Published online 23 July 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 20* ***8689–8702***
doi:10.1093/nar/gkr602

Enhancers regulate progression of development in mammalian cells

Anna-Lena Kranz[1,2], Roland Eils[1,2,*] and Rainer König[1,2,*]

[1]Department of Bioinformatics and Functional Genomics, Institute of Pharmacy and Molecular Biotechnology, and Bioquant, University of Heidelberg, INF 267, 69120 Heidelberg and [2]Theoretical Bioinformatics, German Cancer Research Center, INF 580, 69121 Heidelberg, Germany

Received January 19, 2011; Revised June 17, 2011; Accepted July 6, 2011

ABSTRACT

During development and differentiation of an organism, accurate gene regulation is central for cells to maintain and balance their differentiation processes. Transcriptional interactions between *cis*-acting DNA elements such as promoters and enhancers are the basis for precise and balanced transcriptional regulation. We identified modules of combinations of binding sites in proximal and distal regulatory regions upstream of all transcription start sites (TSSs) *in silico* and applied these modules to gene expression time-series of mouse embryonic development and differentiation of human stem cells. In addition to tissue-specific regulation controlled by combinations of transcription factors (TFs) binding at promoters, we observed that in particular the combination of TFs binding at promoters together with TFs binding at the respective enhancers regulate highly specifically temporal progression during development: whereas 40% of TFs were specific for time intervals, 79% of TF pairs and even 97% of promoter–enhancer modules showed specificity for single time intervals of the human stem cells. Predominantly SP1 and E2F contributed to temporal specificity at promoters and the forkhead (FOX) family of TFs at enhancer regions. Altogether, we characterized three classes of TFs: with binding sites being enriched at the TSS (like SP1), depleted at the TSS (like FOX), and rather uniformly distributed.

INTRODUCTION

Transcription factors (TFs) cooperate with other regulatory co-factors and the complex combinations of multiple cooperative interactions give the necessary specificity for spatio-temporal transcriptional regulation (1). Sets of TFs binding in a defined DNA region are called *cis*-regulatory modules (CRMs). CRMs direct the expression of developmental genes and signaling molecules during development (2) and the combinatorial and temporal binding of CRMs is crucial for metazoan development (3) and for the establishment of tissue specific gene expression (4). Specifically in higher organisms, proximal versus distal regulation needs to be well balanced (5). Whereas promoters are proximal to transcription start sites (TSSs), enhancers can be quite distant from their target genes. Enhancer regions have been suggested to consist of densely clustered TF binding sites (6) and stimulate transcription irrespectively of their position or orientation with respect to the TSS (7). TFs bound at an enhancer interact with co-activators and TFs bound at the promoter. Hence, they increase the concentration of activators at promoters. The large distance between long-range enhancers and proximal promoters can be overcome by chromatin loops, bringing these elements in close proximity (8,9). Thus, enhancers can increase the activity of a promoter considerably, even when located several kilo bases away. For example, deleting the enhancer for immunoglobulin heavy chain (IgH) resulted in loss of gene expression for IgH (10) and deletion of the T-lymphocyte-specific enhancer (E4p) needed for CD4 expression yielded cell populations of which the majority did not show any CD4 expression in T lymphocytes (11).

In addition, enhancers can recruit chromatin modification enzymes (e.g. a histone acetyltransferase) and chromatin remodeling complexes that put up an adequate environment for transcription. Promoter–enhancer interactions depend on regulatory factors binding at promoter-proximal regions, e.g. Krüppel-like factor 1 (erythroid) (KLF1) was suggested to induce the switch between the expression of fetal gamma-globin to adult beta-globin by mediating an interaction between the beta-globin gene promoter and a distal regulatory element (12). In turn, these factors may recruit specific distal enhancers (13,14) depending on the combination of regulatory factors at the

*To whom correspondence should be addressed. Tel: +6221 423600; Fax: +6221 423620; Email: r.koenig@dkfz.de
Correspondence may also be addressed to Roland Eils. Email: r.eils@dkfz.de

proximal promoter (15). Levine and Tjian (15) suggested that a combination of different complexes is needed for a temporal- and tissue-specific regulation of *cis*-DNA elements allowing a vast variety of distinct gene expression patterns.

There exist various examples reporting the involvement of enhancers in the regulation of development for well-studied genes, mainly for invertebrates (16,17) but also for human and mouse in which e.g. conserved distal regulatory regions associated with developmental genes have been identified as enhancers (6,18–21).

We were interested in immanent differences of enhancers and promoters affecting the regulation of genes during critical developmental stages of different tissues and cell types. For this, we set up a statistical analysis. Genes being differentially expressed at specific time intervals of the development of each analyzed tissue (before, during and after its formation) were associated to their regulating TFs. TFs were considered to be temporal-specific if their regulated genes mainly occurred in not more than one time interval. Similarly, we analyzed tissue specificity and considered a TF as tissue-specific if its regulated genes occurred predominantly in one tissue only. In this manner, we systematically compared temporal and tissue specificity of TFs, combinations of TFs binding at promoters, combinations of TFs binding at enhancers, and combinations of TFs binding at promoters and TFs binding at enhancers. Our results not only support tissue specificity of TF pairs [which has been reported previously (4)] but we also show that the combination of TFs in promoter regions together with combinations of *TFs* in *enhancer* regions determines *temporal* specificity.

MATERIALS AND METHODS

Identifying TFs, combinations of TFs, and promoter–enhancer modules

The complete workflow of the analysis is shown in Figure 1. Sequences from 10 000 bp upstream to 1000 bp downstream of the TSS for 32 290 human genes (Build 36.3) as well as for 33 063 genes for mouse (Build 37.1) and 27 110 genes for rat (Build 4.1) were retrieved from the National Center for Biotechnology Information (NCBI, ftp://ftp.ncbi.nlm.nih.gov/). TF annotations and associated position weight matrices (PWMs) were obtained from TRANSFAC (Release 12.1) (22) yielding 549 human TFs (represented by 455 PWMs), 407 TFs for mouse (represented by 410 PWMs) and 366 TFs for rat (represented by 471 PWMs). Each predicted PWM binding site was matched to all TFs associated to this PWM. As in

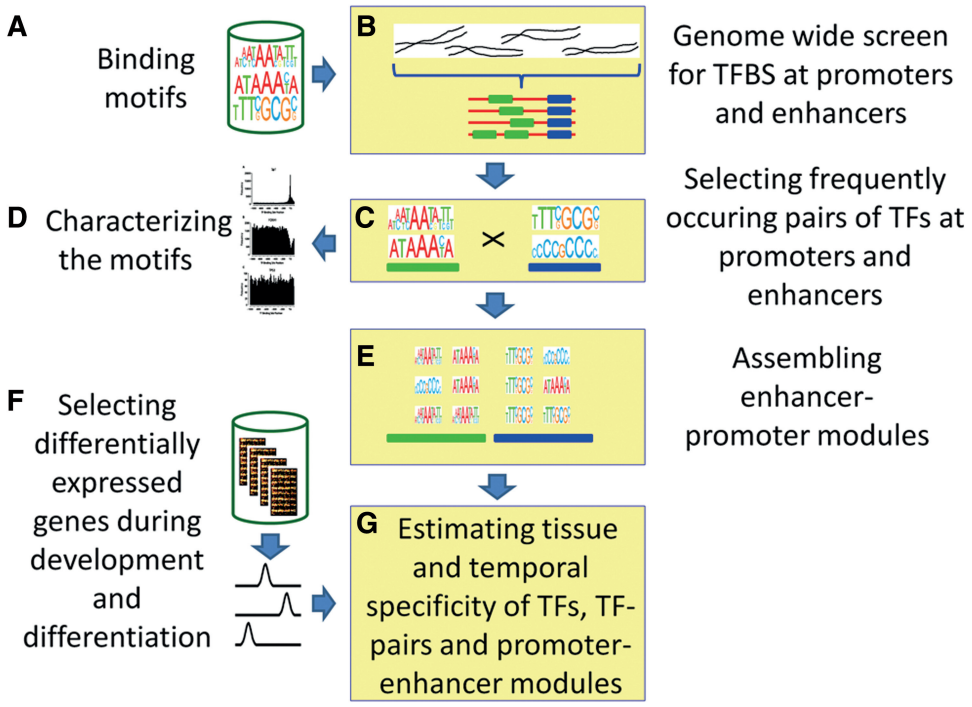

Figure 1. The Workflow. (**A**) Motifs (PWMs) for TF binding sites were collected from a database. (**B**) Upstream sequences were collected for each transcript. Promoters (± 100 bp of the TSS) and enhancers (defined by accumulation of binding motifs and phylogenetic conservation, 2000–10 000 bp upstream of the TSS) were selected. (**C**) Statistical and combinatorial analysis of TF binding sites of promoters and enhancers. (**D**) Characterization of single motifs with respect to their distributions in the observed sequences (0–10 000 bases upstream of the TSS) and network analysis. (**E**) Assembly of promoter–enhancer modules. A promoter–enhancer module consisted of a pair of TFs binding at the promoter and a pair of TFs binding at the enhancer. (**F**) Gene expression data was taken from microarray studies of the development of several mouse tissues and of the differentiation of human stem cells. A time series analysis was performed to identify genes being differentially expressed at distinct (developmental) time intervals and tissues/cell types. (**G**) Differentially expressed genes of each time interval were tested to be enriched of genes with predicted binding sites of single TFs, TF pairs and promoter–enhancer modules. Promoter–enhancer modules were used to predict differential expression of developmental time intervals.

