**METHODS**

**Data.** DNase-seq aligned reads were obtained from ENCODE1. To perform the computational footprint experiments, we obtained data regarding cell types H1-hESC, HeLa-S3, HepG2, Huvec, K562, LNCaP and MCF-7 from Crawford's Lab (labeled with the initials of their institution “DU”) and cell types H7-hESC, HepG2, Huvec, K562 and m3134 from Stamatoyannopoulous' lab (labeled with the initials of their institution “UW”). We also used naked DNA (deproteinized) DNase-seq experiments from cell types MCF-7 and K562 (DU)12 and IMR90 (UW)26. DNase-seq experiments labeled with “DU” follow the single-hit protocol, while the experiments labeled with “UW” follow the double-hit protocol. In addition, to perform the DNase-seq bias estimation clustering, we used all cell types from ENCODE's Tier 1 and Tier 2 cell types1. See **Supplementary Table 2** for a full DNase-seq data description.

Transcription factor (TF) ChIP-seq enriched regions (peaks and summits) were obtained in ENCODE analysis working group (AWG)1 track with exception of the following experiments, in which the enriched regions were obtained using bowtie-227 and MACS28. AR (R1881 treatment) ChIP-seq raw sequences for LNCaP cell type was obtained in gene expression omnibus (GEO) with accession number GSM35364429. ER (40 and 160 minutes after estradiol treatment) ChIP-seq raw sequences for MCF-7 cell type was obtained in GEO with accession number GSE5485530. GR (dexamethasone treatment) ChIP-seq raw sequences for m3134 cell type was obtained in the sequence read archive (SRA) under study number SRP00487131. All organism-specific data (DNase-seq and ChIP-seq) are based on the human genome build 37 (hg19), except the DNase-seq for m3134 and ChIP-seq for GR, which were based on mouse genome build 37 (mm9). Chromosome Y was removed from all analyses. Expression of cells H1-hESC, K562 and GM12878 were obtained from ENCODE (GSE12760 and GSE14863)1.

TF motifs (position frequency matrices; PFMs) were obtained from the Jaspar32, Uniprobe33 and Transfac34 repositories. Non-organism-specific data (PFMs) were obtained for the subphylum Vertebrata. *De novo* PFMs 0458 and 0500 were downloaded from ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/integration\_data\_jan2011/byDataType/footprints/jan2011/de.novo.pwm4. The accession codes for all TF ChIP-seq experiments and PFM IDs are available in the **Supplementary Datasets 1a and 2b-d**.

**Sequence bias correction.** *DNase I hypersensitivity sites.* A first task is the identification of DNase I hypersensitivity sites (DHSs). A nucleotide-resolution genome-wide signal was created for each DNase-seq data set by counting reads mapped to the genome. Here, we considered only the 5' position of the aligned reads (position at which DNase I cleaved the DNA). The genomic signal was created by counting the number of reads that overlapped at each genomic position.

More formally, we define a raw genomic signal as a vector

,

where equals the number of bases in the genome and each is the number of DNase-seq reads in which the 5′ position mapped to position . We also generate strand specific counts , where describes the strand the read was mapped to.

DHSs are estimated based on the DNase I raw signal. First, the F-seq software35 was used to create smoothed DNase-seq signals using Parzen density estimates. Then, the smoothed signal was fit to a gamma distribution,

,

by evaluating and based on mean and standard deviation estimates. Finally, the enriched regions (DHSs) were found by establishing a cutoff based on a *p*-value of 0.011,35. We refer to DHSs as a set of genomic intervals

,

where for and is the total number of DHSs. We ignore for simplicity of notation the fact that intervals are defined on distinct chromosomes or contigs.

*Estimation of DNase-seq sequence bias*. We use two approaches to estimate sequence bias of DNase-seq experiments: (1) aligned reads inside DHSs from DNase-seq experiments (termed “DHS sequence bias”) following He et al.15 and (2) all aligned reads for naked DNA experiments (termed “naked DNA sequence bias”) following Yardımcı et al.12. The observed cleavage score for a *k*-mer corresponds to the number of DNase I cleavage sites centered at . The background cleavage score is defined by the total number of times occurs. Then, the bias estimation is computed as the ratio between the observed and background cleavage scores. Mathematical formalizations of the bias estimation will be made based on the DHS sequence bias approach.

We define as the reference genome sequence with length for strand . indicates the sequence from positions to (including both within the interval). For each *k*-mer with length *k* the observed cleavage score can be calculated as

,

where is an indicator function.

