Published online 4 December 2018

Nucleic Acids Research, 2019, Vol. 47, No. 4 ***e21***
doi: 10.1093/nar/gky1210

A map of direct TF–DNA interactions in the human genome

Marius Gheorghe [1], Geir Kjetil Sandve [2], Aziz Khan [1], Jeanne Chèneby[3], Benoit Ballester [3] and Anthony Mathelier [1,4,*]

[1]Centre for Molecular Medicine Norway (NCMM), University of Oslo, Oslo, Norway, [2]Department of Informatics, University of Oslo, Oslo, Norway, [3]Aix Marseille Université, INSERM, TAGC, Marseille, France and [4]Department of Cancer Genetics, Institute for Cancer Research, Radiumhospitalet, Oslo, Norway

Received August 18, 2018; Revised October 31, 2018; Editorial Decision November 18, 2018; Accepted November 20, 2018

ABSTRACT

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is the most popular assay to identify genomic regions, called ChIP-seq peaks, that are bound *in vivo* by transcription factors (TFs). These regions are derived from direct TF–DNA interactions, indirect binding of the TF to the DNA (through a co-binding partner), nonspecific binding to the DNA, and noise/bias/artifacts. Delineating the *bona fide* direct TF–DNA interactions within the ChIP-seq peaks remains challenging. We developed a dedicated software, ChIP-eat, that combines computational TF binding models and ChIP-seq peaks to automatically predict direct TF–DNA interactions. Our work culminated with predicted interactions covering >2% of the human genome, obtained by uniformly processing 1983 ChIP-seq peak data sets from the ReMap database for 232 unique TFs. The predictions were *a posteriori* assessed using protein binding microarray and ChIP-exo data, and were predominantly found in high quality ChIP-seq peaks. The set of predicted direct TF–DNA interactions suggested that high-occupancy target regions are likely not derived from direct binding of the TFs to the DNA. Our predictions derived co-binding TFs supported by protein-protein interaction data and defined *cis*-regulatory modules enriched for disease- and trait-associated SNPs. We provide this collection of direct TF–DNA interactions and *cis*-regulatory modules through the UniBind web-interface (http://unibind.uio.no).

INTRODUCTION

The transcription of DNA into RNA is mainly regulated through a complex interplay between proteins and the chromatin at *cis*-regulatory regions such as promoters and enhancers. Transcription factors (TFs) are key proteins specifically binding short DNA sequences, known as TF binding sites (TFBSs), to ensure transcription at appropriate rates in the correct cell types (1). Therefore, genome-wide identification of TFBSs is a critical step to decipher transcriptional regulation, and how this process is altered in diseases (2).

Classically, genome-wide *in vivo* TF binding regions are identified through the chromatin immunoprecipitation followed by sequencing (ChIP-seq) assay (3). The genomic regions obtained with ChIP-seq, the so-called ChIP-seq peaks, are usually a few hundred base pairs (bp)-long and should encompass the TFBSs (∼10 bp-long), where direct TF–DNA interactions occur. However, ChIP-seq peaks derive from either direct TF–DNA interactions, protein-protein interactions with other regulators such as co-factors, or unspecific binding. Moreover, ChIP-seq experiments are prone to artifacts and delineating *bona fide* TF-bound regions is still an ongoing challenge (4–6) (Wreczycka *et al.*, bioRxiv, 10.1101/107680).

As TFs specifically recognize DNA sequence motifs, computational tools have been instrumental in the prediction and characterization of direct TF–DNA interactions (7). TFBSs are commonly modelled with position weight matrices (PWMs), which represent the probability of each nucleotide to be present at each position within bona fide TFBSs (7). While PWMs work well (8), more sophisticated approaches have recently been designed to model complex features of TF–DNA interactions captured by next-generation sequencing data (e.g. (9–13)). However, the best performing model varies for different TFs or TF families (8,14,15).

While multiple resources collecting TF binding regions derived from ChIP-seq exist (16–19), a limited number store genome-wide identification of TFBSs (17,20,21). The TFBS Conserved Track of the UCSC Genome Browser combined phylogenetic sequence conservation and PWMs to identify TFBSs (22) while the MANTA resource (23) integrated ChIP-seq peaks from ReMap (16) with PWMs from JASPAR (24) for TFBS predictions. A strong limitation of these approaches is that they use the same pre-defined score

*To whom correspondence should be addressed. Tel: +47 228 40 561; Email: anthony.mathelier@ncmm.uio.no

thresholds for all PWMs and all data sets. The ORegAnno database provides TFBSs obtained through literature curation (21), but the number of TFBSs available for human is limited to ∼8000.

A previous study showed that ChIP-seq data sets fall within one of three categories: (i) data sets enriched for the TF canonical binding motif close to the ChIP-seq peak summit (where the highest number of ChIP-seq reads map), (ii) data sets lacking enrichment for the canonical binding motif close to the peak summit and (iii) data sets having a combination of peaks with and without the TF canonical binding motif proximal to the peak-summit (25). Most ChIP-seq data sets were observed in category (iii). As direct TF–DNA interactions are expected to be enriched at ChIP-seq peak summits (25–30), Worsley Hunt *et al.* developed a heuristic approach specifically based on PWMs to automatically identify, in each ChIP-seq data set, this enrichment zone. The method determines the thresholds on the PWM scores and distances to the peak summits delimiting the enrichment zone that contains direct TF–DNA interactions. However, this method does not work with some more recent TFBS computational models (15,31,32).

In this study, we mapped direct TF–DNA interactions in the human genome in a refined manner by capitalizing on uniformly processed TF ChIP-seq data sets and computational tools modelling TFBSs. We provide (i) a new software to predict direct TF–DNA interactions within ChIP-seq peaks along with (ii) genome-wide predictions of such interactions in the human genome. Using an entropy-based algorithm, we have developed ChIP-eat, a tool that automatically identifies direct TF–DNA interactions using both ChIP-seq peaks and any computational model for TFBSs. We applied ChIP-eat to 1983 human ChIP-seq peak data sets from the ReMap database (16), accounting for 232 distinct TFs. The set of predicted direct TF–DNA interactions derived from PWMs covers >2% of the human genome. To make this resource available to the community, we have created UniBind (http://unibind.uio.no/), a web-interface providing public access to the predictions. We validated *a posteriori* these TFBS predictions using protein binding microarray (33) and ChIP-exo (34) data, and multiple ChIP-seq peak-callers. We used these TFBSs to (i) confirm that hotspots of ChIP-seq peaks (also known as high occupancy target regions (35)) are likely not derived from direct TF–DNA interactions, (ii) predict co-binding TFs and (iii) define *cis*-regulatory modules, which are enriched for disease- and trait-associated SNPs.

MATERIALS AND METHODS

ChIP-seq data

The ChIP-seq data sets considered were retrieved, processed, and classified as part of the last update (2018) of the ReMap database (16) (Supplementary Figure S1).

TF binding profiles

For 1983 ChIP-seq data sets used in the last ReMap update, we were able to manually assign TF binding profiles corresponding to the ChIP'ed TFs as position frequency matrices (PFMs) from the JASPAR (2018) database (24).

Training data sets

To train the TFBS computational models (see below), we considered 101 bp sequences centered around the peak summits as positive training sets. When required for training, negative training sets were obtained by shuffling the positive sequences using the *g* subcommand of the BiasAway (version 0.96) tool to match the %GC composition (25).

TFBS computational models

Position weight matrices. JASPAR PFMs were converted to PWMs as previously described in (36). For each ChIP-seq data set, PWMs were optimized using DiMO (version 1.6; default parameters with a maximum of 150 optimization steps) using the corresponding training sets (37). For TFBS predictions, we considered PWM *relative* scores, which were computed as $relative\ score = 100 \times (absolute\ score - min)/(max - min)$ where *absolute score* corresponds to the PWM absolute/raw score and *min* and *max* to the minimal and maximal absolute/raw PWM scores, respectively.

Binding energy models. JASPAR PFMs were converted to binding energy models (BEMs; (32)) using the implementation from the MARS Tools (https://github.com/kipkurui/MARSTools; Kibet and Machanick, bioRxiv, doi:10.1101/065615). We modified the implementation to return a BEM score corresponding to 1 – (original score) to consider the best site of the DNA sequence as the one with the highest BEM score (instead of the lowest one).

Transcription factor flexible models. First-order transcription factor flexible models (TFFMs) (version 2.0) were initialized with the DiMO-optimized PFMs and trained with default parameters (https://github.com/wassermanlab/TFFM; (31)) on the positive training sets.