TRANSFAC a TF can be associated to several PWMs and vice versa, TFs associated to the same PWM were grouped together and only one representative of such a TF-group was used in the analysis. For example, FOXA1, FOXA2 and FOXA3 were associated to the same PWM and FOXA1 was used as the representative of this TF-group. The grouping resulted in 152 TF-groups for human (Supplementary Table S1A), 139 for mouse (Supplementary Table S1B), and 141 TF-groups for rat which were used for further analysis. The detection of TF binding sites based on the respective PWMs was performed with the software package R (www.r-project.org) as described previously (23,24). Predicted binding sites with a $P > 0.05$ were discarded. The computation of the *P*-value is described in (23). Briefly, a significance value was determined by comparing the obtained score to a score distribution of the motif determined in random sequences generated by a background distribution following the base distribution of the whole genome. It is to note that we treated this value as a parameter to determine a cut-off and not for any significance test. Hence no multiple testing correction was needed. Predicted binding sites for SP1 and FOXA were compared to experimentally identified binding sites. Binding sites for SP1 were compared to chromatin immunoprecipitation sequencing (ChIP-Seq) data from the ENCODE project (25) from a study by the laboratory of Richard M. Myers at the Hudson Alpha Institute for Biotechnology. The data was downloaded from UCSC (26). Binding sites of FOXA were compared to several ChIP-Seq (27–29) and ChIP-chip (30) data sets. Genome coordinates of peak hits were compared to gene annotations (NCBI Build 36.3) and target genes were determined using the same settings as for the *in silico* promoter screen. Binding sites occurring within a range of −10 kb and +1 kb of the annotated TSS of a gene were included in the analysis. The determined genes were then compared to the list of predicted SP1 or FOXA target genes. A Fisher's exact test was conducted to assess the enrichment of experimentally identified binding sites in predicted binding sites. Predicted TF binding sites were combined into pairs of TFs and promoter–enhancer modules. Regions 100 bp upstream and downstream of the TSS were used as promoters, and regions starting 2000 bp and ending 10 000 bp upstream of the TSS were used as potential enhancer regions. The enhancer region was chosen this way as Heintzman *et al.* (31) demonstrated that the majority of predicted enhancers are located >2.5 kb from known TSS. In addition, Blanchette *et al.* (32) identified CRMs and found that the density of modules is lowest in regions starting from 10 kb from the TSS. Combinations of TFs for promoters were obtained by pairing non-overlapping TF binding sites co-occurring in the promoter region of a gene using a sliding window of 20 bp. Only pairs occurring in at least 10 genes were taken into further consideration. To decrease false positives of predicted TF binding sites, only conserved binding sites were analyzed in enhancer regions which increased specificity. To determine the conservation of human, mouse and rat binding sites, we analyzed pair-wise alignments between human and chimp, mouse and rat, and rat and mouse, respectively. Chained and netted pair-wise alignments of human (UCSC version hg18) and chimp (UCSC version panTro2), of mouse (UCSC version mm9) and rat (UCSC version rn4), and of rat (UCSC version rn4) and mouse (UCSC version mm9) were downloaded from UCSC (33) in the axtNet format (ftp://hgdownload.cse.ucsc.edu/). Conserved regions between human and chimp, mouse and rat, and rat and mouse were determined by the given aligned regions in the alignment files. Predicted binding sites were compared to the identified conserved regions and taken if binding sites occurred in these conserved regions. Pairs of non-overlapping co-occurring TFs in enhancer regions were determined using a sliding window of 20 bp (same size as for promoter regions). To analyze enhancer regions with a comparable size to promoter regions, we regarded sequences of a 200 bp sliding window. Regions in which at least 10 binding sites occurred were considered as enhancer regions and TF pairs occurring in at least 10 genes were considered further. Promoter–enhancer modules were constructed by combining two TF pairs occurring at the respective promoter and enhancer regions of a gene. Promoter–enhancer modules occurring in at least 10 genes were taken for further analysis. Hence, promoter–enhancer modules consisted of a combination of a pair of co-occurring TFs at the promoter and a pair of co-occurring TFs at the enhancer region. We also constructed promoter–enhancer modules with different sets of parameters. For this, we defined the promoter region 500 bp upstream and downstream of the TSS, used a sliding window of 50 bp, and TF pairs and promoter–enhancer modules had to occur in at least five genes. To show that combinations of TF pairs of enhancers and promoters show better specificity than combining only TF pairs binding at the promoter, we also constructed the latter combinations. For this, two pairs of co-occurring TFs in the promoter region were combined for each gene and taken for further analyses if the promoter–enhancer module occurred in at least 10 genes.

Gene expression analyses

Gene expression data of mouse embryonic development and differentiation of human stem cells were retrieved from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/). For mouse embryonic development, gene expression data was analyzed comprising early cardiac development (GSE1479), the developing prefrontal cortex (GSE4675), facial prominences (GSE7759), early development of the brain (GSE8091), development of the liver (GSE13149), ovary development (GSE5334), and development of testis (GSE4818). Quality was assessed by manual inspection of probe intensity distributions of each array and discarded if the MA plots showed abnormal distributions. We discarded three samples from the dataset of early brain development, four samples from ovary development, and one sample from testis development. For differentiation of human stem cells, we analyzed gene expression data of cardiomyocytes (GSE13834), chondrogenic differentiation (GSE10315), myoblast differentiation (GSE3780), myelopoiesis (GSE12837), and neural differentiation (GSE9940). Similar to the datasets

for mouse, we discarded data with low quality. We discarded one sample from the cardiomyocytes, four samples from differentiation of chondrogenesis, 24 samples from differentiation of myoblasts, 11 samples from myelopoiesis, and four samples from neural differentiation. The data was analyzed using the affymetrix package (34) of R (www.r-project.org) and normalized with VSN normalization (35). For better comparability, for each gene expression study, time points were grouped into three time intervals: early, mid and late expression, e.g. in the human myelopoiesis data set (GSE12837), the haematopoietic stem/progenitor cells (HSC) were grouped at the early time interval, myeloid precursors at the mid-time interval and terminally differentiated cells at the late time interval. Each data set was tested for differentially expressed genes between the different time intervals using the Rank Product Test (36). Significant genes were determined using a cutoff for percentage of false positives (pfp) <5%. Pfp is an estimate of the false discovery rate, which is determined by a permutation-based procedure of determining the observed value in permutated gene expression values for each sample (36).

Estimating tissue and time specificity for TFs, combinations of TFs and promoter–enhancer modules

For the identified TFs, pairs of TFs at promoter and enhancer regions, and promoter–enhancer modules, we determined if their set of regulated genes was enriched in differentially regulated genes per time interval and tissue. The procedure is explained exemplarily for TFs. For each TF we determined genes with binding sites for the TF identified by our PWM-scans and regarded them as potentially regulated by the specific TF. Using Fisher's exact tests, we tested if these regulated genes were significantly enriched in the list of differentially expressed genes of each time interval for each gene expression study (tissue). We defined this TF to be tissue-specific if such an enrichment occurred only for one tissue (number of tissues = one), otherwise we specified this TF to regulate two or more tissues (number of tissues >1). Similarly, we defined the TF to be time interval-specific if we determined an enrichment of its genes in the list of differentially expressed genes of a tissue at one time interval (number of time intervals = 1), and more than one time interval otherwise (number of time intervals > 1). This enrichment analysis was performed for all TFs. The results were summarized for all TFs and the percentage of TFs per time interval and tissue identified, yielding the results shown in Figure 2A and Supplementary Figure S1A. The same procedure was conducted for pairs of TFs at promoters (Figure 2B and Supplementary Figure S1B), pairs of TFs at enhancers (Figure 2C and Supplementary Figure S1C), and promoter–enhancer modules (Figure 2D and Supplementary Figure S1D). As predicted target genes in differentially expressed genes could differ between promoter–enhancer modules and TF pairs at promoters or enhancers, a promoter–enhancer module could show an enrichment of differentially expressed genes in several tissues even though its containing TF pairs did not get significant enrichments in these tissues. To assess the significance of temporal specificity of promoter–enhancer modules, a Fisher's exact test was performed to test if the number of promoter–enhancer modules specific for a single time interval was enriched compared to the number of TFs, TF pairs at promoters or TF pairs at enhancers specific at a single time interval. In addition, we determined the percentage of genes per promoter–enhancer module that were specific for a single time interval and compared it to the percentage of genes per TF pair at promoters being specific for a single time interval. To validate our results, we also employed permutation tests with 10 000 permutations of randomly assigned differentially regulated genes per tissue and time interval. In addition, we also conducted permutation tests with 10 000 permutations of genes chosen randomly as being regulated by promoter–enhancer modules. We then determined the number of temporal-specific promoter–enhancer modules per permutation using the approach described above. The number of actual temporal-specific permutation-enhancer modules was compared to the distribution of temporal specific promoter–enhancer modules determined by the permutation analyses. In addition, we incorporated binding site predictions including positional preferences of the different TFs (37) into our analysis. Binding site predictions were downloaded from SwissRegulon (38) and genes of promoter–enhancer modules were restricted to predicted genes of the FANTOM study within the promoter region. The described enrichment analysis was then repeated for the restricted promoter–enhancer modules to assess their temporal specificity. We also correlated the expression of each promoter–enhancer module to the expression of its regulated genes per tissue. For this, we calculated the expression profile for each module from the median expression of all four TFs involved. We calculated Pearson's correlation coefficients of this profile and each gene of the predicted regulated gene list (containing the motifs of the TFs from the module). Similarly, we also calculated the correlation to all other genes of the arrays. To test if the predicted regulated gene list correlated significantly better than the rest of the genes, we performed a Student's *t*-test with the absolute values of both distributions (correlations of the predicted gene list versus all other genes).

