Nucleic Acids Research Advance Access published December 13, 2013

Nucleic Acids Research, 2013, 1–12
doi:10.1093/nar/gkt1249

Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments

Pouya Kheradpour[1,2] and Manolis Kellis[1,2,*]

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar St, Cambridge, MA 02139, USA and [2]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02139, USA

Received August 7, 2013; Revised November 6, 2013; Accepted November 7, 2013

ABSTRACT

Recent advances in technology have led to a dramatic increase in the number of available transcription factor ChIP-seq and ChIP-chip data sets. Understanding the motif content of these data sets is an important step in understanding the underlying mechanisms of regulation. Here we provide a systematic motif analysis for 427 human ChIP-seq data sets using motifs curated from the literature and also discovered *de novo* using five established motif discovery tools. We use a systematic pipeline for calculating motif enrichment in each data set, providing a principled way for choosing between motif variants found in the literature and for flagging potentially problematic data sets. Our analysis confirms the known specificity of 41 of the 56 analyzed factor groups and reveals motifs of potential cofactors. We also use cell type-specific binding to find factors active in specific conditions. The resource we provide is accessible both for browsing a small number of factors and for performing large-scale systematic analyses. We provide motif matrices, instances and enrichments in each of the ENCODE data sets. The motifs discovered here have been used in parallel studies to validate the specificity of antibodies, understand cooperativity between data sets and measure the variation of motif binding across individuals and species.

INTRODUCTION

Chromatin immunoprecipitation (ChIP) (1) followed by hybridization to an array (ChIP-chip) (2,3) or sequencing (ChIP-seq) (4) enables the genome-wide identification of the binding locations of transcription factors (TFs) present in a given condition and cell type or tissue. As these technologies have matured, their use has become increasingly widespread. The resolution of these experimental techniques can be as low as 300 bp for ChIP-chip (5) and 50 bp for ChIP-seq (6), depending on the experimental design (e.g. fragment size, paired-end sequencing) and algorithmic processing of the raw data.

The use of these technologies on a variety of factors across many cell types has increasingly highlighted the complex nature of TF activity, often violating the simple model of a factor binding to its recognition pattern (motif) in isolation: binding has been shown to be dynamic across cell types, requiring the coordinated binding of cofactors or specific configurations of the underlying chromatin. Moreover, TF binding frequently occurs in the absence of any discernible motif instance (7,8) or to 'hot-spots' where several factors are simultaneously found (9). Understanding this complex binding necessitates identifying the underlying sequence features responsible. To address this need, we have performed a systematic, motif-centric analysis of hundreds of TF binding experiments made available as part of the human ENCODE project (8,10). As part of this, we provide a collection of motifs for each assayed factor, both taken from the literature and through *de novo* discovery, and also an annotation of motif instances genome-wide, which may be used to pinpoint the specific regulatory bases in regions bound by TFs.

We found that no single algorithm or database comprehensively assays the motifs relevant to the binding diversity surveyed by ENCODE. Therefore, our approach was to collect motifs from several literature sources (11–16) and supplement them with motifs discovered *de novo* on the data sets themselves using five established tools (17–21). Although this general approach of using multiple motif discovery tools is popular [e.g. (22–24)], its application to this number of data sets is unprecedented and permits the identification of TFs that are likely to be interacting or participating in common pathways.

*To whom correspondence should be addressed. Tel: +1 617 253 2419; Fax: +1 617 452 5034; Email: manoli@mit.edu

This work is accompanied by a web interface for browsing the discovered and literature motifs along with their enrichments (Figure 1; http://compbio.mit.edu/encode-motifs). In addition to the browsing interface, we provide several data files including all motif matrices and their matches to the genome, as well as software to compute enrichments and perform unified motif discovery with the five tools we use. Together, these permit both analyses of individual factors (e.g. to identify cooperating TFs) in addition to systematic analysis (e.g. to examine differences between TFs). Moreover, the breadth of data sets available enables systematic comparisons and analyses that are not possible when only one or a few factors are studied in isolation.

Later in the text, we describe the details of how the resource was generated and conduct an initial analysis to provide examples of its usage and to highlight potentially interesting results.

MATERIALS AND METHODS

Our goals were to produce a resource that (i) contains a comprehensive collection of relevant motifs for each factor; (ii) avoids repetitive, weakly enriched motifs that do not contribute to the *in vivo* specificity of the factor or its partners; and (iii) excludes variants of the same motif, particularly among the discovered motifs. With this in mind, we conducted motif discovery separately on each data set using five motif discovery tools and manually placed all its data sets into 'factor groups' on the basis of known motifs and homology (Figure 2). Known motifs from the literature and the top 10 most enriched discovered motifs (excluding duplicates) were collected for each factor group (see Supplementary Methods) and named as TF_known# for known motifs and TF_disc# for discovered motifs, where TF denotes the factor group (e.g. FOXA, CTCF, etc.). Known motifs were ordered arbitrarily, whereas the discovered motifs were ordered in descending order of the enrichment value that was used for their selection.

The 427 ENCODE experiments analyzed correspond to 123 TFs, which we place into 84 factor groups (Figure 3a). We failed to discover an enriched motif for only 12 of the 84 factor groups, of which 9 lack DNA binding domains (BRF, CTBP2, HDAC8, KAT2A, NELFE, SUPT20H, SUZ12, WRNIP1 and XRCC4) as identified by UniProt (27), and 6 have all their data sets flagged as unreliable based on various quality metrics [BRF, KAT2A, NELFE, NR4A, SUPT20H and ZZZ3; see (A. Kundaje, L.Y. Jung, P.V. Kharchenko, B. Wold, A. Sidow, S. Batzoglou and P.J. Park, in preparation)]. Of these factor groups, only NR4A has a previously identified known motif.

We exclude from the discussion below motifs that we consider unlikely to be relevant to our analysis, while maintaining them as part of the overall resource where they may be useful. These include 46 discovered motifs that are either low-complexity (e.g. dinucleotide repeats) or consistently have weak enrichment (<2) and do not match known motifs (Supplementary Table S1). These are likely a consequence of slight biases in the discovery pipeline, or are due to real, but relatively weak, specificity for the factor. We also exclude an additional 36 motifs that have a weak similarity to the known motif for the factor but for which a better matching and enriched motif is also found (Supplementary Table S2). These are most frequently seen for longer motifs that can be broken up into recognizable, but globally dissimilar, patterns that are not captured by our automatic exclusion criteria (see Supplementary Methods). Together, these represent 28% of the 293 discovered motifs.

RESULTS

Using motif similarity metrics, we are able to link the discovered motifs directly to the TFs that recognize them through their known motifs. Here we use these inferred relationships between TFs to make specific biological insights, illustrating the types of analyses that our resource enables. In the interest of clarity, most descriptions of TFs will be omitted, but may be found along with further references at RefSeq (28) and Entrez (29).

Recovery of known specificity for TFs

Most of the known literature motifs we collect are derived from biochemical *in vitro* assays. Thus, they provide a largely independent, although somewhat imperfect way to evaluate the performance of our discovered motifs. Recovery of known motifs varies significantly by method, but taking the most enriched motif (our pipeline) is competitive with the best single method (Figure 3b). Overall, our pipeline found a motif matching a previously characterized literature motif for 41 of the 56 factor groups with a known motif.

One of the most striking observations of this analysis is how frequently other distinct motifs were also found. For 29 of these 41 factor groups other motifs are found, even after manually excluding redundant or repetitive motifs, and for 9 factor groups one or more of these discovered motifs is ranked higher than the motif matching a known motif (see Supplementary Table S3). In the next section, we will analyze the additional motifs we found for these factors, which in many cases identify factors known to interact, either cooperatively or competitively.

For the remaining 15 of 56 factor groups with a known motif (e.g. HSF, NANOG, PBX3, SREBP and TAL1) the known motif is not found at all, including NR4A where no enriched motif is discovered. Frequently this is because the known motif itself is not enriched and may not accurately capture the specificity of the factor *in vivo*. For example, the 'known' EP300 motif from Transfac was likely built on a specific bound region of EP300 and would not accurately capture its binding in all cell types where it interacts with a variety of factors and has no DNA binding domain of its own (we avoided removing such motifs to prevent bias in the database). Likewise, we do not discover a motif that matches the known ZBTB33 specificity, and moreover the known motif itself is not enriched at all in the bound regions.

