Genome Biology

Genome **Biology**
Biology for the post-genomic era

Genome wide prediction of HNF4alpha functional binding sites by the use of local and global sequence context

Genome Biology 2008, **9**:R36 doi:10.1186/gb-2008-9-2-r36

Alexander E Kel (ake@biobase.de)
Monika Niehof (niehof@item.fraunhoferd.de)
Volker Matys (vma@biobase.de)
Rudiger Zemlin (zemlin@item.fraunhofer.de)
Jurgen Borlak (borlak@item.fraunhofer.de)

ISSN	1465-6906
Article type	Method
Submission date	19 July 2007
Acceptance date	21 February 2008
Publication date	21 February 2008
Article URL	http://genomebiology.com/2008/9/2/R36

Genome wide prediction of HNF4α functional binding sites by the use of local and global sequence context

Alexander E Kel*, Monika Niehof†, Volker Matys*, Rüdiger Zemlin†, Jürgen Borlak†

*BIOBASE GmbH, Halchtersche Str. 33, 38304 Wolfenbüttel, Germany

†Fraunhofer Institute of Toxicology and Experimental Medicine, Center for Drug Research and Medical Biotechnology, Nikolai-Fuchs-Str. 1, 30625 Hannover, Germany

Corresponding author:

Prof. Dr. Jürgen Borlak

Fraunhofer Institute of Toxicology and Experimental Medicine

Center for Drug Research and Medical Biotechnology

Nikolai-Fuchs-Str. 1

30625 Hannover

Germany

Tel: +49-511-5350-559

Fax: +49-511-5350-573

Email: borlak@item.fraunhofer.de

Abstract

We report an application of machine learning algorithms that enables prediction of the functional context of transcription factor binding sites in the human genome. We demonstrate that our method allowed de novo identification of HNF4alpha binding sites and significantly improved an overall recognition of faithful HNF4alpha targets. When applied to published findings, an unprecedented high number of false positives were identified. The technique can be applied to any transcription factor.

Background

Regulation of gene expression is accomplished through binding of transcription factors (TFs) to distinct regions of DNA (TF binding sites, TFBSs), and, after anchoring at these sites, transmission of the regulatory signal to the basal transcription complex. Indeed, regions around TF binding sites can be interrogated with regards to binding and interaction with other transcription factors (so-called composite modules) as well as local sequence properties that favor recruitment of TFs, bending and looping of DNA and nucleosome positioning. Some of these TFs are specific for a particular tissue, a definite stage of development, or a given extracellular signal, but most transcription factors are involved in gene regulation under a rather wide spectrum of cellular conditions. It is clear by now that combinations of transcription factors rather than single factors drive gene transcription and define its specificity. Dynamic and function-specific complexes of many different transcription factors, so-called enhanceosomes [1] are formed at gene promoters and enhancers to drive gene expression in a specific manner. At the DNA level, the blueprints for assembling such variable TF complexes on promoter regions may be seen as specific combinations of TFBSs located in close proximity to each other. They are termed “composite modules” (CMs) or “composite regulatory modules" (CRMs) [2] or cis-regulatory modules [3]. There may be several different types of CMs located in the regulatory region of one gene, which may be distant from each other (e.g. liver- and muscle-specific enhancers of one gene) or overlapping. Taking this into account it becomes more and more evident that the “local sequence context”, in the vicinity of the TF binding site, as well as “global context” of the whole promoter/enhancer where the TF site is located, influences binding and functioning of the corresponding TF. Numerous examples of so called composite regulatory elements are reported (see TRANSCompel database [4]), when TF binding and proper functioning of a site is strongly dependent on other sites located in the close vicinity (adjacent or even overlapping sites) or pretty distant from each other (up to 100 and more nucleotides). For instance, for the TF family of nuclear receptors (to which

HNF-4 factors belong) there are experimental evidences showing clear dependence between functioning of HNF-4 factors at their cognate sites and binding of other factors to the neighboring sites, both synergisticly and antagonisticly [4]. There is a need to develop computational models to predict TF binding sites that are functional and are involved in control of gene transcription. Recent developments in the field of machine learning techniques allow us to apply them to build highly sensitive and specific methods for predicting functional TF binding sites in human and other genomes.