DNAshapedTFBS models. The DNA shape-based models were trained on the training sets using the DNAshapedTFBS tool (version 1.0; https://github.com/amathelier/DNAshapedTFBS/; (15)). We trained three types of DNAshapedTFBS models with the following features: (i) DiMO-optimized PWM + DNA shape, (ii) first-order TFFM + DNA shape and (iii) 4-bits encoding + DNA shape following (15). We considered the first and second order DNA shape features helix twist, propeller twist, minor groove width, and roll with values extracted from GBShape (38).

Landscape plots

Each TFBS computational model was applied to each ChIP-seq data set independently. Following the strategy described in (25), we considered 1001 bp sequences centered around the peak summits, obtained using the bedtools (version 2.25) *slop* subcommand (39). The trained computational models were used to extract the best (maximal score) site per 1001 bp ChIP-seq peak region. For each ChIP-seq data set, landscape plots were constructed from the corresponding sites following the TFBS_Visualization tool (25). These scatter plots were also converted into heat maps using the *kde2d* function from the MASS R package (40).

Automated identification of the enrichment zone

To define the enrichment zone for each landscape plot, we automatically identified the thresholds for the TFBS computational model scores and distances to peak summits using the entropy-based algorithm from (41). The algorithm aims at identifying two classes of elements. Given a histogram, the algorithm selects the threshold that maximizes the within-class sum of the Shannon entropies for the elements in two classes (42). The two classes of elements identified are defined by the elements with values (i) above and (ii) below the threshold, respectively. This procedure optimally separates the input elements in two classes. Given a ChIP-seq data set, we applied the algorithm to the histograms of the TFBS computational model scores and distances to peak summits, independently. The maximum entropy implementation of the algorithm available in ImageJ (43) was used with default parameters.

The source code of the ChIP-eat software used to process ChIP-seq peak data sets to predict direct TF–DNA binding events is freely available at https://bitbucket.org/CBGR/chip-eat. Specifically, ChIP-eat trains a TFBS computational model and automatically defines the enrichment zone in the landscape plots to predict the underlying direct TF–DNA interactions. The identification of the enrichment zone has been applied to each TF ChIP-seq peak data set independently, allowing for the automatic detection of the thresholds that are specific to each data set with each TFBS computational model. Note that only the best hit per ChIP-seq peak has been considered to identify the enrichment zones and for all the downstream analyses.

Assessing the robustness of the enrichment zone identification

Random noise. For each ChIP-seq data set, we sampled the set of peaks using the seqtk (version 1.0) (https://github.com/lh3/seqtk) *sample* subcommand. The sequences of the sampled peaks were shuffled using the *fasta-shuffle-letters* subcommand of the MEME suite (version 4.11.4) (44) and added to the original set of ChIP-seq peaks. The automatic thresholding algorithm was applied to this new set. We tested the addition of shuffled peaks representing 10%, 25%, and 50% of the original set peaks.

Window size variability. For each ChIP-seq data set, we considered the region around the peak summit by extending with 300, 400, and 500 bp on each side using the bedtools *slop* subcommand. We considered ChIP-seq data sets where at least one TFBS was predicted within the enrichment zones obtained for all three window sizes.

Comparison with the heuristic approach to predict the enrichment zone. ChIP-eat was compared to the heuristic approach described in (25) and implemented in the TFBS_Visualization tool https://github.com/wassermanlab/TFBS_Visualization using the default parameters. The centrality of the TFBSs within the enrichment zones predicted by ChIP-eat and TFBS_Visualization was assessed using centrality *P*-value computations as described in the CentriMo tool (27). The statistical difference between the centrality *P*-values obtained with the heuristic method and ChIP-eat was assessed using a Mann-Whitney signed-rank test.

Genome coverage. The entire set of predicted TFBSs (within enrichment zones) was concatenated and then sorted using the *cat* and *sort* commands of the Unix operating system. The resulting set of locations was merged using the bedtools *merge* subcommand with default parameters. The genome coverage of the corresponding merged and non-overlapping positions was calculated as the percentage of the total number of nucleotides covered out of the total number of nucleotides in the hg38 version of the human genome.

TF–DNA binding affinity assessment with protein binding microarray data. Protein binding microarray (PBM) (45) data were retrieved from UniProbe (http://the_brain.bwh.harvard.edu/uniprobe/; (46)) for 40 TFs with available ChIP-seq data. For each ChIP-seq data set landscape plot, we extracted the DNA sequences at the sites within and outside of the predicted enrichment zone. The binding affinity of a TF to each site was computed as the median PBM intensity value of all the de Bruijn sequences containing the site sequence. The statistical difference between the distribution of PBM binding affinities from sites within and outside the enrichment zone was assessed using a two samples Mann-Whitney U test (47) implemented in the R package *stats*. A Bonferroni correction was applied to the computed *P*-values. The *P*-value density plot in Figure 3B was generated with the *density* R function with default parameters and the corresponding computed bandwidth was used to plot Supplementary Figure S10.

ChIP-exo data. ChIP-eat was applied with DiMO-optimized PFMs to the ChIP-exo data sets from (48), which were lifted over to hg38 using the liftOver tool (20). As for ChIP-seq peaks, we considered 1 001 bp regions centered around the peak summits.

ChIP-seq peaks from HOMER and BCP peak-callers. We successfully applied the HOMER (version 4.7.2) (49) and BCP (version 1.1) (50) peak-callers to 670 ENCODE ChIP-seq data sets (Supplementary Table S1). ChIP-eat was applied to the corresponding ChIP-seq peak regions with DiMO-optimized PFMs as described above. ChIP-seq peaks predicted to contain a direct TF–DNA interaction or not (using the enrichment zones) from the three peak-callers (MACS2 (51), HOMER, and BCP) were overlapped using the bedtools *intersect* subcommand. Hypergeometric tests were performed to assess the significance of the intersections using the R *phyper* function for every combination of two peak-callers with the following contingency matrix:

number of overlapping peaks **with** TFBSs from two peak-callers - 1	number of peaks **without** TFBSs from the two peak-callers
number of peaks **with** TFBSs from the two peak-callers	number of overlapping peaks from the two peak-callers

HOT/XOT regions. The high occupancy target (HOT) and extreme occupancy target (XOT) regions in all contexts were downloaded through the ENCODE data portal at http://encode-ftp.s3.amazonaws.com/modENCODE_VS_ENCODE/Regulation/Human/hotRegions/maphot_hs_selection_reg_cx_simP05_all.bed and http://encode-ftp.s3.amazonaws.com/modENCODE_VS_ENCODE/Regulation/Human/hotRegions/maphot_hs_selection_reg_cx_simP01_all.bed. ChIP-seq peaks were overlapped with the HOT/XOT regions using the bedtools *intersect* subcommand. The enrichment for overlap was assessed with a hypergeometric test using the R *phyper* function with the following contingency matrix:

number of peaks **without** TFBSs overlapping HOT/XOT regions - 1	number of peaks **with** TFBSs
number of peaks **without** TFBSs	total number of peaks

Identification of TFs with co-localized TFBSs. For each pair of distinct TFs (TF_A, TF_B), we extracted the closest TFBS associated with TF_B for each TFBS associated with TF_A and computed the geometric mean distance between midpoints of the paired TFBSs. With this approach, the geometric mean m_{AB} for the pair (TF_A, TF_B) is different from the geometric mean of the pair (TF_B, TF_A). With 232 TFs available in our analyses, we computed geometric means for 53 592 ordered pairs of TFs.

The colocalization of TFBSs for each TF pair was assessed using a Monte Carlo-based approach as follows. The number of TFBSs per TF ranged from 1 to 404 566, with 455 as the fifth percentile. We uniformly discretized the range [455, 414 172] to consider 50 TFBS set sizes (S_i for i in [1, 50]). We chose 414 172 as the maximum value to be able to compute a P-value for the set of 404 566 TFBSs. For each set size S_i, we created 500 sets of TFBSs by randomly selecting TFBSs from the total pool. Using these random sets, we computed null distributions for 500 Monte Carlo samples of geometric mean distances for each of the 2601 set size combinations. Specifically, this computation led to 2601 distributions of 500 geometric means. For the TF pair (TF_A, TF_B) with N_A and N_B TFBSs, respectively, we extracted the Monte Carlo sample of geometric mean distances M obtained from the random sets with S_A and S_B TFBSs, where $S_A = \min(S_i)$ with $S_i > N_A$ and $S_B = \min(S_i)$ with $S_i > N_B$. The empirical P-value associated with the pair (TF_A, TF_B) was computed as the number of times we observed a geometric mean smaller than m_{AB} from M over the 500 precomputed geometric means; if no smaller geometric mean was observed, the empirical P-value is defined as <0.002 (i.e. $1/500$).

Since the expected geometric mean distance increases with a decreasing number of TFBSs, this P-value computation is conservative (under-estimated significance). The obtained P-values were corrected for multiple testing using the Benjamini–Hochberg method (52), only the TF pairs with a FDR $<5\%$ were considered significant.