Similarly, the background cleavage score can be evaluated as

.

Finally, the cleavage bias for a genomic position , given that , can be calculated as

,

where indicates the total number of reads aligned to strand in DHSs

,

and indicates the total number of *k*-mers in DHS positions

.

The bias score represents how many times the *k*-mer sequence was cleaved by the DNase I enzyme in comparison to its total occurrence in: (1) DHSs (DHS sequence bias approach); (2) the entire genome (naked DNA sequence bias approach). As observed by He et al.15 a *6*-mer bias model captures more information than models and the information added with models are not significant. Therefore, in this study, all analyses were performed using a *6*-mer bias model.

*DNase-seq sequence bias correction.* A “smoothed corrected signal” was calculated using smoothed versions of both raw DNase-seq () and the bias score signal ()15. These smoothed signals were based on a 50 bp window and can be written as

.

With these results we are able to define the smoothed corrected signal as

.

Finally, the bias-corrected DNase-seq genomic signal () can be obtained by applying

. (1)

The corrected DNase-seq signal generated by **equation (1)** may include negative values. Since some posterior statistical analyses required a signal consisting only of positive values, we have shifted the entire signal by adding the global minimum value.

**Computational footprinting methods.** In this section we present an overview of the computational footprinting methods used in this study. Computational resources necessary to the execution of each method were summarized in **Supplementary Table 1**.

*Neph method.* Neph et al.4 used a simplified version of the segmentation method originally proposed in Hesselberth et al.36. Their method consists on applying a sliding window to find genomic regions (6-40 bp) with low DNase I cleavage activity between regions (3-10 bp) with intense DNase I digestion. The footprint score (FS) is evaluated and used to determine the most significant predictions.

We obtained the footprint predictions for cell type K562 (DU) in ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/integration\_data\_jan2011/byDataType/footprints/jan2011/all.footprints.gz4. As predictions were not available for other DNase-seq experiments, we obtained the scripts and parameterization through Neph et al.4 footprinting method code repository at https://github.com/StamLab/footprinting2012. Briefly, we used the DNase I raw signal as input with the parameters from the original publication: flanking component length varied between 3-10 bp and central footprint region length varied between 6-40 bp. Afterwards, the footprints were filtered by an FDR of 1%, which was estimated based on the FS distribution in each cell type4. Finally, we consider only predictions that occurred within DNase-seq hotspots, evaluated using the method first described in Sabo et al.37. We obtained all hotspots generated by Stamatoyannopoulous' lab in ENCODE1 for cell types GM12878 (wgEncodeEH000492; GSM736496 and GSM736620), H1-hESC (wgEncodeEH000496; GSM736582) and K562 (wgEncodeEH000484; GSM736629 and GSM736566). We will refer to this framework as “Neph”.

*Boyle method.* Boyle et al.5 designed a segmentation approach, which is based on using hidden Markov models (HMMs) to predict footprints in specific DNase I cleavage patterns. Briefly, the HMM uses a normalized DNase-seq cleavage signal to find regions with depleted DNase I digestion (footprints) between two peaks of intense DNase I cleavage. Such pattern reflects the inability of the DNase I nuclease to cleave sites where there are proteins bound. As the DNase-seq profiles required a nucleotide-resolution signal, which is usually noisy, the authors used a Savitzky-Golay smoothing filter to reduce noise and to estimate the slope of the DNase-seq signal38. Their HMM had five states, with specific states to identify the decrease/increase of DHS signals around the peak-dip-peak region. Since no source code or software is provided, we used footprint predictions from Boyle et al.5 available at http://fureylab.web.unc.edu/datasets/footprints/. We will refer to this method as “Boyle”.

*Centipede.* Centipede is a site-centric approach, which gathers experimental and genomic information around motif-predicted binding sites (MPBSs). It then uses a Bayesian mixture model approach to label each retrieved site as 'bound' or 'unbound'9. The experimental and genomic data used include DNase-seq, position weight matrix (PWM) bit-score, sequence conservation and distance to the nearest transcription start site (TSS). The experimental data input was generated by fetching the raw DNase-seq signal surrounding a 200 bp window centered on each MPBS. Additionally, to create the genomic data input, we obtained PhastCons conservation score (placental mammals on the 46-way multiple alignment)39 and Ensembl gene annotation from ENCODE1,40 to create the prior probabilities in addition to the PWM bit-score.