Predicting time intervals using promoter–enhancer modules

To further estimate the quality of the identified promoter–enhancer modules, we set up a machine learning system that was trained with the identified modules to predict temporal expression of genes. Additionally, this enabled to select promoter–enhancer modules with a higher predictive value for temporal regulation of gene expression during development and differentiation and to estimate their potential power to regulate distinct gene groups for the progression of development. We employed the method of random forests as a machine learning method (classifier). For training, we chose genes whose expression was associated with a distinct time interval. This way, we selected genes which were differentially expressed at only

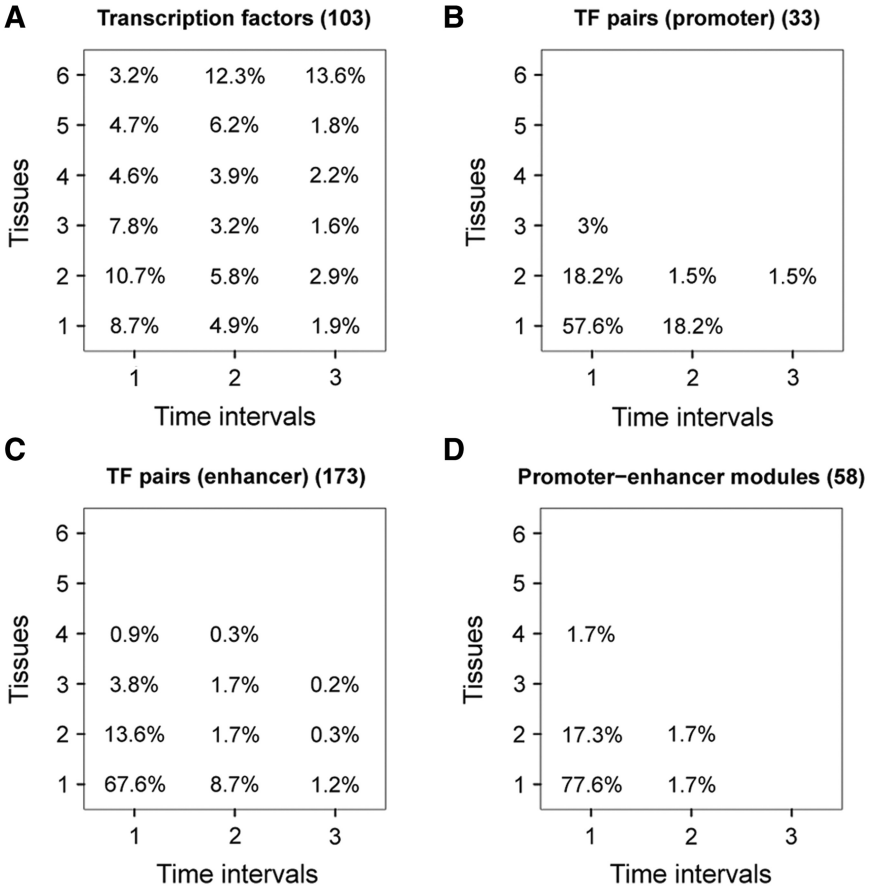

Figure 2. Tissue and temporal specificity for each regulatory element during differentiation of human stem cells. The number of tissues versus the number of time intervals is plotted for (**A**) TFs, (**B**) pairs of co-occurring TFs in promoter regions, (**C**) pairs of co-occurring TFs in enhancer regions and (**D**) promoter–enhancer modules. The percentage of the different regulatory elements is indicated at each entry in the grid (i.e. 9% of TFs are specific for a single tissue and a single time interval). The total number of regulatory elements is given in brackets.

a single time interval in all tissues and determined the set of promoter–enhancer modules associated to these genes. We set up a classification task for two classes. As little data were available for the mid-time interval ($n = 7$) and to simplify classification, we used only two time intervals (early and late). The early time interval constituted the first class and the late time interval the second class. The classifiers were trained to predict the correct time interval for each gene, using only the information which specific promoter–enhancer modules were regulating the respective gene (promoter–enhancer modules served as features for the classifier). We trained 10 000 decision trees yielding an ensemble classifier (random forest) using the package randomForest (39) (http://cran.r-project.org/web/packages/randomForest) in R (www.r-project.org). To identify promoter–enhancer modules with the best discriminative behavior (best separation performance), we applied the gini criterion which minimizes the impurity of the children nodes at each split in the tree (40). To focus on the best discriminators, we used the top 5% of these selected features for classification. A 10 times 10-fold cross-validation was applied to determine the performance of the classifier (yielding accuracy, sensitivity and specificity for the classifier). For comparison, we also trained a random forest using pairs of co-occurring TFs at promoters as features with the same parameters as for promoter–enhancer modules. Similar to promoter–enhancer modules, the best discriminating pairs of TFs at promoters were identified according to the gini criterion. The top 5% of the features were used for predictions and the performance of the classifier was determined employing a 10 times 10-fold cross-validation.

Defining TFs with TSS-enriched, TSS-depleted, and uniformly distributed binding sites

For each TF, the distribution of binding sites was determined with respect to the annotated TSS for all genes. TFs were grouped into three categories: TFs with binding sites predominantly around the TSS (TSS-enriched-BS), TFs with a depletion of binding sites at the TSS (TSS-depleted-BS), and TFs showing a uniform distribution of binding sites (uniformly distributed-BS). For this grouping, a Wilcoxon signed-rank test was conducted for each TF to test if the distribution of binding sites at the TSS (±100 bp around TSS) follows the distribution of the remaining binding sites. To correct for multiple testing, a Benjamini–Hochberg correction (41) was applied. TFs with $P < 0.05$ and a difference of the medians of the distributions of at least 4 bp were classified as TFs preferentially binding at the TSS (TSS-enriched-BS)

or as TFs with binding sites depleted around the TSS (TSS-depleted-BS) depending on the sign of the difference of the medians of the distributions. All other TFs were termed TFs with a uniform distribution of binding sites (uniformly distributed-BS).

Constructing the networks

Using the identified co-occurring TF pairs as links (see 'Identifying combinations of TFs and promoter–enhancer modules' section), two networks were constructed, one for promoters and one for enhancers. To assess if pairs of TFs of the same group (TSS-enriched-BS, TSS-depleted-BS, uniformly distributed-BS) occurred more often than expected by chance, we performed a permutation test with 10 000 permutations of the class labels. Connectivity and betweenness centrality were determined for each node in the network and their maxima were identified for both networks for TFs of the categories TSS-enriched-BS and TSS-depleted-BS. In addition, a protein–protein interaction network of TFs was constructed using physical binding information from a public repository [BIND (42)] and each TF was associated to its corresponding protein in the network. This network was analyzed for the same properties as the described promoter and enhancer networks.

RESULTS

Identifying promoter–enhancer modules

To identify promoter–enhancer modules we performed a genome-wide screen for TF binding sites using position weight matrices (PWMs) for all annotated human genes and TFs (23,24). To validate our predictions, we compared TF binding predictions to experimentally identified binding sites for two central TFs, SP1 and FOXA. Both TFs showed a significantly high overlap between predicted and experimentally identified binding sites. We predicted 9796 genes correctly out of 11 877 genes with experimentally identified binding sites ($P < 2.2E-16$, false positive rate: 0.35) for SP1 and 4954 genes correctly out of 5732 genes with experimentally identified binding sites ($P = 0.002$, false positive rate: 0.7) for FOXA. Figure 1 depicts the workflow of the method. The sequence upstream and downstream (±100 bp) of the annotated TSS was termed promoter region, whereas the studied enhancer region was further upstream of the TSS (2000–10 000 bp upstream). To identify interacting TFs at promoters and enhancers, we selected pairs of co-occurring TF binding sites in a defined window at the promoter and enhancer region for each gene, respectively. We then combined identified pairs of co-occurring TFs at the promoter and enhancer region for each gene to analyze combinations of promoter and enhancer interactions. These combinations were termed promoter–enhancer modules. After filtering ('Materials and Methods' section), we identified 129 promoter–enhancer modules binding at 340 genes. To generalize our investigations, we repeated the analysis and identified promoter–enhancer modules also for mouse and rat. Promoter–enhancer modules for mouse and rat showed similar results when applying the same settings as for human (Table 1).

