# Supplementary Material

## ISMARA: Automated modeling of genomic signals as a democracy of regulatory motifs

Piotr J. Balwierz, Mikhail Pachkov, Phil Arnold, Andreas J. Gruber, Mihaela Zavolan
Erik van Nimwegen

*Biozentrum, the University of Basel, and Swiss Institute of Bioinformatics*
*Klingelbergstrasse 50/70, 4056-CH, Basel, Switzerland,*
*email: erik.vannimwegen@unibas.ch*

January 23, 2014

## Contents

# 1 Supplementary Methods

## 1.1 Human and mouse promoteromes

The central entities whose regulation is modeled by ISMARA are *promoters*. When analyzing expression data, be they micro-array or RNA-seq, ISMARA estimates and models the expression profiles of individual promoters, and when analyzing ChIP-seq data ISMARA models the chromatin state of genomic regions centered on promoters. Thus, the first step in the analysis consists of the construction of reference sets of promoters in human and mouse. To make a comprehensive list of promoters we used two sources of data: deepCAGE data, i.e. next-generation sequencing data of 5' ends of mRNAs [1, 2], and the 5' ends of all known mRNAs listed in GenBank.

Using CAGE data from a considerable set of human and mouse tissues, we recently constructed genome-wide human and mouse 'promoteromes' [3] consisting of a hierarchy of individual transcription start sites (TSSs), transcription start clusters (TSCs) of nearby co-regulated TSSs, and transcription start regions (TSRs), which correspond to clusters of TSCs with overlapping proximal promoter regions. As the basis of our promoter sets we started with the sets of TSCs, i.e. local clusters of TSSs whose expression profiles are proportional to each other to within experimental noise, as identified by deepCAGE.

As the currently available CAGE data do not yet cover all cell types in human and mouse, a substantial number of cell type-specific promoters are not represented within this set of TSCs. We thus supplemented the TSCs with all 5' ends of mRNAs, using the BLAT[4] mappings from UCSC Genome Browser web site[5]. To avoid transcripts whose 5' ends are badly mapped, we filtered out those for which more than $25$ bases at the 5' end of the transcript were unaligned. We then produced reference promoter sets by iteratively clustering the TSCs with the 5' ends of mRNAs as follows: Initially each TSC and each 5' end of an mRNA forms a separate cluster. At each iteration the pair of nearest clusters are clustered, with the constraint that there can be at most one TSC per cluster. That is, we never cluster two TSCs together as our previous analysis in [3] has already established that each TSC is independently regulated. Here the distance between two clusters is defined as the distance between the nearest pair of TSSs of the two clusters, i.e. the distance between the rightmost TSS of the left cluster and leftmost TSS of the right cluster. This iteration is repeated until the distance between the closest pair of clusters is larger than $150$ base pairs. Note that we thus chose the length of sequence wrapped by a single nucleosome, i.e. roughly $150$ base pairs, as an *ad hoc* cut-off length for two TSSs to belong to a common promoter. The reasoning behind this choice of cut-off, is that, on the one hand, we have empirically observed that co-expressed TSSs can spread over roughly this length-scale and, on the other hand, that it is not implausible that the ejection of a single nucleosome near the TSS may be responsible for setting this length scale. In any case, the resulting promoters are not sensitive to the precise setting of this cut-off (data not shown). Finally, inspection of the results showed, especially in ubiquitously expressed genes, many apparent TSSs from Genbank that appear downstream of both the TSSs identified by deepCAGE and the annotated RefSeq transcripts. It is highly likely that many of these apparent TSSs are due to cDNA sequences that were not full length. Indeed, only a small fraction of the transcripts in the database of mRNAs underwent expert curation, and truncated transcripts are likely common. To avoid such spurious apparent TSSs we removed all clusters which did not contain at least one curated transcript (RefSeq) or a TSC. Finally, since a sequence of at least one associated transcript is necessary to estimate a promoter's expression level from either RNA-seq or micro-array data, we also discarded all promoters

that consisted solely of a TSC.

For human, the resulting reference promoter set had $36'383$ promoters, of which $13'265$ contained both a TSC and at least one RefSeq transcript, $14'538$ contained only a TSC together with non-RefSeq transcripts, and $8'580$ had at least one RefSeq transcript and potentially non-RefSeq transcripts, but no TSC. For the mouse genome, the corresponding numbers are: $34'050$ promoters in total, $8'578$ RefSeq-only, $12'303$ TSC-only, and $13'169$ with both a TSC and at least one RefSeq transcript. These reference promoters sets cover almost all known protein-coding genes in human and mouse.

Finally, as we discussed in [3], mammalian promoters clearly fall into two classes associated with high and low content of CpG dinucleotides, and these promoter classes have clearly distinct architectures, i.e. different lengths, different numbers of TSSs per promoters, and different distributions of transcription factor binding sites (TFBSs). We classified all promoters into a high-CpG and low-CpG class based on both the CG content and the CpG content in the proximal promoter, as described in [3]. In the TFBS prediction below we perform separate predictions for high-CpG and low-CpG promoters.

## 1.2 A curated set of regulatory motifs

We use standard position dependent weight matrices (WMs) to represent regulatory motifs, i.e. the sequence specificities of TFs. Each WM is named for the TFs that are annotated to bind its site. For some motifs the names correspond to multiple TFs which are all assumed to bind to the same sites. We used a partly manual curation procedure whose details were first described in [6]. For completeness, we here also give a description of this curation procedure.

For a number of reasons regarding data quality and annotation ambiguities, the construction of a set of position-specific weight matrices (WMs) for mammalian transcription factors is rife with problems that, in our opinion, do not currently have a clean solution. Therefore, our procedures necessarily involve several subjective choices, judgments, and hand-curation, which are certainly far from satisfactory.

Our main objectives were

1. To remove redundancy, we aim to have no more than 1 WM representing any given TF. Whenever multiple TFs have WMs that are statistically indistinguishable or when their DNA binding domains are virtually identical, then we use only one WM for that set of TFs.

2. To associate WMs with TFs based on the sequences of their DNA binding domains. That is, we obtain lists of TFs that can plausibly bind to the sites of a given WM by comparison of DNA binding domain sequences of TFs known to bind to the sites with those of all other TFs.

3. Re-estimation of WMs using genome-wide predictions of regulatory sites in the proximal promoters of CAGE TSSs.

The input data for our WM construction consisted of

1. The collection of JASPAR vertebrate WMs plus, for each WM, the amino acid sequence of the TF that JASPAR associates with the WM [7].

2. The collection of TRANSFAC vertebrate WMs (version 9.4) and the amino acid sequences of all vertebrate TFs in TRANSFAC that are associated with those WMs [8].

3. A list of 1322 human TFs (Entrez gene IDs) and their amino acid sequences (from RefSeq).

4. A list of 483 Pfam IDs corresponding to DNA binding domains and their Pfam profiles [9].

We decided not to include 6 TRANSFAC motifs, which were constructed out of less than 8 sites: M00326 (PAX1, PAX9), M00619 (ALX4), M00632 (GATA4), M00634 (GCM1, GCM2), M00630

(FOXM1), M00672 (TEF). TRANSFAC often associates multiple WMs with a single human TF. Although there undoubtedly are cases where a single TF can have multiple distinct modes of binding DNA, and could therefore be realistically represented by multiple WMs, we believe that for the very large majority of TFs it is more realistic to describe the DNA binding specificity of the TF with a single WM. Indeed, a manual inspection of cases in which TRANSFAC associated multiple WMs with a single TF shows that these WMs are typically highly similar and appear redundant. Therefore, we decided to remove this redundancy and for each TF with multiple WMs in TRANSFAC we choose only a single 'best' WM based on TRANSFAC's own matrix quality annotation, or WM information score when there were multiple WMs with the same quality score. The information score of a WM is given by 2 times the length of the WM minus its entropy in bits.

We next aimed to obtain, for each human TF, a list of WMs from JASPAR/TRANSFAC, that can potentially be associated to this TF. To do this we aim to find, for each TF, which motifs from JASPAR/-TRANSFAC are associated with a TF that has a highly similar DNA binding domain. To this end we we ran Hmmer [10] with the DNA binding domain (DBD) profiles from Pfam to extract the DBDs from all TFs (E-value cut-off $10^{-9}$) associated with either JASPAR or TRANSFAC matrices. We then represented each such TF with the union of its DNA binding domain sequences. Next we used BLAT to map the DBDs of all TFs associated with JASPAR/TRANSFAC matrices against the entire protein sequences of all human TFs. For each human TF we then extracted a list of all JASPAR/TRANSFAC matrices for which the DBDs of at least one associated TF has a significant BLAT hit (default parameters) against the TF sequence. For each human TF the associated WMs were ordered by the percent identity of the hit, i.e. the fraction of all amino acids in the DBDs that map to matching amino acids in the TF.

From this data we created a list of 'necessary WMs' as follows. For each human TF we obtain the JASPAR WM with the highest percent identity in the DBDs of an associated TF. If there is a TRANSFAC WM with a higher percent identity than any JASPAR TF we record this WM as well. Thus, the necessary WMs are those that are the best match for at least one human TF. This list yielded 381 WMs representing 980 human TFs (often the same WM is the best match for multiple TFs). Manual inspection indicated that a lot of redundancy (essentially identical looking WMs) remained in this list. First we often have both a TRANSFAC and a JASPAR WM for the same TF and moreover often there are multiple TFs, each with its own WM, that look essentially identical. We thus want to fuse WMs in the following situations

1. Different WMs for TFs with identical or near identical DBDs.
2. WMs that are statistically indistinguishable, predict highly overlapping sets of sites, and are associated with TFs that have similar DBDs.

For each pair of WMs we obtained three similarity measurements

1. The percent identity of the DBDs of the TFs associated with the WMs. If there are multiple TFs associated with a WM we take the maximum over all TF pairs.
2. The overlap of the binding sites predicted by each WM. We use MotEvo to predict TFBSs in all proximal promoters and we calculate what fraction of predicted TFBS positions are shared between the two WMs.
3. A statistical measure of the similarity of the two WMs. Here we take the two sets of sites that define the two WMs and calculate the likelihood-ratio of these sites assuming they either derive from a single underlying WM or assuming that the set of sites for each WM derives from an independent WM.

For each of these three criteria we set a cut-off: $95\%$ identity of the DBDs, $60\%$ overlap of predicted TFBSs, and a likelihood-ratio of $e^{40}$. Using single-linkage clustering, we cluster all WMs whose similarity is over the cut-off for at least 1 of these three criteria. The resulting clusters were then all checked

manually and whenever the linkage was dubious we split the cluster. That is, we took a conservative attitude towards removing redundancy and only kept clusters when we were convinced the WMs were essentially identical. For each cluster we then constructed a new WM by aligning the WMs in the cluster so as to optimize the information content of the resulting fused WM, which is obtained by simply summing the counts across each column in the alignment.

Finally, we used MotEvo [11] to predict TFBSs for all WMs in the multiple-species alignments of all human proximal promoters. We then constructed new WMs from the list of predicted TFBSs for each WM, weighing each predicted site with its posterior probability (which incorporates position-specific prior probabilities, as described below). The number of top-scoring sites was chosen manually for each motif and was between 100 and 4000 sites, in most cases being 200 or 500 sites.

At this point we excluded one TRANSFAC motif M00395 (HOXA3, HOXB3, HOXD3) which had very low information content and predicted predicted only very low-probability sites. We additionally excluded the motifs M00480 (TOPORS) and M00987 (FOXP1), which were unrealistically specific and (in case of M00987) predicted stretches of poly(T).

For a few TFs we obtained more recent WMs from the literature (SP1, OCT4, NANOG, SOX2, XBP1, PRDM1, and the RXRG dimer) and we used these to replace the corresponding WM in the list.

We improved several motifs by running MotEvo on TF ChIP-seq data: SRF, STAT1/3, REST and ELK1/4/GABPA/GABPB1. Some other motifs were obtained by predicting *de novo* using the Phylogibbs algorithm [12] on ChIP-seq data: SPI1, CTCF, OCT4, SOX2 and NANOG.

For a few motifs JASPAR has recently updated or introduce new motifs which were based on high-throughput data and we included these motifs.This is the case for FOXA2, KLF4, EWSR1-FLI1, FEV, NR4A2. We also removed MA0118, as it had been discarded from the JASPAR data base.

Our final list contains 189 WMs. For each final WM there is an ordered list of associated human TFs, ordered by percent identity of the DBDs of TFs known to bind sites of the WM and the DBDs of the TF. We then checked this list of associations by hand and for each WM cut-off the list of associated human TFs manually. In total 340 human TFs are associated with our 189 WMs. The corresponding mouse orthologous TFs were selected using the MGI data base [13]. The entire set of WMs and mapping to associated TFs is available from the SwissRegulon website (http://www.swissregulon.unibas.ch).

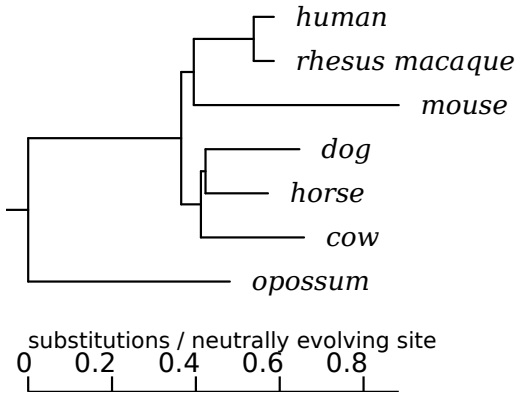
## 1.3 Transcription factor binding site predictions

After creating reference promoter sets and curating a set of mammalian regulatory motifs we next predicted TFBSs in the proximal promoter regions of each promoter. Analysis of sequence conservation in the neighborhood of TSSs (see [3]) and experimentation with TFBS prediction in regions of different lengths around TSSs indicated that a reasonable balance between sensitivity (i.e. including relevant binding sites) and specificity (avoiding too many false positive predictions) can be obtained by predicting TFBSs in a 1 kilobase region around the TSSs of each promoter.

For each promoter, we thus extended the promoter sequence spanned by its cluster of TSSs by 500 bp upstream and 500 bp downstream. We denote this as the *proximal promoter region* of a promoter. We then extracted the sequence of the reference species, i.e. human or mouse and orthologous regions from 6 other mammals (human or mouse, rhesus macaque, cow, dog, horse, and opossum) using pairwise BLASTZ[14] alignments. For each promoter, we multiply aligned the orthologous regions using T-Coffee [15].

To obtain a phylogenetic tree for these mammalian species, with branch lengths corresponding to the expected number of substitutions per neutrally evolving site, we used methods described previously[16]. Briefly, we first obtained the topology of the tree from the UCSC Genome Browser[17]. Then, for each pair of species we made pairwise alignments of the coding regions of orthologous genes and extracted all third positions in fourfold-degenerate codons of amino acids that are conserved between the two species. Using these fourfold-degenerate positions we estimated a pairwise distance for each pair of

species. Finally, we estimated the lengths of the branches in the phylogenetic tree as those that minimize the square-deviations between the implied pairwise distances and the pairwise distances estimated from the fourfold-degenerate positions. The resulting tree structure is shown in Suppl. Fig. 1.




Supplementary Figure 1: The phylogenetic tree used by MotEvo for the transcription factor binding site predictions that are used by ISMARA.

The multiple sequence alignments were then used together with the phylogenetic tree and the collection of WMs as an input for TFBS predictions using the MotEvo algorithm[11]. Given a multiple alignment, MotEvo considers all ways in which the sequence of the reference species can be segmented into 'background' positions, 'binding sites' for one of the supplied WMs, and 'unknown functional elements' (UFEs). The likelihood of alignment columns assigned to background are calculated under a model of neutral evolution along the specified phylogenetic tree. The likelihood of alignment segments assigned to be a site for a given WM are calculated by first estimating which of the species have retained a site for the WM (based on the WM scores of the individual sequences) and then applying an evolutionary model in which substitution rates are set so as to match the sequence preferences of the WM. Finally, segments assigned to be UFEs are assumed to evolve under *unknown* purifying selection constraints on the sequence, which is implemented by treating them as sites for an unknown WM. Each unknown WM column is a nuisance parameter that is integrated out of the likelihood. Finally, MotEvo assigns, at each position of the alignment and for each WM, a posterior probability that a site for the corresponding WM occurs at this position.

Since most motifs show clear positional preferences relative to TSS, we implemented, separately for each motif, a distribution of position-dependent prior probabilities of site occurrence as a function of position relative to the TSS and fitted these distributions by maximum likelihood using expectation-maximization. In addition, since high-CpG and low-CpG promoters have highly distinct configurations of TFBSs, we estimated the position-dependent prior probability distributions separately for high-CpG and low-CpG promoters.

The final result of this analysis is a matrix $\mathbf{N}$, with $N_{pm}$ the total number of predicted sites for motif $m$ in promoter $p$, i.e. the sum of the posterior probabilities of the individual sites. To reduce the probability of spurious predictions, we set $N_{pm} = 0$ whenever the sum of the posteriors of all sites combined was less than $0.1$.

## 1.4 Associating miRNA target sites with each promoter

Apart from incorporating the effects of TFBSs in promoters, ISMARA also integrates the effects of miRNAs in its modeling of expression levels. To this end, we needed to obtain a set of predicted miRNA

target sites for each promoter. We base our predictions on the miRNA target predictions of TargetScan using preferential conservation scoring (aggregate $P_{CT}$) [18] which has shown consistently high performance in various benchmark tests. As opposed to focusing on individual miRNAs, TargetScan groups miRNAs that have identical subsequences at positions 2 through 8 of the miRNA, i.e. the 2-7 seed region plus the 8th nucleotide, and provides predictions for each such seed motif. We will treat these seed motifs exactly like the regulatory motifs (WMs) for TFs, i.e. a miRNA seed motif can be associated with multiple miRNAs. TargetScan provides predictions for 86 mammalian miRNA seed motifs in total.

TargetScan $P_{CT}$ provides a score for each seed motif and each RefSeq transcript. To obtain a 'site count' $N_{pm}$ for the number of sites of miRNA seed motif $m$ associated with promoter $p$ we average the TargetScan $P_{CT}$ scores of all RefSeq transcripts associated with the promoter $p$. Finally, the miRNA seed motif site counts $N_{pm}$ are simply added as columns to the site count matrix $\mathbf{N}$ with site counts of TFBSs.

## 1.5 Expression data processing

When using expression data from oligonucleotide microarrays, the raw probe intensities are corrected for background and unspecific binding using the Bioconductor packages affy[19], oligo[20], and gcrma[21], depending on the type of the particular microarray used. The micro-arrays that are currently supported by ISMARA are listed in supplementary table 1.

| Microarray | Organism | Producer |
|---|---|---|
| HG-U133A | *Homo sapiens* | Affymetrix |
| HG-U133B | *Homo sapiens* | Affymetrix |
| HG-U133_Plus_2 | *Homo sapiens* | Affymetrix |
| HG-U133A_2 | *Homo sapiens* | Affymetrix |
| HuGene-1_0-st-v1 | *Homo sapiens* | Affymetrix |
| HuGene-1_1-st-v1 | *Homo sapiens* | Affymetrix |
| HuGene-2_0-st | *Homo sapiens* | Affymetrix |
| HuGene-2_1-st | *Homo sapiens* | Affymetrix |
| HT_HG-U133A | *Homo sapiens* | Affymetrix |
| HT_HG-U133B | *Homo sapiens* | Affymetrix |
| HT_HG-U133_Plus_PM | *Homo sapiens* | Affymetrix |
| Mouse430_2 | *Mus musculus* | Affymetrix |
| Mouse430A_2 | *Mus musculus* | Affymetrix |
| MOE430A | *Mus musculus* | Affymetrix |
| MOE430B | *Mus musculus* | Affymetrix |
| MoGene-1_0-st-v1 | *Mus musculus* | Affymetrix |
| MoGene-1_1-st-v1 | *Mus musculus* | Affymetrix |
| HT_MG-430A | *Mus musculus* | Affymetrix |
| HT_MG-430B | *Mus musculus* | Affymetrix |
| MG_U74Av2 | *Mus musculus* | Affymetrix |
| MG_U74Bv2 | *Mus musculus* | Affymetrix |
| MG_U74Cv2 | *Mus musculus* | Affymetrix |

Supplementary Table 1: Microarrays currently supported by ISMARA.