Because of our continued interest in the regulation of liver-enriched transcription factors [5, 6], we were particularly interested in identifying novel genes regulated by the hepatic nuclear factor HNF4α. Indeed, HNF4α is a versatile transcription factor, and several investigators reported identification of genes targeted by HNF4α. This included various experimental approaches, including transient transfection of HNF4α into a human hepatoma cell line, a rat insulinoma cell line, and a human kidney cell line [7-9]. Additionally, findings with conditional knock-outs of HNF4α [10] were recently reported. Notably, in the study of Odom et al. a genome-wide identification of binding sites for transcription factors HNF4α, HNF1α, and HNF6 was reported by use of the ChIP-chip assay with a 13,000 human promoter sequence containing microarray [11]. Strikingly, in the case of HNF4α the number of contacted promoters was unexpectedly high, i.e. 1,575 potential HNF4α target genes were identified. In addition, 42% of the genes occupied by RNA polymerase II were also occupied by HNF4α, therefore suggesting that nearly 50% of all liver-expressed genes are regulated by HNF4α alone. Similarly, in another recent ChIP-chip experiment of ENCODE (encyclopedia of DNA elements) genomic regions (about 1% of the human genome) 663 novel HNF4α binding sites were identified in 100 genes [12], which would suggest a large number of HNF4α targets (over 60,000 sites in the vicinity of about 10,000 genes) if extrapolated to the entire genome. This unprecedented high number for HNF4α binding sites revealed by the ChIP-chip method raises the question on the functional role of all these sites in the regulation of gene transcription.

Indeed, the ChIP-chip assay is a much wanted and a highly advanced method for the genome-wide search and identification of TF binding sites. Nonetheless, it suffers from unacceptably high false positive findings. In the study of Odom et al. [11] 16% or 252 false positive binding sites were predicted by the authors. Another problem of this method is that a surprisingly small fraction of identified ChIP fragments possesses the canonical binding motif for the corresponding transcription factor [3]. This limitation needs to be overcome, and it is highly desirable to identify functional binding sites relevant for the regulation of gene transcription. Furthermore, in the current studies there is often no rationale for the selection of promoters spotted on the array, e.g. no bioinformatic approach is applied to identify relevant sequences for the design of the ChIP-chip assay.

Here, we report a computational approach based on a novel machine learning technique, which enabled an identification of genome-wide transcription factor binding sites. This method was applied to search for HNF4α gene targets. A genetic algorithm and an exhaustive feature selection algorithm were trained on 73 known and well characterized HNF4α target sequences in promoters and enhancers of different mammalian genes (see additional file 1). By genome-wide scanning of all human gene promoters we identified novel genes targeted by HNF4α. Then, a subset of predicted binding sites was confirmed by electromobility shift assay. We further interrogated promoter sequences for HNF4α binding sites identified by the ChIP-chip assay. We also analyzed expression of genes targeted by HNF4α and observed a good correlation between computationally annotated HNF4α binding sites and expression of targeted genes. Notably, ChIP-chip experiments tend to report a rather high number of TF binding sites in promoters of genes whose regulation by HNF4α is not observed, whereas our computational method for the prediction of HNF4 regulatory sites enabled improved specificity with the method encompassing rules for the regulation of gene expression.

Overall, we demonstrate the power of our computational approach in identifying novel genes targeted by HNF4α. Our machine learning technique significantly improved an overall recognition and therefore an identification of faithful HNF4α targets. This method enabled refinement of TF site predictions based on the ChIP-chip assay and identification among them the potentially functional sites, as reported herein. Furthermore, our method can easily be applied for genome-wide identification of genes targeted by any mammalian transcription factor and is not limited to promoter sequences alone with an overall success of approx. 80% based on experimental confirmation.

Results

Repeats in HNF4 binding sites

It is generally accepted that hepatocyte nuclear factor-4 (HNF4) regulates gene expression by binding to direct repeat motifs of the RG(G/T)TCA sequence separated by one nucleotide (direct repeat 1, DR1) [13]. We used two “half-site” PWMs (Positional Weight Matrices) taken from the TRANSFAC database (see Methods) in order to identify such repeats in the sequences containing known binding sites of HNF4 (based on TRANSFAC annotation, see additional file 1). We found that, although the DR1 repeat structure is clearly seen in the general consensus and in the full positional weight matrix, actual genomic sites often can be characterized by more complicated structures. The results are presented in Figure 1. We can identify repeats at various distances and with various orientations different from the canonical DR1 structure. As can be shown, that the current common point of view of DR1 repeat as the only characteristic repeat for HNF4 binding sites is not accurate. We can identify repeats at various distances and with various orientations different from the canonical DR1 structure in the sequences experimentally known as true HNF4 binding sites. This, fairly unbiased, analysis of the internal repeat structure of known HNF4alpha binding sites confirms earlier observations that sometimes HNF4alpha factors binds to elements other then DR1s.