Identified promoter–enhancer modules regulate spatio-temporal gene expression in development

To investigate time- and tissue-specific regulatory roles of the identified promoter–enhancer modules in development, we analyzed time series of gene expression profiles of embryonic development in mouse and embryonic stem cell differentiation in human cells. We selected gene expression studies from a broad range of different embryonic mouse tissues and human stem cells of different origin. Each study was regarded as tissue-specific. For better comparison among the different studies, we grouped time points for each gene expression study into three distinct time intervals we termed early, mid and late expression. For each gene expression study, we identified differentially expressed genes at these time intervals and determined their respective regulation by TFs, pairs of co-occurring TFs, and promoter–enhancer modules employing enrichment analyses ('Materials and Methods' section). Differentially expressed genes and significant promoter–enhancer modules per tissue and time interval for human stem cell differentiation and mouse embryonic development are presented in Supplementary Table S2. We compared the number of enriched tissues and time intervals for single TFs, pairs of co-occurring TFs, and promoter–enhancer modules. Strikingly, promoter–enhancer modules showed the highest tissue and temporal specificity. Figure 2 shows the results for human stem cells. Only 16% of TFs were specific for a single tissue whereas 76% of pairs of co-occurring TFs in promoter regions, 77% of pairs of co-occurring TFs in enhancer regions, and 79% of promoter–enhancer modules showed specificity for a single tissue. Temporal specificity was even more distinctive. Whereas only 40%

Table 1. Overview of the number of identified transcriptional regulators for different organisms

	Human		Mouse		Rat	
	Regulatory elements	Genes	Regulatory elements	Genes	Regulatory elements	Genes
TFs	132	32 121	123	33 033	132	27 110
TF pairs at promoters	111	3007	77	1931	74	1891
TF pairs at enhancers	579	11 172	418	10 985	585	8326
Promoter–enhancer modules	129	340	113	311	28	134

of the studied TFs were specific for a single time interval, 79% of pairs of co-occurring TFs in promoter regions, and 85% in enhancer regions and even 97% of the promoter–enhancer modules showed specificity for a single time interval in the data sets of human stem cells (Figure 3). Ravasi *et al.* (4) showed that pairs of TFs rather than single TFs determine tissue specificity. Surprisingly, the additional temporal specificity of promoter–enhancer modules is obtained by pairs of co-occurring TFs at enhancers (97% for promoter–enhancer modules versus 79% for pairs of co-occurring TFs at promoters, significance of the difference: $P = 0.01$, 85% for pairs of co-occurring TFs at enhancers, significance of the difference to promoter–enhancer modules: $P = 0.02$ and 40% for TFs, significance of the difference to promoter–enhancer modules: $P = 2.56\text{E-}14$). In addition, differentially expressed genes regulated by the identified promoter–enhancer modules were also specific for a single time interval (74% for promoter–enhancer modules versus 55% for pairs of co-occurring TFs). Temporal specificity could not be increased when including binding site prediction incorporating positional preferences of TFs of the study by the FANTOM Consortium (37) (80% of promoter–enhancer modules were specific for a single time interval). The expression of each promoter–enhancer module correlated to the expression of its regulated genes in half of the tissues (Supplementary Table S3 lists the results for all tissues). Similar results were obtained for mouse embryonic development (Supplementary Figures S1 and S2). Whereas 34% of the TFs were specific for a single time interval and 11% for a single tissue, 77% and 52% of pairs of co-occurring TFs at promoters, 79% and 58% of pairs at enhancers and 89% and 69% of promoter–enhancer modules showed specificity for a single time interval and tissue, respectively. As seen for both mouse embryonic development and human stem cell differentiation, the combinations of regulatory factors at promoters and enhancers resulted in higher specificity of tissue dependent and temporal regulation during development and differentiation. Concluding, TFs binding at promoters contributed significantly to tissue-specific regulation, whereas regulatory factors at enhancers rather accounted for temporal specificity.

To cross-check the specificity of these promoter–enhancer modules, we constructed promoter modules consisting of combinations of pairs of co-occurring TFs at promoters only and repeated the analysis. Using the same parameter settings, we identified a limited number of promoter modules ($n = 12$) that did not allow any conclusion about tissue and temporal specificity. Even increasing the promoter region by a factor of 10 (which

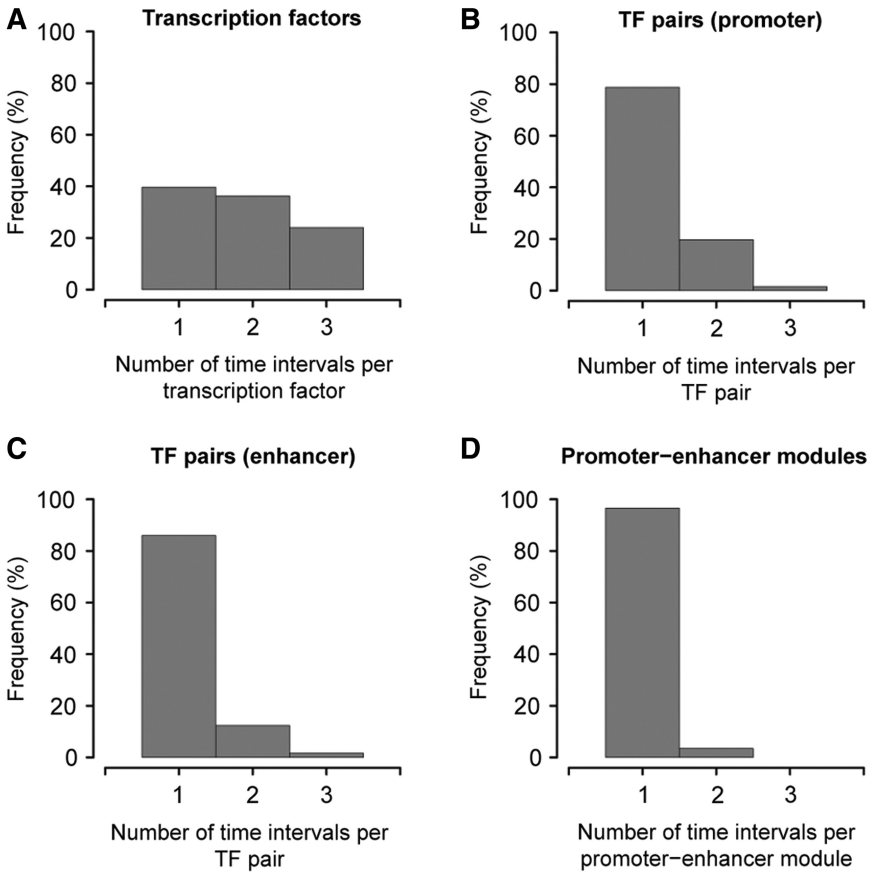

Figure 3. Frequency distribution of the number of time intervals for human stem cell differentiation. The histograms show the frequency distribution of the number of time intervals for (**A**) TFs, (**B**) pairs of co-occurring TFs in promoter regions, (**C**) pairs of co-occurring TFs in enhancer regions, and (**D**) promoter–enhancer modules.

resulted in a sufficient number of promoter modules) revealed 83% of promoter modules as tissue-specific but only 61% as time-specific. To further validate the increased specificity, we also employed permutation tests. For this, differentially expressed genes were randomly assigned for each tissue and time interval. In a second test, randomly chosen genes were assigned to be regulated by a promoter–enhancer module. These permutation tests demonstrated a highly significant temporal regulation of the identified promoter–enhancer modules ($P = 0.0285$ for permuted gene expression and $P = 0.0057$ for permuted genes regulated by promoter–enhancer modules). In addition, we constructed promoter–enhancer modules using different sets of parameters with similar results (Supplementary Table S4). Discarding conservation of TF binding sites slightly reduced temporal specificity of the resulting promoter–enhancer modules (94% compared to 97% using conserved binding sites). These results further support the fact that the combinations of promoter–enhancer interactions establish temporal specificity of gene expression.

For these analyses, we tested a variety of different reasonable parameter settings to select genes with significant motifs and motif combinations. We found quite similar results for these different settings and a setting was chosen for which we got a good temporal precision for all compared sets (single TFs, combinations of TFs, and promoter–enhancer modules; number of genes with binding site for regulatory element ≥ 10, sliding window = 20 bp, promoter region $\pm$ 100 bp of TSS, see 'Identified promoter–enhancer modules regulate spatio-temporal gene expression in development' section). The results for all parameter settings are given in the Supplementary Data (Supplementary Table S4). It is to note that promoter–enhancer modules out-performed TF pairs for all parameter settings.

Promoter–enhancer modules predict temporal gene expression in development

To identify promoter–enhancer modules (combinations of TFs) that predict gene expression at a specific time interval, we further analyzed the active promoter–enhancer modules during human stem cell differentiation. We learned a classifier (of a random forest) to predict the time interval (now simplified for two categories, early and late) of temporal differential expression for each gene based on its promoter–enhancer modules. This way, we were able to predict temporal differential expression of a gene based on the profile of its promoter–enhancer modules. Specifically, with this we yielded the combination of pairs of co-occurring TFs in promoter and enhancer regions represented by the promoter–enhancer modules, which determine the temporal regulation observed during development. The top 10 promoter–enhancer modules explaining best temporal specificity are shown in Table 2. TF-groups SP1 (Sp1 TF), EGR1 (early growth response 1) and E2F1 (E2F TF 1) were the most observed TFs occurring in promoter regions, whereas members of the forkhead box family of TFs (FOXI1, FOXJ1, FOXD3, FOXF1, FOXL1 and FOXA1) and CDX1 (caudal type homeobox 1) were mostly found at enhancer regions. To validate our results, we performed a stratified 10 times 10-fold cross-validation and trained with the top 5% of

Table 2. Top 10 of the list of identified promoter–enhancer modules explaining temporal specificity for the differentiation of human stem cells