For its further analysis, ISMARA uses the logarithms of the probe intensities. For a given sample, the histogram of log-intensities is generally bimodal, with the modes corresponding to probes of non-expressed and expressed genes. The probes are classified as expressed or non-expressed in each sample

separately by fitting a two-component Gaussian mixture model to the log-intensity data using the Mclust R package[22, 23]. Probes that are consistently non-expressed are filtered out from further processing; a probe is considered not to be expressed if in *all* the samples the probability of it belonging to the expressed class is below $0.4$. Subsequently, the intensity values of the remaining probes are quantile normalized across all input samples.

Microarray probes can hybridize to multiple transcripts, belonging to different genes, or different isoforms of one gene, and we decided not to rely on transcript annotations of a micro-array producer. Instead, we comprehensively mapped the probe sequences to the set of all transcripts that are associated with our reference set of promoters. Note that we thus also ignore the annotation of probes into probe sets. To calculate the expression of a promoter we average the log-expression levels of all probes that map to one (or more) of the transcripts associated with the promoter (i.e. the start of the transcript is a member of the cluster of starts that defines the promoter). The expression level of the promoter is then a weighted average of the expression levels of these probes, where a probe that maps to $n$ different transcripts obtains a weight $1/n$. That is, in general, a probe can map to multiple transcripts.

When ISMARA uses RNA-seq for input expression data, it expects the RNA-seq data to be provided as genome alignments of the reads to the hg19 or mm9 genome assembles in BED or BAM format. The loci of the mapped reads are then intersected with the genome alignments of all transcripts that are associated with reference promoters. A read is associated with a particular transcript if its mapping falls entirely into its exons. Note that, more recent RNA-seq data in some cases involved reads that are so long that, frequently, the read overlaps two rather than a single exon of the transcript. To take this into account, recent mapping algorithms allow the start and end of the read to map to different genomic loci. The ISMARA pipe-line associates such a mapping with a given transcript when both the start and end piece map to one of its exons. In the future ISMARA may be extended to include the mapping of raw reads themselves.

To obtain an expression level for each promoter ISMARA calculates a weighted average over all reads mapping to the transcripts associated with the promoter. The weighting results from multiple mappings at two levels. Firstly, a read can map to multiple genomic loci and, secondly, a single locus may intersect multiple transcripts that are associated with multiple promoters. When a read maps to $n$ genomic loci, we assign a weight of $1/n$ to each locus. If that locus intersects transcripts of $m$ different promoters, then this read contributes a final weight of $1/(nm)$ to the expression of the transcript. Each transcript $t$ is assigned a total weight $w_t$ that consist of the sum of the weights of all reads mapping to it. Note that the expected value of $w_t$ is both proportional to the average number of mRNAs per cell this transcript $t$ has as well as proportional to the length $l_t$ of the transcript. The normalized weight $\tilde{w}_t = w_t/l_t$ is proportional to the number of mRNAs per cell of transcript $t$. The expression of a promoter $p$ is measured in terms of the total number of mRNAs deriving from this promoter. Thus, for each promoter $p$, we calculate a total weight $w_p$ by summing $\tilde{w}_t$ over all transcripts $t$ that are associated with the promoter, i.e. $w_p = \sum_t \tilde{w}_t$. We obtain such a weight $w_{ps}$ for each promoter $p$ and each sample $s$. Promoters that have weights $w_{ps} = 0$ in all samples are discarded. There will be some promoters that have zero weights in some, but not all, of the samples. In order to define log-expression values for all promoters we add a small pseudo-count to the weights $w_{ps}$. For each sample $s$, we rank the promoters with nonzero weight by their weight $w_{ps}$ and calculate the 5th percentile $pc_s$. We then add this weight $pc_s$ as a pseudo-count to all weights $w_{ps}$ of promoters, including promoters that had zero weights in sample $s$. Finally, we normalize the $w_{ps}$ and log-transform them as follows:

$$E_{ps} = \log_2 \left[ 10^6 \frac{w_{ps}}{\sum_{p'} w_{p's}} \right]. \tag{1}$$

Note that the resulting expression level $E_{ps}$ corresponds to the logarithm (base 2) of the number of mRNAs deriving from promoter $p$, per million mRNAs in the cell. Note that this weighting procedure for calculating promoter expression levels is robust to redundancy in the transcript sets. For example,

when a promoter is associated with $k$ highly overlapping transcripts, then a read mapping within the exons of these transcripts will get assigned to all these transcripts, with a weight $1/k$ for each. When the total weight $w_{ps}$ of the promoter is calculated, these $k$ are then summed back and will in the end contribute precisely 1 read.

## 1.6 ChIP-seq data processing

Apart from modeling expression dynamics, ISMARA can also process ChIP-seq data to automatically model chromatin state (or TF binding) changes at promoters genome-wide. Examples of such chromatin state data include histone occupancy, histone modifications, TF binding and DNAse1 hypersensitivity in promoter regions. After several experiments, we found that integrating the chromatin signal from a region of 2000 bps centered on the TSS of each promoter gives the most robust results. To obtain a chromatin state level $E_{ps}$ of promoter $p$ in sample $s$, we calculate the sum $r_{ps}$ of the reads that map entirely within this region around promoter $p$ and transform to the log-space after adding a pseudocount:

$$E_{ps} = \log_2 \left( r_{ps} + \frac{N_s l}{L} \right), \tag{2}$$

where the second term is a pseudo-count, $N_s$ is the total number of reads mapped to the genome in sample $s$ (the number of lines in the BED file), $l = 2000$ is the length of the regions, and $L$ is the total length of the genome. Note that this pseudo-count is precisely the number of reads that would be expected if all $N_s$ reads where distributed uniformly over the genome. We set to pseudo-count to this value to make the pseudo-count roughly of the same size as the read-count from background reads in regions where the chromatin mark in question does not appear. The rational is that, in regions where there are only background reads, statistical fluctuations may cause the read-counts $r_{ps}$ to change significantly from sample to sample. By adding a constant pseudo-count of roughly the same size, these fluctuations are effectively dampened. More formally, this pseudo-count results within a Bayesian context if we use a Dirichlet prior with an expected density $l/L$ for each region.

## 1.7 Motif activity fitting.

We model the log-expression (or ChIP-seq signal) value $E_{ps}$ of a promoter $p$ in sample $s$ as a linear function of the site-counts $N_{pm}$ for all motifs $m$ associated with the promoter, i.e. either TFBSs in the proximal promoter region or miRNA binding sites in the 3' UTRs of associated transcripts. In each sample $s$, the contribution of the sites $N_{pm}$ to $E_{ps}$ is given by the (unknown) *motif activity* $A_{ms}$. That is, we fit a model of the form:

$$E_{ps} = \tilde{c}_s + c_p + \sum_m N_{pm} A_{ms} + \text{noise}, \tag{3}$$

where $\tilde{c}_s$ and $c_p$ are sample and promoter-dependent constants.

The last 'noise' term corresponds to the difference between the signal that the model predicts, and the signal $E_{ps}$ that was actually measured. This difference generally results from multiple sources. First, there are measurement errors in $E_{ps}$. Second, there is 'biological noise, i.e. uncontrolled fluctuations in the true state of the biological system. Third, and most importantly, there is the error in the model. Regarding the distribution of measurement errors and biological noise there has been a considerable amount of work in the literature. For microarray measurements, after background correction and normalization, the sum of biological and measurement noise in log-expression levels can be reasonably approximated by a Gaussian. For next-generation sequencing data such as RNA-seq and ChIP-seq data, which is intrinsically digital in nature, we have previously studied the distribution of biological and measurement noise using data from replicate experiments [3]. This analysis showed that on a normal (i.e.

non-logarithmic) scale, the noise distribution can be well approximated by Poisson sampling with a rate that is itself log-normally distribution (with some variance $\sigma^2$). As we showed in [3], in log-scale this distribution can be well-approximated by a normal distribution with a variance that is the sum of the variance $\sigma^2$ of the original log-normal, and a term $1/n$, where $n$ is the total read-count, which results from the Poisson sampling noise. An alternative model of biological and replicate noise that has been used in the literature is the negative binomial distribution [24, 25]. A negative binomial is obtained when there is Poisson sampling noise with a rate that is itself Gamma distributed. Like the distribution derived in [3], this distribution also has the property that, in log-scale, the contribution to the variance due to Poisson sampling decreases with absolute expression level.

However, as mentioned above, besides uncontrolled fluctuations in the state of the biological system and measurement noise, the 'noise' term in equation (3) also contains a contribution from the *error* of the model. That is, even if experimentalists could perfectly control the state of the biological system (i.e. no biological noise) and make measurements without any errors (i.e. no measurement noise) then, because of the simplicity of our model, there would still be a large difference between the predicted signal levels of each promoter, and the true signal levels. Indeed, our model typically only captures a modest fraction of the variance in expression and ChIP-seq levels, meaning that the error in the model is generally much larger than the biological and measurement noise. That is, the noise term in equation (3) is *dominated* by the error in the model. Consequently, the relevant noise distribution is not the distribution of biological and measurement noise, but the distribution of model errors. Since we have no specific information regarding the form of the distribution of modeling errors we will make the assumption that the noise is Gaussian distributed with an unknown variance $\sigma^2$ that is the same for all promoters and in all samples.

Under these assumptions we find the following expression for the likelihood of the expression data given the site-counts, motif activities and sample and promoter-dependent constants:

$$P\left(E \mid A, c, \tilde{c}, N, \sigma\right) \propto \prod_{p,s} \frac{1}{\sigma} \exp\left[-\frac{\left(E_{ps} - \tilde{c}_s - c_p - \sum_m N_{pm} A_{ms}\right)^2}{2\sigma^2}\right] \tag{4}$$

We first maximize this expression with respect to all the constants $c_p$ and $\tilde{c}_s$, and substitute these with their *maximum likelihood* estimates. After doing this we obtain:

$$P\left(E \;\middle|\; A^{'}, N, \sigma\right) \propto \sigma^{-PS} \exp\left[-\frac{\sum_{ps}\left(E^{'}_{ps} - \sum_m N^{'}_{pm} A^{'}_{ms}\right)^2}{2\sigma^2}\right], \tag{5}$$

where $P$ is the number of promoters, $S$ is the number of samples, the $N^{'}_{pm}$ are a motif-normalized site-counts $N^{'}_{pm} = N_{pm} - \langle N_m \rangle$, with $\langle N_m \rangle$ the average site-count per promoter for motif $m$, the $A'_{ms}$ are sample-normalized activities $A^{'}_{ms} = A_{ms} - \langle A_m \rangle$, i.e. with $\langle A_m \rangle$ the average activity of motif $m$ across the samples, and the $E^{'}_{ps}$ are sample- and promoter-normalized expression values $E^{'}_{ps} = E_{ps} - \langle E_p \rangle - \langle E_s \rangle + \langle\langle E \rangle\rangle$. That is the log-expression matrix $E^{'}_{ps}$ is normalized such that all its rows and columns sum to zero, the activities $A^{'}_{ms}$ are normalized such that the average activity over all samples is zero, i.e. $\sum_s A^{'}_{ms} = 0$, and the site-counts $N^{'}_{pm}$ are normalized such that the average count over all promoters is zero, i.e. $\sum_p N^{'}_{pm} = 0$.

To avoid over-fitting we assign a symmetric Gaussian prior to each motif activity, i.e. the joint prior for all activities is given by:

$$P\left(A^{'} \;\middle|\; \lambda, \sigma\right) \propto \prod_{ps} \exp\left[-\frac{\lambda^2}{2\sigma^2} \sum_m A^{'2}_{ms}\right], \tag{6}$$

where the constant $\lambda^2$ sets the width of prior distribution relative to the width of the likelihood function. Using this prior with the likelihood derived above, the posterior distribution of motif activities takes the form:

$$P\left(A^{'} \;\middle|\; E,N,\sigma,\tau\right) \propto \sigma^{-PS} \exp\left[-\frac{\sum_{p,s}\left(\left(E^{'}_{ps}-\sum_m N^{'}_{pm}A^{'}_{ms}\right)^2+\lambda^2\sum_m A^{'2}_{ms}\right)}{2\sigma^2}\right]. \quad (7)$$

Since equation (7) factorizes into independent expressions for the different samples, it is enough to consider one sample at a time. The posterior distribution for the motif activities in a particular sample takes the general form of a multi-variate Gaussian centered around $A^{'*}_{ms}$:

$$P\left(A^{'}_{s} \;\middle|\; E,N,\sigma\right) \propto \sigma^{-P} \exp\left[-\frac{\sum_{m\tilde{m}}\left(A^{'}_{ms}-A^{'*}_{ms}\right)W_{m\tilde{m}}\left(A^{'}_{\tilde{m}s}-A^{'*}_{\tilde{m}s}\right)+\chi^2_s}{2\sigma^2}\right], \quad (8)$$

where the $\chi^2_s$ is the unexplained part of variance in sample $s$

$$\chi^2_s = \sum_p \left(E^{'}_{ps}-\sum_m N^{'}_{pm}A^{'*}_{ms}\right)^2, \quad (9)$$

and the matrix $W$ is given by

$$W_{m\tilde{m}} = \sum_p \left(N^{'}_{pm}N^{'}_{p\tilde{m}}+\lambda^2\delta_{m\tilde{m}}\right). \quad (10)$$

Finally, the *maximum a posteriori* (MAP) estimates $A^{'*}_{ms}$ can be found by minimizing the expression in the numerator of equation (7) using standard numerical procedures for ridge regression. ISMARA performs this calculation by singular value decomposition of the $N^{'}$ matrix.

### 1.7.1 Setting $\lambda$ through cross-validation

Both the MAP estimates $A^{'*}_{ms}$, and the matrix $W_{m\tilde{m}}$ are functions of $\lambda$. The constant $\lambda^2$ represents the ratio between the *a priori* expected variance of activities, to the average squared-deviation of the model from the expression data (which results from both error in the model, noise in the expression measurements, and biological noise). In general $\lambda$ will depend on the measurement platform used, i.e. microarray, RNA-seq, or ChIP-seq, and also on the samples used, because the true variance in motif activities will depend on the variance in the $E_{ps}$ across the samples. Thus, the appropriate value of $\lambda$ will generally not be known in advance and ISMARA therefore includes a method for automatically setting $\lambda$ from the data. To determine the optimal $\lambda$ ISMARA uses a $80/20$ cross-validation scheme. The set of promoters is divided randomly into two sets, with one containing $80\%$ of all promoters (the 'training set') and the other the remaining $20\%$ (the 'test set'). The training set of promoters is used for fitting the motif activities while the quality of the fit is evaluated on the test set. ISMARA then finds the value of $\lambda$ that minimizes the average squared-deviation of the expression levels in the test set from those predicted by the model. We denote this optimal value of $\lambda$ by $\lambda^*$.

### 1.7.2 Error bars on motif activities

Apart from the MAP estimates $A^{'*}_{ms}$ ISMARA also determines the uncertainties associated with these estimates. Since $\sigma$ in Eq. 8 is not known, we integrate it out with a suitable scale-invariant prior $P(\sigma) \propto \frac{1}{\sigma}$.

$$
\begin{aligned}
P\left(A_s^{'} \;\middle|\; E,N,\lambda\right) &= \int_{\sigma=0}^{\infty} P\left(A_s^{'} \;\middle|\; E,N,\sigma,\lambda\right) P(\sigma)\mathbf{d}\sigma \\
&\propto \frac{\Gamma\left(\frac{P}{2}\right)}{\left[\sum_{m\tilde{m}}\left(A'_{ms}-A'^{*}_{ms}\right)W_{m\tilde{m}}\left(A'_{\tilde{m}s}-A'^{*}_{\tilde{m}s}\right)+\chi_s^2\right]^{\frac{P}{2}}} \\
&\propto \exp\left[-\frac{P\sum_{m\tilde{m}}\left(A_{ms}^{'}-A_{ms}^{'*}\right)W_{m\tilde{m}}\left(A_{\tilde{m}s}^{'}-A_{\tilde{m}s}^{'*}\right)}{2\chi_s^2}\right],
\end{aligned}
\tag{11}
$$

where the last proportionality is a very good approximation when the number of promoters is large. Note that this is again a multi-variate Gaussian distribution. The covariance matrix of this Gaussian posterior distribution is given by:

$$
C_{m\tilde{m};s} = \frac{\left(W^{-1}\right)_{m\tilde{m}}\chi_s^2}{P} \tag{12}
$$

As is well known, given this multi-variation Gaussian form, the marginal distribution for a single motif activity $A_{ms}^{'}$ will be Gaussian distributed with standard-deviations $\delta A_{ms}^{'}$ given by the square root of the corresponding diagonal term of the covariance matrix, i.e.

$$
\delta A_{ms}^{'} = \sqrt{C_{mm;s}} \tag{13}
$$

We define the overall *significance* of a motif $m$ as the average squared ratio between fitted activities and their standard deviations ($z$-values)

$$
z_m = \sqrt{\frac{1}{S}\sum_s\left(\frac{A'^{*}_{ms}}{\delta A'_{ms}}\right)^2}. \tag{14}
$$

### 1.7.3 Fitting mean activities

By introducing a promoter-dependent basal expression level $c_p$ in equation (3) we effectively ensure that the average expression of each promoter is accounted for, i.e. only the *changes* in expression of each promoter across the samples are fitted by the motif activities $A_{ms}^{'}$. Consequently, the fitted motif activities all average to zero, i.e. $\sum_s A_{ms}^{'} = 0$. Although, typically, users would indeed be most interested in explaining expression *changes* across the samples, in some cases users might also be interested in knowing to what extent the absolute *average* levels of the promoters across the samples can be fit in terms of 'mean activities' of the motifs, i.e. to learn which motifs are most predictive of consistently high or low absolute expression across the replicates.

To fit mean activities we start from equation (3) and set $c_p = 0$ for all promoters $p$. In addition, we explicitly write the activity in terms of a sample-dependent part that averages to zero, and a mean activity, i.e.

$$
A_{ms} = A_{ms}^{'} + \bar{A}_m. \tag{15}
$$

Defining the sample-corrected average expression values as

$$
\tilde{E}_p = \frac{1}{S}\sum_s\left(E_{ps} - \langle E_s\rangle\right), \tag{16}
$$

and again the motif-normalized site counts $N_{pm}^{'} = N_{pm} - \langle N_m\rangle$, it is straight-forward to show that the mean activities $\bar{A}_m$ are optimized when the expression

$$
\tilde{\chi}^2 = \sum_p\left(\tilde{E}_p - \sum_m N_{pm}^{'}\bar{A}_m\right)^2, \tag{17}
$$

is minimized. We fit the mean activities $\bar{A}_m$ in exact analogy with the fitting of the activities $A'_{ms}$. We introduce a separate Gaussian prior for the mean activities $\bar{A}_m$, with its own parameter $\tilde{\lambda}$, and again set $\tilde{\lambda}$ using $80/20$ cross-validation. We also determine error-bars $\delta\bar{A}_m$ on the mean activities $\bar{A}_m$. Finally, we also define $z$-scores for the mean activities, i.e.

$$\tilde{z}_m = \frac{\bar{A}_m}{\delta\bar{A}_m}. \tag{18}$$

Motifs with the highest positive $\tilde{z}_m$ are the most significant predictors of consistently high expression across the samples, whereas motifs with highly negative $\tilde{z}_m$ are the most significant predictors of consistently low expression across the samples.