Centipede software was obtained at http://centipede.uchicago.edu/ and executed to generate posterior probabilities of regions being bound by TFs. We have previously observed that Centipede is sensitive to certain parameters. Therefore, Centipede parameterization was defined with an extensive computational evaluation described in Gusmao et al.8.

*Cuellar Method.* Cuellar-Partida et al.10 proposed a site-centric method to include DNase-seq data as priors for the detection of active transcription factor binding sites (TFBSs). It is based on a probabilistic classification approach to compute better log-posterior odds score than the ones observed by purely sequence-based approaches. We applied this method as described in Cuellar-Partida et al.10. We created a smoothed DNase-seq input signal by evaluating the number of DNase-seq cleavage based on a 150 bp window with 20 bp steps. We obtained their scripts at http://research.imb.uq.edu.au/t.bailey/SD/Cuellar2011/ and created priors using the smoothed version of the DNase-seq signal. As suggested by the authors, the priors were submitted to the program FIMO41 to obtain the predictions. We will refer to this method as “Cuellar”.

*Wellington.* Wellington is a segmentation approach based on a Binomial test. For a given candidate footprint, it tests the hypothesis that there are more reads in the flanking regions than within the footprint. Following an observation that DNase-seq cuts of the double-hit protocol are strand-specific, Wellington only considers reads mapped to the upstream flanking region of the footprints. Wellington automatically detects the size of footprints (within a user-defined interval) and sets flanking regions at a user-defined length. We have obtained Wellington's source code in http://jpiper.github.com/pyDNase and executed it with default parameters. Briefly, we used a footprint FDR cutoff of –30, footprint sizes varying between 6 and 40 with 1 bp steps and shoulder size (flanking regions) of 35 bp.

*Protein interaction quantification (PIQ).* The protein interaction quantification (PIQ) is a site-centric method, which uses Gaussian process to model and smooth the footprint profiles around candidate MPBSs (± 100 bp)11. Active footprints are estimated with an expectation propagation algorithm. Finally, PIQ indicates the set of motifs which footprint signals are distinguishable from noise to reduce the set of candidate TFs. We obtained PIQ implementation in http://piq.csail.mit.edu and executed it with default parameters, which can be found in the script *common.r*. Briefly, MPBSs were generated with the script *pwmmatch.exact.r*. The DNase-seq signal was created using the script *bam2rdata.r*. And the footprints were detected with the script *pertf.r*.

*Footprint mixture (FLR).* Yardımcı et al.12 proposed a site-centric method based on a mixture of multinomial models to detect active/inactive MPBSs. The method uses an expectation maximization algorithm to find a mixture of two multinomial distributions, representing active (footprints) and inactive (background) MPBSs. The background model is initialized with either naked DNA sequence bias frequencies or estimated *de novo*. After successful estimation, MPBSs are scored with the log odds ratio for the footprint *versus* background model. The model takes DNase-seq cuts within a small window around the candidate profiles (± 25 bp) as input. DNase-seq sequence bias is estimated for *6*-mers based on the DNA sequences extracted within the same regions in which the cuts were retrieved. Method implementation was obtained in https://ohlerlab.mdc-berlin.de/software/FootprintMixture\_109/. We executed the method using naked DNA sequence bias frequencies for initialization of the background models. The width of the window surrounding the TFBS (*PadLen*) was set to the default value of 25 bp. Also, we use the expectation maximization to re-estimate background during training (argument *Fixed* set to *FALSE*). We will refer to this method as “FLR”.

*DNase2TF.* DNase2TF is a segmentation approach based on a binomial *z*-score, which evaluates the depletion of DNase-seq reads around the candidate footprints7. At a second step, DNase2TF interactively merges close candidate footprints whenever they improve depletion scores. DNase2TF corrects for DNase I sequence bias using cleavage statistics for *2*- or *4*-mers. We obtained source code from http://sourceforge.net/projects/dnase2tfr/ and executed DNase2TF with a *4*-mer sequence bias correction. Other parameters were set to their default values: *minw = 6*, *maxw = 30*, *z\_threshold = -2* and *FDR = 10-3*.

*HINT, HINT-BC and HINT-BCN.* Recently, Gusmao et al.8 have proposed the segmentation method HINT (HMM-based identification of transcription factor footprints) as an extension of Boyle method5. HINT is based on eight-state multivariate HMMs and combines DNase-seq and histone modification ChIP-seq profiles at the nucleotide level for the identification of footprints. The pipeline of HINT method starts by normalizing the DNase I cleavage signal using within- and between-dataset normalizations. Then, the slope of the normalized signals is evaluated to identify the DNase-seq signal increase and decrease. Afterwards, an HMM is trained on a supervised manner (maximum likelihood) based on a single manually annotated genomic region. To aid such manual annotation the normalized and slope signals are used in combination with MPBSs for all available PFMs in the repositories Jaspar32 and Uniprobe33. Finally, the Viterbi algorithm is performed on the trained HMMs inside regions consisting of DHSs extended by 5,000 bp upstream and downstream. All parameters were set as described in Gusmao et al.8.

