, 144–147.
33. Maurano,M.T., Humbert,R., Rynes,E., Thurman,R.E., Haugen,E., Wang,H., Reynolds,A.P., Sandstrom,R., Qu,H., Brody,J. *et al.* (2012) Systematic localization of common Disease-Associated variation in regulatory DNA. *Science*, **337**, 1190–1195.
34. McVicker,G., van de Geijn,B., Degner,J.F., Cain,C.E., Banovich,N.E., Raj,A., Lewellen,N., Myrthil,M., Gilad,Y. and Pritchard,J.K. (2013) Identification of genetic variants that affect histone modifications in human cells. *Science*, **342**, 747–749.
35. Kilpinen,H., Waszak,S.M., Gschwind,A.R., Raghav,S.K., Witwicki,R.M., Orioli,A., Migliavacca,E., Wiederkehr,M., Gutierrez-Arcelus,M., Panousis,N.I. *et al.* (2013) Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*, **342**, 744–747.
36. Kasowski,M., Kyriazopoulou-Panagiotopoulou,S., Grubert,F., Zaugg,J.B., Kundaje,A., Liu,Y., Boyle,A.P., Zhang,Q.C., Zakharia,F., Spacek,D.V. *et al.* (2013) Extensive variation in chromatin states across humans. *Science*, **342**, 750–752.
37. Huang,H.W., Mullikin,J.C. and Hansen,N.F. (2015) Evaluation of variant detection software for pooled next-generation sequence data. *BMC Bioinformatics*, **16**, 235.
38. Wu,L., Candille,S.I., Choi,Y., Xie,D., Jiang,L., Li-Pook-Than,J., Tang,H. and Snyder,M. (2013) Variation and genetic control of protein abundance in humans. *Nature*, **499**, 79–82.
39. Majewski,J. and Pastinen,T. (2011) The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet.*, **27**, 72–79.
40. Attanasio,C., Nord,A.S., Zhu,Y., Blow,M.J., Li,Z., Liberton,D.K., Morrison,H., Plajzer-Frick,I., Holt,A., Hosseini,R. *et al.* (2013) Fine tuning of craniofacial morphology by distant-acting enhancers. *Science*, **342**, 1241006.
41. Mathelier,A., Shi,W. and Wasserman,W.W. (2015) Identification of altered cis-regulatory elements in human disease. *Trends Genet.*, **31**, 67–76.
42. Spielmann,M. and Mundlos,S. (2013) Structural variations, the regulatory landscape of the genome and their alteration in human disease. *Bioessays*, **35**, 533–543.

43. Smith,E. and Shilatifard,A. (2014) Enhancer biology and enhanceropathies. *Nat. Struct. Mol. Biol.*, **21**, 210–219.
44. White,M.A., Myers,C.A., Corbo,J.C. and Cohen,B.A. (2013) Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 11952–11957.
45. Albert,F.W. and Kruglyak,L. (2015) The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.*, **16**, 197–212.
46. Whitaker,J.W., Chen,Z. and Wang,W. (2015) Predicting the human epigenome from DNA motifs. *Nat. Methods*, **12**, 265–272.
47. Schaub,M.A., Boyle,A.P., Kundaje,A., Batzoglou,S. and Snyder,M. (2012) Linking disease associations with regulatory information in the human genome. *Genome Res.*, **22**, 1748–1759.
48. Cowie,P., Hay,E.A. and MacKenzie,A. (2015) The noncoding human genome and the future of personalised medicine. *Expert Rev. Mol. Med.*, **17**, e4.
49. Rada-Iglesias,A. (2014) Genetic variation within transcriptional regulatory elements and its implications for human disease. *Biol. Chem.*, **395**, 1453–1460.
50. Friedensohn,S. and Sawarkar,R. (2014) Cis-regulatory variation: significance in biomedicine and evolution. *Cell Tissue Res.*, **356**, 495–505.