Molecular organization of the local context of genomic HNF4α binding sites

We applied the “local context” machine learning technique to the set of known HNF4α binding sites in order to reveal properties of DNA context in the close proximity to the functional HNF4α binding sites. We analyzed frequencies of the short oligonucleotides of length 4, as well as the frequency of short repeating motifs of the lengths 2 and 4. The binding sites for transcription factors of HNF4α are characterized by various repeat structures (see Figure 1a). From our analysis of distribution of half-site motifs above we can see that short additional degenerated motifs resembling parts of the consensus repeat can be seen in the vicinity of the core of the site. Based on these results we decided to perform a thorough contextual analysis of DNA sequences containing HNF4α binding sites. The analysis was done by applying the algorithm of exhaustive search through the space of all possible short oligonucleotides and repeats in various regions of the sites and their flanks. In addition, we searched for non-redundant sets of contextual features as reported previously [14]. Table 1 presents the results of this analysis. We selected a combination of 4 oligonucleotides, 6 dinucleotide pairs and 6 four-nucleotide repeats that are overrepresented or underrepresented in the sequences of genomic HNF4α binding sites and compared the results to background sequences. A linear combination of these local contextual features gives rise to the score of context (*d* in the equation 1, see Method section). Figure 2 depicts two distributions of the score of context which we obtained on a *test* set of HNF4α recognition sites (see additional file 1 and the test and training sets defined therein based) and the *test* background set. Splitting of the site set into the training and test subsets was done by random selection). Note, that the sites from the *test* set were not used in the training phase of the algorithm. As shown in Figure 2, we clearly discriminate real HNF4α sites from false positives in the background. In our further analysis, we used the score of context with the cut-off value 0.55, which minimizes the sum of false negative error (proportion of not recognized real sites to the total number of HNF4α sites in the test set) and false positive error (proportion of false recognition of the background sequences as true sites to the total number of tested sequences in the test background set).

Among selected contextual features, there are some, like the motifs ANGB and MDDR, that fit to different parts of the HNF4 consensus sequence and appeared to be overrepresented in a rather wide area around the center of the binding sites (see Table 1). The motif CDDM is overrepresented in quite a small area corresponding to the central positions of the sites. Very interesting are the “negative” features such as repeats of the motif BNDK, whose positions are at the beginning and at the end of the HNF4 consensus, and repeats NBHV and NVYB with the positions of one part of the repeat just at the left edge of the consensus and the second part of the repeat located at the center of the consensus. Such “negative” features represent some nucleotide combinations which are rarely or never observed at functional binding sites, although such sequence context can be found in background sequences. It important to mention that the background sequences were generated as matching the HNF4 PWM but still have the additional contextual differences that can be found through the local context approach. Therefore, the local context approach can capture contextual rules that cannot be identified by the conventional positional weight matrices, since they distinguish real sites from the false positive hits of the matrix.

In order to validate the contextual features found in our analysis, we run the algorithm several times (3 times) using different samples of 100 background sequences generated in the same way as the first sample. As it was expected (see Method section), the resulted set of identified contextual features was different each time (data not shown), whereas, the oligonucleotides ANGB and CDDM, as well as the repeat $(RBNH)^2$ were identified in all tested cases (although with slightly different “from” and “to” parameters of the sequence window). Overall discrimination of the test distributions using the obtained sets of contextual features was practically the same as obtained in the first run shown in Figure 2. Therefore, in all further analyses we used the set of features obtained in the first run.