We have performed two modifications to the method described in Gusmao et al.8. First, to perform a standardized comparison, we modified HINT to allow only DNase-seq data. The modified HMM model contains five states. The three histone-level states were removed and new transitions were created from the *BACKGROUND* state to the *DNase UP* state and from the *DNase DOWN* state to the *BACKGROUND* state. The second modification concerns the use of bias-corrected DNase-seq signal prior to normalization steps. We will call the method HINT bias-corrected (HINT-BC), for correction based on “DHS sequence bias”, and HINT bias-corrected on naked DNA (HINT-BCN) for the “naked DNA sequence bias” estimation. These modifications required retraining of the HMM models. For this, we used the same manual annotation described in Gusmao et al.8. The novel methods and trained models are available as a command-line tool at www.costalab.org/hint-bc.

*BinDNase.* BinDNase is a site-centric method based on logistic regression to predict active/inactive MBBSs13. The algorithm starts with base pair resolution DNase-seq signal around the MPBSs (± 100 bps) and selects discriminatory features using a backward greedy approach. As a supervised approach, the method requires positive and negative examples, which can be obtained from TF ChIP-seq data. We have used DNase-seq data around MPBSs on chromosome 1 for training. These MPBSs were subsequently removed from the evaluation procedure. The definition of positive and negative examples was the same as in our evaluation data sets. Note that this is the only method evaluated here which requires TF ChIP-seq examples for training. We also point the fact that BinDNase did not successfully executed for 19 TFs of our evaluation data set (POU5F1, REST, RFX5, SP1, SP2, SRF, TCF12 and ZNF143 binding in H1-hESC; ARID3A, CTCF, IRF1, MEF2A, PU1, REST, RFX5, SP1, SP2, STAT2 and ZNF263 binding in K562) given our maximum running time criteria (three weeks). Method implementation was obtained at http://research.ics.aalto.fi/csb/software/bindnase/ and required/provided no parameter selection.

*Footprint score rank (FS-Rank).* He et al.15 used a site-centric MPBS ranking scheme termed “footprint score (FS)”, which is based on a scoring metric from the footprinting methodology proposed in Neph et al.4. The FS statistic is defined as

,

where is the -th MPBS which extends from genomic positions to and . The FS uses the DNase-seq signal in the center () of the MPBS and its upstream () and downstream () flanking regions. These variables can be defined as

(2)

*Tag count rank (TC-Rank).* The site-centric method which we refer to as “tag count (TC)”, corresponds to the number of DNase I cleavage hits in a 200 bp window around predicted TFBS as defined in He et al.15. This can be written as

.

Both TC and FS can be used as quality scores for footprints. However, as a method (termed TC-Rank and FS-Rank) it consists on attributing these quality scores to each MPBS and evaluating the performance at these ranked MPBS. This observation also holds for the PWM-Rank method described below.

**Evaluation.** *Motif-predicted binding sites (MPBSs).* Method evaluation was performed with a site-centric binding site statistics. For this, we generated position weight matrices (PWMs) from PFMs by evaluating the information content of each position and performing background nucleotide frequency correction42. This was performed using Biopython43. Then, we created MPBSs by matching all PWMs against the human (hg19) and mouse (mm9) genomes using the fast performance motif matching tool MOODS44. This procedure produces “PWM bit-scores” for every match. We determined a bit-score cutoff threshold by applying the dynamic programming approach described in Wilczynski et al.45 with a false positive rate (FPR) of 10-4. All site-centric scores were based on the set of MPBSs after the application of the cutoff threshold. Also, the PWM bit-score was used as a baseline method and will be referenced as “PWM-Rank”.