The detailed null distribution values can be downloaded and reproduced at https://hyperbrowser.uio.no/geirksa_sandbox/u/gsandve/h/null-distributions-for-manuscript-a-map-of-direct-tf-dna-interactions-in-the-human-genome. These computations are based on running the static methods 'ConcatenateNullDistributionsTool.execute' and 'ComputeNullDistributionForEachCombinationFromSuiteVsSuiteTool.execute' (with argument values corresponding to parameter settings annotated in the Galaxy (53) history above) in the code provided at https://hyperbrowser.uio.no/geirksa_sandbox/static/hyperbrowser/files/div/hb.zip. The source code for the comparison with null distributions is available at https://bitbucket.org/CBGR/co-binding/.

GeneMANIA. We used the GeneMANIA software (54) to extract known protein–protein interactions from the list of TFs with significant co-localized TFBSs and plot the corresponding network.

Prediction of cis*-regulatory modules.* The TFBSs predicted by ChIP-eat were sorted and merged using the bedtools *sort* and *merge* subcommands. The CREAM tool (Madani Tonekaboni *et al.*, bioRxiv, doi:10.1101/222562) was applied to the merged TFBSs to define *cis*-regulatory modules (CRMs) as genomic regions enriched for clusters of TFBSs.

GWAS trait- and disease-associated single nucleotide polymorphism enrichment analysis. We assessed the enrichment for GWAS trait- and disease-associated single nucleotide polymorphisms (SNPs) at CRMs using the *traseR* R package (version 1.10.0 (55)). CRM genomic positions were lifted over to the hg19 version of the human genome to perform the analyses. The set of SNPs (as of 30 April 2018) considered by *traseR* combined data from dbGaP (56) and NHGRI (57) as described in the corresponding bioconductor package vignette (https://bioconductor.org/packages/release/bioc/vignettes/traseR/inst/doc/traseR.pdf).

Conservation analysis. The hg38 phastCons (58) scores for multiple alignments of 99 vertebrate genomes to the human genome were retrieved as a bigWig file at http://hgdownload.cse.ucsc.edu/goldenpath/hg38/phastCons100way/hg38.phastCons100way.bw. The TFBSs predicted by ChIP-eat were sorted and merged using the bedtools *sort* and *merge* subcommands. The locations overlapping CRMs were obtained using the bedtools *intersect* subcommand. The corresponding genomic locations (for all TFBSs and TFBSs in CRMs) in BED format were decomposed into 1 bp intervals using bedops v.2.4.14 (59) with the –chop 1 option. The phastCons scores at every bp were extracted with the *ex* subcommand of the bwtool (60) using the corresponding BED and phastCons bigWig files.

The UniBind web interface. All the TFBS predictions, corresponding ReMap ChIP-seq peaks, trained TFBS computational models, and CRMs are available through the UniBind database at http://unibind.uio.no/. The UniBind web interface was developed in Python using the model-view-controller framework Django. It uses MySQL to store TFBS metadata and Bootstrap as the frontend template engine. The source code is available at https://bitbucket.org/CBGR/unibind.

Statistical analyses. All statistical analyses were performed in the R environment (version 3.4.4).

RESULTS

Predicting direct TF–DNA interactions in the human genome from ChIP-seq data

Given a set of ChIP-seq peaks and a TFBS computational model such as a PWM, one can extract the best site per peak, which corresponds to the DNA subsequence of the peak with the highest score for the model. The higher the score, the stronger the computational evidence that the site is similar to TFBSs known to be bound by the TF (36). Moreover, it has been shown that the closer the site to the peak summit, the more likely it is to represent a direct TF–DNA interaction with experimental evidence from the ChIP-seq assay (25,27,30). Hence, direct TF–DNA interactions captured by ChIP-seq are enriched for high scores and small distances to the peak summits (Figure 1A,B). These characteristics have previously been used to automatically predict direct TF–DNA interactions by selecting score and distance thresholds defining these enrichment zones using a heuristic approach (25). This approach used pre-defined parameter values and was specifically designed for PWMs, but is not applicable to more recent TFBS computational models such as binding energy models (BEMs) (32), transcription factor flexible models (TFFMs) (31), and DNA shape-based models (DNAshapedTFBS) (15).

We aimed to predict direct TF–DNA interactions (TFBSs) within ChIP-seq peaks and developed the ChIP-eat software that automatically identifies the enrichment zone for any TFBS computational model. It uses a non-parametric, entropy-based algorithm originally designed to separate background/noise from foreground/signal in image processing (41) (Supplementary Figure S2). We applied this algorithm to the distributions of site scores and distance to peak summits independently to separate direct TF–DNA interaction events from other binding subtypes and ChIP-seq artifacts (Figure 1C,D; Materials and Methods). The two thresholds define the enrichment zone, which delimits the sites that are predicted as TFBSs with both experimental and computational evidence of direct TF–DNA interactions. With this approach, we automatically adjust the enrichment zone discovery specifically for each TF ChIP-seq peak data set and for each computational model. The identified enrichment zone defines the thresholds on the TFBS computational model scores and distances to the peak summits in a data set-specific manner.

We retrieved 1983 ChIP-seq peak data sets from ReMap (16), accounting for 232 TFs with a PFM available in the JASPAR database (24). Using DiMO-optimized PWMs, we compared the enrichment zones predicted by ChIP-eat with the ones obtained with the heuristic approach developed in (25). The enrichment zones predicted with ChIP-eat were more stringent than with the heuristic algorithm (Supplementary Figure S3A,B,D,E). The corresponding TFBSs predicted in the enrichment zones were more central to the peak summits with ChIP-eat than with the heuristic method as evaluated with CentriMo (27) (Supplementary Figure S3C, F). Moreover, ChIP-eat does not require any fixed values such as a predefined bin size (25) to predict the enrichment zones. Finally, ChIP-eat is not restricted to work with PWMs only and can be used with any TFBS computational model.

We applied ChIP-eat to the 1983 human ChIP-seq data sets with four types of computational TFBS models: DiMO-optimized PWMs, BEMs, TFFMs, and DNAshapedTFBS. These models were optimized for each ChIP-seq data set, independently (see Materials and Methods). In the following analyses, we focused on the predictions obtained with the DiMO-optimized PWMs (see Materials and Methods). This set of direct TF–DNA interactions (TFBSs) extracted from the enrichment zones covers >2% of the human genome, encompassing 8 304 135 distinct TFBS locations.

Predicted direct TF–DNA interactions are likely *bona fide* TFBSs

Robustness of the enrichment zone identification. The robustness of the method was first evaluated by applying ChIP-eat to genomic regions of ±300, 400, and 500 bp around the peak summits. The median distance threshold to the peak summit shifted from 72 bp using ±500 bp to 64 and 55 using ±400 and 300 bp, respectively. The median PWM scores thresholds were 85, 84.6 and 83.9 with ±500, 400, and 300 bp regions, respectively (see Supplementary Figure S8 for a visual representation using the 10 most frequent ChIP'ed TFs). The variability of the predicted enrichment zone when using different window sizes is similar to the variability between ChIP-seq data sets for the same TF (see below). Further, the number of predicted TFBSs within the enrichment zones were similar when using the different region sizes (Supplementary Figure S9). These analyses confirmed the robustness of the entropy-based thresholding algorithm to the window size considered. As previously used in (25), we considered the ±500 bp regions around the peak summits in the following analyses.

Considering the ChIP-seq data sets for the 10 most frequently ChIP'ed TFs, we observed that the thresholds on the PWM scores and distances to peak summits, defining the enrichment zones, were consistent between data sets for the same TF (Figure 2A,B). Namely, the median pairwise difference between PWM score thresholds for the same TF ranged from 1.7 to 3.7 and the median distance thresholds from 12 to 35 bp. As expected, the thresholds identified for distinct TFs are different (Figure 2C, D). Taken together, these results highlight that the entropy-based algorithm allows for the identification of enrichment zones specific to each TF and ChIP-seq data set, with consistent predictions between data sets for the same TF. Results were consistent with BEM, TFFM, and DNAshapedTFBS models (Supplementary Figures S4–S6).