Although some known motifs were of apparently low quality, we largely found our database of known motifs to

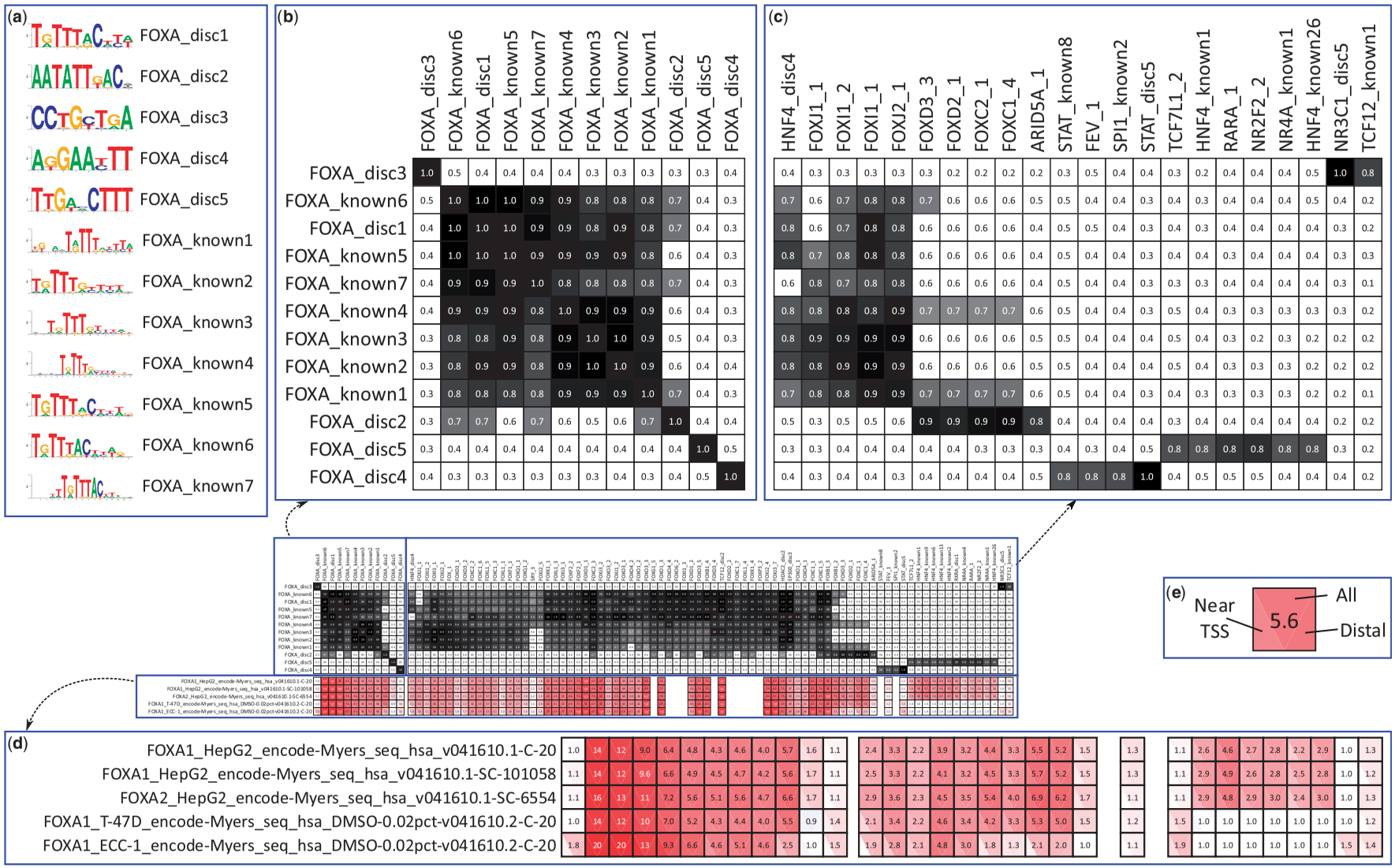

Figure 1. Pipeline output for FOXA factor group, an example that highlights different aspects of the resource (in the interest of space, only selected columns are shown in the enlargements). (**a**) The known and discovered motifs for the factor group, drawn with WebLogo (25). Because the original orientation is arbitrary, motifs are flipped as necessary to increase the similarity of the displayed orientations. (**b** and **c**) Similarity between the motifs for this factor group and the motifs: (b) for this factor group, (c) for all other motifs that match at least one motif for this factor group with similarity ≥ 0.75. Motif similarities are shown in black/white scale with gray starting at 0.65. (**d**) Enrichments of motifs in the data sets for this factor group. Data sets are named indicating the factor, cell type/stage, lab code, followed by values to differentiate data sets and make them unique. Red is used for enrichment, blue is used for depletion (which is rare in this study). Enrichments are not shown for motifs when control motifs could not be generated or there was not sufficient information content to match at a 4^{-8} *P*-value. (**e**) Magnified enrichment heatmap value. The three triangles represent the background regions used for enrichment; the top triangle is all regions, the left triangle is those within 2 kb of an annotated TSS and the right triangle is those outside the 2 kb window (the regions used in the left and right triangles partition that of the top triangle). The number shown is the enrichment in the inclusive background. Here we see an apparent contradiction: a higher enrichment for the union than the parts. This occurs because the higher counts permit a smaller confidence interval around the ratios used to compute the enrichment. Heatmaps are ordered using hierarchical clustering followed by optimal leaf ordering (26). The Supplementary Results include an analysis of the robustness of the motif similarity metric and the enrichment values.

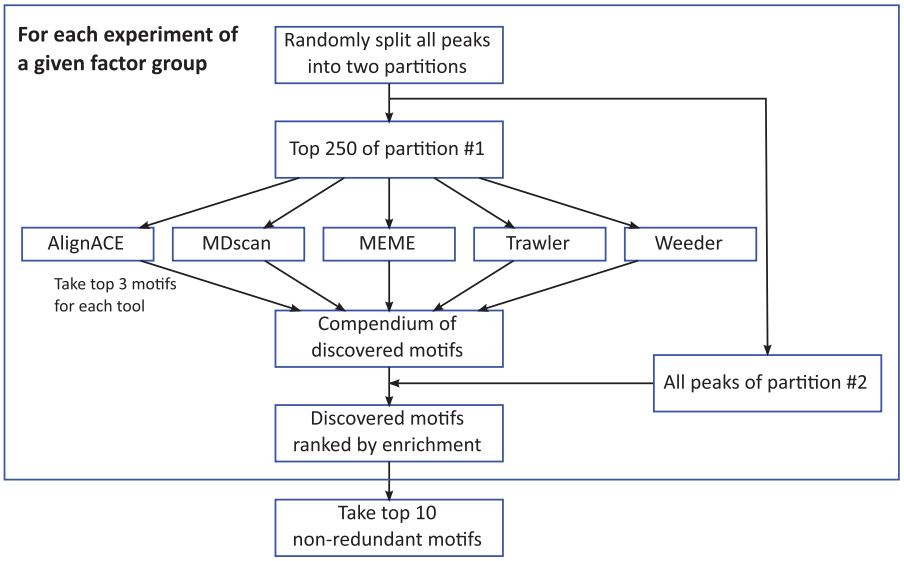

Figure 2. Outline of motif discovery pipeline. Input regions for each data set are randomly partitioned into two groups. The top 250 regions of one of the partitions are scanned for motifs using five *de novo* motif discovery tools. These motifs are evaluated using the peaks from the other partitioned and pooled across data sets for a factor group to produce the final list of discovered motifs for each factor group.

be relatively comprehensive and had difficulty finding matches to novel motifs outside it. An exception is ZNF263_disc1, which does not match a motif in our database, but does roughly match the specificity for ZNF263 indicated in (30) despite only having weak enrichment (1.8-fold).

Although the motifs that match each other (either known or discovered) generally have similar enrichments, in some cases we find substantially higher enrichment for some motif variants over others (Figure 4 and Supplementary Table S3). For example, NFE2_disc1 matches the known NFE2 motif, but has a 76-fold maximal enrichment across NFE2 data sets, compared with 56-fold enrichment for the most enriched known NFE2 motif. Different known motifs for the same factor often show a broad range in enrichment: MEF2 has six motifs described in Transfac, with an enrichment differential of as much as 4-fold consistently across data sets. This enrichment analysis provides a systematic way to choose among variants of a motif.

We also saw varying enrichment of the known motif, depending on the specific data set for a factor group. For example, CTCF_known2 is enriched in CTCF data sets in a range from 30- to 78-fold on identically processed data. This may be a result of varying quality of the samples across data sets or may be a consequence of true biological differences.

Identifying the sequence specificity for factors that were previously uncharacterized is of particular interest. In all, 17 factor groups had no known motif but now have discovered enriched motifs (BCL, BDP1, CCNT2, CHD2, CTCFL, HDAC2, HMGN3, RAD21, SETDB1, SIRT6, SMARC, SMC3, SP2, SIN3A, THAP1, TRIM28 and ZNF263). These discovered motifs may represent the direct or indirect (e.g. through cofactors) DNA binding specificity.