Promoter–enhancer modules[a]	Additional members of the TF-group	Binding preference[b]
SP1 SP1 - FOXI1 FOXJ1	SP2, SP3, SP4 (SP1)	TSS-enriched-BS (SP1)
	FOXD3, FOXF1, FOXF2 (FOXJ1)	TSS-depleted-BS (FOXI1,FOXJ1)
SP1 SP1 - FOXJ1 FOXJ1	SP2, SP3, SP4 (SP1)	TSS-enriched-BS (SP1)
	FOXD3, FOXF1, FOXF2 (FOXJ1)	TSS-depleted-BS (FOXJ1)
SP1 SP1 – CDX1 FOXA1	SP2, SP3, SP4 (SP1)	TSS-enriched-BS (SP1)
	CDX2 (CDX1)	TSS-depleted-BS (CDX1,FOXA1)
	FOXA2, FOXA3 (FOXA1)	
SP1 SP1 - FOXI1 FOXA1	SP2, SP3, SP4 (SP1)	TSS-enriched-BS (SP1)
	FOXA2, FOXA3 (FOXA1)	TSS-depleted-BS (FOXI1,FOXA1)
EGR1 SP1 - FOXI1 FOXA1	EGR2, EGR3, EGR4 (EGR1)	TSS-enriched-BS (EGR1,SP1)
	SP2, SP3, SP4 (SP1)	
	FOXA2, FOXA3 (FOXA1)	TSS-depleted-BS (FOXI1,FOXA1)
SP1 SP1 - FOXJ1 FOXJ2	SP2, SP3, SP4 (SP1)	TSS-enriched-BS (SP1)
	FOXD3, FOXF1, FOXF2 (FOXJ1)	TSS-depleted-BS (FOXJ1,FOXJ2)
SP1 SP1 - FOXL1 FOXL1	SP2, SP3, SP4 (SP1)	TSS-enriched-BS (SP1)
		TSS-depleted-BS (FOXL1)
E2F1 E2F1 - FOXL1 FOXA1	E2F2, E2F3, E2F4, E2F5, E2F7, TFDP1 (E2F1)	TSS-enriched-BS (E2F1)
	FOXA2, FOXA3 (FOXA1)	TSS-depleted-BS (FOXL1,FOXA1)
E2F1 EGR1 - FOXJ2 FOXL1	E2F2, E2F3, E2F4, E2F5, E2F7, TFDP1 (E2F1)	TSS-enriched-BS (E2F1,EGR1)
	EGR2, EGR3, EGR4 (EGR1)	TSS-depleted-BS (FOXJ1,FOXL1)
EGR1 SP1 - CDX1 FOXI1	EGR2, EGR3, EGR4 (EGR1)	TSS-enriched-BS (EGR1,SP1)
	SP2, SP3, SP4 (SP1)	
	CDX2 (CDX1)	TSS-depleted-BS (CDX1,FOXI1)

[a]The first two TFs were identified at promoters, the last two at enhancers.
[b]TSS-enriched-BS, TFs with binding sites predominantly around the TSS; TSS-depleted-BS, TFs with a depletion of binding sites at the TSS.

promoter–enhancer modules yielding a considerably good prediction performance (70% accuracy, 73% sensitivity, 69% specificity). In comparison, pairs of co-occurring TFs at promoters were not sufficient to predict temporal gene expression and failed to detect differences between the time intervals (43% accuracy, 21% sensitivity, 76% specificity). These results support the specificity of the identified promoter–enhancer modules for temporal gene expression.

TFs show distinct binding site distributions for promoter and enhancer regions

To identify differences among TFs binding preferentially either at promoter or enhancer regions of the identified promoter–enhancer modules, we analyzed the distributions of binding sites for all TFs with respect to the annotated TSS. Interestingly, we identified three different binding site distributions for the analyzed TFs. Figure 4 shows exemplarily the distributions for the TF-groups SP1, FOXA1 and TP53 (tumor protein p53). The distribution of SP1 showed an enrichment of binding sites close to the TSS (Figure 4A). Binding sites with these distributions were termed TSS-enriched-BS, whereas FOXA1 exhibited a depletion of binding sites at the TSS (TSS-depleted-BS, Figure 4B). We also observed rather uniform distributions, e.g. TP53 (uniformly distributed-BS, Figure 4C). Binding preferences for all analyzed TFs are shown in Supplementary Table S5. We investigated the motifs of these three groups and found that TFs with TSS-enriched-BS had binding sites with a higher GC content compared to the other TFs ($P = 8.22E\text{-}14$). This is consistent with reports that sequences at TSS are often GC rich (5,43,44). All TFs occurring at promoters of the identified promoter–enhancer modules had TSS-enriched-BS, and 91% of the TFs at enhancers had TSS-depleted-BS (Supplementary Figure S4 for additional analysis). Notably, this tendency was even stronger for the promoter–enhancer modules selected by the classification algorithm (100% TSS-enriched-BS for the promoter pairs and 100% TSS-depleted-BS for the enhancer pairs of promoter–enhancer modules). When applying the analysis to all TFs analyzed, the majority of TFs (53%) showed a uniform distribution of binding sites with no preferential binding position (uniformly distributed-BS). 21% of TFs were determined to preferentially bind close to the TSS (TSS-enriched-BS), whereas 26% TFs showed a depletion of binding sites around the TSS (TSS-depleted-BS). Comparing identified TFs with TSS-enriched-BS to previously determined TFs with enrichments of binding sites close to the TSS (45,46) revealed a high overlap (e.g. SP1, NF-Y, YY1, TBP, REST, NRF-1, ELK-1, ATF3, SREBP-1, MAZ, CREBP). It is to note that although previous studies identified TFs with preferential binding close to the TSS (5,44–48), TFs showing a depletion of binding sites around the TSS or a uniform binding site distribution have been noted (44) but have not been quantified so far.

To further analyze characteristics of TFs with different binding site distributions in promoter and enhancer regions, we constructed two networks. The topological

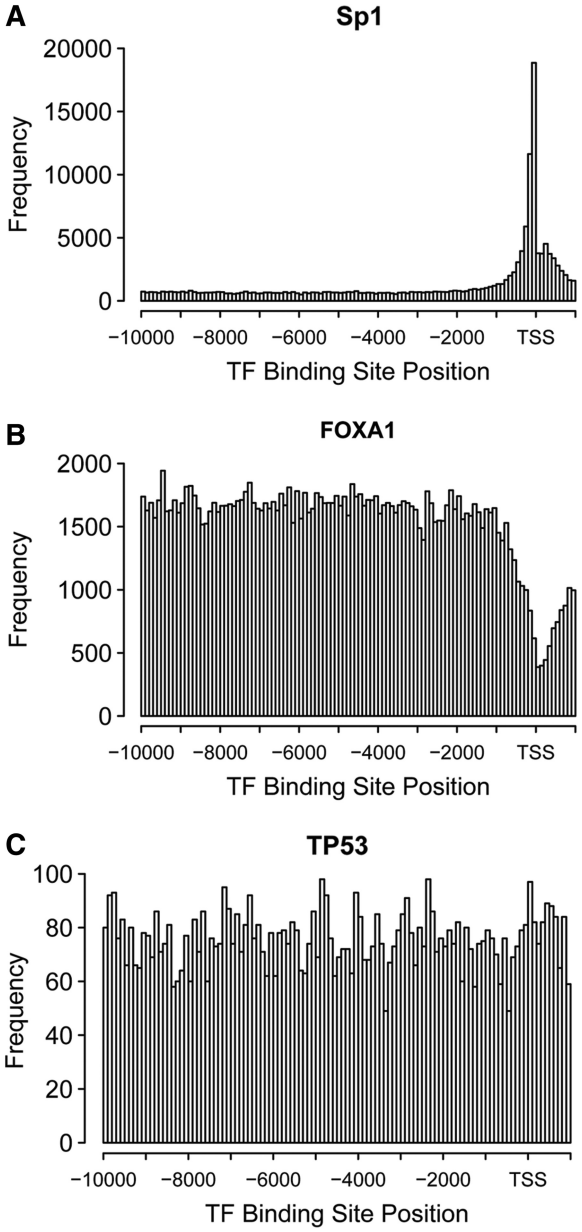

Figure 4. Distribution of binding sites for different groups of TFs. For different groups of TFs the distribution of binding sites with respect to the TSS is shown exemplarily for the TFs SP1, FOXA1, and TP53: (**A**) The distribution for SP1 which represents the distribution of binding sites for TFs preferentially binding at the TSS (TSS-enriched-BS), (**B**) the distribution of FOXA1 which represents the distribution of binding sites for TFs with a depletion of binding sites at the TSS (TSS-depleted-BS), and (**C**) the distribution for TP53 representing uniformly distributed binding sites (uniformly distributed-BS).

structure of a network can reveal significant biological properties (49). A link in the networks was set for each pair of co-occurring TFs identified at promoters for the promoter network and enhancer regions for the enhancer network. Interestingly, in both networks, the majority of TFs with TSS-enriched-BS was adjacent to TFs of the same entity (TSS-enriched-BS) [significant ($P = 0.002$) for the promoter network and tendency ($P = 0.1$) for the enhancer network]. Similarly, TFs with TSS-depleted-BS

were preferentially adjacent to TFs with TSS-depleted-BS ($P = 0.002$ for the promoter network and $P = 0.001$ for the enhancer network). Figure 5 shows the networks and Supplementary Figures S3 and S4 the distributions of TFs in the promoter and enhancer networks. As shown in Table 3, most TFs in the promoter network had TSS-enriched-BS (65%) whereas only 24% of TFs had TSS-depleted-BS. To further characterize distinct roles for TFs with different binding site distributions and to estimate their functional importance in the network, we determined connectivity and betweenness centrality for each node in the networks (Table 3). Whereas betweenness centrality measures the traffic load through a node, connectivity indicates the significance of a node in the network as essential nodes are often so called hubs in a network (49–54). The TF SP1 with TSS-enriched-BS had the highest connectivity and highest centrality in the promoter network with a connectivity of 50 and betweenness centrality of 351.5. In contrast, the highest connectivity of a TF with TSS-depleted-BS was 12 and the betweenness centrality was zero for all TFs with TSS-depleted-BS. These results supported the fact that TFs with TSS-enriched-BS played a central role in the promoter network. These TFs constituted the main component of the network (Figure 5A) while TFs with TSS-depleted-BS formed a rather small and separated component. In contrast, TFs with TSS-depleted-BS played a central role in the enhancer network. These TFs constituted the core of the enhancer network (Figure 5B) with other TFs at its periphery. Only 44% of the TFs in the enhancer network had TSS-depleted-BS whereas only 29% of the TFs had TSS-enriched-BS. The fork head TF FOXA1 (forkhead box A1) with TSS-depleted-BS had the highest connectivity (124) and centrality (1062.4) in the enhancer network. In contrast, the highest connectivity of a TF with TSS-enriched-BS was 64 and the highest betweenness centrality was 349.4. In addition, we constructed a network of known TF