51. Cowie,P., Ross,R. and MacKenzie,A. (2013) Understanding the Dynamics of gene regulatory Systems; Characterisation and clinical relevance of cis-Regulatory polymorphisms. *Biology (Basel)*, **2**, 64–84.
52. Ward,L.D. and Kellis,M. (2012) Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.*, **30**, 1095–1106.
53. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
54. Robertson,G., Hirst,M., Bainbridge,M., Bilenky,M., Zhao,Y., Zeng,T., Euskirchen,G., Bernier,B., Varhol,R., Delaney,A. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.
55. Chen,X., Xu,H., Yuan,P., Fang,F., Huss,M., Vega,V.B., Wong,E., Orlov,Y.L., Zhang,W., Jiang,J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.
56. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
57. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.
58. Song,L., Zhang,Z., Grasfeder,L.L., Boyle,A.P., Giresi,P.G., Lee,B.K., Sheffield,N.C., Graf,S., Huss,M., Keefe,D. *et al.* (2011) Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.*, **21**, 1757–1767.
59. Crawford,G.E., Holt,I.E., Whittle,J., Webb,B.D., Tai,D., Davis,S., Margulies,E.H., Chen,Y., Bernat,J.A., Ginsburg,D. *et al.* (2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.*, **16**, 123–131.
60. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.
61. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
62. Belton,J.M., McCord,R.P., Gibcus,J.H., Naumova,N., Zhan,Y. and Dekker,J. (2012) Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods*, **58**, 268–276.
63. Nagalakshmi,U., Wang,Z., Waern,K., Shou,C., Raha,D., Gerstein,M. and Snyder,M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.
64. Consortium,T.E.P. (2004) The ENCODE (ENCyclopedia Of DNA Elements) project. *Science*, **306**, 636–640.
65. ENCODE. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.
66. Stamatoyannopoulos,J.A., Snyder,M., Hardison,R., Ren,B., Gingeras,T., Gilbert,D.M., Groudine,M., Bender,M., Kaul,R., Canfield,T. *et al.* (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol.*, **13**, 418.
67. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L., Ecker,J.R. *et al.* (2010) The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.*, **28**, 1045–1048.
68. Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J., Ziller,M.J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
69. GTEx Consortium. (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
70. Consortium,G. (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.
71. Lizio,M., Harshbarger,J., Shimoji,H., Severin,J., Kasukawa,T., Sahin,S., Abugessaisa,I., Fukuda,S., Hori,F., Ishikawa-Kato,S. *et al.* (2015) Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.*, **16**, 22.
72. Zabidi,M.A., Arnold,C.D., Schernhuber,K., Pagani,M., Rath,M., Frank,O. and Stark,A. (2015) Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, **518**, 556–559.
73. Arner,E., Daub,C.O., Vitting-Seerup,K., Andersson,R., Lilje,B., Drablos,F., Lennartsson,A., Ronnerblad,M., Hrydziuszko,O., Vitezic,M. *et al.* (2015) Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science*, **347**, 1010–1014.
74. Forrest,A.R., Kawaji,H., Rehli,M., Baillie,J.K., de Hoon,M.J., Haberle,V., Lassmann,T., Kulakovskiy,I.V., Lizio,M., Itoh,M. *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462–470.
75. Won,K.J., Chepelev,I., Ren,B. and Wang,W. (2008) Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinformatics*, **9**, 547.
76. Won,K.J., Agarwal,S., Shen,L., Shoemaker,R., Ren,B. and Wang,W. (2009) An integrated approach to identifying cis-regulatory modules in the human genome. *PLoS One*, **4**, e5501.
77. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
78. Ernst,J. and Kellis,M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, **28**, 817–825.
79. Hoffman,M.M., Buske,O.J., Wang,J., Weng,Z., Bilmes,J.A. and Noble,W.S. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, **9**, 473–476.