We further evaluated the robustness of the method to noise by adding 10%, 25%, and 50% of shuffled sequences to the initial set of ChIP-seq peaks for all ChIP-seq peak data sets (see Materials and Methods). The median threshold on the distances to peak summits shifted from 73 bp in the initial set of ChIP-seq peaks to 70 bp with 10% noise, 67 bp with 25% noise, and to 63 bp when adding 50% noise. The median PWM score threshold was 85.2 for the initial set of ChIP-seq peaks and shifted to 85 when adding 10%

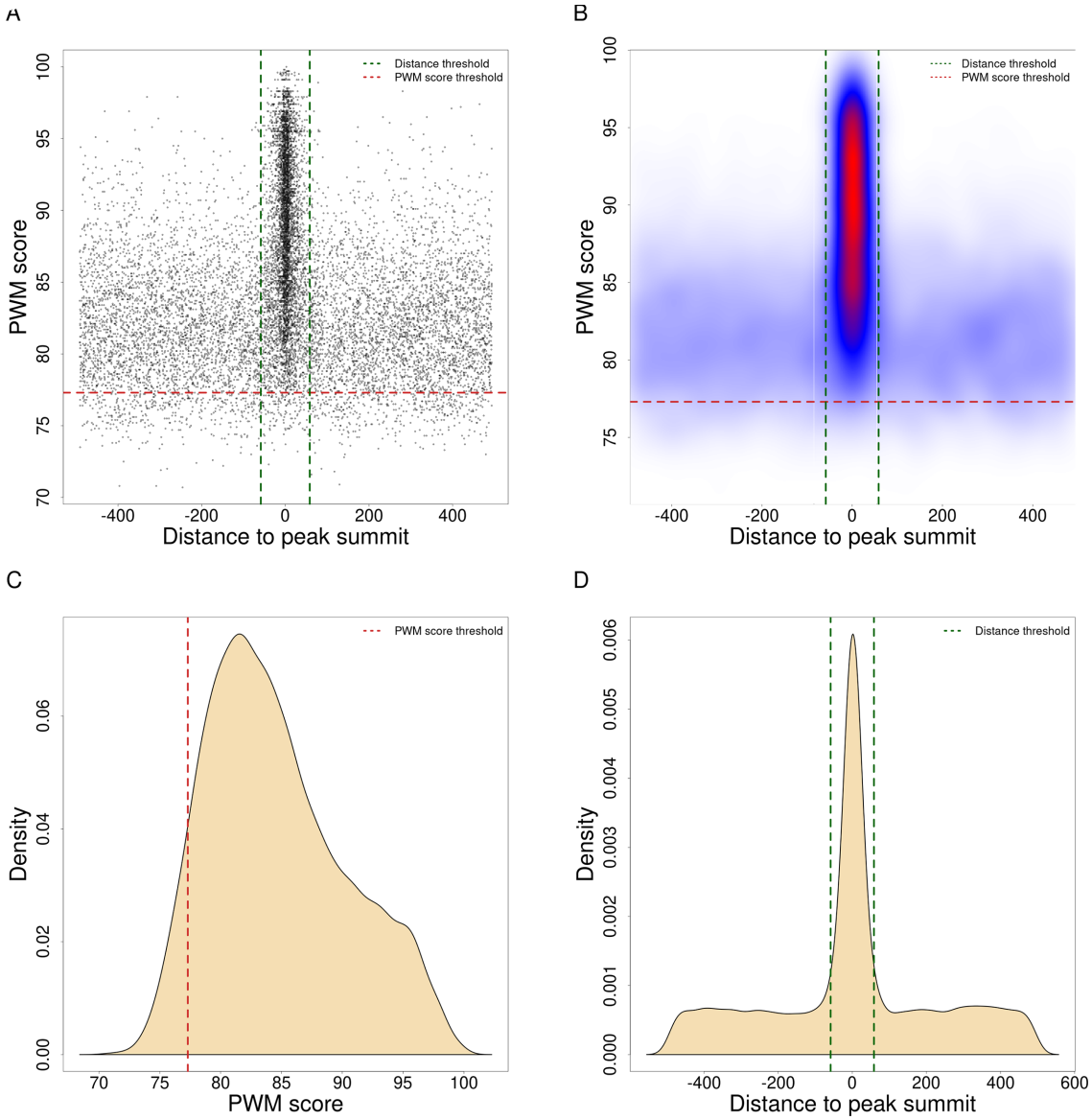

Figure 1. Automatic detection of the TFBS enrichment zone. Landscape plots (25) obtained with SRF ChIP-seq peaks using the DiMO-optimized PWM MA0083.3 from JASPAR are presented as scatter (**A**) and heatmap (**B**) plots. The enrichment zone (defined within the red and green dashed line boundaries, A-B) is automatically obtained by ChIP-eat with thresholds on PWM scores (red dashed lines; **C**) and distances to peak summits (green dashed lines; **D**). The enrichment zone provides TFBSs in ChIP-seq peaks (points in A) with supporting evidence for direct TF–DNA binding from the ChIP-seq assay (close distance to peak-summits, A-B, x-axis) and the computational model (PWM score, A-B, y-axis). Distances to peak summits in A, B and D are provided using a base pair unit.

of noise, to 84.8 when adding 25% of noise, and to 84.4 when adding 50% of noise. A visual representation for the 10 most frequently ChIP'ed TFs is available in Supplementary Figure S7. The variability of the thresholds defining the enrichment zones when adding noise is limited, within the range of variability between ChIP-seq peak data sets for the same TF (Figure 2). Taken together, these results show that the entropy-based thresholding algorithm delimiting the enrichment zones, as implemented in ChIP-eat, provides consistent results between data sets for the same ChIP'ed TF and is robust to the window sizes considered and random noise.

Validation using in vitro DNA binding affinities. To confirm *a posteriori* the high quality of our set of TFBS predictions, we assessed the TF binding affinity to DNA sequences derived experimentally from protein binding microarrays (PBM) (61). The PBM assay quantifies the binding affinity of a protein to all possible combinations of 8-mer DNA sequences. We retrieved PBM data from the UniPROBE database (46) for 40 different TFs present in our collection, corresponding to 249 ChIP-seq data sets (Supplementary Table S2). Note that the JASPAR PFMs for the ATF1, ATF3, and FOXJ2 TFs were originally derived from PBM data. For each ChIP-seq data set, we tested if the sites located in the enrichment zone presented higher binding affinity than sites outside (see Materials and Methods). The