## 1.8 Processing of replicates

Careful studies typically involve experimental replicates to account for the part of variability in the readout which is not under direct experimental control. ISMARA allows users to indicate which samples correspond to replicates and will automatically calculate averaged motif activities and error bars across these replicates. To perform this analysis the user should first upload all samples and perform the standard ISMARA analysis. On the results page ISMARA provides a link to a page where users can interactively annotate which samples are replicates. In addition, if the replicates came in clearly defined batches, for example, when a time-course was performed multiple times, then the user can also indicate this. Once all samples are annotated ISMARA can then perform motif activity averaging across the replicates. Note that this approach can easily be extended beyond replicates, i.e. the user can arbitrarily divide the samples into groups and ISMARA will automatically calculate average motif activities and associated standard-deviations for each group of samples.

Here we describe how activities within a group are averaged. For a given group $G$ of samples and a particular motif, we assume that its activities $A_s$ in samples $s \in G$ are given by a mean activity $\bar{A}^g$ plus some deviation $\delta_s$, i.e

$$A_s = \bar{A}^g + \delta_s, \tag{19}$$

where we assume that the prior probability of $\delta_s$ is Gaussian distributed with (unknown) standard-deviation $\sigma_g$, i.e

$$P(\delta_s|\sigma_g) = \frac{1}{\sqrt{2\pi}\sigma_g} \exp\left[-\frac{1}{2}\frac{\delta_s^2}{\sigma_g^2}\right]. \tag{20}$$

Thus, given the mean activity $\bar{A}^g$ in the group, the prior probability to have activity $A_s$ in a particular sample $s$ from the group is

$$P(A_s|\bar{A}^g, \sigma_g) = \frac{1}{\sqrt{2\pi}\sigma_g} \exp\left[-\frac{1}{2}\frac{\left(A_s - \bar{A}^g\right)^2}{\sigma_g^2}\right]. \tag{21}$$

Using the input data, ISMARA has inferred the motif activity $A_s$ to have expected value $A_s^*$ with standard-error $\delta A_s$ for each sample $s$. That is, once the dependence on all other activities is integrated out, the probability of the expression data $D$ conditioned on the motif activity $A_s$ is a Gaussian with standard-deviation $\delta A_s$, i.e.

$$P(D|A_s) = \frac{1}{\sqrt{2\pi}\delta A_s} \exp\left[-\frac{1}{2}\frac{(A_s - A_s^*)^2}{(\delta A_s)^2}\right]. \tag{22}$$

Using the expressions for $P(D|A_s)$ and $P(A_s|\bar{A}^g, \sigma_g)$ we can calculate the probability of the data $D$ given the mean activity $\bar{A}^g$ and standard-deviation $\sigma_t$ by integrating over all unknown $A_s$:

$$P(D|\bar{A}^g, \sigma_g) = \prod_{s\in G}\left[\int_{-\infty}^{\infty} P(D|A_s)P(A_s|\bar{A}^g, \sigma_g)\mathbf{d}A_s\right]. \tag{23}$$

These integrals can be performed analytically and we obtain

$$P(D|\bar{A}^g, \sigma_g) = \prod_{s \in G} \frac{1}{\sqrt{2\pi(\sigma_g^2 + \sigma_s^2)}} \exp\left[-\frac{(A_s^* - \bar{A}^g)^2}{2(\sigma_g^2 + \sigma_s^2)}\right]. \tag{24}$$

Although, formally, we should integrate this expression over the unknown standard-deviation $\sigma_g$ as well, this integral unfortunately cannot be performed analytically. Therefore, we estimate the integral simply by finding the value $\sigma_g^*$ that maximizes $P(D|\bar{A}^g, \sigma_g)$. Assuming a uniform prior for the mean activity $\bar{A}^g$ of the samples in the group, we then finally obtain an expression for the posterior probability $P(\bar{A}^g|D)$ which we characterize by its mean $\langle \bar{A}^g \rangle$ and standard-deviation $\delta\bar{A}^g$. That is, $\langle \bar{A}^g \rangle$ is the inferred average motif activity for the samples within the group, and $\delta\bar{A}^g$ is the error-bar on this average activity. This mean and error-bar of the activity for the 'group' of samples are given by

$$\langle \bar{A}^g \rangle = \frac{\sum_{s \in G} \frac{A_s^*}{(\sigma_g^*)^2 + \sigma_s^2}}{\sum_{s \in G} \frac{1}{(\sigma_g^*)^2 + \sigma_s^2}}, \tag{25}$$

and

$$\delta\bar{A}^g = \sqrt{\frac{1}{\sum_{s \in G} \frac{1}{(\sigma_g^*)^2 + \sigma_s^2}}}. \tag{26}$$

Finally, we assign significances $z_m$ to each motif completely analogously as before, but now averaging over all groups, i.e.

$$z_m = \sqrt{\frac{1}{|G|} \sum_g \left(\frac{\langle \bar{A}^g \rangle}{\delta\bar{A}^g}\right)^2}, \tag{27}$$

where $|G|$ is the number of groups. A motif will have a high significance $z_m$ when its motif activities vary relatively little within each group, and vary by a large amount across groups.

## 1.9 Target predictions

In order to infer motif activities $A_{ms}$, ISMARA assumes that all promoters with predicted target sites for a motif $m$ will respond to changes in motif activity, i.e. in proportion to the predicted number of sites $N_{pm}$. This is a reasonable assumption when inferring motif activities, as the activities $A_{ms}$ depend on the statistics of all promoters with sites for motif $m$. However, in a given condition or system, it is likely that only a subset of the promoters with sites for a motif $m$ are in fact regulated by this regulator. This might be due to a limited accessibility, dependence on particular co-factors, weaker affinity of a site, and other context-dependent factors. Thus, when we aim to predict individual target promoters of a given motif $m$, we not only use the binding site predictions $N_{pm}$, but also evaluate at which promoters the activities $A_{ms}$ contribute to explaining the profiles $E_{ps}$.

To quantify if a given promoter $p$ is targeted by a motif of interest $m$ we first demand that there exists a TFBS prediction, i.e $N_{pm} > 0$. Second, we quantify the contribution of $m$ to the fit of the expression/chromatin state profile $E_{ps}$. The most rigorous approach to quantifying the effect of motif $m$ on promoter $p$ is to calculate both the probability of the entire data set, i.e. the profiles $E_{ps}$ across all promoters and samples, with the original site-count matrix $\mathbf{N}$, and a site-count matrix $\tilde{\mathbf{N}}$ where only the sites for motif $m$ in promoter $p$ are set to zero. To calculate this probability we treat all the unknown motif activities $A_{ms}$ as well as the standard-deviation $\sigma$ as nuisance parameters that are integrated out of the likelihood. That is, we formally want to calculate the ratio of probabilities

$$R_{pm} = \frac{\int_{-\infty}^{\infty} \mathbf{d}A \int_0^{\infty} \mathbf{d}\sigma P(E|N, A, \sigma)}{\int_{-\infty}^{\infty} \mathbf{d}A \int_0^{\infty} \mathbf{d}\sigma P(E|\tilde{N}, A, \sigma)}, \tag{28}$$

where the integrals are over all motif activities $A_{ms}$, and over the standard-deviations $\sigma$. Note that, when we set $N_{pm} = 0$ for promoter $p$ and motif $m$, we make a very small change to the site-count matrix. That is, as there are tens of thousands of promoters and close to 200 motifs, we are changing only one of the millions of entries in the matrix. As a consequence, the inferred motif activities $A^{'*}_{ms}$ that result from the mutated matrix $\tilde{N}$ are generally very close to those that result from the original matrix $N$. Similarly, the inverse covariance matrix $W$ of the mutated matrix is also very close to that of the original matrix and, finally, the optimal values of the constants $c_p$, $\tilde{c}_s$, and the prior constant $\lambda^*$ will also change very little under mutation of the matrix. To make the calculation more tractable we will make the approximation that all these quantities are *unchanged* upon mutation of the matrix. Under that approximation we have

$$P(E|A,N,\sigma,\lambda^*) \propto \sigma^{-PS} \exp\left[-\frac{\sum_{s,m,\tilde{m}} \left(A^{'}_{ms} - A^{'*}_{ms}\right) W_{m\tilde{m}} \left(A^{'}_{\tilde{m}s} - A^{'*}_{\tilde{m}s}\right) + \sum_{p,s} \chi^2_{ps}}{2\sigma^2}\right], \tag{29}$$

where $\chi^2_{ps}$ is the squared-deviation between the observed value $E'_{ps}$ and the predicted value, i.e.

$$\chi^2_{ps} = \left(E^{'}_{ps} - \sum_m N^{'}_{pm} A^{'*}_{ms}\right)^2. \tag{30}$$

For the probability of the data with the mutated site-count matrix we have

$$P(E|A,\tilde{N},\sigma,\lambda^*) = P(E|A,N,\sigma,\lambda^*) \exp\left[-\frac{\sum_s (\chi^2_{psm} - \chi^2_{ps})}{2\sigma^2}\right], \tag{31}$$

where $\chi^2_{psm}$ is the squared-deviation for promoter $p$ and sample $s$ when motif $m$ is removed, i.e.

$$\chi^2_{psm} = \left(E^{'}_{ps} - \sum_{m'} \tilde{N}^{'}_{pm'} A^{'*}_{m's}\right)^2. \tag{32}$$

In this form the integrals over the motif activities and $\sigma$ can be easily performed and we find for the ratio of the probabilities

$$R_{pm} = \left(\frac{\sum_{p',s} \chi^2_{p's}}{\sum_{p',s} \chi^2_{p's} - \sum_s (\chi^2_{psm} - \chi^2_{ps})}\right)^{S(P-M)}, \tag{33}$$

where $M$ is the total number of motifs. Since $P \gg M$ we approximate $P - M \approx P$ and we find approximately

$$R_{pm} = \exp\left[\frac{\sum_s (\chi^2_{psm} - \chi^2_{ps})}{\langle \chi^2 \rangle}\right], \tag{34}$$

where we have defined the average squared-deviation per sample/promoter combination

$$\langle \chi^2 \rangle = \frac{1}{PS} \sum_{p,s} \chi^2_{ps}, \tag{35}$$

and made use of the fact that $[1 - x/(SP)]^{-SP} \approx e^x$ for large $SP$.

In the results shown in the web-server we show, for each predicted target, the logarithm of the likelihood ratio, i.e. the score $S_{pm}$ for motif $m$ targeting promoter $p$ is

$$S_{pm} = \frac{\sum_s \chi^2_{psm} - \chi^2_{ps}}{\langle \chi^2 \rangle}. \tag{36}$$

Note that this result has a straightforward interpretation: The difference $\chi^2_{psm} - \chi^2_{ps}$ is the amount by which the square deviation between the predicted and observed signal increases when the sites for motif $m$ are removed from promoter $p$, and the ratio $(\chi^2_{psm} - \chi^2_{ps})/\langle\chi^2\rangle$ is the relative increase in square-deviation, i.e. relative to the average squared-deviation between the predicted and observed signals. The score $S_{pm}$ is obtained by summing this relative change in $\chi^2$ over all samples. By default ISMARA reports all target promoters for which this score is positive, i.e. where removing the motif from the promoter reduces the quality of the fit.

### 1.9.1 Enriched Gene Ontology categories

To analyze whether there are any Gene Ontology categories whose genes are over-represented among the targets of a motif, we use the "GO::TermFinder" Perl module[26]. The ontology files and associations between genes and categories were taken from the Gene Ontology (GO) Consortium web-site[27]. As a set of target genes for motif $m$ we include all genes associated with promoters that have a target score $S_{pm} > 0$. For microarray chips we create a background set from all the genes which have probes present on the microarray, i.e. according to our mappings of the probes (see Expression data processing). For RNA-seq data we take as a background set all genes associated with promoters which have mapped reads. In the web results we display all GO categories with a $p$-value of $0.05$ or less. These $p$-values are corrected for multiple testing using a simple Bonferroni correction, i.e. multiplied by the number of tests performed.

## 1.10 Principal component analysis of the activities explaining chromatin mark levels

We first performed standard ISMARA analysis on the $n = 10$ data sets measuring expression and 9 different chromatin marks (ChIP-seq), across $S = 8$ cell types [28]. For each motif $m$, and each mark $i$, we thus obtained estimated activities $A^i_{ms}$.

We performed principal component analysis (PCA) of the expression and chromatin mark levels across all promoters, separately for each cell type. For a given sample $s$, let $E_{pi}$ denote the level of mark $i$ at promoter $p$ (suppressing the label $s$ for notational simplicity). We have here already column normalized these levels, i.e.

$$\sum_p E_{pi} = 0, \tag{37}$$

for all marks $i$.

Using singular value decomposition, the matrix $E = U \cdot D \cdot V^T$ can be uniquely decomposed into an orthonormal matrix $U$ (of size $P \times n$), a diagonal positive-semidefinite matrix $D$ (of size $n \times n$), and an orthonormal matrix $V$ (of size $n \times n$) as:

$$E_{pi} = \sum_{k=1}^{n} U_{pk} D_{kk} V_{ik}, \tag{38}$$

where $k$ denotes the index of each component, the column vectors $\vec{V}_k$ with components $V_{ik}$ contain the principal components, and $D^2_{kk}$ is the fraction of the variance in the $E_{pi}$ values, i.e.

$$\mathrm{var}(E) = \frac{1}{nP} \sum_{p,i} (E_{pi})^2 , \tag{39}$$

that is explained by component $k$.

The first principal component $\vec{V}_1$, shown in the top panels of Suppl. Fig. 23, is virtually identical in all cell types and captures approximately $60\%$ of the collective behavior of the expression and 9 chromatin

marks (8 histone modification and CTCF binding) across promoters in each sample. As discussed in the main text, this first principal component appears to capture the combination of chromatin mark levels associated with the general 'activity' of a promoter. As a consequence, the effect of a given TF on a specific chromatin mark is confounded by its effect on general promoter activity and we therefore decided to subtract it from the activity profiles of all TFs.

For the purpose of removing the first principal component from the motif activities, we will treat each motif $m$ separately and ignore the covariances in the inferred motif activities, i.e. as we assumed previously when calculating the error bars on the motif activities in (13). We perform the removal one sample (cell line) at a time. A careful probabilistic analysis must be performed in order to calculate the error bars.

Let's focus on a given motif $m$ in sample $s$ and denote by $A$ the vector of activities across the marks, i.e. $A_i$ is the activity associated with mark $i$. In addition, let $\delta A_i$ denote the standard-deviation (error-bar) of this activity. The posterior distribution $P(A|D)$ of this activity vector given the data is given by a Gaussian, i.e. as in (12), of the form

$$P(A|D) \propto \exp\left[-\frac{1}{2}\sum_i \frac{(A_i - A_i^*)^2}{\delta A_i^2}\right], \tag{40}$$

where $A_i^*$ is the MAP estimate of the motif activity of mark $i$. If we introduce a diagonal matrix containing the inverse of the standard-deviation, we can write this expression in matrix-vector form:

$$P(A|D) \propto \exp\left[-\frac{1}{2}(A - A^*)^T \cdot \mathrm{diag}\left(\frac{1}{\delta A^2}\right) \cdot (A - A^*)\right], \tag{41}$$

where $A^*$ is a $n \times 1$ vector of the MAP estimates and $\mathrm{diag}\left(\frac{1}{\delta A^2}\right)$ is a $10 \times 10$ diagonal precision matrix which elements are set to the inverses of motif activity variances.

Using principal components $V$ of $E$ (38) and their orthonormality $V \cdot V^T = \mathbb{1}$ this distribution can be rewritten as

$$P(A|D) \propto \exp\left[-\frac{1}{2}(A - A^*)^T \cdot V \cdot V^T \cdot \mathrm{diag}\left(\frac{1}{\delta A^2}\right) \cdot V \cdot V^T \cdot (A - A^*)\right]. \tag{42}$$

We can rewrite the activities in the basis of the principal vectors as $B \equiv V^T \cdot (A - A^*)$ and the precision matrix in the same basis as $M \equiv V^T \cdot \mathrm{diag}\left(\frac{1}{\delta A^2}\right) \cdot V$. In this basis the probability distribution takes the form:

$$P(B|D) \propto \exp\left[-\frac{1}{2}B^T \cdot M \cdot B\right]. \tag{43}$$

Note that in this basis, the inverse covariance matrix $M$ contains off-diagonal terms.

We want to integrate out the activities along the first principal component, therefore we separate elements of $B$ and $M$ in the following way

$$B = \begin{pmatrix} b_1 \\ \begin{pmatrix} b_2 \\ \vdots \\ b_n \end{pmatrix} \end{pmatrix} \equiv \begin{pmatrix} b_1 \\ B_y \end{pmatrix} \tag{44}$$

$$M = \begin{pmatrix} m_{11} & \begin{pmatrix} m_{12} & \cdots & m_{1n} \end{pmatrix} \\ \begin{pmatrix} m_{21} \\ \vdots \\ m_{n1} \end{pmatrix} & \begin{pmatrix} m_{22} & \cdots & m_{2n} \\ \vdots & \ddots & \vdots \\ m_{n2} & \cdots & m_{nn} \end{pmatrix} \end{pmatrix} \equiv \begin{pmatrix} m_{11} & M_y^T \\ M_y & M_w \end{pmatrix}, \tag{45}$$

and the last equivalency holds because the matrix $M$ is symmetric.

Using these definitions, eq. (43) can be expanded and rewritten to obtain:

$$\begin{aligned} P(B|D) &\propto \exp\left[-\frac{1}{2}\left(b_1^2 m_{11} + 2b_1 B_y^T \cdot M_y + B_y^T \cdot M_w \cdot B_y\right)\right] \\ &= \exp\left[-\frac{1}{2}\left(m_{11}\left(b_1 + \frac{B_y^T \cdot M_y}{m_{11}}\right)^2 + B_y^T \cdot M_w \cdot B_y - \frac{B_y^T \cdot M_y \cdot M_y^T \cdot B_y}{m_{11}}\right)\right] \end{aligned} \tag{46}$$

Where we reordered terms and completed the square to bring out that this posterior is proportional to a Gaussian with respect to $b_1$. It is now straightforward to integrate this probability distribution along the first principal direction:

$$\begin{aligned} P(B_y|D) = \int_{b_1=-\infty}^{\infty} P(B|D)\mathbf{d}b_1 &\propto \exp\left[-\frac{1}{2}\left(B_y^T \cdot M_w \cdot B_y - \frac{B_y^T \cdot M_y \cdot M_y^T \cdot B_y}{m_{11}}\right)\right] \cdot \\ &\quad \cdot \int_{b_1=-\infty}^{\infty} \exp\left[-\frac{1}{2} m_{11}\left(b_1 + \frac{B_y^T \cdot M_y}{m_{11}}\right)^2\right] \mathbf{d}b_1 \\ &\propto \exp\left[-\frac{1}{2} B_y^T \cdot \left(M_w - \frac{M_y \cdot M_y^T}{m_{11}}\right) \cdot B_y\right], \end{aligned} \tag{47}$$

The last proportionality holds because the Gaussian integral yields a constant (with respect to $B_y$). Since the covariance matrix is the inverse of the precision matrix, the covariance matrix $W$ in the reduced $(n-1)$-dimensional space (i.e. without the first principal direction) has the form:

$$W = \left(M_w - \frac{M_y \cdot M_y^T}{m_{11}}\right)^{-1} \tag{48}$$

Finally, this covariance matrix $W$ needs to be transformed back from the principal component basis to the original basis. To this end we use the principal components contained in columns 2 through $n$ of the $V$ matrix. We obtain for the final covariance matrix $K$ in the original basis

$$K_{ij} = \sum_{k,l=2}^{n} V_{ik} W_{kl} V_{jl}. \tag{49}$$

The standard deviation of activities of the $i^{\text{th}}$ mark is given by square root of the corresponding diagonal element of this matrix

$$\delta \tilde{A}_i = \sqrt{K_{ii}}. \tag{50}$$

The corrected MAP activities are obtained by first defining

$$B^* = V^T \cdot A^*, \tag{51}$$

and then transforming back to the original basis using only the components along principal vectors 2 through $n$:

$$\tilde{A} = \sum_{k=2}^{n} V_{ik} B_k^*. \tag{52}$$

The reported z-value of the $i^{\text{th}}$ mark (we introduce back the indices for motif $m$ and sample $s$ omitted previously) is given by

$$z^i_{ms} = \frac{\tilde{A}_i}{\delta \tilde{A}_i} \tag{53}$$

After removing the contribution of the first principal component to the motif activities, we re-calculated significance $z$-values $z^i_m$ for each motif $m$ and each mark $i$ (x-axis in the Suppl. Fig. 24)

$$z^i_m = \sqrt{\frac{\sum_{s'} \left(z^i_{ms'}\right)^2}{S}}. \tag{54}$$

In addition, we calculated a specificity $s^i_m$ which measures the fraction of the overall significance that is associated with mark $i$ (y-axis in the Suppl. Fig. 24)

$$s^i_m = \frac{z^2_{mk}}{\sum_{k'} z^2_{mk'}}. \tag{55}$$

That is, a motif $m$ will be highly specific for mark $i$ if it has a high $z$-value $z^i_m$, and low $z$-values for all other marks.