Shared motifs suggest interacting relationships

We find that most factors have motifs for other factors enriched in their binding sites (summarized in Supplementary Table S4). This may occur due to (i) cooperative binding of the two factors to the same locations; (ii) interfering binding between factors where one binds near the other to prevent binding; (iii) some similarity in motif specificity; (iv) the two factors functioning on a similar set of genes (e.g. ones specific to one tissue), without directly interacting; or (v) the factors binding to similar genomic regions (e.g. near genes). Our analysis does not directly rule out any of these possibilities; however, (iii) is generally verifiable using our motif similarity metrics and (v) can be examined by inspecting only the TSS-proximal enrichment.

The motif most enriched in multiple data sets was the TPA DNA response element (TRE; TGA[C/G]TCA), which is recognized by the AP1 TF when it is formed by FOS/JUN dimers (31) and other factors including MAF and NFE2. The enrichment of the TRE in a data set is often stronger than that of even the known *in vitro* sequence specificity and may arise from a number of phenomena, including (i) a cooperatively interaction with AP1, (ii) competition with AP1 for the same binding sites, leading to a potentially repressive role for the TF or (iii) reuse of binding sites due to, for example, accessibility of chromatin. We find a motif matching the TRE motif for 20 factor groups (AP1_disc3, AP2_disc1, BATF_disc1, BCL_disc2, CTCF_disc8-9, EP300_disc1, GATA_disc2, HMGN3_disc1, IRF_disc2, MAF_disc1, MEF2_disc3, MYC_disc3, NFE2_disc1, NR3C1_disc2, PRDM1_disc2, RXRA_disc3, SMARC_disc1, STAT_disc2, TCF7L2_disc1 and TRIM28_disc1).

We found that the enrichment of the TRE to be particularly notable for a few factors. GATA and AP1 have

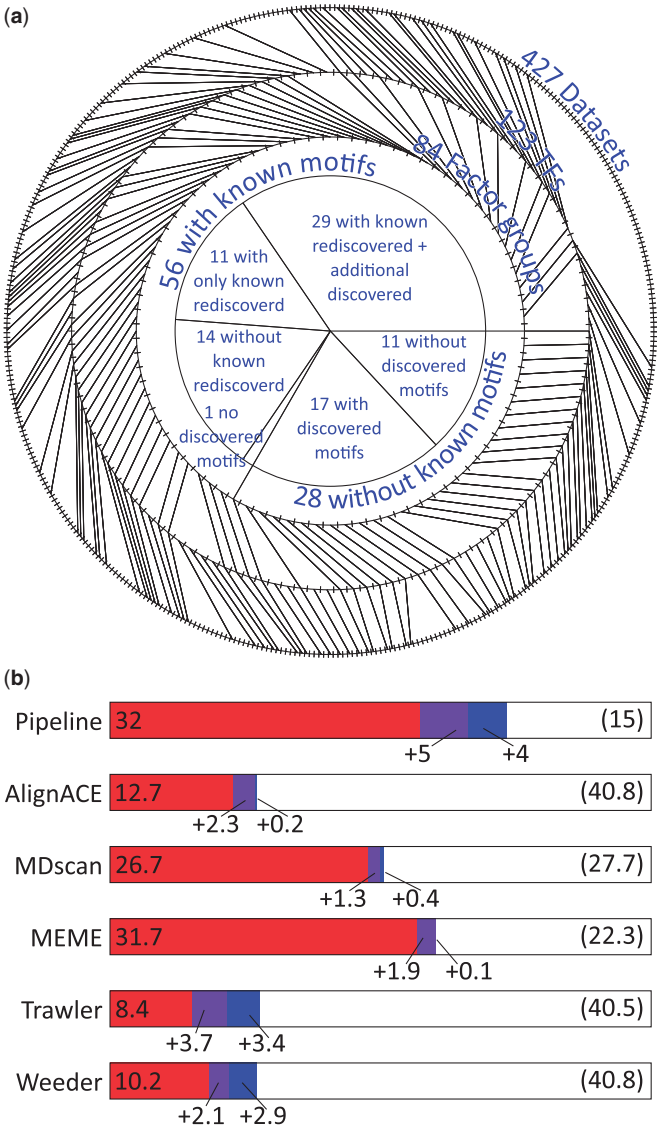

Figure 3. (**a**) Summary of input data used. The outside ring indicates the experimental data sets (one tick for each of 427), which are separated into 123 transcription factors (second ring). The TFs are further grouped into 84 factor groups (third ring). We are able to find a matching discovered motif for 41 of the 56 factor groups with a known motif; 29 of these 41 factor groups have additional discovered motifs that may be associated with cofactors. For all but 1 of the 15 factor groups where the known motif is not recovered we still find enriched discovered motifs. We also discovered enriched motifs for 17 of the 28 factor groups without a known motif. (**b**) Recovery of known motifs by each of the discovery tools. Performance of discovery in terms of number of factor groups for which the known motif was recovered. A motif is considered a match if it matches any of the known motifs for a factor group (see Supplementary Methods for details on how matches are computed). The number of additional factors that have a match is shown with each additional motif (only three motifs are taken from each individual method, whereas we have up to 10 for the pipeline). The number of factor groups with no motif match is shown in parenthesis. When multiple data sets exist for a factor group, the fraction that matches is used in computing its contribution for computing the performance of the individual tools.

known cooperative binding (32). TFs in the SMARC factor group are members of the SWI/SNF chromatin remodeling complex (33), which is necessary for proper regulation by FOS/JUN dimers (34); and TCF7L2_

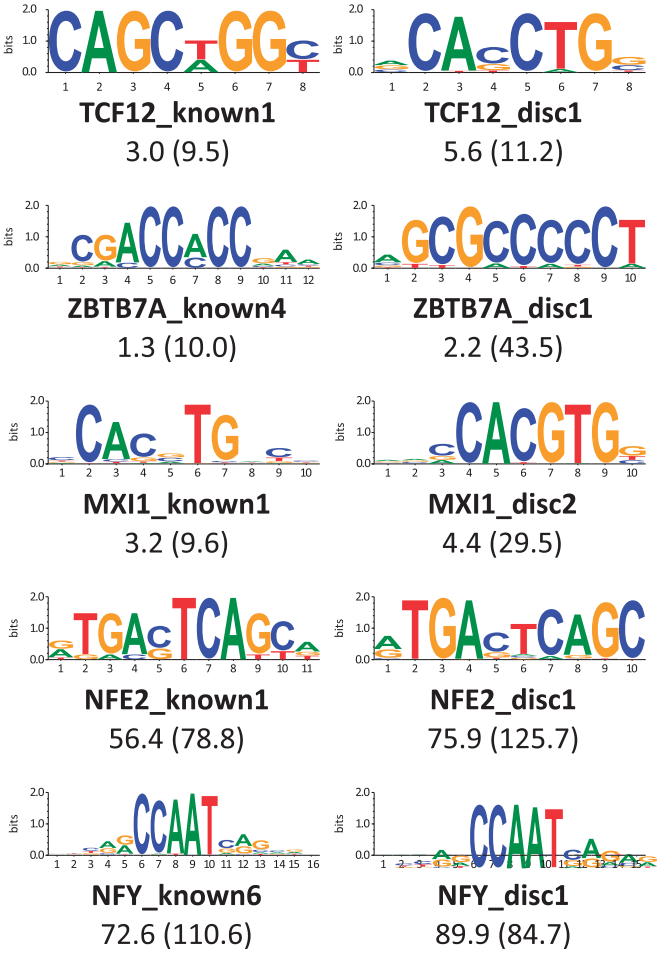

Figure 4. Comparison of known versus discovered motifs (selected where discovered better enriched than known; all factor groups with a discovered motif matching a known motif in Supplementary Table S3). Displayed is the known and discovered motif with the maximum enrichment across all data sets for a factor group. Only the discovered motifs that match a known motif for a factor group are considered. The maximum enrichment is indicated for each factor and, in parenthesis, the 'raw' enrichment for the same data set without the use of the shuffle motifs for correction.

disc1, which matches the TRE, is more enriched than the known TCF7L2 motif (TCF7L2_disc2) in only the TCF7L2 colorectal cancer cell line HCF-116 data set, consistent with the known interaction of JUN and TCF7L2 during intestinal cancer development (35).

AP1 also binds to the cAMP response element (CRE; TGACGTCA) when the dimer is formed by ATF3/JUN (31) and this is the motif we find as AP1_disc1. However, AP1_disc3 (which matches the TRE) is the most enriched motif in FOS data sets. Interestingly, ATF3_disc1 is not the CRE, but rather the E-box (see later in text). We do, however, find a variant of the CRE (with additional specificity) as ATF3_disc2. The most enriched discovered motif for E2F, E2F_disc1 also matches the CRE and is highly enriched in all data sets.