80. Hoffman,M.M., Ernst,J., Wilder,S.P., Kundaje,A., Harris,R.S., Libbrecht,M., Giardine,B., Ellenbogen,P.M., Bilmes,J.A., Birney,E. *et al.* (2013) Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.*, **41**, 827–841.
81. Firpi,H.A., Ucar,D. and Tan,K. (2010) Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics*, **26**, 1579–1586.
82. Rajagopal,N., Xie,W., Li,Y., Wagner,U., Wang,W., Stamatoyannopoulos,J., Ernst,J., Kellis,M. and Ren,B. (2013) RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput. Biol.*, **9**, e1002968.
83. Villarroel,M.C., Rajeshkumar,N.V., Garrido-Laguna,I., De Jesus-Acosta,A., Jones,S., Maitra,A., Hruban,R.H., Eshleman,J.R., Klein,A., Laheru,D. *et al.* (2011) Personalizing cancer treatment in the age of global genomic analyses: PALB2 gene mutations and the response to DNA damaging agents in pancreatic cancer. *Mol. Cancer Ther.*, **10**, 3–8.

84. Kleftogiannis,D., Kalnis,P. and Bajic,V.B. (2015) DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Res.*, **43**, e6.
85. Ghandi,M., Lee,D., Mohammad-Noori,M. and Beer,M.A. (2014) Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.*, **10**, e1003711.
86. Kwasnieski,J.C., Fiore,C., Chaudhari,H.G. and Cohen,B.A. (2014) High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.*, **24**, 1595–1602.
87. Dogan,N., Wu,W., Morrissey,C.S., Chen,K.B., Stonestrom,A., Long,M., Keller,C.A., Cheng,Y., Jain,D., Visel,A. *et al.* (2015) Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenet. Chromatin*, **8**, 16.
88. Niu,M., Tabari,E.S. and Su,Z. (2014) De novo prediction of cis-regulatory elements and modules through integrative analysis of a large number of ChIP datasets. *BMC Genomics.*, **15**, 1047.
89. Rosenbloom,K.R., Sloan,C.A., Malladi,V.S., Dreszer,T.R., Learned,K., Kirkup,V.M., Wong,M.C., Maddren,M., Fang,R., Heitner,S.G. *et al.* (2013) ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.*, **41**, D56–D63.
90. Visel,A., Minovitsky,S., Dubchak,I. and Pennacchio,L.A. (2007) VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic Acids Res.*, **35**, D88–D92.
91. Bailey,T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**, 1653–1659.
92. Zhang,S., Xu,M., Li,S. and Su,Z. (2009) Genome-wide de novo prediction of cis-regulatory binding sites in prokaryotes. *Nucleic Acids Res.*, **37**, e72.
93. Zhang,S., Li,S., Pham,P.T. and Su,Z. (2010) Simultaneous prediction of transcription factor binding sites in a group of prokaryotic genomes. *BMC Bioinformatics*, **11**, 397.
94. Zhang,S., Jiang,L., Du,C. and Su,Z. (2012) A novel information content-based similarity metric for comparing transcription factor binding site motifs. *IEEE 6th International Conference on Systems Biology (ISB)*, 32–36.
95. van Dongen,S. (2000) *National Research Institute for Mathematics and Computer Science in the Netherlands*, Amsterdam.
96. Frietze,S., O'Geen,H., Blahnik,K.R., Jin,V.X. and Farnham,P.J. (2010) ZNF274 recruits the histone methyltransferase SETDB1 to the 3′ ends of ZNF genes. *PLoS One*, **5**, e15082.
97. Hou,C., Dale,R. and Dean,A. (2010) Cell type specificity of chromatin organization mediated by CTCF and cohesin. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 3651–3656.
98. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
99. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C.Y., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R. *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.