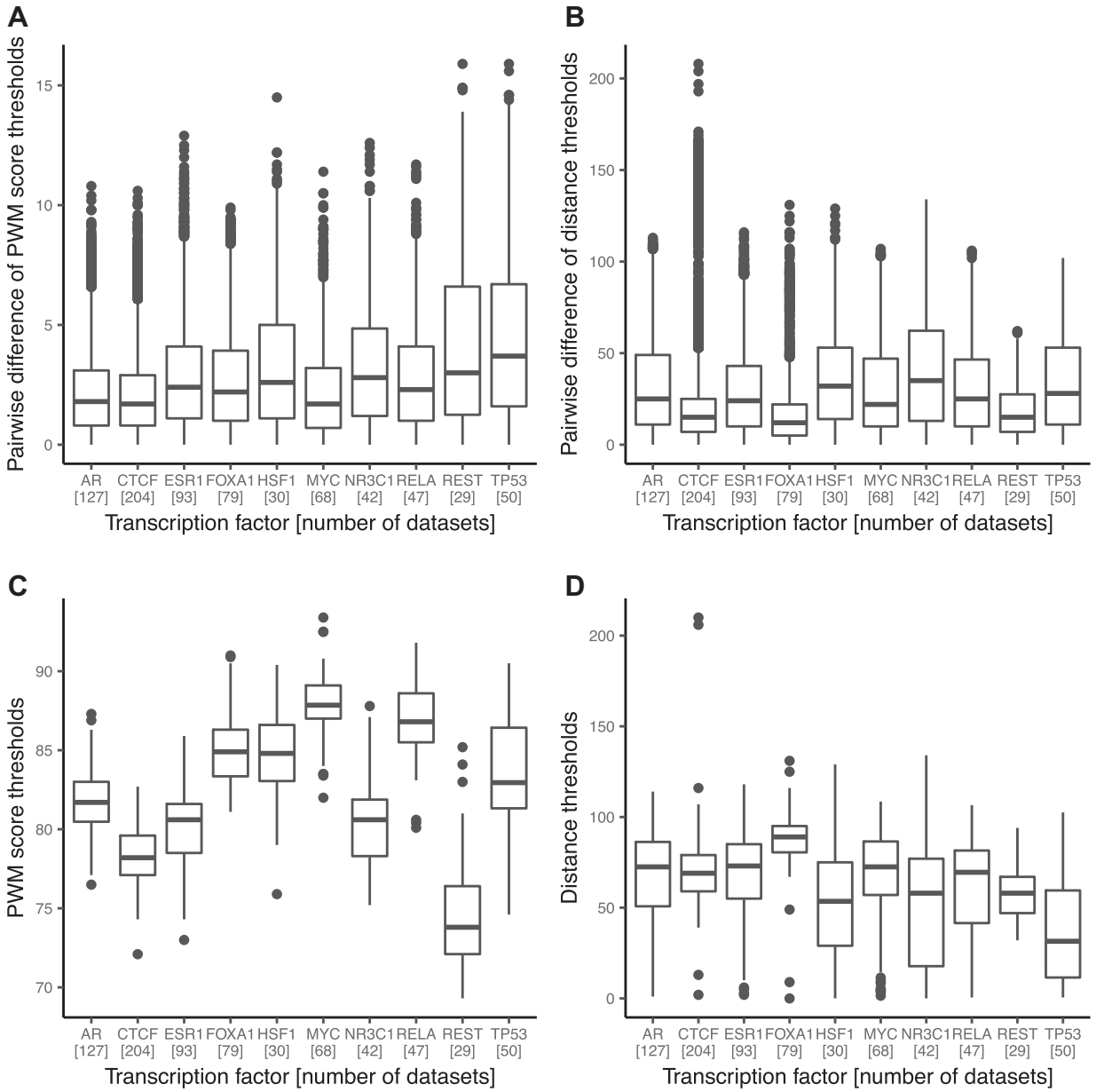

Figure 2. Assessment of the thresholds predicted by ChIP-eat across data sets. Boxplots of the pairwise differences for DiMO-optimized PWM score thresholds and distances to peak summits thresholds between ChIP-seq data sets for the same TF are provided in panels (**A**) and (**B**), respectively. Absolute variations of DiMO-optimized PWM score thresholds and distances to the peak summits within all data sets for the same TF are provided in panels (**C**) and (**D**), respectively. The ten TFs with the highest number of data sets were selected; the number of data sets for each TF is provided between brackets.

distributions of the binding affinity scores for sites within and outside the enrichment zones were compared using a Mann-Whitney U test (Figure 3A; Materials and Methods). Predicted direct TF–DNA interactions (sites within the enrichment zone) had significantly higher binding affinity than the other sites for 75% of the data sets with *P*-value <0.01 and 81% with *P*-value <0.05 (Figure 3B). Similar results were obtained when considering BEM, TFFM, and DNAshapedTFBSs computational models (Supplementary Figure S10). This analysis emphasizes that the sites predicted in the defined enrichment zones are likely to correspond to direct TF–DNA interactions.

Predicted direct TF–DNA interactions are found in high confidence ChIP-seq peaks. We hypothesized that the ChIP-seq signal at ChIP-seq peaks containing a predicted direct TF–DNA interaction were more likely to be higher than at the other peaks. To test this hypothesis, we looked at (i) the quality of the peaks based on *P*-values assigned to the peaks by the MACS2 peak-caller and (ii) the reproducibility of calling these peaks with multiple peak-callers (MACS2, HOMER, and BCP; see Materials and Methods).

We observed that the distribution of *P*-values assigned by MACS2 to the peaks containing a predicted TFBS were significantly (*P*-value < 0.01; Mann–Whitney *U* test) lower than for the rest of the peaks for 1862 (96%) data sets (Fig-

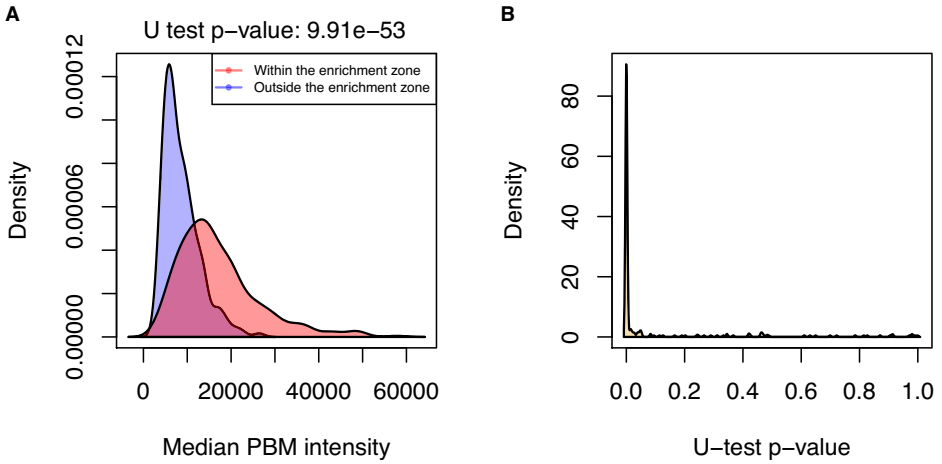

Figure 3. Binding affinity assessment for the predicted direct TF–DNA interactions. (**A**) Distribution of the median PBM intensity scores for the ENCSR000BMX GATA3 ChIP-seq data set between sequences at TFBSs (i.e. sites within the enrichment zone; in red) and sites outside the enrichment zone (in blue). (**B**) Distribution of Mann–Whitney *U* test *P*-values across the 249 data sets, showing distinct distributions of PBM intensity scores between sites within and outside the enrichment zones.

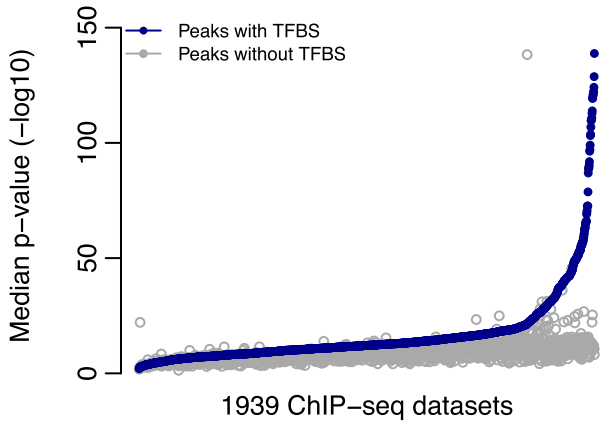

Figure 4. Quality assessment of the ChIP-seq peaks derived from direct TF–DNA interactions. Distribution of the median MACS2 *P*-values (y-axis) across all data sets. Values for peaks containing a predicted TFBS are provided in blue and values for the other peaks in grey. 1939 ChIP-seq data sets were predicted to contain direct TF–DNA interactions (x-axis).

ure 4). The other 77 data sets contained a reduced number of peaks (median of 837 compared to 18 968 for the complete set of ChIP-seq data sets), which can explain the lack of statistical significance. These results confirm that the predictions of direct TF–DNA interactions were found in ChIP-seq peaks of higher quality as assessed by MACS2.

To test ChIP-seq peak-calling reproducibility, we used two other peak-callers (HOMER and BCP) on 670 ChIP-seq data sets from ENCODE. Our choice of peak-callers was motivated by their distinct statistical approaches for peak prediction. While MACS2 and HOMER are based on an empirical model supported by a Poisson distribution, BCP uses a Bayesian approach implementing infinite-state hidden Markov models. We applied ChIP-eat to the ChIP-seq peaks to predict TFBSs. For each pair of peak-callers, we assessed whether the peaks predicted to contain a direct TF–DNA interaction were more prevalent (*P*-value < 0.01, hypergeometric test) in the set of peaks called by both peak-callers. This was observed for 63% of the data sets for MACS2 and BCP, 70% for MACS2 and HOMER, and 66% for HOMER and BCP. The data sets without significant enrichment had a median number of peaks predicted to be derived from direct TF–DNA interactions that was ∼7 fold smaller (e.g. 3358 compared to 22 499 between MACS2 and BCP) than for the data sets with significant enrichment, and a median number of peaks without TFBS ∼2 fold larger (e.g. 40 050 compared to 21 256 between MACS2 and BCP) (Supplementary Table S3). Moreover, the median quality scores assigned by the peak-callers to the peaks from the enriched data sets were significantly (*P*-value < 0.01, Mann–Whitney *U* test) higher than for the peaks in the other data sets (Supplementary Figure S11). It suggests that the data sets enriched for reproducible peaks containing predicted direct TF–DNA interactions are of better quality than the rest of the data sets.

Taken together, these results highlight that the ChIP-seq peaks in which ChIP-eat predicts direct TF–DNA interactions are of higher quality than the other peaks. Note that the ChIP-eat tool does not consider the peak quality when predicting direct TF–DNA interactions. These observations reinforce the confidence in the predicted TFBSs by ChIP-eat.