MYC is a critical regulator, which recognizes the E-box sequence. To aid in comparisons, we include MAX, which forms complexes with MYC, and USF1/2, which also recognizes the E-box sequence, in the MYC factor group.

We find multiple motifs enriched in MYC binding sites, highlighting the multifunctional role MYC and the other E-box recognizing proteins play. We found a version of the E-box with additional specificity (MYC_disc1) that was highly enriched in USF1/2 bound regions (max 98-fold for USF2 versus <9-fold enrichment for MYC/MAX). This motif was more enriched than the known E-box motifs, including known USF motifs, in many USF data sets. We find a second, less specific E-box motif (MYC_disc2), which shows more even enrichment across factors. We also find discovered motifs of other factors matching the E-box, including SIN3A_disc2 (discussed later in text), NFE2_disc2-3 and SIRT6_disc1. It is notable that although SIRT6 is a chromatin-associated protein without a known DNA binding domain (36), the only discovered motif matches the E-box (with 16-fold enrichment in SIRT6 bound regions), suggesting that MYC or another E-box recognizing factor may play an important, but indirect, chromatin-related role.

Motif enrichment is able to identify both positive and negative interactions for the same factor. For example, SIN3A, a corepressor known to interact with a number of proteins, has discovered motifs matching REST (SIN3A_disc1 and more weakly disc3–4) and MYC (SIN3A_disc2). These are consistent with SIN3A's known involvement in repression by REST (37) and SIN3A being a known antagonist for MYC (38).

Morever, MYC_disc4 matches RFX5 and is enriched particularly for MAX-bound regions in H1-hESC and GM12878, and MYC_disc5 matches the CEBPB known motif and is enriched in MYC regions bound in unstimulated K562 cells. MXI1, which was not included in the MYC factor group although it does interact with MAX to bind to MYC-MAX sites (39), has MXI1_disc1 that matches RFX5 in both the K562 and HeLa-S3 cell lines.

We analyzed six IRF family data sets: IRF1 binding in K562 cells stimulated by IFNa (viral innate response) or IFNg (viral, bacterial and tumor control); IRF3 in HepG2, GM12878 and HeLa-S3; and IRF4 in GM12878. The most strongly enriched motif (IRF_disc1, matching NFY) is highly enriched (>20-fold) for all three IRF3 data sets and IRF1 in K562 under IFNg stimulation. This suggests that binding of IRF to NFY sites occurs only under specific conditions and by only some IRF members and potentially expands on the previously documented interaction of NFY and IRF2 at a single promoter (40). IRF_disc4, which matches SP1, is enriched in the same cell types, albeit at much lower levels. IRF_disc3, which matches the known IRF consensus, shows weak-to-no enrichment in these data sets, but shows an enrichment of 8.8-fold for IRF1 bound regions in K562 cells under IFNa stimulation and 3.1-fold enrichment for IRF4 bound regions in GM12878. IRF_disc2, which matches the TRE, is enriched primarily in GM12878 regions bound by IRF4. The known SPI1 motif matches IRF_disc5, and reciprocally SPI1_disc2 matches the IRF motif, consistent with the importance of SPI1 in hematopoietic development (41).

Beyond the discovered motif for IRF, several other discovered motifs (AP1_disc2, CEBP_disc2, E2F_disc4, PBX3_disc1, RFX5_disc2 and SP1_disc1-2) match the known NFY specificity (CCAAT). These discovered motifs are consistent with several known interactions of NFY. RFX5 promotes the cooperative binding between RFX and NFY (42), CEBPB and NFY interact in at least one promoter (43) and SP1 and NFY are known to interact (44). E2F_disc4 has particularly high enrichment in E2F4 data sets, consistent with the cooperative role E2F4 and NFY play in cell cycle regulation (45).

STAT factors are involved in regulating number of growth-related functions. We analyze STAT1, STAT2 and STAT3 here in the context of GM12878, HeLa-S3, MCF10A-Er-Src and K562 cells. We find relatively consistent enrichment of the STAT full site (TTCCNGGAA), which STAT_disc1 matches, while finding weak enrichment for just the half-site (TTCC). We also find motifs involved in other proliferative functions including STAT_disc2, which is particularly enriched in STAT3 data sets and matches the TRE, consistent with STAT3 being one of the many interaction partners for AP1 (46). STAT_disc3 matches the IRF consensus and has enrichment that is particularly high in STAT1 and STAT2 data sets stimulated by IFNa, highlighting the cooperativity of STAT factors and IRF in immune functions. STAT_disc4 is a match to the CEBPB motif and is found enriched in STAT3 data sets, consistent with the known cooperative role for these two factors (47).

TFs with ETS domains are highly conserved and involved in several cellular processes [reviewed in (48)]. A number of TFs have discovered motifs that match the ETS consensus, including EGR1_disc2, GATA_disc3, MEF2_disc2, NRF1_disc2, NR2C2_disc1 and PAX5_disc4. These discovered motifs are supported by known interactions between GATA and ETS in sea squirts (49), MEF2 and the ETS factor PEA3 (50) and NR2C2 with the ETS factor ELK4 (51). Moreover, PAX5 and ETS factors have shared roles in the development of B-cells (52,53). Looking at the discovered ETS motifs, we find that ETS_disc8 matches the known motif for MYB and the two have been known to cooperate, a relationship that is important in the context of certain cancers (54).

THAP1 has two discovered motifs, both of which match the known YY1 motif (the first with additional specificity added by an apparent HNF4 motif). To our knowledge, the relationship between THAP1 and YY1 has not been directly observed; however, THAP1 has been known to associate with the coactivator HCF-1 (55), and YY1 and HCF-1 are known to interact (56). Our result suggests that THAP1 and YY1, possibly with the addition of HNF4, may interact at least in the K562 cell line for which we have THAP1 binding data. RAD21_disc3 also matches YY1, suggesting an additional interaction.

NANOG, an important pluripotency TF, has a known motif that is only weakly enriched (1.3-fold) in the bound regions and not discovered by our pipeline. We see much stronger enrichment for the known POU5F1 and POU2F2 motifs, for which we also find similar motifs (NANOG_disc2 and NANOG_disc4, respectively), consistent with their shared roles in pluripotency (57,58). The interaction of these factors is further supported by

POU5F1_disc2 matching the known POU2F2 motif. Additionally, NANOG_disc2 and disc3 match the known motifs for TCF7L2 and TCF12, respectively, again consistent with the important role TCF proteins play in stem cells (59).

CTCF plays a variety of vital roles in the organization of chromatin architecture (60) and the motifs we discover matching the known CTCF specificity (RAD21_disc1, SMC3_disc1,2-4, CTCFL_disc1,10, ZBTB7A_disc1,2, SP2_disc3 and RXRA_disc2,5; some weakly) are largely compatible with this role. RAD21 is a highly conserved protein involved in DNA double-strand repair (61) known to co-localize with CTCF (62). Cohesin, of which SMC3 is a subunit, is brought to the chromatin by CTCF (63). Further, although the function of the CTCF paralog CTCFL is not completely known, it does appear to be involved in imprinting through interaction with a histone methyltransferase (64).

Combinations of motifs

A few of the discovered motifs contain additional specificity or have distinct segments matching multiple motifs. For example, EGR1_disc4 appears to be a combination of multiple motifs (EGR1, IKZF1 and a homeobox motif), and SETDB1_disc1 contains the ZNF143 core sequence with significant additional specificity. The appearance of these motifs suggests highly specific 'grammars' for these motifs that may require specific spacing and orientation of binding sites for functionality.

We find several additional enrichments of potential interest. PBX3_disc2 matches the known MEIS1 motif, consistent with the known cooperative binding of MEIS1 and PBX (65). TAL1_disc1 matches GATA, with the potential connection that GATA and TAL1 are known to be important in hematopoesis and vascular development (66,67). HSF_disc1 matches the known CEBP motif and has much higher enrichment in HSF data sets (31-fold) compared with the known motifs for HSF (<9-fold). Additionally, EGR1_disc5, HNF4_disc5, NRF1_disc3, PAX5_disc2, RXRA_disc4/PAX5_disc3 and SREBP_disc1 match the known motifs for ZIC, SOX, SP1, PAX2/PAX3, IRF and RFX5, respectively, suggesting additional previously uncharacterized interactions. Lastly, we find some motifs that show more ambiguous matches: SMARC_disc2 shows weak similarity to homeobox TGTAGT motif, NR2C2_disc2–3 weakly matches the known HNF4 motif and EGR1_disc3/SETDB1_disc2 matches the repetitive NRF1 motif.