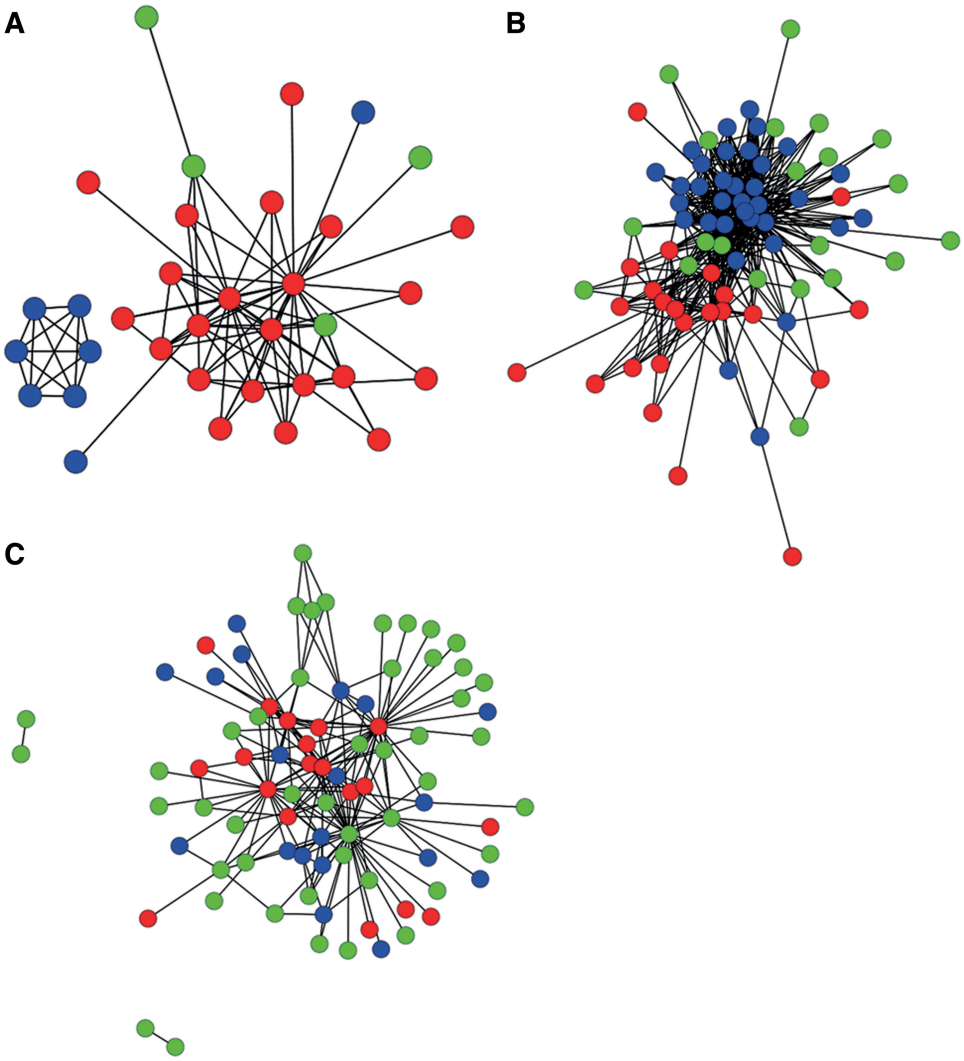

Figure 5. Networks of TF pairs (human). The network of pairs of co-occurring TFs are shown for (**A**) the promoter regions and (**B**) the enhancer regions. (**C**) A network of TFs mapped onto a PPI network (42). TFs showing preferential binding around the TSS (TSS-enriched-BS) are marked in red, TFs with a depletion of binding sites around the TSS (TSS-depleted-BS) in blue (dark) and TFs showing no preferential binding (uniformly distributed-BS) in green (light).

Table 3. Overview of network properties

	Promoter network		Enhancer network		PPI network	
	TSS-enriched	TSS-depleted	TSS-enriched	TSS-depleted	TSS-enriched	TSS-depleted
Quantity	17	8	23	34	19	19
Connectivity[a]	50	12	124	64	62	20
Betweenness centrality[a]	351.5	0	349.4	1062.4	2055.2	401.3

[a]For each TF group, the maximum is shown.

interactions [physical binding of pairs of TFs, obtained from a public repository (42)]. Interestingly, the number of TFs with TSS-enriched-BS and TSS-depleted-BS was balanced (both 23%) and these TFs were located rather at the core of the network (Figure 5C). TBP (TATA box binding protein) with TSS-enriched-BS showed the highest connectivity (62) and betweenness centrality (2055.2) in the network compared to TFs with TSS-depleted-BS which had a maximum connectivity of 20 and a maximum centrality of 401.3.

DISCUSSION

It was suggested previously that combinations of different complexes of TFs offer a plethora of specific gene expression profiles (15). Here, we identified *in silico* promoter–enhancer modules consisting of combinations of TFs binding at promoters and enhancers that determine specific tissue dependent and temporal regulation of gene expression during development and differentiation. In addition to tissue-specific regulation established by pairs of TFs as shown previously (4), we now also show that promoter–enhancer modules consisting of pairs of TFs at promoters and enhancers regulate the *progression* of gene expression patterns during development and differentiation. Furthermore, we found that these enhancer sites were rather depleted of Guanin–Cytosin (CpG). It was shown recently, that methylation-modifications of CpG-regions are a major regulation mechanism during development (55–57). Enhancer regions therefore may contribute to a more constitutive regulation program during development which is rather independent from these methylation-modifications. To systematically analyze which TFs might be employed for such a mechanism, we analyzed the distributions of putative binding sites from 100 bp downstream to 10 000 bp upstream of the TSS for every family of TF binding sites. Indeed, we identified three classes comprising binding sites being enriched at the TSS, depleted at the TSS, and rather uniformly distributed. The first class is in line with previous studies identifying TFs with preferential binding close to the TSS (5,44,48). In addition to this, we found TFs showing a depletion of binding sites around the TSS and TFs with a uniform binding site distribution which have been noted (44) but have not been quantified so far.

Analyzing the identified promoter–enhancer modules revealed a number of TFs binding preferentially either at promoters or at enhancers. For human stem cell differentiation, we identified SP1 to preferentially bind at promoters. Although SP1 is ubiquitously expressed and regulates gene expression of many constitutively expressed genes (58,59), its expression was shown to change at different developmental stages and in different cell types, suggesting specific roles in distinct developmental processes (60). As SP1-null mice died prenatal, SP1 was shown to be essential for mouse embryonic development (61). In contrast, members of the FOX (forkhead box) family of TFs had binding sites preferentially located at enhancers of the identified promoter–enhancer modules providing temporal specificity during human stem cell differentiation. FOX TFs have been identified to bind at enhancers in a number of studies (62–67). FOX proteins are substantial in a variety of cellular processes including development, differentiation, proliferation, apoptosis, and migration (68). As FOX proteins are regulators for a multitude of biological processes, their deregulation can contribute significantly to tumorigenesis and cancer progression (68). Various members of this family have been identified previously to be implicated in development (69–81). It is to note that we analyzed motif distributions and statistics for families of TFs sharing the same binding preferences. For example, the family of forkhead TFs FOXA1, FOXA2 and FOXA3 share the same binding motif even though they can exhibit quite distinct roles during developmental processes. It was shown that specifically FOXA3 differs from FOXA1 and FOXA2 in several developmental processes (70). In the data we analyzed, we also found a rather unalike gene expression pattern for FOXA3 in comparison to FOXA1 and FOXA2 (e.g. in cardiomyocytes Pearson's correlation: FOXA1-FOXA2: 0.78, FOXA1-FOXA3: 0.15, FOXA2-FOXA3: 0.63, Supplementary Table S6 shows the median expression of these TFs for all tissues). For future projects, it seems worthwhile studying such abundances of the particular family members in more detail employing gene and even better protein expression and phosphorylation data.

The CDX family was another group of TFs we identified at enhancers of our promoter–enhancer modules. *cdx* genes are closely related to genes of the Hox family and are expressed during embryonic development and gut morphogenesis (82). CDX2 is specifically required during early development and CDX2-null mice are nonviable as the blastocyst fails to implant into the uterus (83). It is to note that our results lack of TFs previously associated to differentiation and development such as the HOX family of TFs, Oct, Sox or Nanog. This can be explained by either lack of binding motifs (Nanog) or rather unspecific binding motifs (Hox, Sox, Oct). Due to unspecific motifs, these TFs were omitted by our stringent filtering settings. Further improvements of existing binding motifs will allow the identification of these TFs in promoter–enhancer modules when applying our method.

We used the identified promoter–enhancer modules to predict temporal expression of genes during human stem cell differentiation and achieved considerably good prediction accuracy (70%). Zinsen *et al*. (3) used binding sites of five TFs determined by ChIP-on-chip assays at

consecutive time points during *Drosophila* mesoderm development to predict temporal expression of CRMs using SVMs and achieved a similar good accuracy of 71.4%. In contrast, we analyzed the regulation of *human* cells and used the prediction of TF pairs at promoter and enhancer regions to predict temporal expression of genes.