Predictions of direct TF–DNA interactions in ChIP-exo data

The ChIP-exo assay has been developed to provide a higher resolution than ChIP-seq to identify TFBSs *in vivo* (34). We aimed at assessing the performance of ChIP-eat on predicting direct TF–DNA interactions using ChIP-exo data. The ChExMix tool has recently been introduced to characterize protein-DNA binding event subtypes from ChIP-exo peak (48). ChExMix predicted different binding event subtypes for ChIP-exo data obtained for the TFs ESR1 and FOXA1, one of these subtypes corresponding to direct TF–DNA interactions (48). We applied ChIP-eat on the same ESR1 and FOXA1 ChIP-exo data sets. We compared the set of peaks identified to contain direct TF–DNA interactions

predicted by ChExMix and ChIP-eat in these two data sets. We found that 93.6% (for ESR1) and 91.3% (for FOXA1) of the peaks predicted to contain TFBSs by ChIP-eat were also predicted as direct binding events by ChExMix (Supplementary Table S4). The high overlaps between the predictions from ChExMix and ChIP-eat were confirmed by Jaccard similarity indexes of 63.7% and 68.7% for ESR1 and FOXA1, respectively. The similar results obtained with the two tools suggest that ChIP-eat, designed for the more noisy and less precise ChIP-seq data, is able to capture direct binding events from ChIP-exo data.

High-occupancy target regions are likely not derived from direct TF–DNA interactions

High-occupancy target (HOT) and extreme-occupancy target (XOT) regions are genomic regions where ChIP-seq peaks were observed for a large number of distinct ChIP'ed TFs (35,62,63). These regions are observed across species (63) and contain an unusually high frequency of ChIP-seq peaks (35,62,63). We used our set of high quality TFBS predictions to confirm that HOT/XOT regions were depleted of direct TF–DNA interactions. Indeed, we found that ChIP-seq peaks that do not contain a predicted TFBS were significantly enriched at HOT/XOT regions (odds ratio = 1.43 for HOT and 1.44 for XOT, *P*-value < 2.2e–16, hypergeometric test, Supplementary Table S5). Similar results were obtained when considering the three other computational models (BEM, TFFM, and DNAshapedTFBSs; Supplementary Table S5). This observation, combined with a previous study describing that HOT/XOT regions are likely to be derived from ChIP-seq artifacts (Wreczycka *et al.*, bioRxiv, 10.1101/107680), suggests that HOT/XOT regions are not derived from the direct binding of the ChIP'ed TFs.

Predicted direct TF–DNA interactions reveal co-binding TFs and cis-regulatory modules enriched for disease- and trait-associated SNPs

TFs are known to collaborate through specific co-binding at *cis*-regulatory modules (CRMs) to achieve their function (1,36). Hence, identifying co-binding TFs is critical to decipher transcriptional regulation of gene expression. We aimed at using our predicted direct TF–DNA interactions to reveal co-binding TFs and CRMs. We hypothesized that the distances between TFBSs of cooperating TFs are smaller than expected by chance. We tested this hypothesis for all pairs of TFs for which we predicted TFBSs (232 TFs, 53 592 pairs tested; see Materials and Methods). For each TF pair, we used a conservative Monte Carlo-based approach to compare the geometric mean of the distances between their TFBSs to the geometric mean distance expected by chance for a similar number of TFBSs randomly selected from the complete pool of TFBSs (see Materials and Methods). This approach predicted 150 pairs of TFs (accounting for 112 distinct TFs) with TFBSs closer in the genome than expected by chance (FDR < 5%; Supplementary Table S6). For 82% of the predicted TF pairs, we confirmed that the corresponding TFs physically interact using the protein-protein interaction networks from the GeneMANIA tool (54) (Supplementary Figure S12). This analysis further supports the biological relevance of the TFBSs predicted by ChIP-eat.

Next, we aimed to automatically identify CRMs, which correspond to clusters of direct TF–DNA interactions, using the clustering of genomic regions analysis method (CREAM; (Madani Tonekaboni *et al.*, bioRxiv, doi:10.1101/222562)). When considering our complete set of TFBSs, CREAM detected 61 934 CRMs in the human genome, encompassing 2 474 587 distinct TFBS locations. We found that the predicted CRMs were significantly enriched (FDR-corrected *P*-value = $2.9e^{-150}$) for disease- and trait-associated SNPs using traseR (55). Further, we observed that the TFBSs lying within the CRMs were more conserved than the TFBSs predicted outside (Supplementary Figure S13). Taken together, these results indicate a potentially functional role of the CRMs identified as clusters of direct TF–DNA interactions.

The UniBind web interface to access our collection of direct TF–DNA interactions

We catalogued the complete set of TFBS predictions from each prediction model, trained models, original ChIP-seq peaks from ReMap, and computed CRMs, and made them publicly available through UniBind at http://unibind.uio.no/. UniBind provides an interactive web interface with easy browsing, searching, and downloading for all our predictions (Figure 5). For instance, users can search for predictions for specific TFs, cell lines, and conditions.

The data can be searched by using the case insensitive search option available on the homepage. The database can be searched for each of the four TF binding models, cell/tissue type, and TF name using the 'Advanced Options', available on the homepage (Figure 5A). Search results are presented in a responsive and paginated table along with metadata information (Figure 5B), which can be clicked to view the detailed information and download TFBSs, summary plots, and ReMap ChIP-seq peaks (Figure 5C-D). All the metadata in the responsive tables can be downloaded as CSV files. UniBind displays by default the results obtained with the DiMO-optimized PWMs, but results obtained from all TFBS computational models along with the trained models are available for browsing and/or download.

DISCUSSION

To summarize, we have uniformly processed 1983 ChIP-seq peak data sets to predict high quality direct TF–DNA binding interactions in the human genome. The predictions were obtained using a non-parametric, entropy-based algorithm that automatically detects thresholds for TFBS computational model scores and distances to peak summits for each ChIP-seq data set. This new approach identified TFBSs supported by strong experimental and computational evidences for direct TF–DNA interactions. The accuracy of the predictions was *a posteriori* validated using the PBM *in vitro* assay, ChIP-exo data, and multiple ChIP-seq peak-calling algorithms. Our set of direct TF–DNA interactions confirmed that HOT genomic regions are likely not derived from direct binding of the TFs to the DNA. We used