General factors enriched in cell line-specific key regulators

Factors directly responsible for the establishment of enhancers, chromatin restructuring or polymerase recruitment frequently exhibit binding that is highly cell type specific. Because most of these factors do not have their own sequence specificity, their binding is often correlated with that of regulators important for the specific cell line. We analyze several such factors (BCL, BDP1, CCNT2, EP300, FOXA, HDAC2, HMGN3, TATA, TCF12 and TRIM28) and find that key cell line regulators can be identified by examining enrichments in cell lines-specific data sets.

As a transcriptional coactivator, EP300 interacts with numerous TFs [reviewed in (68)] and has been shown to have binding that can identify tissue-specific enhancers (69). Conversely, FOXA has a DNA binding domain and plays an important role in liver development and function (70) and is a pioneer factor responsible for priming chromatin for the binding of other factors (reviewed in (71)]. Other proteins involved in chromatin restructuring include HDAC2, which transcriptionally represses through histone deacetylation (72) and HMGN3 (73). Further, two factor groups are directly involved in transcription including three RNA Pol3 subunits (BDP1, RPC155 and TFIIIC-110) and CCNT2, which is involved in the elongation of Pol2 (74).

Eight of these ten factor groups have at least one data set in K562 (erythroleukemia cells), and for four of these we discover motifs that match the GATA consensus, which is then enriched specifically in the K562 data sets (BCL_disc5, CCNT2_disc1, HDAC2_disc1 and HMGN3_disc2). GATA has a known important role in K562 (75), and we also have previously found an association with GATA motifs and chromatin state-derived enhancers for K562 cells (76). We also find three additional motifs that have enrichment specific to the factor group's K562 data set: BDP1_disc1, a 23-nt motif that contains the STAT consensus; HMGN3_disc1, which matches the TRE; and TRIM28_disc2, which matches no known motif and may be associated with an uncharacterized regulator active in this cell line.

Likewise, for GM12878, an EBV-mediated lymphoblastoid cell line, we find three discovered motifs (BCL_disc4, EP300_disc5 and TCF12_disc4) that match the known IRF consensus. IRF4 has been shown to be important in the establishment of these cell lines (77), and the family is an important player in immune cells (78). This enrichment is also consistent with our previous study using epigenetic marks (76), where we found IRF to be the strongest enriched motif in GM12878-specific enhancers. We also find GM12878-specific enrichment for motifs matching NFKB (BCL_disc6) and POU2F2 (TATA_disc9), consistent with the known biology of these factors (79,80).

The motifs we find specifically enriched in HepG2 (liver carcinoma) data sets match the known motifs for FOXA (EP300_disc3, HDAC2_disc2, and TCF12_disc2), HNF4 (FOXA_disc5 and HDAC2_disc5) and CEBP (EP300_disc2,6), three key liver regulators (70,81). We find motifs with enrichments specific to H1-hESC, which include matches to the pluripotency factor POU2F2 (TATA_disc9), the near universally expressed repressor REST (BCL_disc3 and HDAC2_disc4) and key metabolic regulator NRF1 (HDAC2_disc4). We find additional cell line-specific enrichments for FOXA_disc3 (TCF12) in ECC-1, FOXA_disc4 (STAT) in both T-47D and ECC-1 and EP300_disc2,6 (CEBP) and EP300_disc4 (ETS) with enrichment in the HeLa-S3 data set.

Even for these factors, we find motifs that are consistently enriched across assayed cell lines for a given factor. FOXA_disc1, for example, matches the known FOXA

motif, indicating that FOXA's own motif also plays an important role in its specificity. Most of the motifs we identify for RNA Pol2 machinery (TAF1, GTF2B, GTF2F1 and TBP) are enriched in all cell lines, including the known TATAAA motif (TATA_known2). Also, TATA_disc1, disc6 and disc8 have consistent enrichment and match the known motifs for YY1 (which is known to be important in establishing transcription) (82), NFY and ETS. The top discovered motif BCL_disc1 matches the known ETS motif and is also enriched across data sets.

Interestingly, we find that the TRE motif is found and enriched in a cell line-specific manner for several factors, but for different cell lines. For example, HMGN3_disc1 is enriched in K562, BCL_disc2 has the highest enrichment in GM12878, TRIM28_disc1 is only enriched in the HEK2932 and U2OS cell lines and EP300_disc7 has enrichment in the neuroblastoma cell line SK-N-SH-RA and HeLa-S3. This suggests that perhaps AP1 or other factors recognizing TRE are selectively interacting with these proteins depending on the cell line.

Novel motifs raise possibility of unknown regulators

Although we are able to putatively explain the majority of the motifs we discover as either matches to previously known motifs or low complexity sequences, we do identify 30 putative novel motifs (Figure 5). We placed these into eight groups on the basis of their similarity: Novel1 (BRCA1_disc1, CHD2_disc1, ETS_disc3,6, NR3C1_disc3 and ZBTB33_disc1-4), Novel2 (EGR1_disc4, ETS_disc1,5,7, SETDB1_disc1, SIX5_disc1-3, SMARC_disc2 and ZNF143_disc1-3), Novel3 (SP2_disc3, TCF12_disc3 and ZBTB7A_disc2), Novel4 (RFX5_disc3), Novel5 (BDP1_disc2), Novel6 (TATA_disc5,7), Novel7 (TRIM28_disc2) and Novel8 (E2F_disc6).

Novel1 (using ZBTB33_disc1) is highly enriched in at least one data set for each of the factor groups for which it is found (BRCA1, CHD2, ETS, NR3C1 and ZBTB33). All five factor groups except CHD2 have at least one known motif, and for each of these data sets Novel1 is more enriched in at least one data set than any known motif [the result for NR3C1 is questionable because only one data set has enrichment and that data set has been independently flagged as problematic; see http://www.encodeproject.org/encode/qualityMetrics.html]. The shared role of BRCA1 and CHD2 in DNA damage repair (83,84) suggests that Novel1 may be involved in this or other shared roles for these factors and highlights the utility in shared motif enrichment even outside of motifs directly tied to a factor.

Similarly, for SIX5, we see only weak enrichment of the known SIX5 motif and fail to discover a motif similar to it. However, Novel2 (using SIX5_disc1) shows over 100-fold enrichment for all three data sets (K562, GM12878 and H1-hESC). Novel2 also shows high enrichment in data sets for which it was not rediscovered, including ATF3 (all data sets have >20-fold enrichment with GM12878 having 106-fold) and NRF1 (all data sets have >30-fold enrichment). Moreover, the known ZNF143 motif, which is 4-fold enriched in the one ZNF143 data set, is also not recovered, but Novel2 is 24-fold enriched. The breath of data sets sharing this motif suggests it may be recognized by an important yet unknown or under-characterized regulator.

Like the known ZBTB7A motif, Novel3 (using SP2_disc3) is largely poly-G, which causes us to underestimate its enrichment due to our shuffling process. Despite

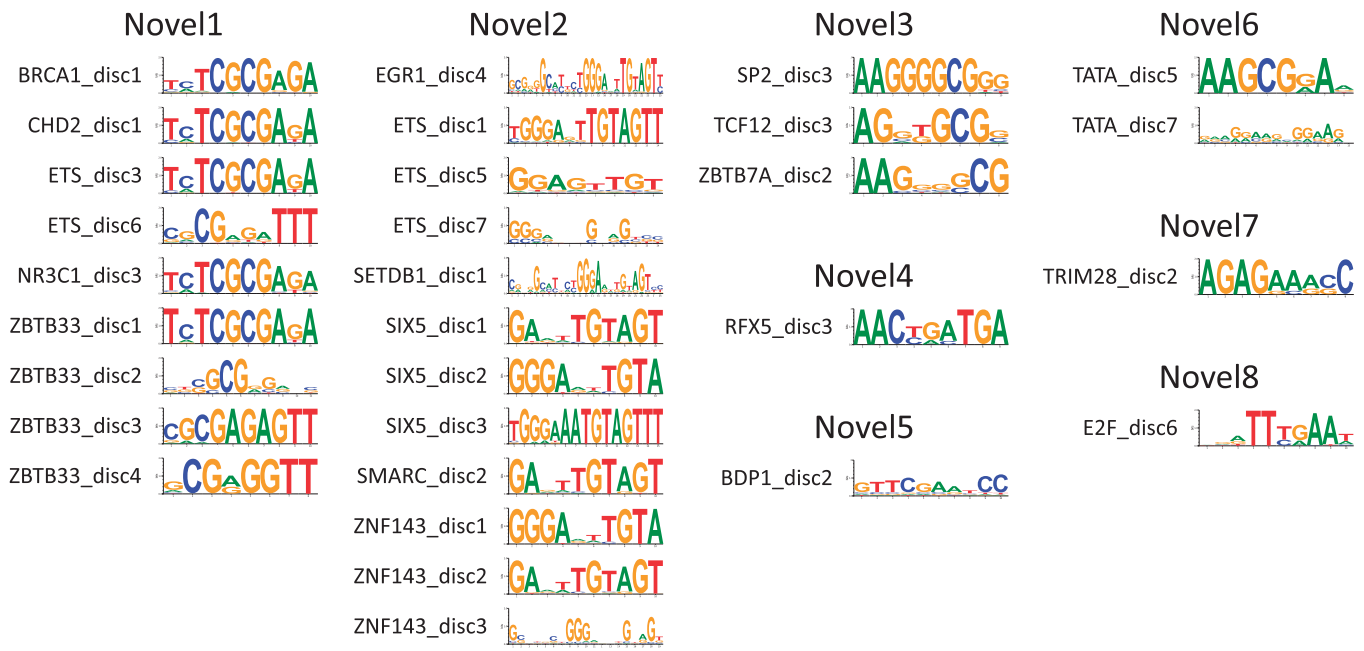

Figure 5. Putative novel motifs. We find eight motifs that are not represented in the literature motifs we collected, three of which are found for at least two factor groups. These patterns may represent the binding specificity of the factors for which they are discovered or for other factors that cooperate with them.

this, however, it does show enrichment in several data sets, including for the factor groups for which it was identified. This motif shows similarity to other poly-G motifs, such as known SP1 motifs, but appears to be distinct due to its other bases.