100. Gupta,S., Stamatoyannopoulos,J.A., Bailey,T.L. and Noble,W.S. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.
101. Davydov,E.V., Goode,D.L., Sirota,M., Cooper,G.M., Sidow,A. and Batzoglou,S. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, **6**, e1001025.
102. Wang,J., Zhuang,J., Iyer,S., Lin,X., Whitfield,T.W., Greven,M.C., Pierce,B.G., Dong,X., Kundaje,A., Cheng,Y. *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.*, **22**, 1798–1812.
103. Wang,J., Zhuang,J., Iyer,S., Lin,X.Y., Greven,M.C., Kim,B.H., Moore,J., Pierce,B.G., Dong,X., Virgil,D. *et al.* (2013) Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res.*, **41**, D171–D176.
104. Machanick,P. and Bailey,T.L. (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, **27**, 1696–1697.
105. Kheradpour,P. and Kellis,M. (2014) Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.*, **42**, 2976–2987.
106. Manson,M.A. and Church,G.M. (2000) Predicting regulons and their cis-regulatory motifs by comparative genomics. *Nucleic Acids Res.*, **28**, 4523–4530.
107. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
108. Pavesi,G., Mauri,G. and Pesole,G. (2001) An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, **17**(Suppl. 1), S207–S214.
109. Pavesi,G., Mereghetti,P., Mauri,G. and Pesole,G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.*, **32**, W199–W203.
110. Ettwiller,L., Paten,B., Ramialison,M., Birney,E. and Wittbrodt,J. (2007) Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nat. Methods*, **4**, 563–565.
111. Fauteux,F., Blanchette,M. and Stromvik,M.V. (2008) Seeder: discriminative seeding DNA motif discovery. *Bioinformatics*, **24**, 2303–2307.
112. Martin,D., Pantoja,C., Fernandez Minan,A., Valdes-Quezada,C., Molto,E., Matesanz,F., Bogdanovic,O., de la Calle-Mustienes,E., Dominguez,O., Taher,L. *et al.* (2011) Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes. *Nat. Struct. Mol. Biol.*, **18**, 708–714.
113. Chaumeil,J. and Skok,J.A. (2012) The role of CTCF in regulating V(D)J recombination. *Curr. Opin. Immunol.*, **24**, 153–159.
114. Phillips,J.E. and Corces,V.G. (2009) CTCF: master weaver of the genome. *Cell*, **137**, 1194–1211.
115. Chong,J.A., Tapia-Ramirez,J., Kim,S., Toledo-Aral,J.J., Zheng,Y., Boutros,M.C., Altshuller,Y.M., Frohman,M.A., Kraner,S.D. and Mandel,G. (1995) REST: a mammalian silencer protein that restricts sodium channel gene expression to neurons. *Cell*, **80**, 949–957.
116. Charpentier,M.S., Christine,K.S., Amin,N.M., Dorr,K.M., Kushner,E.J., Bautch,V.L., Taylor,J.M. and Conlon,F.L. (2013) CASZ1 promotes vascular assembly and morphogenesis through the direct regulation of an EGFL7/RhoA-mediated pathway. *Dev. Cell*, **25**, 132–143.
117. Takahashi,K., Liu,F.C., Hirokawa,K. and Takahashi,H. (2008) Expression of Foxp4 in the developing and adult rat forebrain. *J. Neurosci. Res.*, **86**, 3106–3116.
118. Bowers,J.M. and Konopka,G. (2012) The role of the FOXP family of transcription factors in ASD. *Dis. Markers*, **33**, 251–260.
119. Moses,A.M., Pollard,D.A., Nix,D.A., Iyer,V.N., Li,X.Y., Biggin,M.D. and Eisen,M.B. (2006) Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput. Biol.*, **2**, e130.
120. Wray,G.A. (2007) The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.*, **8**, 206–216.
121. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.
122. Wilson,D., Charoensawan,V., Kummerfeld,S.K. and Teichmann,S.A. (2008) DBD–taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res.*, **36**, D88–D92.