# 2 Fraction of variance explained by the fit

The total variance $V$ in a data set is given by the sum of the squared normalized expression values

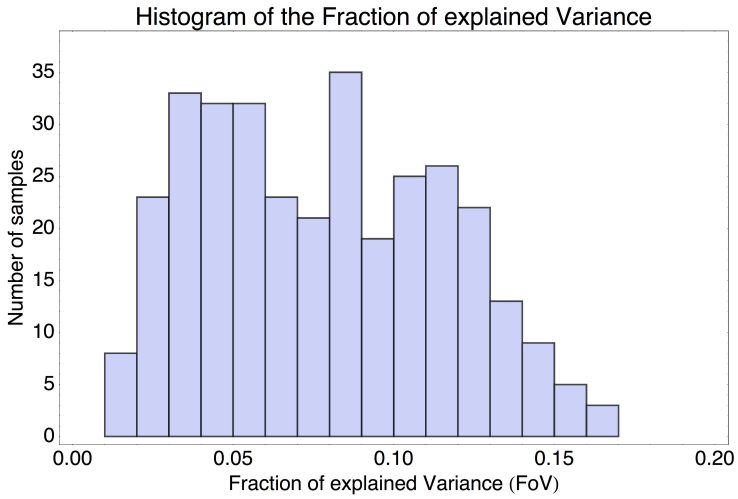
$$V = \frac{1}{PS} \sum_{p,s} (E'_{ps})^2. \tag{56}$$

After fitting the model, the average squared deviation left unexplained is given by the average of $\chi^2_{ps}$ across all promoters and samples, i.e. as defined by equations (30) and (35). The fraction of the variance $f$ explained by the fit is thus

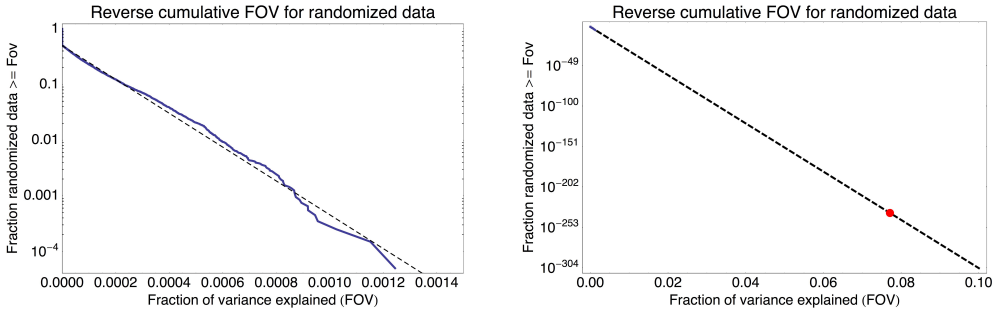
$$f = 1 - \frac{\langle \chi^2 \rangle}{V}. \tag{57}$$

For the data-sets that we analyze in this study, the fraction of explained variance ranges from slightly less than $2\%$ to almost $17\%$, with a median of $7.7\%$. Suppl. Fig. 2 shows a histogram of the fraction of variance explained across all samples.

For the first data-set, the Illumina Body Map 2, we find that $7.71\%$ of the variance is explained by the model. To assess the statistical significance of this fraction, we performed $10'000$ randomization experiments in which we randomized the association between promoter expression profiles $E_{ps}$ and site-counts $N_{pm}$, i.e. we randomly shuffled the rows of the matrix $\mathbf{N}$ while leaving the matrix $\mathbf{E}$ unchanged. For each of the $10'000$ randomizations, we then fitted the model, including fitting the parameter of the Gaussian prior through cross-validation so as to maximize the fraction of explained variance on the test-set. The left panel of Suppl. Fig. 3 shows the distribution of fraction of explained variance $f$ for the $10'000$ randomizations. As the figure shows, there is a roughly exponential distribution of $f$ and the highest observed $f$ was $f = 0.0012$. If we extent the exponential fit to the distribution of $f$ values in randomization experiments (right panel of Suppl. Fig. 3) we see that the observed fraction of variance $f = 0.0771$ on the unshuffled promoters corresponds roughly to a $p$-value of $1.3 * 10^{-235}$.




Supplementary Figure 2: Histogram of the fraction of variance explained for all the gene expression samples analyzed in this study (data-sets $1$ through $5$ and data-set $6.1$).

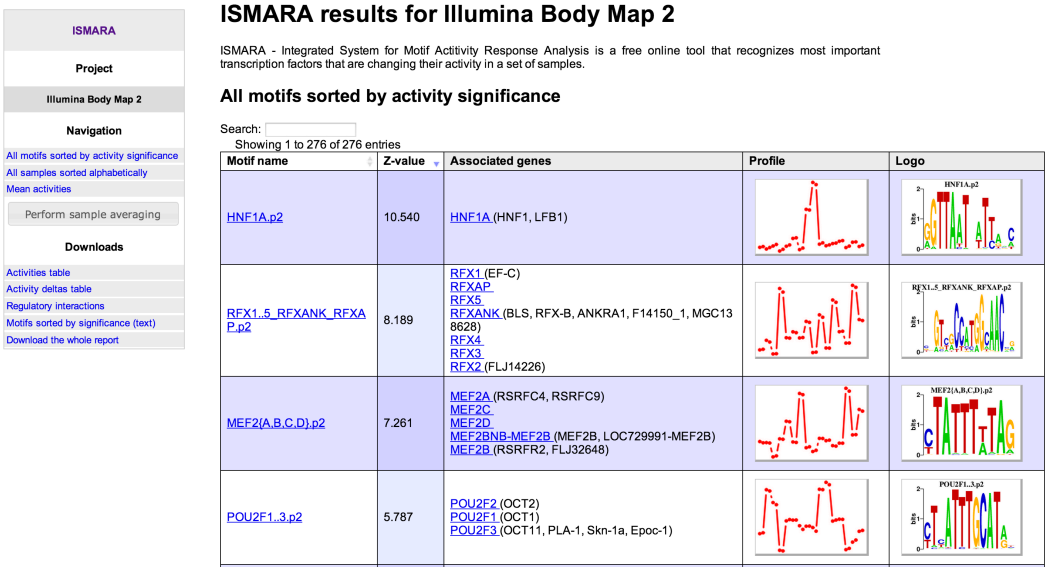



Supplementary Figure 3: **Left panel:** Reverse-cumulative distribution of the fraction of explained variance $f$ on the Illumina Body Map 2 data-set on $10'000$ randomizations between promoter expression and site-counts (solid line). The dashed line shows an exponential fit. **Right panel:** The exponential fit of the left panel extended to the observed fraction of explained variance ($f = 0.0771$, red dot) of the original, i.e. non-randomized, data-set. The estimated $p$-value of the observed fraction of variance is approximately $1.3 * 10^{-235}$.

# 3 Overview of results presented in the web-interface

To illustrate the results that ISMARA provides, we here present a number of figure that show examples of results on the RNA-seq data of the Illumina Body Map 2 [29]. Note that almost all of these figures are screen shots from the actual web-interface. All the full results for the Illumina Body Map are available at `http://ismara.unibas.ch/supp/dataset1_IBM/ismara_report/`.

The main page of results that ISMARA provides for a given data set centers around a list of motifs, sorted by their significance, showing for each motif its significance, the associated TFs, a sequence logo of the motif, and a thumbnail image of its inferred activity across the samples. Supplementary Fig. 4 shows an excerpt from this list of motifs.



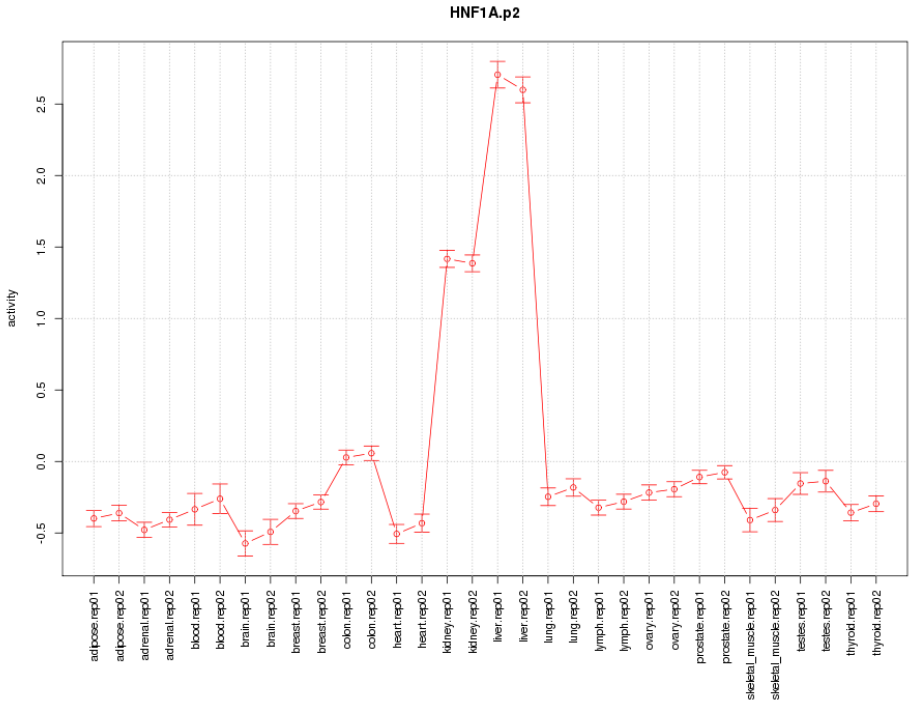

Supplementary Figure 4: Fragment of the list of regulatory motifs sorted by their significance ($z$-score). The motifs are sorted from top to bottom. Shown for each motif are, from left to right, the name of the motif (which is a link to a separate page with results for the motif), its $z$-score, a list of associated TFs (links to NCBI pages for these genes), a thumbnail of the inferred motif activity profile, and the sequence logo of the motif.

Each motif name in this list is in fact a link to a separate page with much more extensive results for the motif. Among these more extensive results is, first of all, a figure showing the inferred motif activity (and error bars) across all samples, where the samples are ordered according from left to right, according to the user's input.

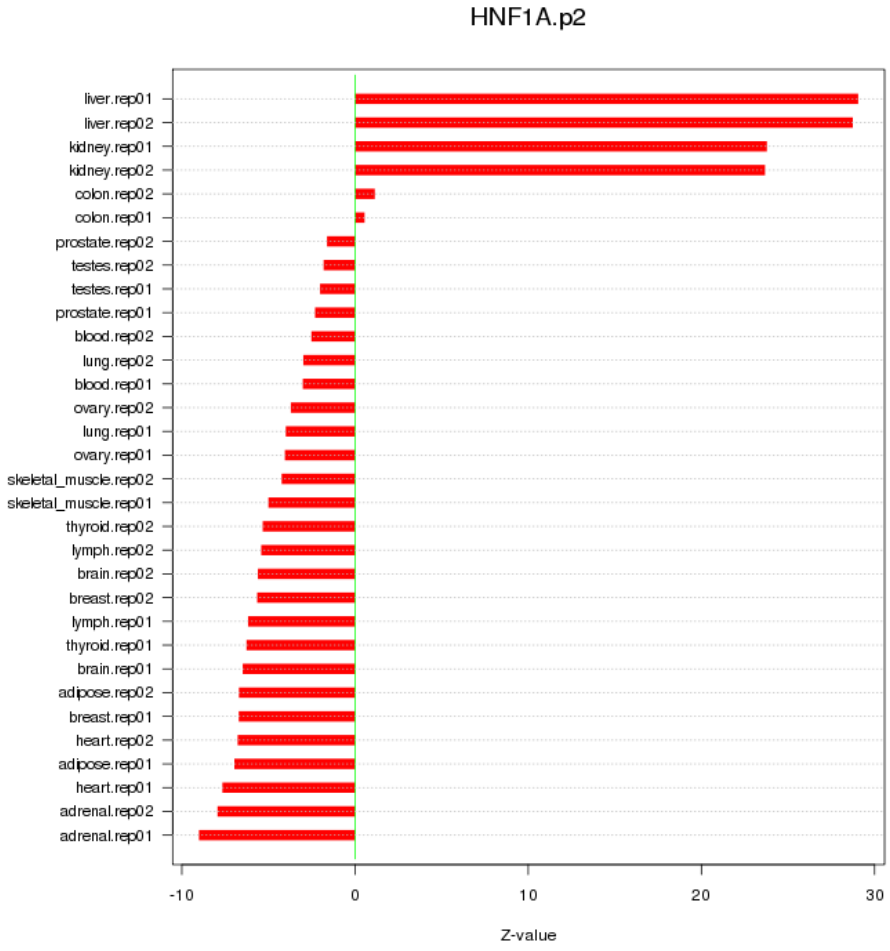
Supplementary Fig. 5 shows the activity profile of the HNF1A motif across the Illumina Body Map 2 samples. Note that such a lexicographic ordering of motif activities across samples is especially helpful when the samples come from a time course, in which case the graph shows the motif activity across time.

However, in many cases, including the Illumina Body Map analyzed here, there is no preferred natural ordering of the samples. In those cases it is more natural to present the motif activities with samples sorted from those in which the motif is most significantly upregulated, to those where it is most significantly downregulated. ISMARA provides such a list of motif $z$-values, with samples sorted from largest to smallest $z$-value, as shown in Suppl. Fig. 6 for the HNF1A motif. In this case, HNF1A activity is highly specific to liver and kidney.

For many of the motifs incorporated into the ISMARA analysis, there is more than one TF that can




Supplementary Figure 5: Inferred activities of the HNF1A motif on the tissues of the Illumina Body Map 2. The samples are ordered, from left to right, in lexicographic order, according to sample names input by the user. Note that this causes the two replicate samples from the same tissue to appear next to each other. The red circles show the estimated activities $A^*_{ms}$ and the error-bars $\delta A_{ms}$ are shown as red vertical bars. Samples names are indicated on the bottom.




Supplementary Figure 6: Sorted list of $z$-values for the HNF1A motif across all samples of the Illumina Body Map 2. Note that the replicate samples from liver and kidney have much higher $z$-value then all other samples.

potentially bind to sites for the motif. As a consequence, it is not always clear which individual TFs are responsible for the observed motif activity in a particular system. To help determine which TFs are most likely involved in the activity of a given motif, ISMARA provides an analysis of the correlation of motif activity and mRNA expression of the associated TFs. In particular, a table is provided showing the Pearson correlation between the motif's activity profile and the mRNA expression profiles of each of the TFs that can bind to the sites of the motif. The TFs in the list are sorted by their $p$-value. Supplementary Fig. 7 shows the list of correlations for the POU2F TFs.

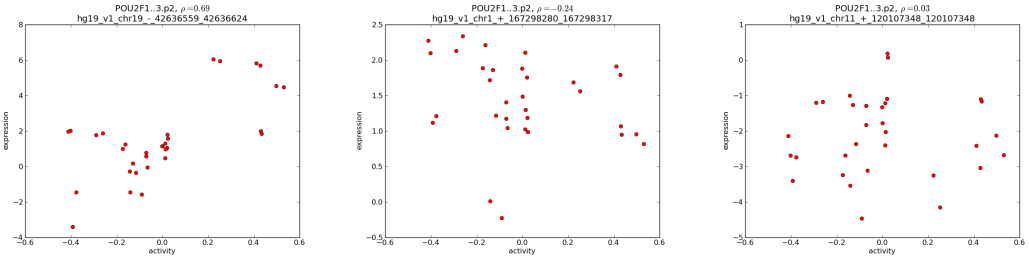
**Activity-expression correlation:**

| Gene Symbol | Promoter | Pearson corr. coef. | P-value | Plot |
|---|---|---|---|---|
| POU2F2 | hg19_v1_chr19_-_42636559_42636624 | 0.69 | 1.2e-05 | Click! |
| POU2F1 | hg19_v1_chr1_+_167298280_167298317 | -0.24 | 1.9e-01 | Click! |
| POU2F3 | hg19_v1_chr11_+_120107348_120107348 | 0.03 | 8.5e-01 | Click! |

Supplementary Figure 7: Correlations between the HNF1A motif activity and mRNA expression profiles of TFs that can bind to sites of the motif. The table shows the names of the associated TF genes, the IDs of the associated promoters of these genes, the Pearson correlation coefficient, the $p$-value for the correlation, and a link to a figure showing a scatter of the motif activity and mRNA expression levels across the samples (Suppl. Fig. 8) below.

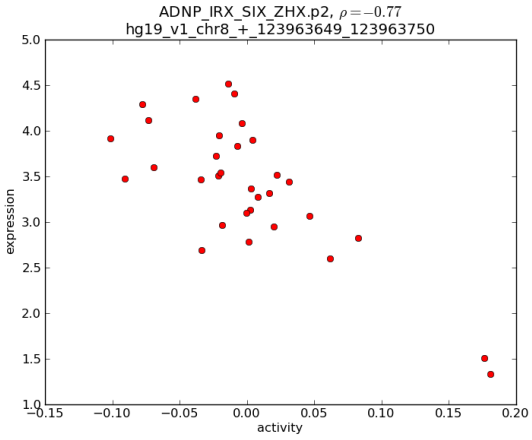
For each of the correlations a link is also provided to a scatter plot showing the mRNA expression levels and motif activities across the samples. Supplementary Fig. 8 shows example scatter plots for the TFs POU2F1, POU2F2, and POU2F3. Note that only POU2F2's expression is significantly correlated with the motif activity, suggesting that it is this TF that is mainly responsible for the motif activity in these samples. In addition, the fact that the TF's mRNA expression correlates *positively* with motif activity strongly suggests that this TF act as an activator, i.e. as its mRNA levels go up, the expression of target genes is affected positively.

To show an example of the opposite behavior, Suppl. Fig. 9 shows the mRNA expression levels of the TF ZHX2 against its inferred motif activities across the Illumina Body Map 2 samples. The clear negative correlation strongly suggests that ZHX2 acts as a *repressor* of its targets, and this matches what has been reported in the literature [30].




Supplementary Figure 8: Example scatter plots showing the correlations between HNF1A motif activity and the mRNA expression of POU2F2 left panel), POU2F1 (middle panel), and POU2F3 (right panel) TFs, across the samples of the Illumina Body Map 2. Each dot corresponds to one sample. The estimated expression levels correspond to the $log_2$ of the number of mRNAs per million mRNAs. At the top of the panel the Pearson correlation coefficient $\rho$ and the ID of the promoter are shown.