Novel4 (RFX5_disc3) shows moderate, but consistent (2- to 6-fold) enrichment across the RFX5 data sets. The consensus is composed of two of the same components as the known motifs (AAC and TGA), but ordered differently. Consequently, it may represent the binding specificity of, for example, an alternative isoform of RFX5. The remaining motifs (Novel5-8), were found for factors that show cell line-specific enrichments. Consequently, these may represent specificities for regulators that are previously unidentified.

Experimental and evolutionary validation of novel motifs

Following the motif discovery and selection of these putative novel motifs, a study released hundreds of new motifs generated using high-throughput SELEX (16). Two of the putative novel motifs described in this section match motifs generated by (16): Novel1 matches the motif for ETV6 and Novel6 matches ZBED1. Although we have incorporated these SELEX motifs into our resource, we continue to include Novel1 and Novel6 as putative novel motifs because they were identified without knowledge of these new specificities and thereby strengthen the evidence for the remaining novel motifs.

Four of these putatively novel motif groups (Novel1–3, 6) match motifs that were previously identified using conservation signals across four mammals (85) (Supplementary Table S5). Therefore, this study provides additional support for these conservation-based motifs and, conversely, the motifs identified here gain comparative evidence. The relatively few distinct novel patterns that are found in this study and the comparative support for many of the few that are found suggests that there may be a limited number of human TF motifs with many instances and which interact with one of the assayed factors that remain unknown.

DISCUSSION

In this article, we provide a systematic and comprehensive collection of motifs for hundreds of human TF binding data sets. TF binding can be complex, with a factor recognizing several or motifs or binding in the apparent absence of any motif [reviewed in (86)]. We also show that it is possible to identify cofactors that may be partially responsible for binding or function.

This motif resource has already been used in several articles while this article was in preparation, demonstrating its value for high-throughput analyses. Our motifs are being matched at low stringency to identify peaks that are void of any motif to understand the mechanism through which motif-less peaks are generated (8). The collection of known motifs and enrichment techniques we present here was also used as a secondary validation of peaks (87). Because having the motifs allows for more precisely determining the bases responsible for binding, these motifs enable analyzes involving population data (88) and for interpreting GWAS data (89). Two other ENCODE articles also perform motif discovery: (90) produce a non-redundant list of discovered motifs but do not perform an extensive analysis of the relationships between factors and (91) use DNaseI footprinting data to identify relevant motifs.

Having a motif catalog is also the first step in identifying high-quality computational targets of factors, which may allow the identification of binding sites that were, for example, not found in the conditions assayed. Two popular strategies are used for this purpose. One is using clustering of motif instances for factors known to cooperate to form *cis*-regulatory modules (92,93). This resource is well suited for this purpose because it naturally provides sets of motifs that are likely to cooperate.

A second approach is the use of conservation on many closely related species (85,94–97). This can be performed readily on these motif instances because a dense tree of mammalian species has been sequenced readily permitting their alignment and measuring selection of a near-nucleotide level. Because changes in the underlying motif matches are largely responsible for changes in binding across species (98), evolutionary-based approaches on the motif instances may be a means to deal with the high rate of non-functional binding (99–101).

AVAILABILITY

A web interface, along with data files and accompanying software, is available at http://compbio.mit.edu/encode-motifs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online, including [102–110].

ACKNOWLEDGEMENTS

The authors thank Ewan Birney, Christopher Bristow, Luke Ward, Jason Ernst, Anshul Kundaje, Gerald Quon and other members of the Kellis Laboratory for helpful discussions.

FUNDING

National Institutes of Health (NIH) [HG004037, HG007000 and HG006991]. Funding for open access charge: NIH [HG004037, HG007000 and HG006991].

Conflict of interest statement. None declared.

REFERENCES

1. Solomon,M.J., Larsen,P.L. and Varshavsky,A. (1988) Mapping proteinDNA interactions *in vivo* with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell*, **53**, 937–947.
2. Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E. *et al.*

(2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.
3. Iyer,V.R., Horak,C.E., Scafe,C.S., Botstein,D., Snyder,M. and Brown,P.O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.
4. 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.
5. Qi,Y., Rolfe,A., MacIsaac,K.D., Gerber,G.K., Pokholok,D., Zeitlinger,J., Danford,T., Dowell,R.D., Fraenkel,E., Jaakkola,T.S. *et al.* (2006) High-resolution computational models of genome binding events. *Nat. Biotechnol.*, **24**, 963–970.
6. Guo,Y., Papachristoudis,G., Altshuler,R.C., Gerber,G.K., Jaakkola,T.S., Gifford,D.K. and Mahony,S. (2010) Discovering homotypic binding events at high spatial resolution. *Bioinformatics*, **26**, 3028–3034.
7. Li,X.Y., MacArthur,S., Bourgon,R., Nix,D., Pollard,D.A., Iyer,V.N., Hechmer,A., Simirenko,L., Stapleton,M., Hendriks,C.L. *et al.* (2008) Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biol.*, **6**, e27.
8. The ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
9. Moorman,C., Sun,L.V., Wang,J., deWit,E., Talhout,W., Ward,L.D., Greil,F., Lu,X., White,K.P., Bussemaker,H.J. *et al.* (2006) Hotspots of transcription factor colocalization in the genome of *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA*, **103**, 12027–12032.
10. Gerstein,M.B., Kundaje,A., Hariharan,M., Landt,S.G., Yan,K.-K., Cheng,C., Mu,X.J., Khurana,E., Rozowsky,J., Alexander,R. *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature*, **489**, 91–100.
11. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC(R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
12. Sandelin,A., Alkema,W., Engstrm,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
13. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.
14. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A., Chen,X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
15. Berger,M.F., Badis,G., Gehrke,A.R., Talukder,S., Philippakis,A.A., Pea-Castillo,L., Alleyne,T.M., Mnaimneh,S., Botvinnik,O.B., Chan,E.T. *et al.* (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, **133**, 1266–1276.
16. 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.
17. Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of Cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205–1214.
18. Liu,X.S., Brutlag,D.L. and Liu,J.S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–839.
19. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Int. Syst. Mol. Biol.*, **2**, 28–36.
20. Pavesi,G., Mauri,G. and Pesole,G. (2001) An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, **17**, S207–S214.
21. 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.
22. Che,D., Jensen,S., Cai,L. and Liu,J.S. (2005) BEST: binding-site estimation suite of tools. *Bioinformatics*, **21**, 2909–2911.
23. Romer,K.A., Kayombya,G. and Fraenkel,E. (2007) WebMOTIFS: automated discovery, filtering and scoring of DNA sequence motifs using multiple programs and Bayesian approaches. *Nucleic Acids Res.*, **35**, W217–W220.
24. Sun,H., Yuan,Y., Wu,Y., Liu,H., Liu,J.S. and Xie,H. (2010) Tmod: toolbox of motif discovery. *Bioinformatics*, **26**, 405–407.
25. Crooks,G.E., Hon,G., Chandonia,J. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
26. Bar-Joseph,Z., Gifford,D.K. and Jaakkola,T.S. (2001) Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, **17**, S22–S29.
27. Bairoch,A. (2004) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
28. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
29. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.
30. Frietze,S., Lan,X., Jin,V.X. and Farnham,P.J. (2010) Genomic targets of the KRAB and SCAN domain-containing zinc finger protein 263. *J. Biol. Chem.*, **285**, 1393–1403.
31. Karin,M., Liu,Z.g. and Zandi,E. (1997) AP-1 function and regulation. *Curr. Opin. Cell Biol.*, **9**, 240–246.
32. Kawana,M., Lee,M.E., Quertermous,E.E. and Quertermous,T. (1995) Cooperative interaction of GATA-2 and AP1 regulates transcription of the endothelin-1 gene. *Mol. Cell. Biol.*, **15**, 4225–4231.
33. Wang,W., Xue,Y., Zhou,S., Kuo,A., Cairns,B.R. and Crabtree,G.R. (1996) Diversity and specialization of mammalian SWI/SNF complexes. *Genes Dev.*, **10**, 2117–2130.
34. Ito,T., Yamauchi,M., Nishina,M., Yamamichi,N., Mizutani,T., Ui,M., Murakami,M. and Iba,H. (2001) Identification of SWI.SNF complex subunit BAF60a as a determinant of the transactivation potential of Fos/Jun dimers. *J. Biol. Chem.*, **276**, 2852–2857.
35. Nateri,A.S., Spencer-Dene,B. and Behrens,A. (2005) Interaction of phosphorylated c-Jun with TCF4 regulates intestinal cancer development. *Nature*, **437**, 281–285.
36. Mostoslavsky,R., Chua,K.F., Lombard,D.B., Pang,W.W., Fischer,M.R., Gellon,L., Liu,P., Mostoslavsky,G., Franco,S., Murphy,M.M. *et al.* (2006) Genomic instability and aging-like phenotype in the absence of mammalian SIRT6. *Cell*, **124**, 315–329.
37. Huang,Y., Myers,S.J. and Dingledine,R. (1999) Transcriptional repression by REST: recruitment of Sin3A and histone deacetylase to neuronal genes. *Nat. Neurosci.*, **2**, 867–872.
38. Nascimento,E.M., Cox,C.L., Macarthur,S., Hussain,S., Trotter,M., Blanco,S., Suraj,M., Nichols,J., Kbler,B., Benitah,S.A. *et al.* (2011) The opposing transcriptional functions of Sin3a and c-Myc are required to maintain tissue homeostasis. *Nat. Cell Biol.*, **13**, 1395–1405.
39. Zervos,A.S., Gyuris,J. and Brent,R. (1993) Mxi1, a protein that specifically interacts with Max to bind Myc-Max recognition sites. *Cell*, **72**, 223–232.
40. Li-Weber,M., Davydov,I., Krafft,H. and Krammer,P. (1994) The role of NF-Y and IRF-2 in the regulation of human IL-4 gene expression. *J. Immunol.*, **153**, 4122–4133.
41. Scott,E., Simon,M., Anastasi,J. and Singh,H. (1994) Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages. *Science*, **265**, 1573–1577.
42. Villard,J., Peretti,M., Masternak,K., Barras,E., Caretti,G., Mantovani,R. and Reith,W. (2000) A functionally essential domain of RFX5 mediates activation of major histocompatibility complex class II promoters by promoting cooperative binding between RFX and NF-Y. *Mol. Cell. Biol.*, **20**, 3364–3376.
43. Yu,L., Wu,Q., Yang,C.P. and Horwitz,S.B. (1995) Coordination of transcription factors, NF-Y and C/EBP beta, in the regulation of the mdr1b promoter. *Cell Growth Differ.*, **6**, 1505–1512.