*Method comparison.* Methods were evaluated using a site-centric approach10, which combines MPBSs with ChIP-seq data for every TF. In this scheme, MPBSs with ChIP-seq evidence (located within 100 bp from the ChIP-seq peak summit) are considered “true” TFBSs; while MPBSs without ChIP-seq evidence are considered “false” TFBSs. Every TF prediction that overlaps a true TFBS is considered a correct prediction (true positive; TP) and every prediction that overlaps with a false TFBS is considered an incorrect prediction (false positive; FP). Therefore, true negatives (TN) and false negatives (FN) are, respectively, false and true TFBSs without overlapping predictions. To assess the accuracy of digital genomic footprinting methods we created receiver operating characteristic (ROC) curves. Briefly, ROC curves describe the sensitivity (recall) increase as we decrease the specificity of the method. The area under the ROC curve (AUC) metric was evaluated at 100%, 10% and 1% false positive rates (FPRs). We also evaluated the area under the precision-recall curve (AUPR). This metric is indicated for problems with imbalanced data sets (distinct number of positive and negative examples)20,46.

Segmentation approaches (Boyle, DNase2TF, HINT, Neph and Wellington) provide footprint predictions that do not necessarily encompass all MPBSs. To create full ROC curves for these methods, we first ranked all predicted sites by their DNase I cleavage tag count followed by all non-predicted sites ranked by their tag count. In order to present a fair comparison, this approach was also applied to all site-centric methods (Centipede, Cuellar, FLR and PIQ). For that, we considered distinct probability thresholds of (0.8, 0.85, 0.9, 0.95, 0.99) for detection of footprints on all site-centric methods. We performed additional experiments to select the best threshold per method (see **Supplementary Fig. 8**).

Our TF ChIP-seq based comparative experiments comprise the following three evaluation scenarios. All evaluation statistics and method performances are available at the **Supplementary Dataset 1**.

*He Dataset:* To replicate the analysis performed by He et al.15, we analyzed DNase-seq from cell types K562 (UW), LNCaP (DU) and m3134 (UW) on 36 TFs and we evaluated the methods PWM, FS, TC, HINT, HINT-BC and HINT-BCN.

*Benchmarking Dataset:* For comparative analysis of several competing methods, we selected the two cell types with highest number of TF ChIP-seq data sets evaluated in our study: K562 (DU) with 59 TFs and H1hesc (DU) with 29 TFs. We can therefore make use of predictions provided by Gusmao et al.8 and Boyle et al.5, which includes evaluation of PWM, Boyle, Cuellar, Centipede, HINT and Neph methods. For this data set, we have estimated novel footprints for FS, TC, HINT-BC, HINT-BNC, DNase2TF, PIQ, Wellington and FLR methods, which were not previously evaluated.

*Comprehensive dataset:* Lastly, we have compiled a comprehensive data set containing 233 combinations of cells and TFs with matching cellular background. This data set was built from a catalog of 144 TF ChIP-seq and 13 DNase-seq data sets. This data is used to evaluate the effects of bias correction and TF binding time. In this scenario we evaluated the methods PWM, FS, TC, HINT, HINT-BC and HINT-BCN.

*Expression-based evaluation (FLR-Exp).* As shown in Yardımcı et al.12, ChIP-seq evaluation of putative TFBSs may present biases regarding the fact that ChIP-seq data alone is not able to distinguish direct from indirect binding events. Consequently, we performed an evaluation procedure which combines MPBSs with differentially expressed genes from two cell types. The method evaluates the association of the quality of footprints overlapping particular motifs and the expression of the TF.

We used limma47 to perform between-array normalization on expression of H1-hESC, K562 and GM12878 cells and obtain fold change estimates. Then, we retrieved all non-redundant PFMs from Jaspar in which gene symbol is a perfect match with genes present in the array platform. This leads us to 143 PFMs (see **Supplementary Datasets 2b-d**). We applied a genome-wide motif matching using these PFMs.

Afterwards, we evaluated the FLR12 score, TC15 and FS15 for the footprints of each evaluated method, which intersects with MPBSs of a particular motif. We only considered the footprints within DHSs that are in common between the cell type pair being evaluated, as described in Yardımcı et al.12. We expect that TFs expressed in cell type A would present higher values regarding these metrics (FLR, TC and FS) with DNase-seq from cell type A in comparison with these metrics evaluated with DNase-seq from cell type B, and vice-versa. We used a two-sample Kolmogorov-Smirnov (KS) test to assess the difference between each metrics' distribution between the two cell types being evaluated. The KS statistic, which varies from 0 to 1, is used to indicate the difference between two distributions; higher values indicate higher differences. As the KS score do not indicate the direction of the changes in distribution, we obtained a signed version by multiplying KS statistic by –1 in cases where the median of A < median of B. We calculate the Spearman correlation between the signed KS test statistic and the expression fold change for each TF (see **Supplementary Fig. 1 and 2**). Positive values indicate an association between expression of TFs and quality of footprint predictions. We will call this correlation “FLR-Exp”. Results for FLR-Exp analysis are summarized in **Supplementary Dataset 2a**.