The next important information provided for each motif, is a predicted list of target promoters. ISMARA provides the target promoters $p$ for a motif $m$ sorted by their target score $S_{pm}$ (see section 1.9). As an example, the list of top targets for the HNF1A motif is shown in Suppl. Fig. 10. Each row in




Supplementary Figure 9: Scatter plots showing the correlation between the ADNP_IRX_SIX_ZHX motif activity and the mRNA expression of the ZHX2 TF, across the samples of the Illumina Body Map 2. Each dot corresponds to one sample. The expression levels are shown on a logarithmic scale. At the top of the panel the Pearson correlation coefficient $\rho$ and the ID of the promoter are shown.

the table corresponds to one target promoter and information shown includes the promoter ID, its score $S_{pm}$, associated transcripts and Entrez gene symbol, and the gene's name. Note that all these pieces of information are links that take the user to additional information on the promoter, the associated transcripts, and the gene. To keep the page easily viewable, by default only the top 20 targets are shown. However, the user can interactively change the number of targets shown in the list. In addition, a search box allows the user to search whether a particular promoter, transcript, or gene of interest occurs within the full list of targets.

Of particular interest is the additional information provided about each promoter, through the links with the promoter IDs. Following this link takes the user to the genome browser of our SwissRegulon database [31], showing the section containing the proximal promoter region (500 base pairs up-stream and down-stream of the major TSS of the promoter). In this browser the user is shown all the predicted TFBSs that are used by ISMARA in its modeling of expression or ChIP-seq data. This thus allows the user to determine the precise locations of the TFBSs on the genome, through which a particular TF is predicted to target a given promoter. Supplementary Fig. 11 shows, as an example, the promoter of the Albumin gene, which is the among the top 10 targets of HNF1A and is in fact a well-known target gene of HNF1A.

Beyond a list of individual targets, a user would typically like to gain some intuition of the pathways and particular biological processes that are targeted by a particular motif. One way of visualizing the functional structure of the predicted targets of a motif, is to represent these as a network, with links between pairs of genes that are known to be functionally related. The STRING database [32] maintains a curated collection of functional links between proteins, where 'functional link' can range from direct physical interaction, to over-representation of the protein pair within abstracts of scientific articles. For any set of proteins, STRING provides visualizations of the network of known functional interactions between these proteins, which visually brings out groups of proteins known to be functionally related. ISMARA provides such a STRING network picture for the targets of each motif (for visibility at most the top 200 targets are shown). Supplementary Fig. 12 shows the STRING network for the predicted targets of HNF1A. Note that the picture is itself a link to the STRING database, where the figure is interactive and allows the user more detailed information on each of the proteins in the network and each

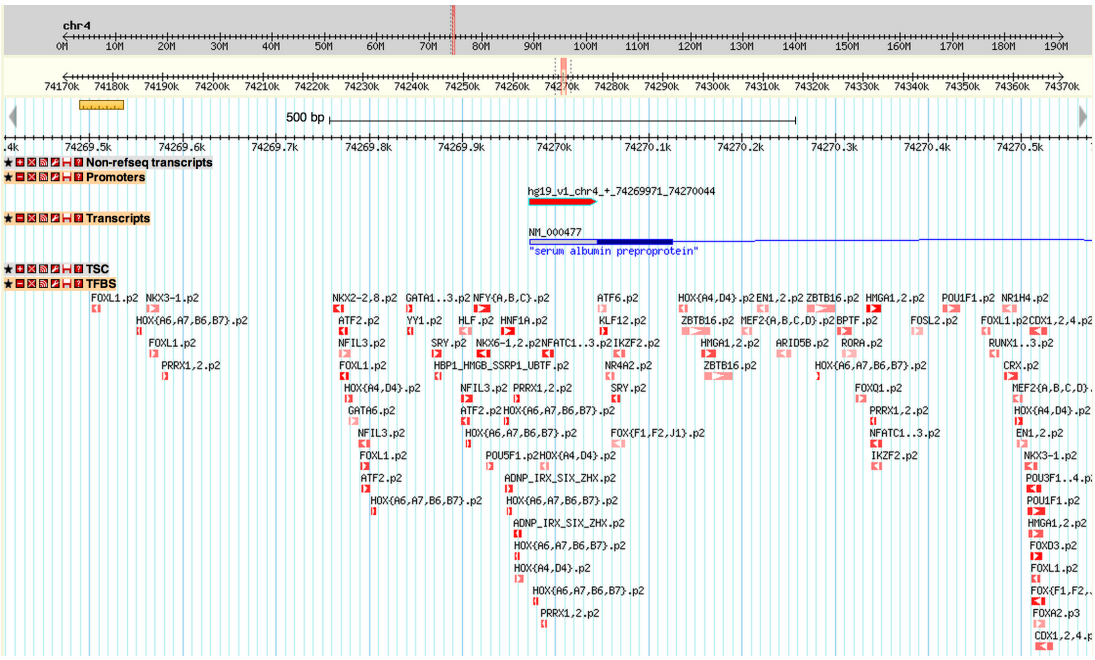
Search:

Show 20 entries

Showing 1 to 20 of 200 entries

| Promoter | Score | Refseq | Gene Symbol | Gene Name |
|---|---|---|---|---|
| chr4_-_72649734 | 161.545 | NM_000583 | GC | group-specific component (vitamin D binding protein) |
| chr1_-_159684274 | 158.324 | NM_000567 | CRP | C-reactive protein, pentraxin-related |
| chr4_+_74347461 | 145.745 | NM_001133 | AFM | afamin |
| chr4_-_155511838 | 131.965 | NM_000508<br>NM_021871 | FGA | fibrinogen alpha chain |
| chr4_+_155484131 | 129.741 | NM_001184741<br>NM_005141 | FGB | fibrinogen beta chain |
| chr17_-_64225496 | 119.840 | NM_000042 | APOH | apolipoprotein H (beta-2-glycoprotein I) |
| chr4_+_74269971 | 115.688 | NM_000477 | ALB | albumin |
| chr1_+_159557615 | 106.774 | NM_001639 | APCS | amyloid P component, serum |
| chr17_+_41052813 | 95.277 | NM_000151 | G6PC | glucose-6-phosphatase, catalytic subunit |
| chr2_+_234668914 | 93.719 | NM_000463 | UGT1A1 | UDP glucuronosyltransferase 1 family, polypeptide A1 |
| chr5_-_147211141 | 91.777 | | SPINK1 | serine peptidase inhibitor, Kazal type 1 |
| chr2_+_234601511 | 89.928 | NM_001072 | UGT1A6 | UDP glucuronosyltransferase 1 family, polypeptide A6 |
| chr19_-_36303766 | 89.070 | | PRODH2 | proline dehydrogenase (oxidase) 2 |
| chrX_-_105282714 | 88.849 | NM_000354 | SERPINA7 | serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 7 |
| chr5_-_147211235 | 88.367 | NM_003122 | SPINK1 | serine peptidase inhibitor, Kazal type 1 |
| chr8_-_17752847 | 88.288 | NM_147203<br>NM_201553<br>NM_004467<br>NM_201552 | FGL1 | fibrinogen-like 1 |
| chr14_-_94854910 | 86.867 | | SERPINA1 | serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1 |
| chr6_+_161123224 | 86.262 | NM_000301<br>NM_001168338 | PLG | plasminogen |
| chr14_-_94789641 | 82.717 | NM_001756 | SERPINA6 | serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 6 |
| chr14_-_94855120 | 81.964 | NM_000295 | SERPINA1 | serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1 |

Supplementary Figure 10: Top target promoters of the HNF1A motif for the Illumina Body Map 2. Targets are sorted by the log-likelihood score $S_{pm}$. Shown for each target promoter are the promoter ID (a link to the SwissRegulon web-browser page showing the promoter on the genome), the target score $S_{pm}$, associated RefSeq transcripts, associated gene symbols (links to NCBI pages), and gene names (which typically provide a short description of the gene's function). By default the top 20 targets are shown, but this can be changed using the drop-down menu at the top of the table. A search box allows users to search for genes or transcripts within the entire target list.




Supplementary Figure 11: Example of a promoter region as displayed in the SwissRegulon genome browser. The region shown corresponds to the proximal promoter of the Albumin gene (the 7th highest target of the HNF1A motif) and this is the region that will be displayed when following the link to the promoter displayed in Suppl. Fig. 10. The genome browser shows the RefSeq transcript, the promoter, the associated annotated transcript start cluster (TSC) based on the CAGE data, and all the predicted TFBSs. Here the intensity of the color indicates the posterior probability assigned to each site, and the name of the cognate motif is written above each side. The arrows inside the TFBSs indicate on which strand the motif occurs. Note that an HNF1A site occurs just upstream of the TSC.

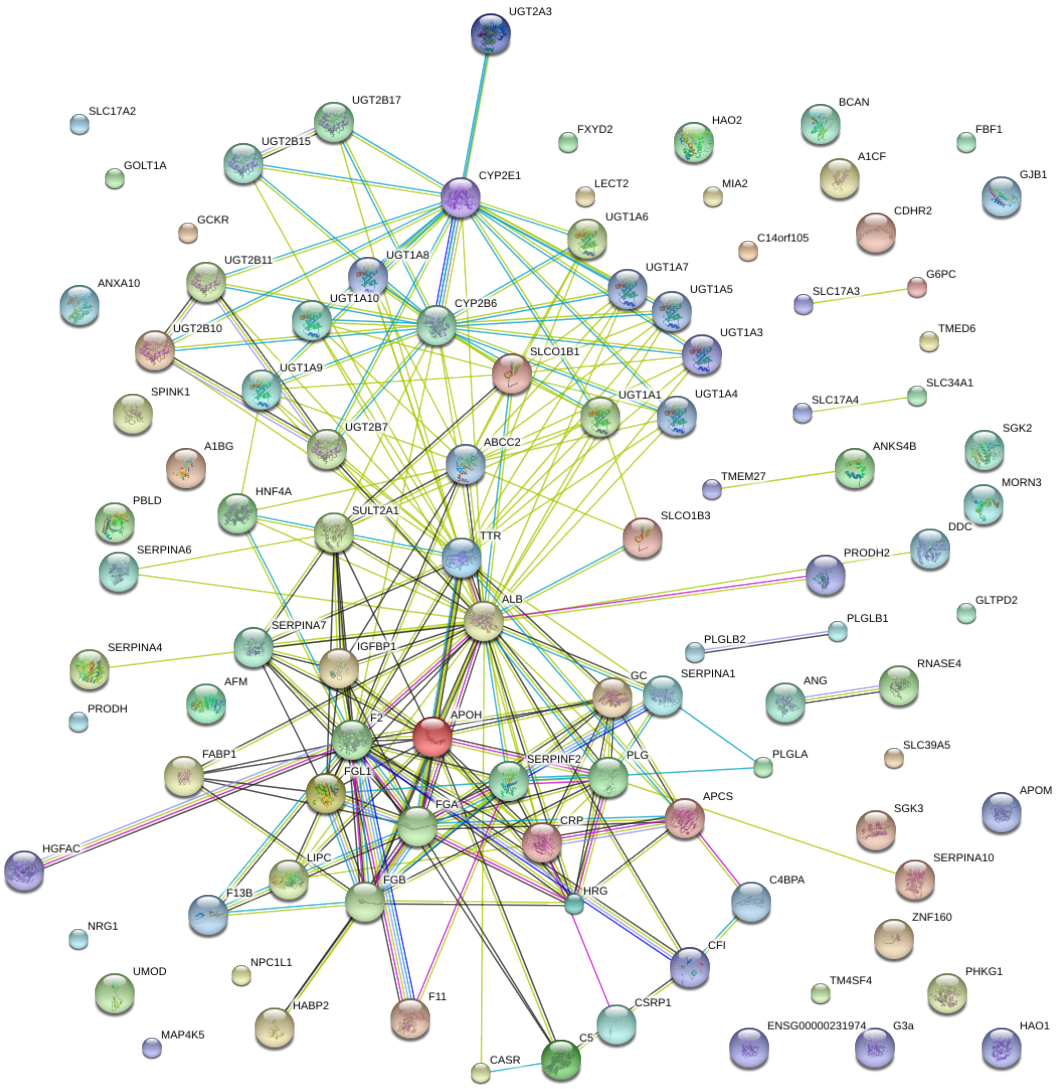
functional link between the proteins.

Apart from the STRING network, ISMARA also provides list of Gene Ontology categories that are enriched among the predicted targets of a motif. Lists are provided for the 'biological process', 'cellular component', and 'molecular function' hierarchies. A $p$-value for enrichment is calculated using a simple hypergeometric test and only categories with a $p$-value below $0.05$ are shown. The categories can be sorted either by the fold-enrichment of targets relative to what would be expected by chance or by the $p$-value of the enrichment. As an example, Suppl. Fig. 13 shows the most significantly enriched categories of the biological process hierarchy for the HNF1A motif.

One of our our aims is to understand the causal structure of the transcription regulatory network, and a first step in that direction are predictions of direct regulatory interactions between the motifs. For each motif, we check its list of predicted targets for promoters of TFs that are associated with other motifs. Using this we build a regulatory network where nodes correspond to motifs and a directed edge from motif $m$ to motif $m'$ occurs whenever a promoter of at least one of the TFs associated with motif $m'$ is a predicted target of motif $m$. On the page with results of a given motif, a part of this regulatory network centered around the motif in question is shown, i.e. all edges from or to the motif in question as well as edges between the direct neighbors of the motif. Supplementary Fig. 14 shows the most significant interactions of this network for the HNF1A motif. Note that a slider on the left-hand side of the network allows the user to vary a cut-off on the target score $S_{pm}$, i.e. showing only nodes and edges over the cut-off. In addition, placing the mouse pointer over a node brings up a pop-up with the $z$-value of the motif, and placing the mouse pointer on an edge will bring up a pop-up with the target score of the link.

Note that ISMARA predicts that HNF1A targets HNF4A, FOXA2, NR5A2, and its own promoter. In addition, HNF4A and FOXA2 are predicted to target the HNF1A promoter as well. A literature search shows that, in fact, all these direct regulatory interactions have independent experimental support [33, 34, 35, 36, 37, 38], demonstrating that the top predicted direct regulatory interactions between regulators can be highly accurate.

Finally, as described in section 1.7.3, we also fit the average expression level $\tilde{E}_p$ of each promoter in terms of mean motif activities $\bar{A}_m$. For each motif, a $z$-score $\tilde{z}_m$ quantifies the significance of the motif in explaining the mean expression level of the promoter, i.e. highly positive $\tilde{z}_m$ indicates that the occurrence of the motif is predictive for a high average expression level of the promoter, whereas a highly negative $\tilde{z}_m$ indicates that the occurrence of the motif is predictive for low average expression of the promoter.



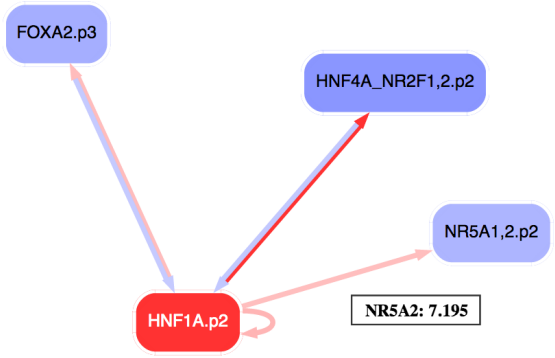

Supplementary Figure 12: Network of target genes of the HNF1A motif as displayed by the STRING database [32]. Each node corresponds to a predicted target gene of the HNF1A motif (in the Illumina Body Map 2, i.e. data set 1). Links are drawn by STRING whenever there is any evidence that the two genes may interact or be functionally linked, where evidence may range from measured direct protein-protein interaction to significant co-occurrence of the gene names within abstracts of articles.

**Gene overrepresentation in process category:**

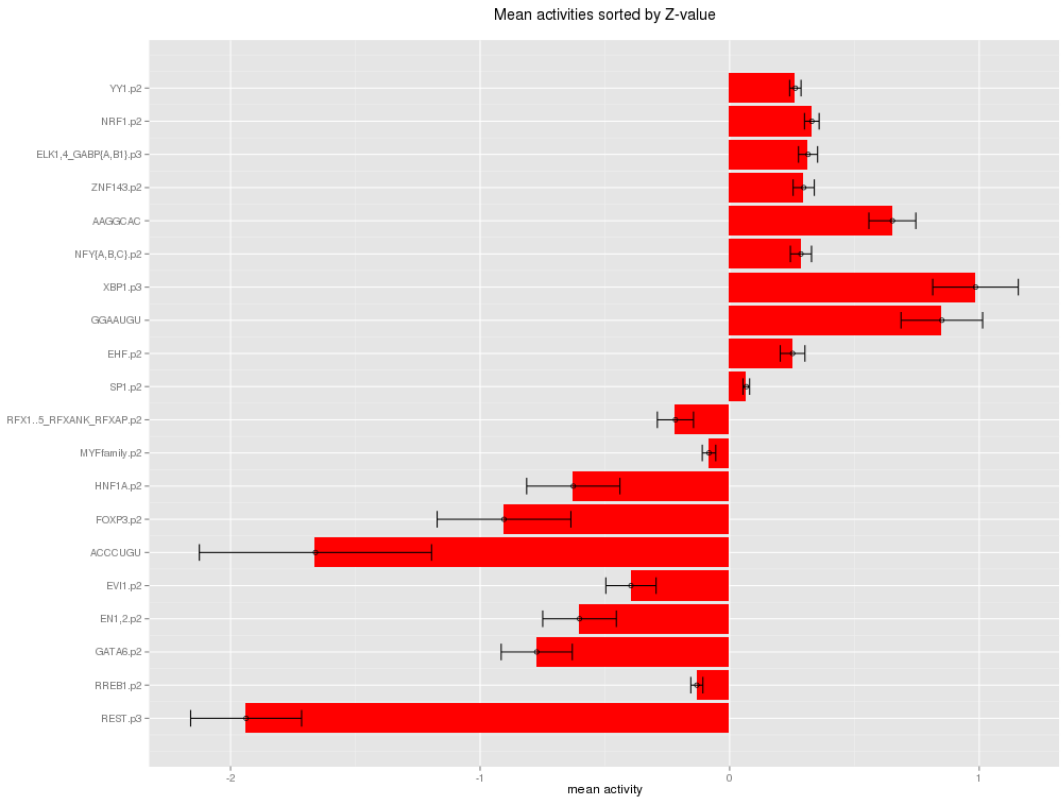
Search:
Show 10 entries
Showing 1 to 20 of 67 entries

| Enrichment | P-value | GO Accession | GO Term |
|---|---|---|---|
| 12.10 | 6.48e-15 | GO:0006805 | xenobiotic metabolic process |
| 12.10 | 6.48e-15 | GO:0071466 | cellular response to xenobiotic stimulus |
| 12.01 | 7.66e-15 | GO:0009410 | response to xenobiotic stimulus |
| 8.02 | 3.15e-12 | GO:0008202 | steroid metabolic process |
| 4.22 | 8.94e-11 | GO:0006082 | organic acid metabolic process |
| 3.40 | 7.82e-09 | GO:0006629 | lipid metabolic process |
| 8.73 | 4.04e-08 | GO:0006820 | anion transport |
| 3.77 | 4.66e-08 | GO:0042180 | cellular ketone metabolic process |
| 3.74 | 1.18e-07 | GO:0019752 | carboxylic acid metabolic process |
| 3.74 | 1.18e-07 | GO:0043436 | oxoacid metabolic process |
| 5.23 | 6.40e-07 | GO:0032787 | monocarboxylic acid metabolic process |
| 2.20 | 4.08e-06 | GO:0065008 | regulation of biological quality |
| 2.30 | 1.04e-05 | GO:0044281 | small molecule metabolic process |
| 7.49 | 1.88e-05 | GO:0006814 | sodium ion transport |
| 21.33 | 2.19e-05 | GO:0017144 | drug metabolic process |
| 10.92 | 2.22e-05 | GO:0015711 | organic anion transport |
| 4.09 | 2.34e-05 | GO:0071702 | organic substance transport |
| 3.07 | 2.50e-05 | GO:0055085 | transmembrane transport |
| 2.09 | 2.70e-05 | GO:0042221 | response to chemical stimulus |
| 15.00 | 4.37e-05 | GO:0030193 | regulation of blood coagulation |

Supplementary Figure 13: Top over-represented categories from the Gene Ontology hierarchy of biological processes among the predicted targets of the HNF1A motif. The categories are sorted by the significance of their enrichment (second column), and the first column shows the fold-enrichment relative to random expectation. The third and fourth columns in the table show the GO identifier and a description of the categories and these are again links to pages with more extensive information on the GO category. Finally, the user can interactively change the number of top categories shown using the drop-down menu or search for keywords.