44. Roder,K., Wolf,S., Larkin,K. and Schweizer,M. (1999) Interaction between the two ubiquitously expressed transcription factors NF-Y and Sp1. *Gene*, **234**, 61–69.
45. Caretti,G., Salsi,V., Vecchi,C., Imbriano,C. and Mantovani,R. (2003) Dynamic recruitment of NF-Y and histone acetyltransferases on cell-cycle promoters. *J. Biol. Chem.*, **278**, 30435–30440.
46. Ivanov,V.N., Bhoumik,A., Krasilnikov,M., Raz,R., Owen-Schaub,L.B., Levy,D., Horvath,C.M. and Ronai,Z. (2001) Cooperation between STAT3 and c-jun suppresses fas transcription. *Mol. Cell*, **7**, 517–528.
47. Choi,S., Cho,Y., Kim,H. and Park,J. (2007) ROS mediate the hypoxic repression of the hepcidin gene by inhibiting C/EBPalpha and STAT-3. *Biochem. Biophys. Res. Commun.*, **356**, 312–317.
48. Sementchenko,V.I. and Watson,D.K. (2000) Ets target genes: past, present and future. *Oncogene*, **19**, 6533–6548.
49. Rothbcher,U., Bertrand,V., Lamy,C. and Lemaire,P. (2007) A combinatorial code of maternal GATA, Ets and beta-catenin-TCF transcription factors specifies and patterns the early ascidian ectoderm. *Development*, **134**, 4023–4032.
50. Taylor,J.M., Dupont-Versteegden,E.E., Davies,J.D., Hassell,J.A., Houl,J.D., Gurley,C.M. and Peterson,C.A. (1997) A role for the ETS domain transcription factor PEA3 in myogenic differentiation. *Mol. Cell. Biol.*, **17**, 5550–5558.
51. O'Geen,H., Lin,Y., Xu,X., Echipare,L., Komashko,V.M., He,D., Frietze,S., Tanabe,O., Shi,L., Sartor,M.A. *et al.* (2010) Genome-wide binding of the orphan nuclear receptor TR4 suggests its general role in fundamental biological processes. *BMC Genomics*, **11**, 689.
52. Adams,B., Drfler,P., Aguzzi,A., Kozmik,Z., Urbnek,P., Maurer-Fogy,I. and Busslinger,M. (1992) Pax-5 encodes the transcription factor BSAP and is expressed in B lymphocytes, the developing CNS, and adult testis. *Genes Dev.*, **6**, 1589–1607.
53. Fitzsimmons,D., Hodsdon,W., Wheat,W., Maira,S.M., Wasylyk,B. and Hagman,J. (1996) Pax-5 (BSAP) recruits Ets proto-oncogene family proteins to form functional ternary complexes on a B-cell-specific promoter. *Genes Dev.*, **10**, 2198–2211.
54. Dudek,H., Tantravahi,R.V., Rao,V.N., Reddy,E.S. and Reddy,E.P. (1992) Myb and Ets proteins cooperate in transcriptional activation of the mim-1 promoter. *Proc. Natl Acad. Sci. USA*, **89**, 1291–1295.
55. Mazars,R., Gonzalez-de-Peredo,A., Cayrol,C., Lavigne,A., Vogel,J.L., Ortega,N., Lacroix,C., Gautier,V., Huet,G., Ray,A. *et al.* (2010) The THAP-zinc finger protein THAP1 associates with coactivator HCF-1 and O-GlcNAc transferase: a link between DYT6 and DYT3 dystonias. *J. Biol. Chem.*, **285**, 13364–13371.
56. Yu,H., Mashtalir,N., Daou,S., Hammond-Martel,I., Ross,J., Sui,G., Hart,G.W., Rauscher,F.J.R., Drobetsky,E., Milot,E. *et al.* (2010) The ubiquitin carboxyl hydrolase BAP1 forms a ternary complex with YY1 and HCF-1 and is a critical regulator of gene expression. *Mol. Cell. Biol.*, **30**, 5071–5085.
57. Looijenga,L.H., Stoop,H., deLeeuw,H.P., deGouveia Brazao,C.A., Gillis,A.J., vanRoozendaal,K.E., vanZoelen,E.J., Weber,R.F., Wolffenbuttel,K.P., vanDekken,H. *et al.* (2003) POU5F1 (OCT3/4) identifies cells with pluripotent potential in human germ cell tumors. *Cancer Res.*, **63**, 2244–2250.
58. Loh,Y., Wu,Q., Chew,J., Vega,V.B., Zhang,W., Chen,X., Bourque,G., George,J., Leong,B., Liu,J. *et al.* (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.*, **38**, 431–440.
59. Yi,F. and Merrill,B.J. (2007) Stem cells and TCF proteins: a role for beta-catenin-independent functions. *Stem Cell Rev.*, **3**, 39–48.
60. Phillips,J.E. and Corces,V.G. (2009) CTCF: master weaver of the genome. *Cell*, **137**, 1194–1211.
61. McKay,M.J., Troelstra,C., vander,P., Kanaar,R., Smit,B., Hagemeijer,A., Bootsma,D. and Hoeijmakers,J.H. (1996) Sequence conservation of therad21 *Schizosaccharomyces* pombeDNA double-strand break repair gene in human and mouse. *Genomics*, **36**, 305–315.
62. Wendt,K.S., Yoshida,K., Itoh,T., Bando,M., Koch,B., Schirghuber,E., Tsutsumi,S., Nagae,G., Ishihara,K., Mishiro,T. *et al.* (2008) Cohesin mediates transcriptional insulation by CCCT C-binding factor. *Nature*, **451**, 796–801.
63. Rubio,E.D., Reiss,D.J., Welcsh,P.L., Disteche,C.M., Filippova,G.N., Baliga,N.S., Aebersold,R., Ranish,J.A. and Krumm,A. (2008) CTCF physically links cohesin to chromatin. *Proc. Natl Acad. Sci. USA*, **105**, 8309–8314.
64. Jelinic,P., Stehle,J. and Shaw,P. (2006) The testis-specific factor CTCFL cooperates with the protein methyltransferase PRMT7 in H19 imprinting control region methylation. *PLoS Biol.*, **4**, e355.
65. Bischof,L.J., Kagawa,N., Moskow,J.J., Takahashi,Y., Iwamatsu,A., Buchberg,A.M. and Waterman,M.R. (1998) Members of the Meis1 and Pbx homeodomain protein families cooperatively bind a cAMP-responsive sequence (CRS1) from BovineCYP17. *J. Biol. Chem.*, **273**, 7941–7948.
66. Kappel,A., Schlaeger,T.M., Flamme,I., Orkin,S.H., Risau,W. and Breier,G. (2000) Role of SCL/Tal-1, GATA, and ets transcription factor binding sites for the regulation of flk-1 expression during murine vascular development. *Blood*, **96**, 3078–3085.
67. Mouthon,M.A., Bernard,O., Mitjavila,M.T., Romeo,P.H., Vainchenker,W. and Mathieu-Mahul,D. (1993) Expression of tal-1 and GATA-binding proteins during human hematopoiesis. *Blood*, **81**, 647–655.
68. Chan,H.M. and La Thangue,N.B. (2001) p300/CBP proteins: HATs for transcriptional bridges and scaffolds. *J. Cell Sci.*, **114**, 2363–2373.
69. Visel,A., Blow,M.J., Li,Z., Zhang,T., Akiyama,J.A., Holt,A., Plajzer-Frick,I., Shoukry,M., Wright,C., Chen,F. *et al.* (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**, 854–858.
70. Costa,R.H., Kalinichenko,V.V., Holterman,A.L. and Wang,X. (2003) Transcription factors in liver development, differentiation, and regeneration. *Hepatology*, **38**, 1331–1347.
71. Zaret,K.S. and Carroll,J.S. (2011) Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.*, **25**, 2227–2241.
72. Johnson,C.A. and Turner,B.M. (1999) Histone deacetylases: complex transducers of nuclear signals. *Semin. Cell Dev. Biol.*, **10**, 179–188.
73. Furusawa,T. and Cherukuri,S. (2010) Developmental function of HMGN proteins. *Biochim. Biophys. Acta*, **1799**, 69–73.
74. Peng,J., Zhu,Y., Milton,J.T. and Price,D.H. (1998) Identification of multiple cyclin subunits of human P-TEFb. *Genes Dev.*, **12**, 755–762.
75. Partington,G.A. and Patient,R.K. (1999) Phosphorylation of GATA-1 increases its DNA-binding affinity and is correlated with induction of human K562 erythroleukaemia cells. *Nucleic Acids Res.*, **27**, 1168–1175.
76. 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.
77. Xu,D., Zhao,L., Del Valle,L., Miklossy,J. and Zhang,L. (2008) Interferon regulatory factor 4 is involved in *Epstein-Barr virus*-mediated transformation of human B lymphocytes. *J. Virol.*, **82**, 6251–6258.
78. Paun,A. and Pitha,P.M. (2007) The IRF family, revisited. *Biochimie*, **89**, 744–753.
79. Corcoran,L.M., Karvelas,M., Nossal,G.J., Ye,Z.S., Jacks,T. and Baltimore,D. (1993) Oct-2, although not required for early B-cell development, is critical for later B-cell maturation and for postnatal survival. *Genes Dev.*, **7**, 570–582.
80. Baeuerle,P.A. and Henkel,T. (1994) Function and activation of NF-kappa B in the immune system. *Annu. Rev. Immunol.*, **12**, 141–179.
81. Lee,C.S., Friedman,J.R., Fulmer,J.T. and Kaestner,K.H. (2005) The initiation of liver development is dependent on Foxa transcription factors. *Nature*, **435**, 944–947.
82. Seto,E., Shi,Y. and Shenk,T. (1991) YY1 is an initiator sequence-binding protein that directs and activates transcription *in vitro*. *Nature*, **354**, 241–245.
83. Nagarajan,P., Onami,T.M., Rajagopalan,S., Kania,S., Donnell,R. and Venkatachalam,S. (2009) Role of chromodomain helicase DNA-binding protein 2 in DNA damage response signaling and tumorigenesis. *Oncogene*, **28**, 1053–1062.