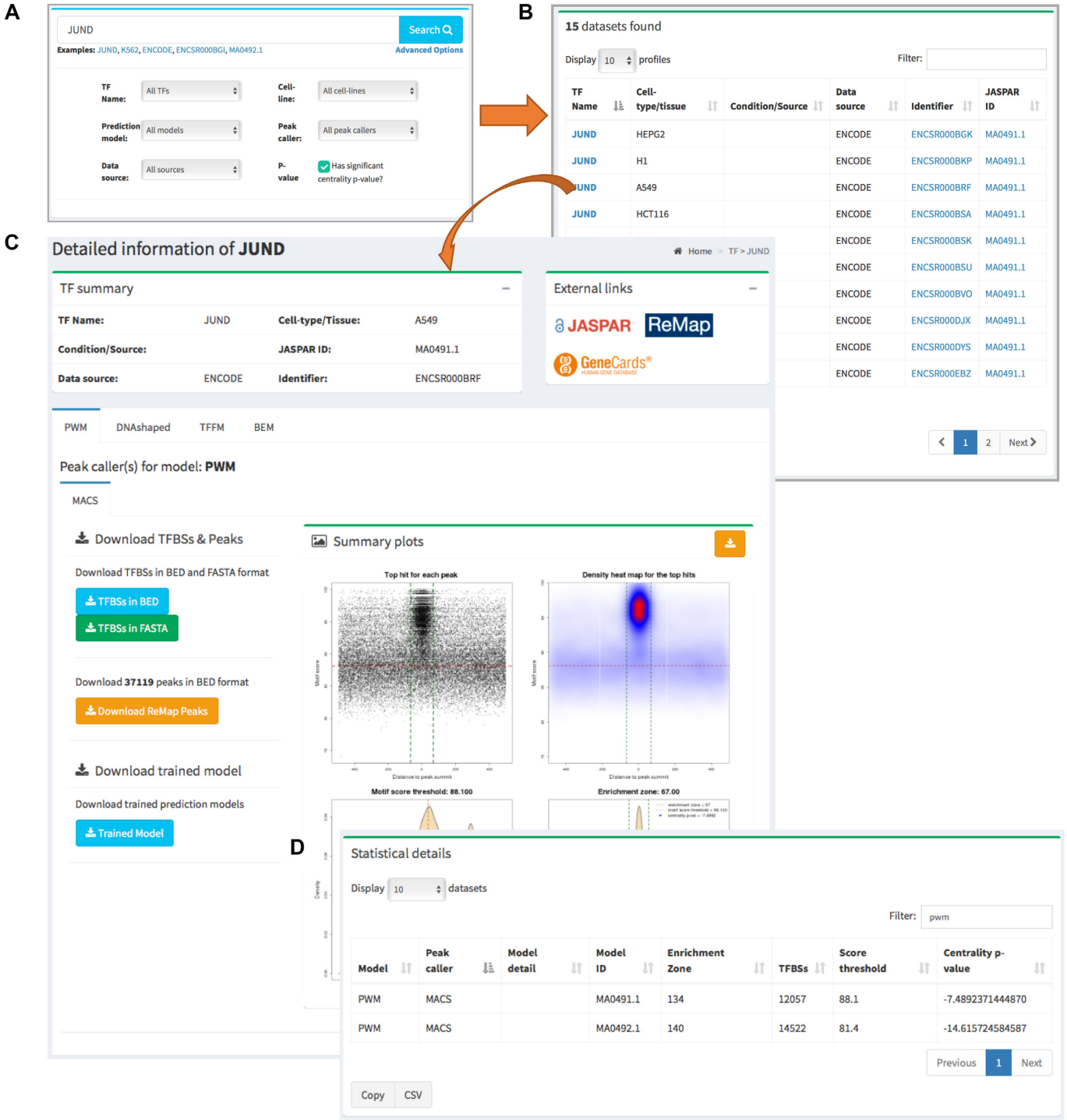

Figure 5. Overview of the UniBind user interface with interactive searching activity. (**A**) A quick and detailed search feature on the homepage. (**B**) A responsive table lists the searched data set(s), which can be clicked to view the details. (**C**) A detailed page shows the analysis for the JUND TF in cell-line A549, which is divided into sub-panels including the TF summary, external links, summary plots, and download options for each computational TFBS model. (**D**) Statistical details of the results.

our TFBSs to predict TFs with proximal binding events in the human genome, which could cooperate to achieve specific functions. Further, we defined *cis*-regulatory modules, which are clusters of TFBSs, that were enriched for disease- and trait-associated SNPs from GWAS. The complete set of predictions is publicly and freely available through the UniBind web-interface (http://unibind.uio.no/), in an effort to provide the community with an unprecedented collection of high quality direct TF–DNA interaction events in the human genome.

The output of ChIP-seq assays is generally composed of direct protein-DNA interactions, indirect binding of the protein to the DNA (through a co-binding partner), nonspecific protein binding to the DNA, and noise/bias/artifacts (4–6). Here, we specifically aimed at identifying direct TF–DNA interaction events by using an entropy-based algorithm (41). This algorithm was originally developed to discriminate between foreground and background in image processing. Hence, it assumes the presence of background (or noise) in the data. As a consequence, our approach is limited by the assumption that there is background/noise in the ChIP-seq data sets analyzed. We assume that this noise represents indirect binding of TFs, nonspecific binding, or ChIP-seq experimental artifacts. Moreover, our approach considered the best site per ChIP-seq peak (defined using TFBS computational models), which represents the best candidate. We recognize that other sites with lower scores could represent direct TF–DNA interactions. These limitations denote that our approach is stringent for the prediction of direct TF–DNA interactions, favoring specificity over sensitivity. The ChIP-seq peaks that our method did not predict to contain direct TF–DNA binding events could be further analyzed to discriminate other mechanisms for protein-DNA interactions from background noise, as proposed in the ChExMix tool established for ChIP-exo data (48).

The ChIP-eat pipeline developed for this study used four TFBS computational models to predict TF–DNA binding events. These models were specifically trained for each ChIP-seq data set to improve the quality of the predictions, as the best-performing computational model varies for different TFs or TF families (8,14,15). As a consequence, we advocate that a 'one-fits-all' TFBS prediction model is not optimal and that one should compare results from multiple models. With the predictions available through UniBind, users can assess which model would perform better for each data set. Of course, it requires to use a specific metric to compare performance. As our methods aimed at identifying enrichment zones centered around ChIP-seq peak summits, we suggest to rely on a centrality measure as implemented in the CentriMo method (27). In UniBind, we provide centrality *P*-values computed following (27) for the predictions from each model in each ChIP-seq data set. Moreover, the ChIP-eat pipeline is generalizable and users can incorporate other TFBS computational models to predict direct TF–DNA interactions and compare them to the ones already stored in UniBind.

While studies alike focus on determining where TFs directly interact with DNA, our understanding of how these TF–DNA interactions influence expression is limited. Surely, it is critical to decipher the relationship between TF–DNA interactions and transcriptional regulation (64). It is expected that a large portion of the TFBSs identified in our study are not functional, as suggested by the futility theorem (36). Nevertheless, functional TF binding events are likely to be clustered (65–68) and associated with stronger ChIP-seq peak signals (12,69). We expect that the direct TF–DNA interactions predicted in *cis*-regulatory modules and stored in UniBind are more likely to be enriched for functional events. Determining the specific set of functional TF–DNA interactions would require dedicated computational models and experiments.

DATA AVAILABILITY

Source code of the ChIP-eat software is available at https://bitbucket.org/CBGR/chip-eat and of UniBind at https://bitbucket.org/CBGR/unibind. The source code used for the identification of co-localized TFs is available at https://bitbucket.org/CBGR/co-binding. Users can browse and/or download the data through the UniBind web interface at http://unibind.uio.no/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

As research parasites (70), we would like to thank all the researchers who deposited their data. We thank Georgios Magklaras and his team for systems support, Manuela Zucknick and Andrea Cremaschi for statistical insights, Elisa Bjørgo and Ingrid Kjelsvik for management support, and Roza Berhanu Lemma, Jaime Castro-Mondragon, Oriol Fornes and Phillip Richmond for comments on the manuscript draft.

FUNDING

Norwegian Research Council (project #187615), Helse Sør-Øst, and the University of Oslo through the Centre for Molecular Medicine Norway (NCMM) (to A.M., A.K., M.G.); Ph.D. fellowship from the French Ministry of Higher Education and Research (to J.C.). Funding for open access charge: Norges Forskningsråd.
Conflict of interest statement. None declared.

REFERENCES

1. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The human transcription factors. *Cell*, **172**, 650–665.
2. Mathelier,A., Shi,W. and Wasserman,W.W. (2015) Identification of altered cis-regulatory elements in human disease. *Trends Genet.*, **31**, 67–76.
3. 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.
4. Teytelman,L., Thurtle,D.M., Rine,J. and van Oudenaarden,A. (2013) Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 18602–18607.
5. Jain,D., Baldi,S., Zabel,A., Straub,T. and Becker,P.B. (2015) Active promoters give rise to false positive 'Phantom Peaks' in ChIP-seq experiments. *Nucleic Acids Res.*, **43**, 6959–6968.