Chromatin loops can overcome large distances between long-range enhancers and proximal promoters and may lead to false positive hits at promoter regions when screening promoters with ChIP-chip assays (8,9). So far, most ChiP sequencing studies neglect to assess indirect binding of TFs. However, it was shown for hepatocyte nuclear factor 4, alpha (HNF4A) that identified binding sites at the promoter occurred mainly at distal regulatory elements (84). Our results support the fact that key TFs bind preferentially at either promoters or enhancers and that the interactions between those elements are crucial for specific gene regulation. Therefore, indirect binding is a central issue that cannot be neglected in prospective TF binding site analyses and specifically when analyzing ChIP-chip and ChIP-Seq experiments.

Time- and tissue-specific regulation of gene expression is central not only during development but also in all processes of a cell in an organism. Here, we identified promoter–enhancer modules that determined specific gene regulation during development and revealed distinct binding site distributions for TFs binding preferentially at promoter or enhancer regions. The *in silico* identification of combinations of TFs binding at promoters and enhancers yielded generic insights into the temporal regulation of gene expression and improved our understanding of enhancer function.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Rolf Kabbe and Karlheinz Groß for IT support, and Thomas Wolf for fruitful discussions on the statistical analysis.

FUNDING

Helmholtz Alliance on Systems Biology (SB Cancer, D.141100/07.997); BMBF-FORSYS consortium Viroquant (#0313923); the Nationales Genom-Forschungs-Netz (NGFN+) for the neuroblastoma project, ENGINE (#01GS0898). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the article. Funding for open access charge: Helmholtz Alliance on Systems Biology.

Conflict of interest statement. None declared.

REFERENCES

1. Makeev,V.J., Lifanov,A.P., Nazina,A.G. and Papatsenko,D.A. (2003) Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. *Nucleic Acids Res.*, **31**, 6016–6026.
2. Howard,M.L. and Davidson,E.H. (2004) cis-Regulatory control circuits in development. *Dev. Biol.*, **271**, 109–118.
3. Zinzen,R.P., Girardot,C., Gagneur,J., Braun,M. and Furlong,E.E. (2009) Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, **462**, 65–70.
4. Ravasi,T., Suzuki,H., Cannistraci,C.V., Katayama,S., Bajic,V.B., Tan,K., Akalin,A., Schmeier,S., Kanamori-Katayama,M., Bertin,N. *et al.* (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, **140**, 744–752.
5. Koudritsky,M. and Domany,E. (2008) Positional distribution of human transcription factor binding sites. *Nucleic Acids Res.*, **36**, 6795–6805.
6. Woolfe,A., Goodson,M., Goode,D.K., Snell,P., McEwen,G.K., Vavouri,T., Smith,S.F., North,P., Callaway,H., Kelly,K. *et al.* (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.*, **3**, e7.
7. Banerji,J., Rusconi,S. and Schaffner,W. (1981) Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, **27**, 299–308.
8. Horike,S., Cai,S., Miyano,M., Cheng,J.F. and Kohwi-Shigematsu,T. (2005) Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. *Nat. Genet.*, **37**, 31–40.
9. Murrell,A., Heeson,S. and Reik,W. (2004) Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. *Nat. Genet.*, **36**, 889–893.
10. Grosschedl,R. and Marx,M. (1988) Stable propagation of the active transcriptional state of an immunoglobulin mu gene requires continuous enhancer function. *Cell*, **55**, 645–654.
11. Chong,M.M., Simpson,N., Ciofani,M., Chen,G., Collins,A. and Littman,D.R. (2010) Epigenetic propagation of CD4 expression is established by the Cd4 proximal enhancer in helper T cells. *Genes Dev.*, **24**, 659–669.
12. Perkins,A.C., Gaensler,K.M. and Orkin,S.H. (1996) Silencing of human fetal globin expression is impaired in the absence of the adult beta-globin gene activator protein EKLF. *Proc. Natl Acad. Sci. USA*, **93**, 12267–12271.
13. Calhoun,V.C., Stathopoulos,A. and Levine,M. (2002) Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex. *Proc. Natl Acad. Sci. USA*, **99**, 9243–9247.
14. Su,W., Jackson,S., Tjian,R. and Echols,H. (1991) DNA looping between sites for transcriptional activation: self-association of DNA-bound Sp1. *Genes Dev.*, **5**, 820–826.
15. Levine,M. and Tjian,R. (2003) Transcription regulation and animal diversity. *Nature*, **424**, 147–151.
16. Davidson,E.H., Rast,J.P., Oliveri,P., Ransick,A., Calestani,C., Yuh,C.H., Minokawa,T., Amore,G., Hinman,V., Arenas-Mena,C. *et al.* (2002) A genomic regulatory network for development. *Science*, **295**, 1669–1678.
17. Oliveri,P. and Davidson,E.H. (2004) Gene regulatory network controlling embryonic specification in the sea urchin. *Curr. Opin. Genet. Dev.*, **14**, 351–360.
18. Bejerano,G., Pheasant,M., Makunin,I., Stephen,S., Kent,W.J., Mattick,J.S. and Haussler,D. (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.
19. Visel,A., Prabhakar,S., Akiyama,J.A., Shoukry,M., Lewis,K.D., Holt,A., Plajzer-Frick,I., Afzal,V., Rubin,E.M. and Pennacchio,L.A. (2008) Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet.*, **40**, 158–160.
20. Pennacchio,L.A., Ahituv,N., Moses,A.M., Prabhakar,S., Nobrega,M.A., Shoukry,M., Minovitsky,S., Dubchak,I., Holt,A., Lewis,K.D. *et al.* (2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, **444**, 499–502.

21. Nobrega,M.A., Ovcharenko,I., Afzal,V. and Rubin,E.M. (2003) Scanning human gene deserts for long-range enhancers. *Science*, **302**, 413.
22. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al*. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*., **34**, D108–D110.
23. Rahmann,S., Muller,T. and Vingron,M. (2003) On the power of profiles for transcription factor binding site detection. *Stat. Appl. Genet. Mol. Biol*., **2**, Article7.
24. Westermann,F., Muth,D., Benner,A., Bauer,T., Henrich,K.O., Oberthuer,A., Brors,B., Beissbarth,T., Vandesompele,J., Pattyn,F. *et al*. (2008) Distinct transcriptional MYCN/c-MYC activities are associated with spontaneous regression or malignant progression in neuroblastomas. *Genome Biol*., **9**, R150.
25. Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigo,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T., Thurman,R.E. *et al*. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
26. Rosenbloom,K.R., Dreszer,T.R., Pheasant,M., Barber,G.P., Meyer,L.R., Pohl,A., Raney,B.J., Wang,T., Hinrichs,A.S., Zweig,A.S. *et al*. (2010) ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res*., **38**, D620–D625.
27. Wallerman,O., Motallebipour,M., Enroth,S., Patra,K., Bysani,M.S., Komorowski,J. and Wadelius,C. (2009) Molecular interactions between HNF4a, FOXA2 and GABP identified at regulatory DNA elements through ChIP-sequencing. *Nucleic Acids Res*., **37**, 7498–7508.
28. Wederell,E.D., Bilenky,M., Cullum,R., Thiessen,N., Dagpinar,M., Delaney,A., Varhol,R., Zhao,Y., Zeng,T., Bernier,B. *et al*. (2008) Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res*., **36**, 4549–4564.
29. Motallebipour,M., Ameur,A., Reddy Bysani,M.S., Patra,K., Wallerman,O., Mangion,J., Barker,M.A., McKernan,K.J., Komorowski,J. and Wadelius,C. (2009) Differential binding and co-binding pattern of FOXA1 and FOXA3 and their relation to H3K4me3 in HepG2 cells revealed by ChIP-seq. *Genome Biol*., **10**, R129.
30. Lupien,M., Eeckhoute,J., Meyer,C.A., Wang,Q., Zhang,Y., Li,W., Carroll,J.S., Liu,X.S. and Brown,M. (2008) FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell*, **132**, 958–970.
31. Heintzman,N.D., Stuart,R.K., Hon,G., Fu,Y., Ching,C.W., Hawkins,R.D., Barrera,L.O., Van Calcar,S., Qu,C., Ching,K.A. *et al*. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet*., **39**, 311–318.
32. Blanchette,M., Bataille,A.R., Chen,X., Poitras,C., Laganiere,J., Lefebvre,C., Deblois,G., Giguere,V., Ferretti,V., Bergeron,D. *et al*. (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res*., **16**, 656–668.
33. Rhead,B., Karolchik,D., Kuhn,R.M., Hinrichs,A.S., Zweig,A.S., Fujita,P.A., Diekhans,M., Smith,K.E., Rosenbloom,K.R., Raney,B.J. *et al*. (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res*., **38**, D613–D619.
34. Gautier,L., Cope,L., Bolstad,B.M. and Irizarry,R.A. (2004) Affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.
35. Huber,W., von Heydebreck,A., Sultmann,H., Poustka,A. and Vingron,M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18(Suppl. 1)**, S96–S104.
36. Breitling,R. and Herzyk,P. (2005) Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. *J. Bioinform. Comput. Biol*., **3**, 1171–1189.
37. Suzuki,H., Forrest,A.R., van Nimwegen,E., Daub,C.O., Balwierz,P.J., Irvine,K.M., Lassmann,T., Ravasi,T., Hasegawa,Y., de Hoon,M.J. *et al*. (2009) The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat. Genet*., **41**, 553–562.
38. Pachkov,M., Erb,I., Molina,N. and van Nimwegen,E. (2007) SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Res*., **35**, D127–D131.
39. Liaw,A. and Wiener,M. (2002) Classification and Regression by randomForest. *R News*, **2**, 18–22.
40. Breiman,L. (1996) Technical note: Some properties of splitting criteria. *Mach. Learn*, **24**, 41–47.
41. Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. B*, **57**, 289–300.
42. Bader,G.D., Betel,D. and Hogue,C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*., **31**, 248–250.
43. Heintzman,N.D. and Ren,B. (2007) The gateway to transcription: identifying, characterizing and understanding promoters in the eukaryotic genome. *Cell Mol. Life Sci*., **64**, 386–400.
44. FitzGerald,P.C., Shlyakhtenko,A., Mir,A.A. and Vinson,C. (2004) Clustering of DNA sequences in human promoters. *Genome Res*., **14**, 1562–1574.
45. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
46. Yokoyama,K.D., Ohler,U. and Wray,G.A. (2009) Measuring spatial preferences at fine-scale resolution identifies known and novel cis-regulatory element candidates and functional motif-pair relationships. *Nucleic Acids Res*., **37**, e92.
47. Vardhanabhuti,S., Wang,J. and Hannenhalli,S. (2007) Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. *Nucleic Acids Res*., **35**, 3203–3213.
48. Tharakaraman,K., Bodenreider,O., Landsman,D., Spouge,J.L. and Marino-Ramirez,L. (2008) The biological function of some human transcription factor binding motifs varies with position relative to the transcription start site. *Nucleic Acids Res*., **36**, 2777–2786.
49. Zhu,X., Gerstein,M. and Snyder,M. (2007) Getting connected: analysis and principles of biological networks. *Genes Dev*., **21**, 1010–1024.
50. Estrada,E. (2006) Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics*, **6**, 35–40.
51. Hahn,M.W. and Kern,A.D. (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol*., **22**, 803–806.
52. Plaimas,K., Eils,R. and König,R. (2010) Identifying essential genes in bacterial metabolic networks with machine learning methods. *BMC Syst. Biol*., **4**, 56.
53. Plaimas,K., Mallm,J.P., Oswald,M., Svara,F., Sourjik,V., Eils,R. and König,R. (2008) Machine learning based analyses on metabolic networks supports high-throughput knockout screens. *BMC Syst. Biol*., **2**, 67.
54. Rahman,S.A. and Schomburg,D. (2006) Observing local and global properties of metabolic pathways: 'load points' and 'choke points' in the metabolic networks. *Bioinformatics*, **22**, 1767–1774.
55. Lei,H., Oh,S.P., Okano,M., Juttermann,R., Goss,K.A., Jaenisch,R. and Li,E. (1996) De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. *Development*, **122**, 3195–3205.
56. Okano,M., Bell,D.W., Haber,D.A. and Li,E. (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, **99**, 247–257.
57. Borgel,J., Guibert,S., Li,Y., Chiba,H., Schubeler,D., Sasaki,H., Forne,T. and Weber,M. (2010) Targets and dynamics of promoter DNA methylation during early mouse development. *Nat. Genet*., **42**, 1093–1100.
58. Imataka,H., Sogawa,K., Yasumoto,K., Kikuchi,Y., Sasano,K., Kobayashi,A., Hayami,M. and Fujii-Kuriyama,Y. (1992) Two regulatory proteins that bind to the basic transcription element (BTE), a GC box sequence in the promoter region of the rat P-4501A1 gene. *EMBO J*., **11**, 3663–3671.
59. Gidoni,D., Kadonaga,J.T., Barrera-Saldana,H., Takahashi,K., Chambon,P. and Tjian,R. (1985) Bidirectional SV40 transcription