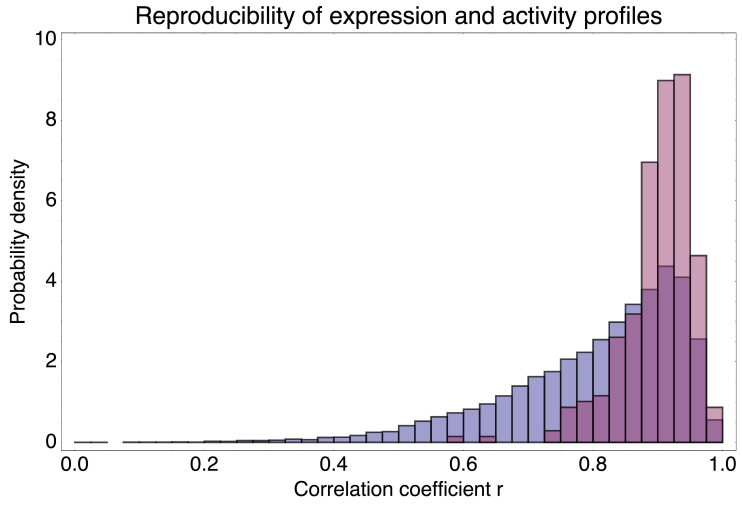

Supplementary Figure 14: The top predicted direct regulatory interactions between HNF1A and other motifs. An edge from motif $m$ to $m'$ is drawn whenever a promoter $p$, associated with motif $m'$, is a predicted target of motif $m$, with target score $S_{pm}$ larger than a given cut-off $c$. In the web browser, the user can interactively change the cut-off $c$ using the slider on the left of the figure. In this example the cut-off was set at 7.195. When the cursor is placed on an edge the target score $S_{pm}$ is shown, e.g. in this example the score of HNF1A targeting the NR5A2 promoter is shown. The intensity of the color of each motif corresponds to its $z$-score. Finally, only the direct network neighborhood of the motif in question (HNF1A) is shown, i.e. edges that are directly linked to HNF1A, or that link between motifs that directly link to HNF1A.




Supplementary Figure 15: Regulatory motifs most predictive for high or low average absolute expression across the IBM2 samples. For each promoter an average expression was calculated and for each motif $m$ a mean activity $\bar{A}_m$ (red bar), and its standard-error $\delta\bar{A}_m$ (error-bar) was calculated. The $z$-value of a motif's mean activity is defined as the ratio $\tilde{z}_m = \bar{A}_m/\delta\bar{A}_m$ and the table shows the motifs with the most positive and most negative $z$-values.

# 4 Reproducibility of motif activities

The inferred motif activities depend both on our binding site predictions, and on the assumed simple linear relationship between predicted numbers of sites and mRNA expression. As explained in the main text, there are many reasons why such a 'cartoon' model is very unlikely to produce an accurate quantitative model of genome-wide expression profiles. As a consequence, one may wonder how robust the inferred motif activities are. However, as shown in Suppl. Fig. 16, the motif activities inferred from the two replicates of the human GNF atlas are typically more reproducible across these replicates than the expression levels of the individual promoters which are used to infer the motif activities. The reason for this is that the motif activity is inferred from the behavior of the hundreds to thousands of predicted targets of the motif. Thus, although at each individual promoter the expression is likely a complex function of the regulatory sites and the linear model is likely a poor approximation, these complications are effectively averaged out when inferring motif activities from the joint behavior of all targets.




Supplementary Figure 16: Reproducibility of the inferred motif activities and the expression profiles of promoters. For each motif, and each promoter, we calculated the Pearson correlation coefficient of the activity/expression profiles for the two replicates of the samples in the human GNF atlas [39]. The figure shows the distribution of observed correlation coefficients for the motif activities (red) and promoter expression profiles (blue). The motif activities are generally considerably more reproducible than the expression profiles of the promoters from which they are inferred.

# 5 Motifs dis-regulated in tumor cells

To identify motifs whose motif activities are consistently dis-regulated in tumors, we first separated all samples $s$ from the GNF and NCI-60 data sets into the set of tumor samples $T$ and non-tumor samples $N$. Next, we used the replicate averaging described in section 1.8 to calculate, for each motif, an average activity $\langle \bar{A}^T \rangle$ in tumor samples, an associated error-bar $\delta \bar{A}^T$, an average activity in non-tumor samples $\langle \bar{A}^N \rangle$, and an error-bar $\delta \bar{A}^N$ associated with the average activity in non-tumor samples. From these, we calculate a $z$-value $z_m$ for each motif $m$ that quantifies the significance of the difference in the average activities in tumor and non-tumor samples. Tables 2 and 3 show the motifs with highest and lowest $z$-values, respectively. That is, these are the motifs most significantly dis-regulated in tumor cells.

| Motif | $z$-values |
|---|---|
| blah_family.p2 | 2.398858 |
| HIF1A.p2 | 2.230493 |
| E2F1..5.p2 | 2.140652 |
| ARNT_ARNT2_BHLHB2_MAX_MYC_USF1.p2 | 2.071274 |
| BPTF.p2 | 1.977484 |
| NFY{A,B,C}.p2 | 1.920594 |
| FOXD3.p2 | 1.915846 |
| TFDP1.p2 | 1.901083 |
| ELF1,2,4.p2 | 1.874818 |
| ZNF143.p2 | 1.802732 |
| ATF4.p2 | 1.786143 |
| YY1.p2 | 1.735238 |
| EHF.p2 | 1.718308 |
| NRF1.p2 | 1.674024 |
| ELK1,4_GABP{A,B1}.p3 | 1.667680 |
| CCUUCAU (hsa-miR-205) | 1.525379 |
| PAX5.p2 | 1.500615 |
| UCAAGUA (hsa-miR-26a, hsa-miR-26b, hsa-miR-1297, hsa-miR-4465) | 1.404557 |
| BACH2.p2 | 1.371868 |
| GUAACAG (hsa-miR-194) | 1.349047 |
| HES1.p2 | 1.317505 |

Supplementary Table 2: Motifs that are most consistently upregulated in tumor samples of the NCI-60 and GNF data sets, relative to healthy (non-tumor) tissues in the GNF data set. The motifs are sorted by their $z$-value (shown in the second column).

Many of the TFs that ISMARA identifies as disregulated in cancer are well-known in cancer biology, including HIF1A[40] (Suppl. Fig. 17), MYC[41], E2F1..5[42], NF-Y[43], YY1[44], TFCP2[45], and the SMAD TFs[46]. However, our brief survey of the literature also suggests that several other TFs that ISMARA identifies as consistently disregulated in cancers are currently not recognized as major players in cancer biology, although there is some evidence in the literature that these TFs may play a role in cancer. These TFs include HAND1,2[47], KLF12[48], BPTF[49], FOXD3[50], and ZNF143[51].

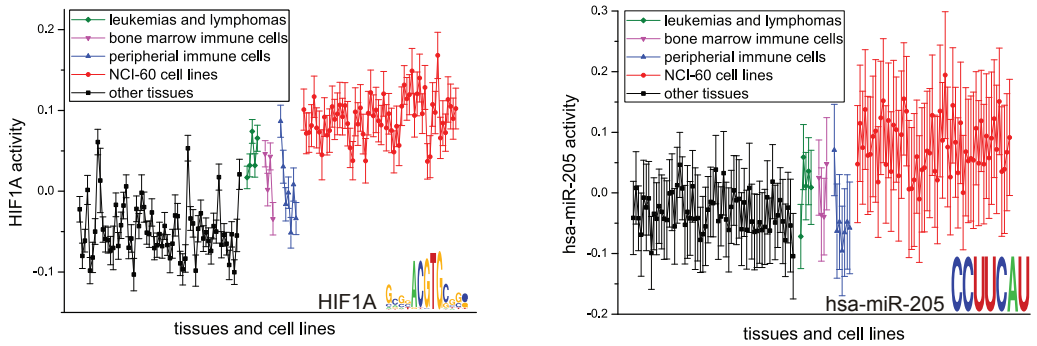
ISMARA also identifies a number of miRNAs whose targets are either consistently upregulated in tumors, e.g. hsa-miR-205 (Suppl. Fig. 17) and hsa-miR-26, or consistently down-regulated, e.g. hsa-miR-24 and the hsa-miR-17/93/106 seed family. Indeed, multiple studies have found hsa-miR-205 to be down-regulated in a number of different cancers, and hsa-miR-205 has been shown to have tumor suppressor function[52, 53, 54, 55, 56]. It has also been shown that hsa-miR-26a delivery suppresses hepatic tumors in mouse[57], supporting the downregulation of this miRNA in cancer. Conversely, hsa-miR-17 is a known oncogene[58], consistent with the downregulation of its targets in cancer. The literature on hsa-miR-24 function in cancer is more ambiguous[59]. Some evidence has been provided that hsa-miR-24 acts as repressor of apoptosis and is upregulated in certain cancers[60]. On the other hand, another study found that hsa-miR-24 can inhibit proliferation[61]. Notably, the latter study suggested that hsa-miR-24 acts through seedless target sites, which by construction are not detected by TargetScan. In summary, in this system ISMARA successfully identified oncogenes and tumor suppressors *ab initio*.

# 6 Example of species-specific targeting

The MotEvo algorithm that we use for predicting TFBSs in all promoters operates on multiple alignments and incorporates information on binding site conservation using an explicit model of TFBS evolution. This does not mean, however, that MotEvo only predicts binding sites that are well-conserved across orthologous promoters in mammals. Although evidence of conservation increases the posterior proba-

| Motif | $z$-values |
|---|---|
| SMAD1..7,9.p2 | -2.194113 |
| HAND1,2.p2 | -2.185943 |
| TGIF1.p2 | -2.117814 |
| MAZ.p2 | -2.076224 |
| TFCP2.p2 | -2.071225 |
| KLF12.p2 | -1.958392 |
| GGCUCAG (hsa-miR-24) | -1.918863 |
| FOX{D1,D2}.p2 | -1.839199 |
| TBX4,5.p2 | -1.805228 |
| FOXP3.p2 | -1.740035 |
| EVI1.p2 | -1.701934 |
| HBP1_HMGB_SSRP1_UBTF.p2 | -1.688854 |
| AAAGUGC (hsa-miR-17, hsa-miR-20a, hsa-miR-20b, hsa-miR-93, hsa-miR-106a, hsa-miR-106b, hsa-miR-519d) | -1.628037 |
| GAGAUGA (hsa-miR-143, hsa-miR-4770) | -1.619611 |
| HIC1.p2 | -1.607936 |
| NANOG{mouse}.p2 | -1.576193 |
| FEV.p2 | -1.574951 |
| MYOD1.p2 | -1.565920 |
| NR1H4.p2 | -1.562673 |
| POU1F1.p2 | -1.556216 |
| TCF4_dimer.p2 | -1.536692 |
| MYFfamily.p2 | -1.514719 |
| TAL1_TCF{3,4,12}.p2 | -1.499900 |
| POU5F1.p2 | -1.480033 |
| NR3C1.p2 | -1.473553 |
| HOX{A5,B5}.p2 | -1.440485 |
| STAT1,3.p3 | -1.417964 |
| GTF2A1,2.p2 | -1.416557 |
| RORA.p2 | -1.391819 |
| CAGCAGG (hsa-miR-214, hsa-miR-761, hsa-miR-3619-5p) | -1.356781 |
| ETS1,2.p2 | -1.337667 |
| EN1,2.p2 | -1.337051 |
| AR.p2 | -1.330996 |
| RREB1.p2 | -1.330444 |
| CUCCCAA (hsa-miR-150) | -1.318296 |
| CACAGUG (hsa-miR-128) | -1.318135 |
| JUN.p2 | -1.313498 |

Supplementary Table 3: Motifs that are most consistently down-regulated in tumor samples of the NCI-60 and GNF data sets, relative to healthy (non-tumor) tissues in the GNF data set. The motifs are sorted by their $z$-value (shown in the second column).




Supplementary Figure 17: Motif activities (with error-bars) across the human GNF and NCI-60 samples for an example TF (HIF1A, left panel) and miRNA motif (hsa-miR-205, right panel) that are disregulated in cancer. Note that different subsets of samples are colored differently as indicated in the legend.

bility assigned to a given TFBS, species-specific TFBSs that are predicted to have high-affinity for the regulator can also attain high posterior probability. Consequently, ISMARA will typically also identify targets that are species-specific or specific to a subclade of closely-related species, e.g. primate-specific targets. An example of a primate-specific target is ISMARA's prediction that, in the innate immune response time course in HUVEC cells, the IRF motif targets the promoter of the ATF5 transcription factor. As shown in Suppl. Fig. 18, the corresponding TFBS for IRF in the ATF5 promoter is primate-specific, i.e. only conserved in Rhesus Macaque.

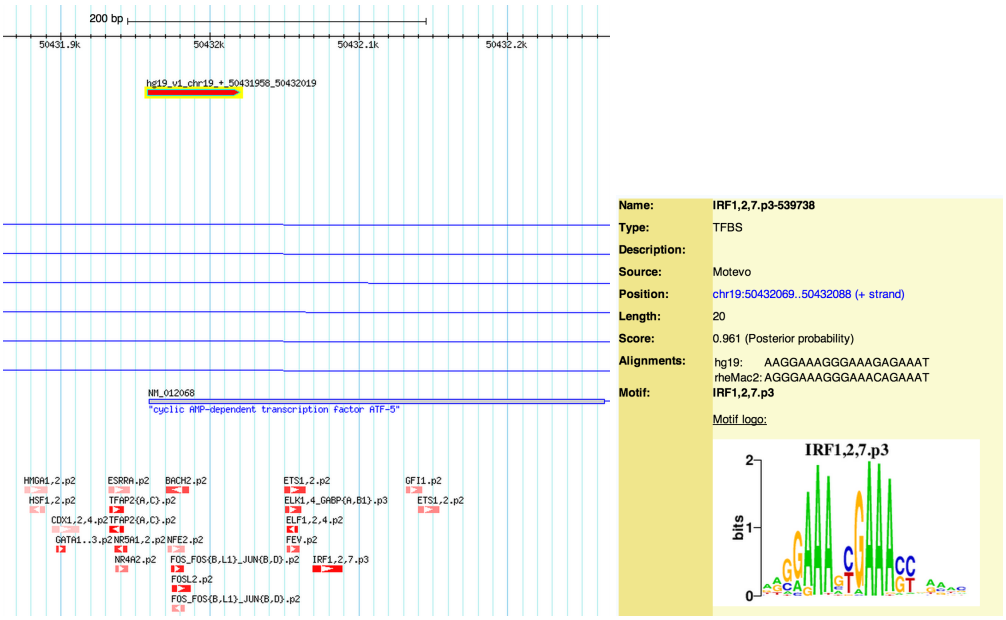
# 7 Validation of predicted NF$\kappa$B targets using ChIP-seq data

To assess the accuracy of the genome-wide targets that ISMARA predicts we compared the predicted targets of NF$\kappa$B in the innate immune response time course in which HUVEC cells were treated with TNF$\alpha$ with NF$\kappa$B targets based on ChIP-seq experiments.

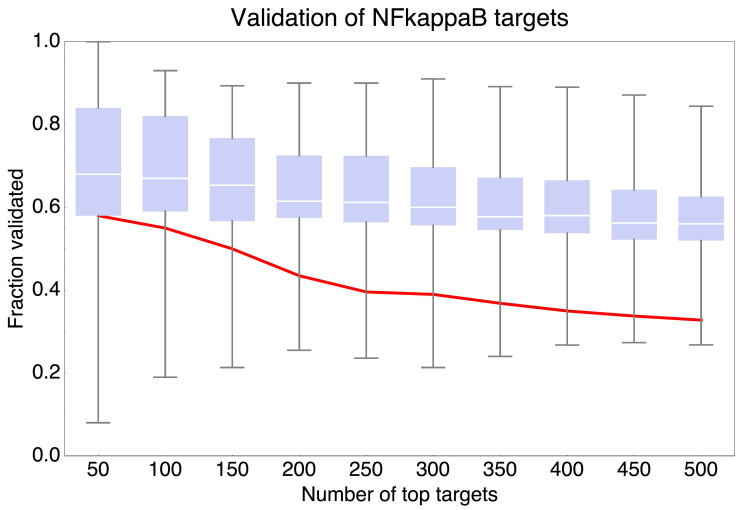
We collected data on NF$\kappa$B binding sites in 10 lymphoblastoid cell lines derived from 10 different individuals of African, European, and Asian ancestry [62]. From this study we obtained predicted peaks for 33 ChIP-seq samples (10 different individuals with between 2 and 5 replicates per individual), with each peak's significance quantified by a $z$-value. Because ISMARA's predictions are exclusively associated with promoters, we focused on ChIP-seq peaks associated with each of the promoters in our promoter set. For each human promoter, and each of the 33 ChIP-seq data-sets, we calculated a binding score by summing the $z$-values of all peaks whose center fall within 1 kilobase of the center of the promoter. Then, for each human promoter, we calculated a final binding score by averaging the binding scores across the 33 ChIP-seq data-sets. Using a cut-off score of $z = 4.5$, a little over 8% of promoters (2969 of 35821) are then classified as showing significant evidence of binding.

There we 2636 promoters that had a predicted regulatory site for NF$\kappa$B. Sorting these 2636 human promoters by their ISMARA target score for NF$\kappa$B, we then calculated the fraction of the top 50, 100, 150, 200, 250, 300, 350, 400, 450, and 500 targets that show binding according to the ChIP-seq data (red line in Fig. 19). Almost two-thirds of the top 50 targets are validated by ChIP-seq binding, more than half of the top 150 targets, and about 40% of the top 300 targets.

To compare this validation of predicted targets with the reproducibility of the ChIP-seq data themselves across replicates and individuals, we 'validated' the binding promoters as measured by each ChIP-seq data-set by the average of all other ChIP-seq data-sets. Specifically, for each ChIP-seq data-set, we sorted all promoters by its binding score, and then calculated what fraction of top targets have an average




Supplementary Figure 18: Example of a primate-specific target prediction of ISMARA. ISMARA predicts that the IRF motif targets the ATF5 promoter in the innate immune response time course of HUVEC cells. **Left panel:** Close-up of the predicted TFBSs in the ATF5 promoter, as displayed in the Swiss-Regulon genome browser [31]. The predicted IRF site occurs roughly 60 base pairs downstream of the promoter. **Right panel:** Detailed information on the IRF site in the ATF5 promoter. Besides human, an orthologous IRF site is only found in Rhesus Macaque. However, because of the site's strong match to the IRF motif, the site is still assigned a high posterior probability.