**Protection score.** We propose a measure to detect TF-specific footprint protection for a given DNase-seq experiment and MPBSs of a given motif/TF. As previously indicated in Sung et al.7, fewer DNase-seq cuts (protection) surrounding the binding site characterizes TFs with shorter binding times. More formally, the protection score for a set of is defined as:

where is set of binding sites for a given motif, is the genomic location of the -th binding site and , and are the number of DNase-seq reads in the binding site, upstream and downstream flanking positions, respectively (see **equation (2)** for details).

In short, the protection score indicates the average difference of DNase-seq counts in the flanking region and the DNase-seq counts within the MPBS. Positive values will indicate protection in the flanking regions, while values close to zero or negative indicate no protection. The protection score is a similar statistic as the FS15. The main difference is that the FS score measures the ratio between reads in flanking versus binding sites, while the protection score measures the difference. Finally, since we are interested in using the protection score as a measure of quality for a given TF and set of footprint predictions, we only evaluate MPBSs overlapping with footprints for a given cell type. The DNase-seq count values are previously corrected for DHS sequence bias and coverage differences. Results for protection scores are provided in **Supplementary Dataset 1**.

**Statistical methods.** The non-parametric Friedman-Nemenyi hypothesis test48 was used to compare the AUC and AUPR of the methods regarding all data set combinations (TFs versus cell types). Such test provides a rank of the methods as well as the statistical significance of whether a particular method was outperformed. All correlations are based on Spearman values. All reported *p*-values have been corrected with the Benjamini and Hochberg method49.

**Code Availability.** Software, custom code, benchmarking data, DNase-seq sequence bias estimates and further graphical results are available at [www.costalab.org/hint-bc](http://www.costalab.org/hint-bc). The HINT, HINT-BC and HINT-BCN softwares can be directly accessed through the regulatory genomics toolbox website at [www.regulatory-genomics.org/hint/](http://www.regulatory-genomics.org/hint/).

26 Lazarovici, A. et al. Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc. Natl. Acad. Sci. USA* **110**, 6376–6381 (2013).

27 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

28 Zhang, Y. et al. Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9**, R137+ (2008).

29 Yu, J. et al. An integrated network of androgen receptor, polycomb and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell* **17**, 443–454 (2010).

30 Guertin, M.J, Zhang, X., Coonrod, S.A. & Hager, G.L. Transient estrogen receptor binding and p300 redistribution support a squelching mechanism for estradiol-repressed genes*. Mol. Endocrinol.* **28**, 1522–1533 (2014).

31 John, S. et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43**, 264–268 (2011).

32 Mathelier, A. et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42**, D142–D147 (2014).

33 Robasky, K. & Bulyk, M. L. UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* **39**, D124–D128 (2011).

34 Matys, V. et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).

35 Boyle, A.P. Guinney, J., Crawford, G.E. & Fury, T.S. F-seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**, 2537-2538 (2008).

36 Hesselberth, J. R. et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods* **6**, 283–289 (2009).

37 Sabo, P. J. et al. Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc. Natl. Acad. Sci. USA* **101**, 16837–16842 (2004).

38 Madden, H.H. Comments on the Savitzky-Golay convolution method for least-squares fit smoothing and differentiation of digital data. *Anal. Chem.* **50**, 1383–1386 (1978).

39 Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).

40 Hubbard, T. et al. The ensembl genome database project. *Nucleic Acids Res.* **30**, 38–41 (2002).

41 Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

42 Stormo, G.D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16–23 (2000).

43 Cock, P.J.A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

44 Korhonen, J., Martinmäki, P., Pizzi, C. Rastas, P. & Ukkonen, E. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics* **25**, 3181–3182 (2009).

45 Wilczynski, B., Dojer, N., Patelak, M. & Tiuryn, J. Finding evolutionarily conserved cis-regulatory modules with a universal set of motifs. *BMC Bioinform.* **10**, 82+ (2009).

46 Fawcett, T. An introduction to ROC analysis. *Pattern Recogn. Lett.* **27**, 861–874 (2006).

47 Ritchie, M. E. et al. *limma* powers differential expression analyses for RNA-sequencing and microarray studies*. Nucleic Acids Res*. **43**, e47 (2015).

48 Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7**, 1–30 (2006).

49 Benjamini, Y. & Hochberg, Y. et al. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).