84. Deng,C. (2003) Roles of BRCA1 in DNA damage repair: a link between development and cancer. *Hum. Mol. Genet.*, **12**, 113R–123R.
85. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3[prime] UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
86. Farnham,P.J. (2009) Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.*, **10**, 605–616.
87. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
88. Spivakov,M., Akhtar,J., Kheradpour,P., Beal,K., Girardot,C., Koscielny,G., Herrero,J., Kellis,M., Furlong,E.E. and Birney,E. (2012) Analysis of variation at transcription factor binding sites in *Drosophila* and humans. *Genome Biol.*, **13**, R49.
89. Ward,L.D. and Kellis,M. (2011) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.*, **40**, D930–D934.
90. 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.
91. Neph,S., Vierstra,J., Stergachis,A.B., Reynolds,A.P., Haugen,E., Vernot,B., Thurman,R.E., John,S., Sandstrom,R., Johnson,A.K. *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**, 83–90.
92. Berman,B.P., Nibu,Y., Pfeiffer,B.D., Tomancak,P., Celniker,S.E., Levine,M., Rubin,G.M. and Eisen,M.B. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA*, **99**, 757–762.
93. Schroeder,M.D., Pearce,M., Fak,J., Fan,H., Unnerstall,U., Emberly,E., Rajewsky,N., Siggia,E.D. and Gaul,U. (2004) Transcriptional control in the segmentation gene network of *Drosophila*. *PLoS Biol.*, **2**, e271.
94. Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.
95. Moses,A., Chiang,D., Pollard,D., Iyer,V. and Eisen,M. (2004) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol.*, **5**, R98.
96. Kheradpour,P., Stark,A., Roy,S. and Kellis,M. (2007) Reliable prediction of regulator targets using 12 *Drosophila* genomes. *Genome Res.*, **17**, 1919–1931.
97. Lindblad-Toh,K., Garber,M., Zuk,O., Lin,M.F., Parker,B.J., Washietl,S., Kheradpour,P., Ernst,J., Jordan,G., Mauceli,E. *et al.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, **478**, 476–482.
98. Schmidt,D., Wilson,M.D., Ballester,B., Schwalie,P.C., Brown,G.D., Marshall,A., Kutter,C., Watt,S., Martinez-Jimenez,C.P. *et al.* (2010) Five-vertebrate ChIP-seq Reveals the evolutionary dynamics of transcription factor binding. *Science*, **328**, 1036–1040.
99. Boyer,L.A., Lee,T.I., Cole,M.F., Johnstone,S.E., Levine,S.S., Zucker,J.P., Guenther,M.G., Kumar,R.M., Murray,H.L., Jenner,R.G. *et al.* (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**, 947–956.
100. Lee,T.I., Jenner,R.G., Boyer,L.A., Guenther,M.G., Levine,S.S., Kumar,R.M., Chevalier,B., Johnstone,S.E., Cole,M.F., Isono,K.I. *et al.* (2006) Control of developmental regulators by polycomb in human embryonic stem cells. *Cell*, **125**, 301–313.
101. MacArthur,S., Li,X., Li,J., Brown,J., Chu,H.C., Zeng,L., Grondona,B., Hechmer,A., Simirenko,L., Keranen,S. *et al.* (2009) Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol.*, **10**, R80.
102. Pietrokovski,S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.*, **24**, 3836–3845.
103. Gray,K.A., Daugherty,L.C., Gordon,S.M., Seal,R.L., Wright,M.W. and Bruford,E.A. (2013) Genenames.org: the HGNC resources in 2013. *Nucleic Acids Res.*, **41**, D545–D552.
104. Kharchenko,P.V., Tolstorukov,M.Y. and Park,P.J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotech.*, **26**, 1351–1359.
105. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
106. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.
107. Touzet,H. and Varre,J. (2007) Efficient and accurate *P*-value computation for position weight matrices. *Algorithms Mol. Biol.*, **2**, 15.
108. Wilson,E.B. (1927) Probable Inference, the Law of Succession, and Statistical Inference. *J. Am. Stat. Assoc.*, **22**, 209–212.
109. Mahony,S., Auron,P.E. and Benos,P.V. (2007) DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS Comput. Biol.*, **3**, e61.
110. Sandelin,A. and Wasserman,W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol.*, **338**, 207–215.