6. Worsley Hunt,R. and Wasserman,W.W. (2014) Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. *Genome Biol.*, **15**, 412.
7. Stormo,G.D. (2013) Modeling the specificity of protein-DNA interactions. *Quant Biol.*, **1**, 115–130.
8. Weirauch,M.T., Cote,A., Norel,R., Annala,M., Zhao,Y., Riley,T.R., Saez-Rodriguez,J., Cokelaer,T., Vedenko,A., Talukder,S. *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–134.
9. Kulakovskiy,I., Levitsky,V., Oshchepkov,D., Bryzgalov,L., Vorontsov,I. and Makeev,V. (2013) From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. *J. Bioinform. Comput. Biol.*, **11**, 1340004.
10. Eggeling,R., Roos,T., Myllymäki,P. and Grosse,I. (2015) Inferring intra-motif dependencies of DNA binding sites from ChIP-seq data. *BMC Bioinformatics*, **16**, 375.
11. Siebert,M. and Söding,J. (2016) Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.*, **44**, 6055–6069.
12. Slattery,M., Zhou,T., Yang,L., Dantas Machado,A.C., Gordân,R. and Rohs,R. (2014) Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.*, **39**, 381–399.
13. Keilwagen,J. and Grau,J. (2015) Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Res.*, **43**, e119.
14. Yang,L., Orenstein,Y., Jolma,A., Yin,Y., Taipale,J., Shamir,R. and Rohs,R. (2017) Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Mol. Syst. Biol.*, **13**, 910.
15. Mathelier,A., Xin,B., Chiu,T.-P., Yang,L., Rohs,R. and Wasserman,W.W. (2016) DNA shape features improve transcription factor binding site predictions in vivo. *Cell Syst.*, **3**, 278–286.
16. Chèneby,J., Gheorghe,M., Artufel,M., Mathelier,A. and Ballester,B. (2018) ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.*, **46**, D267–D275.
17. Yevshin,I., Sharipov,R., Valeev,T., Kel,A. and Kolpakov,F. (2017) GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res.*, **45**, D61–D67.
18. Zhou,K.-R., Liu,S., Sun,W.-J., Zheng,L.-L., Zhou,H., Yang,J.-H. and Qu,L.-H. (2017) ChIPBase v2.0: decoding transcriptional regulatory networks of non-coding RNAs and protein-coding genes from ChIP-seq data. *Nucleic Acids Res.*, **45**, D43–D50.
19. Mei,S., Qin,Q., Wu,Q., Sun,H., Zheng,R., Zang,C., Zhu,M., Wu,J., Shi,X., Taing,L. *et al.* (2017) Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.*, **45**, D658–D662.
20. Hinrichs,A.S., Karolchik,D., Baertsch,R., Barber,G.P., Bejerano,G., Clawson,H., Diekhans,M., Furey,T.S., Harte,R.A., Hsu,F. *et al.* (2006) The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.*, **34**, D590–D598.
21. Montgomery,S.B., Griffith,O.L., Sleumer,M.C., Bergman,C.M., Bilenky,M., Pleasance,E.D., Prychyna,Y., Zhang,X. and Jones,S.J.M. (2006) ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics*, **22**, 637–640.
22. Kent,W.J. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
23. Fornes,O., Gheorghe,M., Richmond,P.A., Arenillas,D.J., Wasserman,W.W. and Mathelier,A. (2018) MANTA2, update of the Mongo database for the analysis of transcription factor binding site alterations. *Sci Data*, **5**, 180141.
24. Khan,A., Fornes,O., Stigliani,A., Gheorghe,M., Castro-Mondragon,J.A., van der Lee,R., Bessy,A., Chèneby,J., Kulkarni,S.R., Tan,G. *et al.* (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46**, D1284.
25. Worsley Hunt,R., Mathelier,A., Del Peso,L. and Wasserman,W.W. (2014) Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment. *BMC Genomics*, **15**, 472.
26. Guo,Y., Mahony,S. and Gifford,D.K. (2012) High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput. Biol.*, **8**, e1002638.
27. Bailey,T.L. and Machanick,P. (2012) Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.*, **40**, e128.
28. Kulakovskiy,I.V., Boeva,V.A., Favorov,A.V. and Makeev,V.J. (2010) Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics*, **26**, 2622–2623.
29. Jothi,R., Cuddapah,S., Barski,A., Cui,K. and Zhao,K. (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.*, **36**, 5221–5231.
30. Wilbanks,E.G. and Facciotti,M.T. (2010) Evaluation of algorithm performance in ChIP-Seq peak detection. *PLoS One*, **5**, e11471.
31. Mathelier,A. and Wasserman,W.W. (2013) The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.*, **9**, e1003214.
32. Zhao,Y., Ruan,S., Pandey,M. and Stormo,G.D. (2012) Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics*, **191**, 781–790.
33. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. 3rd and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.
34. Rhee,H.S. and Pugh,B.F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, **147**, 1408–1419.
35. Yip,K.Y., Cheng,C., Bhardwaj,N., Brown,J.B., Leng,J., Kundaje,A., Rozowsky,J., Birney,E., Bickel,P., Snyder,M. *et al.* (2012) Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.*, **13**, R48.
36. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.
37. Patel,R.Y. and Stormo,G.D. (2014) Discriminative motif optimization based on perceptron training. *Bioinformatics*, **30**, 941–948.
38. Chiu,T.-P., Yang,L., Zhou,T., Main,B.J., Parker,S.C.J., Nuzhdin,S.V., Tullius,T.D. and Rohs,R. (2015) GBshape: a genome browser database for DNA shape annotations. *Nucleic Acids Res.*, **43**, D103–D109.
39. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
40. Venables,W.N. and Ripley,B.D. (2002) *Modern Applied Statistics with S Springer*, NY.
41. Kapur,J.N., Sahoo,P.K. and Wong,A.K.C. (1985) A new method for gray-level picture thresholding using the entropy of the histogram. *Comput. Vis. Graph. Image Process.*, **29**, 140.
42. Shannon,C.E. (1948) A Mathematical Theory of Communication. *Bell Syst. Tech. J.*, **27**, 623–656.
43. Schneider,C.A., Rasband,W.S. and Eliceiri,K.W. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods*, **9**, 671–675.
44. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
45. Bulyk,M.L., Gentalen,E., Lockhart,D.J. and Church,G.M. (1999) Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nat. Biotechnol.*, **17**, 573–577.
46. Hume,M.A., Barrera,L.A., Gisselbrecht,S.S. and Bulyk,M.L. (2015) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **43**, D117–D122.
47. Mann,H.B. and Whitney,D.R. (1947) On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.*, **18**, 50–60.
48. Yamada,N., Lai,W.K.M., Farrell,N., Pugh,B.F. and Mahony,S. (2018) Characterizing protein-DNA binding event subtypes in ChIP-exo data. *Bioinformatics*, doi:10.1093/bioinformatics/bty703.
49. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell*, **38**, 576–589.
50. Xing,H., Mo,Y., Liao,W. and Zhang,M.Q. (2012) Genome-wide localization of protein-DNA binding and histone modification by a

Bayesian change-point method with ChIP-seq data. *PLoS Comput. Biol.*, **8**, e1002613.
51. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
52. Hochberg,Y. and Benjamini,Y. (1990) More powerful procedures for multiple significance testing. *Stat. Med.*, **9**, 811–818.
53. Afgan,E., Baker,D., Batut,B., van den Beek,M., Bouvier,D., Cech,M., Chilton,J., Clements,D., Coraor,N., Grüning,B.A. *et al.* (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.*, **46**, W537–W544.
54. Warde-Farley,D., Donaldson,S.L., Comes,O., Zuberi,K., Badrawi,R., Chao,P., Franz,M., Grouios,C., Kazi,F., Lopes,C.T. *et al.* (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, **38**, W214–W20.
55. Chen,L. and Qin,Z.S. (2015) traseR: an R package for performing trait-associated SNP enrichment analysis in genomic intervals. *Bioinformatics*, **32**, 1214–1216.
56. Mailman,M.D., Feolo,M., Jin,Y., Kimura,M., Tryka,K., Bagoutdinov,R., Hao,L., Kiang,A., Paschall,J., Phan,L. *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.*, **39**, 1181–1186.
57. Welter,D., MacArthur,J., Morales,J., Burdett,T., Hall,P., Junkins,H., Klemm,A., Flicek,P., Manolio,T., Hindorff,L. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.
58. Siepel,A. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
59. Neph,S., Kuehn,M.S., Reynolds,A.P., Haugen,E., Thurman,R.E., Johnson,A.K., Rynes,E., Maurano,M.T., Vierstra,J., Thomas,S. *et al.* (2012) BEDOPS: high-performance genomic feature operations. *Bioinformatics*, **28**, 1919–1920.
60. Pohl,A. and Beato,M. (2014) bwtool: a tool for bigWig files. *Bioinformatics*, **30**, 1618–1619.
61. Berger,M.F. and Bulyk,M.L. (2006) Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *Methods Mol. Biol.*, **338**, 245–260.
62. Xie,D., Boyle,A.P., Wu,L., Zhai,J., Kawli,T. and Snyder,M. (2013) Dynamic trans-Acting factor colocalization in human cells. *Cell*, **155**, 713–724.
63. Boyle,A.P., Araya,C.L., Brdlik,C., Cayting,P., Cheng,C., Cheng,Y., Gardner,K., Hillier,L.W., Janette,J., Jiang,L. *et al.* (2014) Comparative analysis of regulatory information and circuits across distant species. *Nature*, **512**, 453–456.
64. Whitfield,T.W., Wang,J., Collins,P.J., Christopher Partridge,E., Aldred,S., Trinklein,N.D., Myers,R.M. and Weng,Z. (2012) Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.*, **13**, R50.
65. Hnisz,D., Abraham,B.J., Lee,T.I., Lau,A., Saint-André,V., Sigova,A.A., Hoke,H.A. and Young,R.A. (2013) Super-enhancers in the control of cell identity and disease. *Cell*, **155**, 934–947.
66. Wilczyński,B. and Furlong,E.E.M. (2010) Dynamic CRM occupancy reflects a temporal map of developmental progression. *Mol. Syst. Biol.*, **6**, 383.
67. Whyte,W.A., Orlando,D.A., Hnisz,D., Abraham,B.J., Lin,C.Y., Kagey,M.H., Rahl,P.B., Lee,T.I. and Young,R.A. (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**, 307–319.
68. He,Q., Bardet,A.F., Patton,B., Purvis,J., Johnston,J., Paulson,A., Gogol,M., Stark,A. and Zeitlinger,J. (2011) High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species. *Nat. Genet.*, **43**, 414–420.
69. Fisher,W.W., Li,J.J., Hammonds,A.S., Brown,J.B., Pfeiffer,B.D., Weiszmann,R., MacArthur,S., Thomas,S., Stamatoyannopoulos,J.A., Eisen,M.B. *et al.* (2012) DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 21330–21335.
70. Longo,D.L. and Drazen,J.M. (2016) Data sharing. *N. Engl. J. Med.*, **374**, 276–277.