mediated by tandem Sp1 binding interactions. *Science*, **230**, 511–517.
60. Saffer,J.D., Jackson,S.P. and Annarella,M.B. (1991) Developmental expression of Sp1 in the mouse. *Mol. Cell. Biol.*, **11**, 2189–2199.
61. Marin,M., Karis,A., Visser,P., Grosveld,F. and Philipsen,S. (1997) Transcription factor Sp1 is essential for early embryonic development but dispensable for cell growth and differentiation. *Cell*, **89**, 619–628.
62. De Val,S., Chi,N.C., Meadows,S.M., Minovitsky,S., Anderson,J.P., Harris,I.S., Ehlers,M.L., Agarwal,P., Visel,A., Xu,S.M. *et al.* (2008) Combinatorial regulation of endothelial gene expression by ets and forkhead transcription factors. *Cell*, **135**, 1053–1064.
63. Yamagishi,H., Maeda,J., Hu,T., McAnally,J., Conway,S.J., Kume,T., Meyers,E.N., Yamagishi,C. and Srivastava,D. (2003) Tbx1 is regulated by tissue-specific forkhead proteins through a common Sonic hedgehog-responsive enhancer. *Genes Dev.*, **17**, 269–281.
64. Cirillo,L.A., Lin,F.R., Cuesta,I., Friedman,D., Jarnik,M. and Zaret,K.S. (2002) Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol. Cell*, **9**, 279–289.
65. Chaya,D., Hayamizu,T., Bustin,M. and Zaret,K.S. (2001) Transcription factor FoxA (HNF3) on a nucleosome at an enhancer complex in liver chromatin. *J. Biol. Chem.*, **276**, 44385–44389.
66. Gualdi,R., Bossard,P., Zheng,M., Hamada,Y., Coleman,J.R. and Zaret,K.S. (1996) Hepatic specification of the gut endoderm in vitro: cell signaling and transcriptional control. *Genes Dev.*, **10**, 1670–1682.
67. McPherson,C.E., Shim,E.Y., Friedman,D.S. and Zaret,K.S. (1993) An active tissue-specific enhancer and bound transcription factors existing in a precisely positioned nucleosomal array. *Cell*, **75**, 387–398.
68. Myatt,S.S. and Lam,E.W. (2007) The emerging roles of forkhead box (Fox) proteins in cancer. *Nat. Rev. Cancer*, **7**, 847–859.
69. Perreault,N., Katz,J.P., Sackett,S.D. and Kaestner,K.H. (2001) Foxl1 controls the Wnt/beta-catenin pathway by modulating the expression of proteoglycans in the gut. *J. Biol. Chem.*, **276**, 43328–43333.
70. Friedman,J.R. and Kaestner,K.H. (2006) The Foxa family of transcription factors in development and metabolism. *Cell Mol. Life Sci.*, **63**, 2317–2328.
71. Lee,C.S., Friedman,J.R., Fulmer,J.T. and Kaestner,K.H. (2005) The initiation of liver development is dependent on Foxa transcription factors. *Nature*, **435**, 944–947.
72. Wan,H., Dingle,S., Xu,Y., Besnard,V., Kaestner,K.H., Ang,S.L., Wert,S., Stahlman,M.T. and Whitsett,J.A. (2005) Compensatory roles of Foxa1 and Foxa2 during lung morphogenesis. *J. Biol. Chem.*, **280**, 13809–13816.
73. Lai,E., Prezioso,V.R., Tao,W.F., Chen,W.S. and Darnell,J.E. Jr (1991) Hepatocyte nuclear factor 3 alpha belongs to a gene family in mammals that is homologous to the Drosophila homeotic gene fork head. *Genes Dev.*, **5**, 416–427.
74. Ang,S.L. and Rossant,J. (1994) HNF-3 beta is essential for node and notochord formation in mouse development. *Cell*, **78**, 561–574.
75. Weinstein,D.C., Ruiz i Altaba,A., Chen,W.S., Hoodless,P., Prezioso,V.R., Jessell,T.M. and Darnell,J.E. Jr (1994) The winged-helix transcription factor HNF-3 beta is required for notochord development in the mouse embryo. *Cell*, **78**, 575–588.
76. Besnard,V., Wert,S.E., Kaestner,K.H. and Whitsett,J.A. (2005) Stage-specific regulation of respiratory epithelial cell differentiation by Foxa1. *Am. J. Physiol.*, **289**, L750–L759.
77. Hulander,M., Wurst,W., Carlsson,P. and Enerback,S. (1998) The winged helix transcription factor Fkh10 is required for normal development of the inner ear. *Nat. Genet.*, **20**, 374–376.
78. Labosky,P.A. and Kaestner,K.H. (1998) The winged helix transcription factor Hfh2 is expressed in neural crest and spinal cord during mouse development. *Mech. Dev.*, **76**, 185–190.
79. Brody,S.L., Yan,X.H., Wuerffel,M.K., Song,S.K. and Shapiro,S.D. (2000) Ciliogenesis and left-right axis defects in forkhead factor HFH-4-null mice. *Am. J. Resp. Cell Mol. Biol.*, **23**, 45–51.
80. Chen,J., Knowles,H.J., Hebert,J.L. and Hackett,B.P. (1998) Mutation of the mouse hepatocyte nuclear factor/forkhead homologue 4 gene results in an absence of cilia and random left-right asymmetry. *J. Clin. Invest.*, **102**, 1077–1082.
81. Tuteja,G. and Kaestner,K.H. (2007) SnapShot: forkhead transcription factors I. *Cell*, **130**, 1160.
82. Beck,F. and Stringer,E.J. The role of Cdx genes in the gut and in axial development. *Biochem. Soc. Trans.*, **38**, 353–357.
83. Strumpf,D., Mao,C.A., Yamanaka,Y., Ralston,A., Chawengsaksophak,K., Beck,F. and Rossant,J. (2005) Cdx2 is required for correct cell fate specification and differentiation of trophectoderm in the mouse blastocyst. *Development*, **132**, 2093–2102.
84. Rada-Iglesias,A., Wallerman,O., Koch,C., Ameur,A., Enroth,S., Clelland,G., Wester,K., Wilcox,S., Dovey,O.M., Ellis,P.D. *et al.* (2005) Binding sites for metabolic disease related transcription factors inferred at base pair resolution by chromatin immunoprecipitation and genomic microarrays. *Hum. Mol. Genet.*, **14**, 3435–3447.