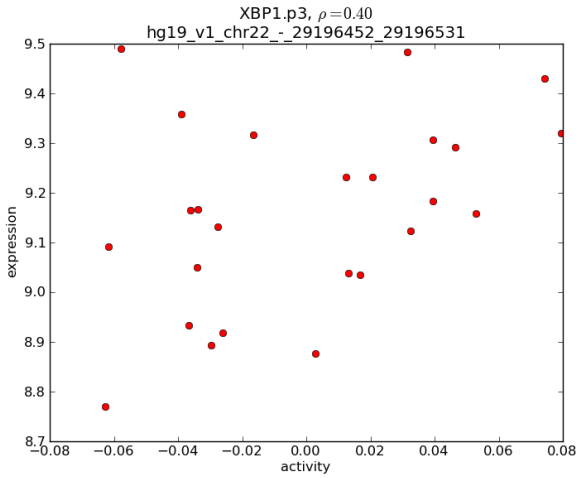



Supplementary Figure 19: Validation of ISMARA's predicted NF$\kappa$B targets from the innate immune response time course in HUVEC cells using ChIP-seq data from lymphoblastoid cell lines in 10 different individuals [62]. The red line shows the percentage of top predicted target promoters that have a ChIP-seq binding peak as a function of the number of top predicted promoters. The box-plot indicates the variation in ChIP-seq binding across samples from the different individuals. In particular, for each of 33 ChIP-seq samples, its target promoters are 'validated' by comparison with the other 32 ChIP-seq samples exactly in the same way as for the ISMARA targets. The box-plot shows the 5, 25, 50, 75, and 95 percentiles of the distribution of percentages of validated targets across the 33 samples.

binding score over the cut-off according to all *other* ChIP-seq data-sets. Doing this for all 33 samples we obtained a distribution of the fraction of validated top $x$ targets and calculated median, inter-quartile range, and 5 and 95 percentile for each value of $x \in \{50, 100, 150, 200, 250, 300, 350, 400, 450, 500\}$ (the box-whisker plots in Suppl. Fig. 19). We observed that the validation rate for ChIP-seq targets is higher, i.e. typically between 60 and 70 percent, than for the ISMARA targets. To some extent this may result from the fact that all ChIP-seq data were obtained in the same cell type, which was different from the HUVEC cells used in the innate immune response time course. However, there was considerable variability in the validation rates of ChIP-seq samples, and some samples had lower validation rates than ISMARA targets. This result shows that the accuracy of ISMARA's target predictions are comparable to targets obtained through ChIP-seq.

# 8 XBP1 motif activity and mRNA expression

The XPB1 motif is the third most significant motif in the innate immune response time course in which HUVEC cells were treated with TNF$\alpha$. The motif is upregulated during the time course. However, as shown in Suppl. Fig. 20, the mRNA expression of the XPB1 gene is almost constant across the time course, and not significantly correlated with the motif's activity. In fact, it has been established that XBP1's activity is regulated post-transcriptionally, i.e. through alternative splicing [63, 64].




Supplementary Figure 20: Scatter plot showing the correlation between the inferred activity of the XBP1 motif and the mRNA expression of the XBP1 gene for the innate immune response time course. The mRNA expression is shown on a logarithmic scale (base 2) along the vertical axis. Note the small range in expression variation.

# 9 Epithelial-Mesenchymal Transition: including microRNAs in core regulatory networks

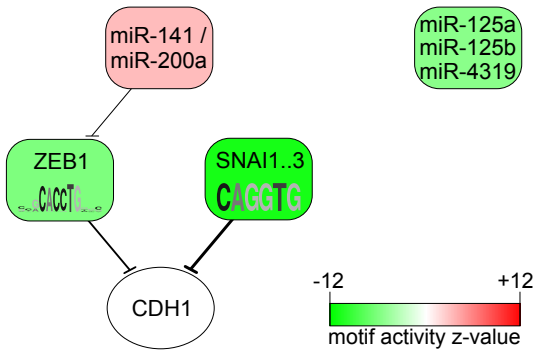
To illustrate ISMARA's ability to integrate the role of both TFs and miRNAs in the gene regulatory network, we took advantage of data from a system in which miRNAs are known to play important regulatory roles: the epithelial-to-mesenchymal transition (EMT). Recently, mRNA expression measurements were performed in duplicate on epithelial and 3 independently-isolated mesenchymal subpopulations within immortalized mammary epithelial cells[65]. After running ISMARA on this data (results at `http://ismara.unibas.ch/supp/dataset5/ismara_report/`), we used replicate-averaging to identify regulators that most consistently and significantly explain the mRNA expression differences between epithelial and mesenchymal cells (results at `http://ismara.unibas.ch/supp/dataset5/averaged_report/`).

Remarkably, much of what is known about EMT (reviewed by Polyak and Weinberg[66]) is inferred automatically by ISMARA using only the gene expression data. In particular, among the top regulators that ISMARA infers in this system are SNAI1..3, ZEB1, and a family of miRNAs consisting of hsa-miR-141 and hsa-miR-200a (sharing the same seed sequence), that have been shown to form a regulatory network essential for EMT. The predicted activity changes of these regulators are consistent with the extant literature. Namely, the decrease in SNAI1..3 and ZEB1 activity (which indicates a reduced level of their predicted targets) in mesenchymal subpopulations is consistent with the fact that both of them are mainly acting as repressors and are transcriptionally up-regulated in the mesenchymal state. The hsa-miR-141 and hsa-miR-200a miRNAs are known to be down-regulated in the mesenchymal state, causing the mRNA levels of their targets to increase, which matches the positive change in activity predicted by ISMARA. Known regulatory interactions between these factors are also uncovered by ISMARA. For instance, *ZEB1* is the top predicted target of the hsa-miR-141/200a miRNAs and existing literature confirms that the direct regulation of *ZEB1* by hsa-miR-200 is critical in EMT[67, 68, 69]. Similarly, promoters of E-cadherin (*CDH1*) gene are the 3rd and 4th top target promoters of the ZEB1 and SNAI1..3 motifs, respectively, and indeed this gene is an epithelial marker known to be targeted by both SNAIL

| Cell | Description |
|---|---|
| GM12878 | B-lymphocyte, lymphoblastoid |
| HepG2 | hepatocellular carcinoma |
| HMEC | mammary epithelial cells |
| HSMM | skeletal muscle myoblasts |
| Huvec | umbilical vein endothelial cells |
| K562 | chronic myelogenous leukemia |
| NHEK | epidermal keratinocytes |
| NHLF | lung fibroblasts |

Supplementary Table 4: Human tissues and cell lines used as the source of experimental material in the ENCODE data sets for which we analyzed ChIP-seq data of chromatin marks. We used all available samples for which a consistent measurement platform was used.

transcription factors[70] and by ZEB1[71]. These key predictions by ISMARA are summarized in Fig. 21.




Supplementary Figure 21: Core TF and miRNA regulatory interactions in the epithelial-to-mesenchymal transition, as predicted by ISMARA. Each rectangular node corresponds to a regulatory motif with its color indicating the significance of the change in activity when going from the epithelial to mesenchymal state ($z$-value defined as $z = (A_{m,\mathrm{mes}} - A_{m,\mathrm{epi}})/\sqrt{\delta A^2_{m,\mathrm{mes}} + \delta A^2_{m,\mathrm{epi}}}$). Green/Red indicates targets of the motif are down/up-regulated in the mesenchymal state. Both Zeb1 and Snail are predicted to target the E-cadherin (CDH1) promoter. Note that all interactions shown are repressive.

The activity of the family containing the hsa-miR-125a/b and hsa-miR-4319 miRNAs is the most significantly reduced miRNA family in EMT. This suggests that these miRNAs play a role in mesenchymal cells, consistent with observations that hsa-miR-125b promotes invasive tumor characteristics[72].

# 10 Analysis of the ENCODE ChIP-seq data

To illustrate ISMARA's performance on ChIP-seq data we used data from the ENCODE Project in which expression and 9 different chromatin modifications were measured across 8 different cell types[28]. Supplementary table 4 shows the list of cell types used together with their description and Suppl. table 5 shows a list of all the signals that were measured. For simplicity, we will refer to all 10 signals (which include expression and the binding of the CTCF transcription factor) as 'marks' in our description below.

| Profiling | Platform |
|---|---|
| expression | Affymetrix HT Human Genome U133A Array |
| H3K4me3 | Illumina Genome Analyzer II |
| H3K27me3 | Illumina Genome Analyzer II |
| H3K27ac | Illumina Genome Analyzer II |
| H3K9ac | Illumina Genome Analyzer II |
| H3K36me3 | Illumina Genome Analyzer II |
| H3K4me1 | Illumina Genome Analyzer II |
| CTCF | Illumina Genome Analyzer II |
| H3K4me2 | Illumina Genome Analyzer II |
| H4K20me1 | Illumina Genome Analyzer II |

Supplementary Table 5: List of the signals (i.e. expression, histone modifications, and the binding of one TF) and corresponding measurement platforms from the ENCODE data sets, that we used to demonstrate ISMARA's performance on ChIP-seq data sets. We used available BED and CEL files from the GSE26386 and GSE26312 GEO series.

| Data Set | ISMARA URL |
|---|---|
| Illumina body map 2 | ismara.unibas.ch/supp/dataset1_IBM/ismara_report |
| GNF SymAtlas + NCI-60 cancer cell lines, human [39, 73] | ismara.unibas.ch/supp/dataset2/ismara_report |
| Inflammatory response time course, HUVEC [74] | ismara.unibas.ch/supp/dataset3/ismara_report |
| Mucociliary differentiation, bronchial epithelial cells, human [75] | ismara.unibas.ch/supp/dataset4/ismara_report |
| Epithelial-Mesenchymal Transition, human [65] | ismara.unibas.ch/supp/dataset5/ismara_report |
| ENCODE cell lines, expression [28] | ismara.unibas.ch/supp/dataset6.1_ENCODE_expression/ismara_report |
| ENCODE cell lines, H3K4me3 [28] | ismara.unibas.ch/supp/dataset6.2_ENCODE_H3K4me3/ismara_report |
| ENCODE cell lines, H3K27me3 [28] | ismara.unibas.ch/supp/dataset6.3_ENCODE_H3K27me3/ismara_report |
| ENCODE cell lines, H3K27ac [28] | ismara.unibas.ch/supp/dataset6.4_ENCODE_H3K27ac/ismara_report |
| ENCODE cell lines, H3K9ac [28] | ismara.unibas.ch/supp/dataset6.5_ENCODE_H3K9ac/ismara_report |
| ENCODE cell lines, H3K36me3 [28] | ismara.unibas.ch/supp/dataset6.6_ENCODE_H3K36me3/ismara_report |
| ENCODE cell lines, H3K4me1 [28] | ismara.unibas.ch/supp/dataset6.7_ENCODE_H3K4me1/ismara_report |
| ENCODE cell lines, CTCF [28] | ismara.unibas.ch/supp/dataset6.8_ENCODE_CTCF/ismara_report |
| ENCODE cell lines, H3K4me2 [28] | ismara.unibas.ch/supp/dataset6.9_ENCODE_H3K4me2/ismara_report |
| ENCODE cell lines, H4K20me1 [28] | ismara.unibas.ch/supp/dataset6.10_ENCODE_H4K20me1/ismara_report |

Supplementary Table 6: URLs with the results of ISMARA's analyses of the data sets discussed in this paper.

We first ran ISMARA separately on the data sets for each of the 10 signals. For all the ChIP-seq data we thus modeled the occurrence of each of the marks at promoters in terms of the predicted TFBSs at the promoters. Supplementary table 6 lists all the data sets that were analyzed in this paper and shows, including references to the original publications, and lists for each data set the URL at which ISMARA's results for the corresponding data set can be found. Note that, for data sets 1, 2, and 5, there are also replicate averaged results available. These can be found by replacing 'ismara_report' at the end of the URL with 'averaged_report'.

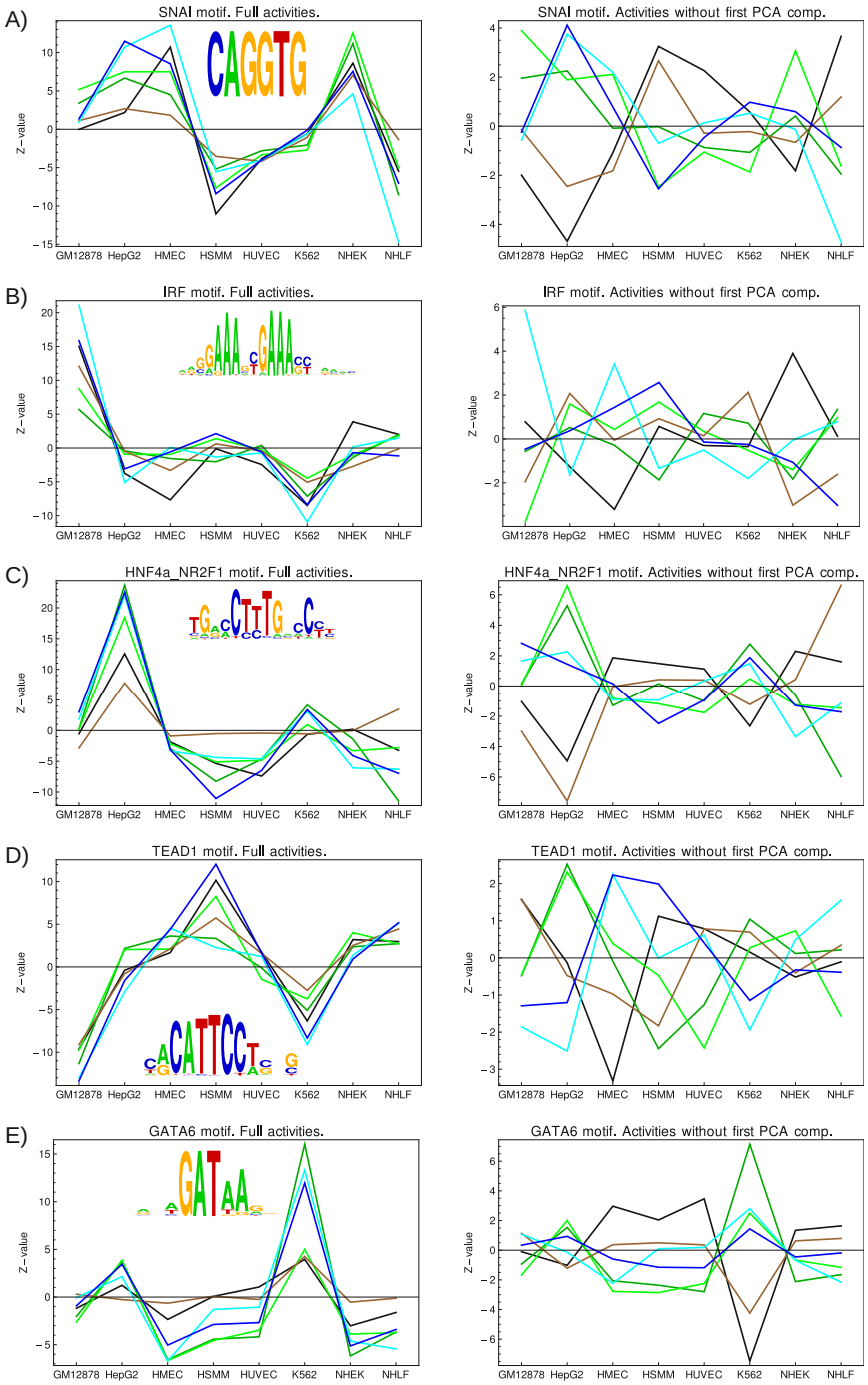
## 10.1 PCA analysis

We first performed principal component analysis of the 10 marks across all promoters genome-wide, separately for each of the 8 cell types, as described in section 1.10. As shown in Suppl. Fig. 23, we find that the first principal component explains approximately 60% of the variation in each of the 8 cell types. In addition, the first principal component is almost identical in each of the cell types. This strongly suggests that this first principal component is a general feature of the distribution of chromatin

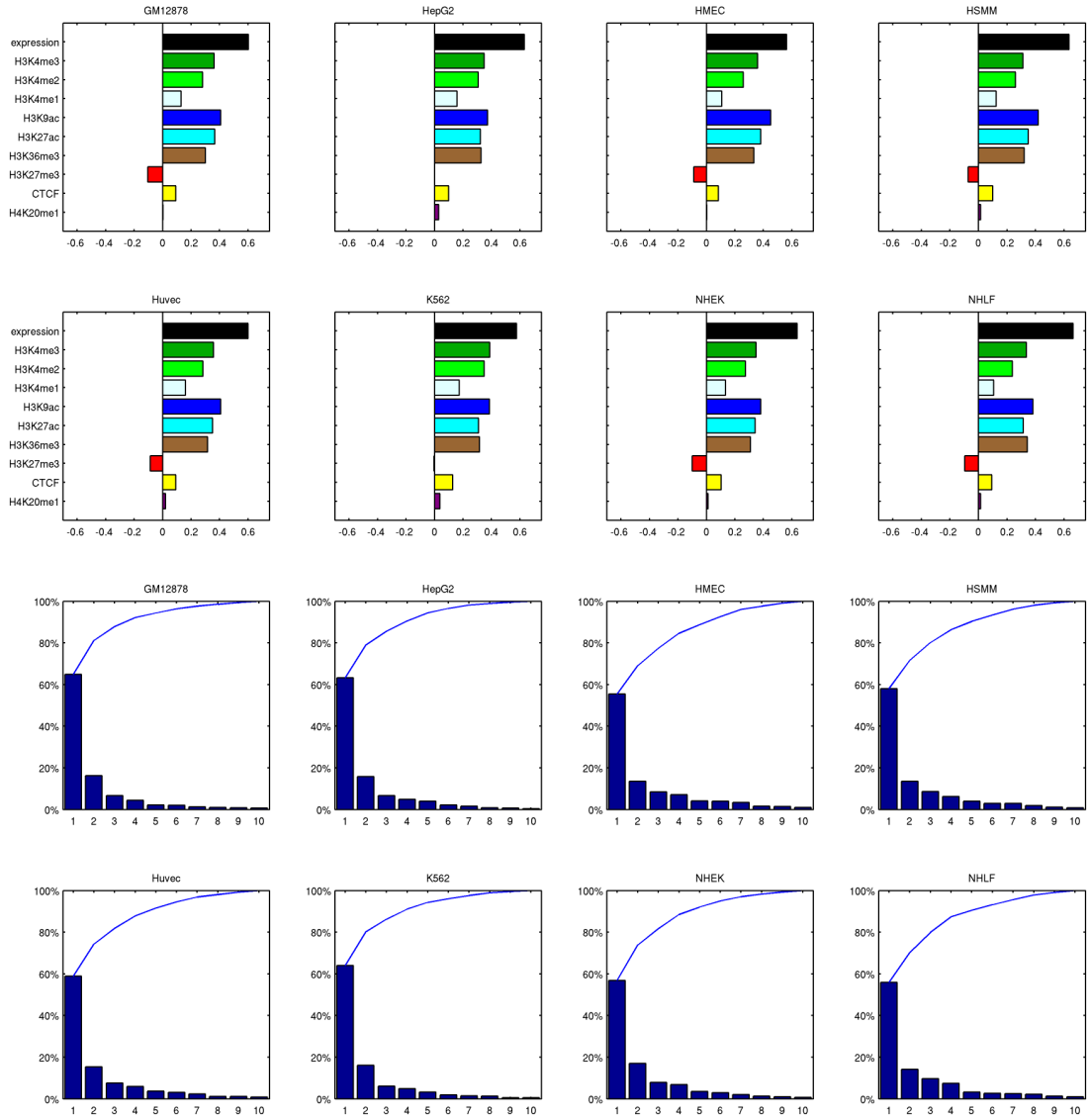
marks. Moreover, the fact that this component aligns positively with expression and activity-associated chromatin marks, suggests that this first component reflects general promoter activity. We then pooled the data from all samples and performed principal component analysis on this complete data set, i.e. treating each promoter sample combination $(p, s)$ as if it were a separate promoter. The resulting first principal component is shown in Fig. 6B of the main article.

Next, as described in section 1.10, we took the inferred motif activities for all marks and removed the component along the first principal component. That is, we removed the contribution to the motif activities that comes from the general 'promoter activity'. As an illustration, Suppl. Fig. 22 shows the inferred motif activities for 5 motifs (SNAI, IRF, HNF4a_NR2F1, TEAD1, and GATA6) both before (left panels) and after (right panels) the contribution from general promoter activity has been removed, for expression and the activation associated marks H3K4me3, H3K4me2, H3K9ac, H3K27ac, and H3K36me3. As the figure shows, before removal of the first PCA component, the activities for all marks are highly correlated, but this correlation disappears when the first PCA component is removed. This confirms that the highly correlated motif activities and the activation-associated chromatin marks is accounted for by the first PCA component that captures the relative chromatin mark levels associated with the general activity of a promoter. The remaining activities (right panels) thus provide a clearer insight in the specific role of a motif for specific marks across the cell-types. For example, for the SNAI motif the two acetylation marks are highly positive in HepG2 cells, whereas expression and H3K36me3 are clearly negative. Thus, promoters carrying SNAI sites tend to have higher histone acetylation levels than expected based on their general activity, and lower gene expression and H3K36me3 levels than expected based on the general activity.

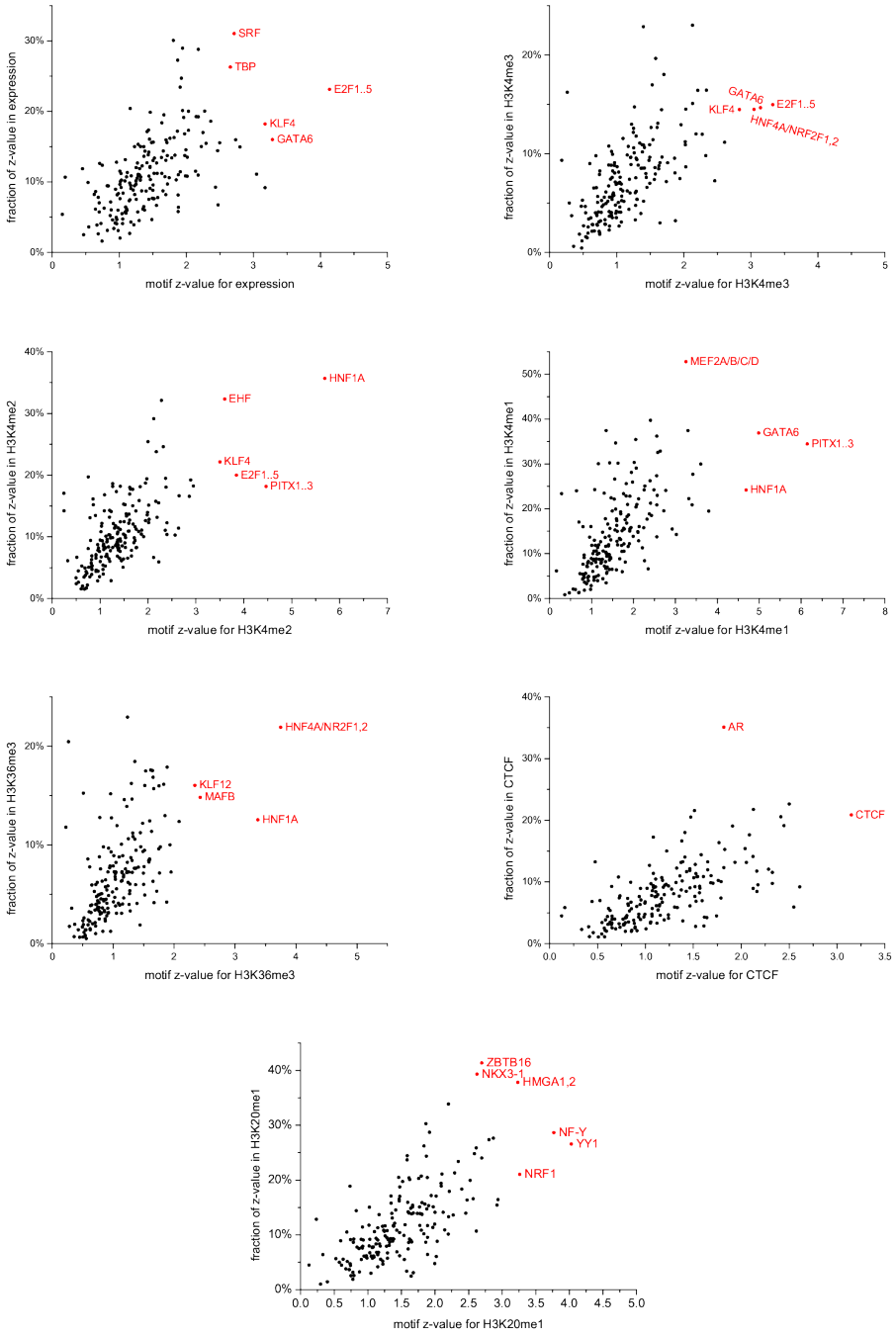
As described in section 1.10, after removing the contribution of the first principal component to the motif activities, we re-calculated significance $z$-values $z_m^i$ for each motif $m$ and each mark $i$. In addition, we calculated a specificity $s_m^i$ which measures the fraction of the overall that is associated with mark $i$. That is, a motif $m$ will be highly specific for mark $i$ if it has a high $z$-value $z_m^i$, and low $z$-values for all other marks. To identify motifs that are either most significant or highly specific for particular marks, we plotted scatter plots showing the significance and specificity for each motif (Suppl. Fig. 24). In each of the scatters we have indicated in red those motifs that had either very high significance or high specificity for the motif. Interestingly, we often find that the motifs with highest significance for a particular mark also have high specificity. For example, HNF1a is both most significant and most specific for H3K4me2 levels in promoters. Not surprisingly, the occurrence of CTCF motifs is the most significant determinant of the observed levels of bound CTCF.




Supplementary Figure 22: Inferred motif activities for 5 example motifs on the ENCODE ChIP-seq data sets measuring chromatin [28]. Each row (labeled A through E) shows the activities for explaining expression (black), H3K4me3 (dark green), H3K4me2 (light green), H3K9ac (dark blue), H3K27ac (light blue), and H3K36me3 (brown) levels, for one motif. The left panels show the motif activities as inferred from the original data the right panel the motif activities after the contribution along the first principal component has been subtracted. The names of the motifs are indicated above each panel and sequence logos are shown as insets. Note that the motif activities for the different marks go from highly correlated to essentially uncorrelated as the first principal component is removed.




Supplementary Figure 23: First principal component explaining the largest amount of chromatin mark and expression levels associated with each promoter, separately for each of the 8 cell types (top 8 panels). The bars indicate the relative contributions of expression and each of the chromatin marks to the first principal component. Note that the first principal component is virtually identical in each cell type. The bottom 8 panels show the fraction of the total variance explained by each subsequent principal component (bars) and the cumulative fraction of variance explained by consecutive components. Note that, for each cell type, close to 60% of the variance in expression and the 9 chromatin marks is explained by the first component.




Supplementary Figure 24: Significances and specificities of the motifs for explaining variations in different chromatin marks. Each panel corresponds to one mark (as indicated on the axes) and each dot corresponds to one motif. The significance of each motif is quantified by a $z$-value of the motif's activity for a given mark, after motif activities along the first principal component have been removed (see section 1.10). The specificity of a motif for a given mark is the fraction of all significance associated with a given mark (its $z$-value squared relative to the sum of all $z$-values squared, see section 1.10). the most significant and/or specific motifs for each mark are indicated in red.

# References


[1] Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**, 626–635 (2006).

[2] R, K. *et al.* Cage: cap analysis of gene expression. *Nat Methods* **3**, 211–22 (2006). PMID: 16489339.

[3] Balwierz, P. J. *et al.* Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biol* **10**, R79 (2009). URL `http://dx.doi.org/10.1186/gb-2009-10-7-r79`.

[4] Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Research* **12**, 656–664 (2002).

[5] Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res* **31**, 51–54 (2003).

[6] FANTOM Consortium & RIKEN Omics Science Center. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet* **41**, 553–562 (2009). URL `http://dx.doi.org/10.1038/ng.375`.

[7] Portales-Casamar, E. *et al.* JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **38**, D105D110 (2010).

[8] Matys, V. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31**, 374–378 (2003).

[9] Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Research* **36**, D281–D288 (2008).

[10] Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).

[11] Arnold, P., Erb, I., Pachkov, M., Molina, N. & van Nimwegen, E. MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. *Bioinformatics* **28**, 487–494 (2012). URL `http://dx.doi.org/10.1093/bioinformatics/btr695`.

[12] Siddharthan, R., Siggia, E. D. & van Nimwegen, E. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol* **1**, e67 (2005). URL `http://dx.doi.org/10.1371/journal.pcbi.0010067`.

[13] Eppig, J. T. *et al.* The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Res.* **40**, D881–886 (2012).

[14] Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res* **13**, 103–107 (2003). URL `http://dx.doi.org/10.1101/gr.809403`.

[15] Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205–217 (2000). URL `http://dx.doi.org/10.1006/jmbi.2000.4042`.

[16] Molina, N. & van Nimwegen, E. Universal patterns of purifying selection at noncoding positions in bacteria. *Genome Res* **18**, 148–160 (2008). URL `http://dx.doi.org/10.1101/gr.6759507`.

[17] Miller, W. *et al.* 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res* **17**, 1797–1808 (2007). URL `http://dx.doi.org/10.1101/gr.6761107`.

[18] Friedman, R. C., Farh, K. K.-H., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19**, 92–105 (2009). URL http://dx.doi.org/10.1101/gr.082701.108.

[19] Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307–315 (2004).

[20] Carvalho, B. S. & Irizarry, R. A. A Framework for Oligonucleotide Microarray Preprocessing. *Bioinformatics* (2010).

[21] Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F. & Spencer, F. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association* **99**, 909 (2004).

[22] Fraley, C. & Raftery, A. E. Model-Based Clustering, Discriminant Analysis and Density Estimation. *Journal of the American Statistical Association* **97**, 611–631 (2002).

[23] Fraley, C. & Raftery, A. E. MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering. Technical Report 504, University of Washington, Department of Statistics (2006, revised in 2009).

[24] Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

[25] Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

[26] Boyle, E. I. *et al.* GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20**, 3710–3715 (2004). URL http://dx.doi.org/10.1093/bioinformatics/bth456.

[27] Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **1**, 25–29 (2000).

[28] Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).

[29] Bodymap2. Illumina human body map 2.0 project. Geo Accession GSE30611 (2011). Http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30611.

[30] Kawata, H. *et al.* Zinc-fingers and homeoboxes (ZHX) 2, a novel member of the ZHX family, functions as a transcriptional repressor. *Biochem. J.* **373**, 747–757 (2003).

[31] Pachkov, M., Balwierz, P. J., Arnold, P., Ozonov, E. & van Nimwegen, E. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res.* (2012).

[32] Jensen, L. J. *et al.* STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* **37**, D412–D416 (2009). URL http://dx.doi.org/10.1093/nar/gkn760.

[33] Piaggio, G. *et al.* LFB1/HNF1 acts as a repressor of its own transcription. *Nucleic Acids Res.* **22**, 4284–4290 (1994).

[34] Boj, S. F., Parrizas, M., Maestro, M. A. & Ferrer, J. A transcription factor regulatory circuit in differentiated pancreatic cells. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 14481–14486 (2001).

[35] Bartoov-Shifman, R. *et al.* Activation of the insulin gene promoter through a direct effect of hepatocyte nuclear factor 4 alpha. *J. Biol. Chem.* **277**, 25914–25919 (2002).

[36] Tomaru, Y. *et al.* Identification of an inter-transcription factor regulatory network in human hepatoma cells by Matrix RNAi. *Nucleic Acids Res.* **37**, 1049–1060 (2009).

[37] Bochkis, I. M. *et al.* Genome-wide location analysis reveals distinct transcriptional circuitry by paralogous regulators Foxa1 and Foxa2. *PLoS Genet.* **8**, e1002770 (2012).

[38] Molero, X. *et al.* Gene expression dynamics after murine pancreatitis unveils novel roles for Hnf1 $\alpha$ in acinar cell homeostasis. *Gut* **61**, 1187–1196 (2012).

[39] Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**, 6062–6067 (2004). URL `http://dx.doi.org/10.1073/pnas.0400782101`.

[40] Semenza, G. L. Defining the role of hypoxia-inducible factor 1 in cancer biology and therapeutics. *Oncogene* **29**, 625–634 (2010).

[41] Meyer, N. & Penn, L. Z. Reflecting on 25 years with MYC. *Nat. Rev. Cancer* **8**, 976–990 (2008).

[42] Chen, H. Z., Tsai, S. Y. & Leone, G. Emerging roles of E2Fs in cancer: an exit from cell cycle control. *Nat. Rev. Cancer* **9**, 785–797 (2009).

[43] Dolfini, D. & Mantovani, R. Targeting the Y/CCAAT box in cancer: YB-1 (YBX1) or NF-Y? *Cell Death Differ.* **20**, 676–685 (2013).

[44] Castellano, G. *et al.* The involvement of the transcription factor Yin Yang 1 in cancer development and progression. *Cell Cycle* **8**, 1367–1372 (2009).

[45] Yoo, B. K. *et al.* Transcription factor Late SV40 Factor (LSF) functions as an oncogene in hepatocellular carcinoma. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 8357–8362 (2010).

[46] Samanta, D. & Datta, P. K. Alterations in the Smad pathway in human cancers. *Front Biosci (Landmark Ed)* **17**, 1281–1293 (2012).

[47] Martinez Hoyos, J. *et al.* HAND1 gene expression is negatively regulated by the High Mobility Group A1 proteins and is drastically reduced in human thyroid carcinomas. *Oncogene* **28**, 876–885 (2009).

[48] Nakamura, Y. *et al.* Krppel-like factor 12 plays a significant role in poorly differentiated gastric cancer progression. *Int. J. Cancer* **125**, 1859–1867 (2009).

[49] Buganim, Y. *et al.* A novel translocation breakpoint within the BPTF gene is associated with a pre-malignant phenotype. *PLoS ONE* **5**, e9657 (2010).

[50] Basile, K. J., Abel, E. V. & Aplin, A. E. Adaptive upregulation of FOXD3 and resistance to PLX4032/4720-induced cell death in mutant B-RAF melanoma cells. *Oncogene* **31**, 2471–2479 (2012).

[51] Izumi, H. *et al.* Role of ZNF143 in tumor growth through transcriptional regulation of DNA replication and cell-cycle-associated genes. *Cancer Sci.* **101**, 2538–2545 (2010).

[52] Gandellini, P. *et al.* miR-205 Exerts tumor-suppressive functions in human prostate through down-regulation of protein kinase Cepsilon. *Cancer Res.* **69**, 2287–2295 (2009).

[53] Majid, S. *et al.* MicroRNA-205-directed transcriptional activation of tumor suppressor genes in prostate cancer. *Cancer* **116**, 5637–5649 (2010).

[54] Dar, A. A. *et al.* miRNA-205 suppresses melanoma cell proliferation and induces senescence via regulation of E2F1 protein. *J. Biol. Chem.* **286**, 16606–16614 (2011).

[55] Wu, H., Zhu, S. & Mo, Y. Y. Suppression of cell growth and invasion by miR-205 in breast cancer. *Cell Res.* **19**, 439–448 (2009).

[56] Liu, S. *et al.* Loss of microRNA-205 expression is associated with melanoma progression. *Lab. Invest.* **92**, 1084–1096 (2012).

[57] Kota, J. *et al.* Therapeutic microRNA delivery suppresses tumorigenesis in a murine liver cancer model. *Cell* **137**, 1005–1017 (2009).

[58] He, L. *et al.* A microRNA polycistron as a potential human oncogene. *Nature* **435**, 828–833 (2005).

[59] Chhabra, R., Dubey, R. & Saini, N. Cooperative and individualistic functions of the microRNAs in the miR-23a 27a 24-2 cluster and its implication in human diseases. *Mol. Cancer* **9**, 232 (2010).

[60] To, K. H., Pajovic, S., Gallie, B. L. & Theriault, B. L. Regulation of p14ARF expression by miR-24: a potential mechanism compromising the p53 response during retinoblastoma development. *BMC Cancer* **12**, 69 (2012).

[61] Lal, A. *et al.* miR-24 Inhibits cell proliferation by targeting E2F2, MYC, and other cell-cycle genes via binding to "seedless" 3'UTR microRNA recognition elements. *Mol. Cell* **35**, 610–625 (2009).

[62] Kasowski, M. *et al.* Variation in transcription factor binding among humans. *Science* **328**, 232–235 (2010).

[63] Yoshida, H., Matsui, T., Yamamoto, A., Okada, T. & Mori, K. XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. *Cell* **107**, 881–891 (2001).

[64] Calfon, M. *et al.* IRE1 couples endoplasmic reticulum load to secretory capacity by processing the XBP-1 mRNA. *Nature* **415**, 92–96 (2002).

[65] Scheel, C. *et al.* Paracrine and autocrine signals induce and maintain mesenchymal and stem cell states in the breast. *Cell* **145**, 926–940 (2011). URL `http://dx.doi.org/10.1016/j.cell.2011.04.029`.

[66] Polyak, K. & Weinberg, R. A. Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. *Nat Rev Cancer* **9**, 265–273 (2009). URL `http://dx.doi.org/10.1038/nrc2620`.

[67] Xiong, M. *et al.* The miR-200 family regulates TGF-1-induced renal tubular epithelial to mesenchymal transition through Smad pathway by targeting ZEB1 and ZEB2 expression. *Am J Physiol Renal Physiol* **302**, F369–F379 (2012). URL `http://dx.doi.org/10.1152/ajprenal.00268.2011`.

[68] Burk, U. *et al.* A reciprocal repression between ZEB1 and members of the miR-200 family promotes EMT and invasion in cancer cells. *EMBO Rep* **9**, 582–589 (2008). URL `http://dx.doi.org/10.1038/embor.2008.74`.

[69] Gregory, P. A. *et al.* The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nat Cell Biol* **10**, 593–601 (2008). URL http://dx.doi.org/10.1038/ncb1722.

[70] Hajra, K. M., Chen, D. Y.-S. & Fearon, E. R. The SLUG zinc-finger protein represses E-cadherin in breast cancer. *Cancer Res* **62**, 1613–1618 (2002).

[71] Grooteclaes, M. L. & Frisch, S. M. Evidence for a function of CtBP in epithelial gene regulation and anoikis. *Oncogene* **19**, 3823–3828 (2000). URL http://dx.doi.org/10.1038/sj.onc.1203721.

[72] Tang, F. *et al.* MicroRNA-125b Induces Metastasis by Targeting STARD13 in MCF-7 and MDA-MB-231 Breast Cancer Cells. *PLoS ONE* **7(5)**, e35435 (2012).

[73] Ross, D. T. *et al.* Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* **24**, 227–235 (2000). URL http://dx.doi.org/10.1038/73432.

[74] Wada, Y. *et al.* A wave of nascent transcription on activated human genes. *Proc Natl Acad Sci U S A* **106**, 18357–18361 (2009). URL http://dx.doi.org/10.1073/pnas.0902573106.

[75] Ross, A. J., Dailey, L. A., Brighton, L. E. & Devlin, R. B. Transcriptional profiling of mucociliary differentiation in human airway epithelial cells. *Am J Respir Cell Mol Biol* **37**, 169–185 (2007). URL http://dx.doi.org/10.1165/rcmb.2006-0466OC.