Molecular organization of the global context of HNF4α binding sites

In order to study the “global context”, we retrieved the flanking sequences of the length +/-500 bp around known HNF4 binding sites (sites shown in the additional file 1) and put them into the Y_{Global} set. The background set N_{Global} is constructed based on randomly chosen intergenic fragments of DNA from various human chromosomes applying the same strategy as for N_{Local} (described in Methods), but with the 500-bp flanks around the assumed false positive match of the HNF4 matrix (642 sequence fragments were randomly chosen scattered through intergenic regions on all human chromosomes).

We analyzed these sets using the CMA (Composite Module Analyst) program (see Method section) that allowed us to study combinations of TF binding sites in the interrogated sequences. Input for CMA is a set of DNA sequences under study (foreground set), e.g. the set of HNF4 functional sites, and a set of background sequences. By comparison of two sequence sets, CMA identifies through an iterative genetic algorithm a specific combination of TF matrices (PWMs) which are common for the foreground set of sequences and differ them from the background sequences [15]. Results are given in Table 2. The CMA algorithm identified 6 single TF matrices and 8 pairs of TF matrices characterized by variable distances between sites in each pair (dmax: 100, 200 and 500). Figure 3 represents the results of comparison of the distributions of the CM score in the two sets: Y_{Global} set – “HNF4alpha sites (+/-500bp)” (solid bars) and N_{Global} set – “Genome PWM matches (+/-500bp)” in random genomic positions (empty bars). One can see the clear discrimination between these sets. The average CM score for real HNF4 sites equals 0.499, whereas for random genome PWM matches equals 0.050 (ratio = 9.98, t-test p-value = $1.4896*10^{-26}$).

The obtained significant combination of matrices determines the “global context” that is characteristic for the regulatory regions around functional HNF4α binding sites in the genome. The biological interpretation of found composite modules is based on the concept of “enhanceosome”, postulating that for a proper performance of regulatory function, a transcription factor, while binding to the DNA target sites, should participate in many protein-protein interactions with other transcription factors binding in the neighborhood of the sites. As can be demonstrated, the algorithm selected HNF4 matrices three times, e.g. as a single element, as well as parts of matrix pairs with another HNF4 matrix and with the V$EFC matrix. Note that the algorithm additionally selected transcription factor matrices corresponding to recognition motifs of, for instance, MAZ, ER, FOX, CREB, Elk1 (Ets domain factor), COUP-TF, RFX1 and some others. Strikingly, it is known that HNF4α transcription factors cooperate with ER [16] and build synergistic composite elements with CREB [17, 18] and antagonistic composite elements with COUP-TF [19] (see also the TRANSCompel® entries: C00369, C00129, C00124). Interaction and cooperation between some other factors listed in the composite module is also known, e.g., COUP-TF with ER

[20] and CREB with Ets [21]. Thus, the found composition around known HNF4α binding sites represents potential interaction partners of HNF4 factors, therefore providing functionality in the regulation of HNF4α target genes. Note that in the case of computing the “global context” there was no test set available, i.e. all known sites were used to train the algorithm. In order to validate the computed composite module, we performed a series of data shuffling experiments (10 shuffling experiments). Each time, the assignments of positive and negative sets were randomly shuffled among the sequences and the CMA was applied in order to find a matrix combination that would discriminate best these sets of sequences. No good discrimination was obtained in such shuffling iterations. The maximum ratio achieved between the mean values was 1.6 with t-test p-values of 10^{-5}, which is much higher than in the unshuffled case (see Figure 3).

Complex criteria for determining functional HNF4α binding sites

We requested the following complex recognition criteria for a sequence of length 1,000 bp to be a potential target for HNF4α TFs:

1) the maximal matrix score of HNF4 site in the sequence should be: $q_{max} > 0.8$
2) the maximal local context score should be: $d_{max} > 0.28$
3) the maximal global context score (composite module): $v_{max} > 0.18$
4) the sum of matrix scores of all HNF4 sites found in the sequence: $q_{Sum} > 10.0$
5) the TFBS with the maximal score should be considered as the binding site for HNF4alpha, whereas the 1000bp regions provide the functional context for this site.

This rather complex criterion was derived through an iterative computation of different combinations of each individual threshold with a goal of achieving a method, which would be characterized by approximately 90% sensitivity and would efficiently use individual criteria of the local and global context. Finally, we obtained a criterion, which yields 87% of sensitivity on the set Y_{Global} (known functional sites for HNF4 factors with 500 bp flanks) and thresholds of the local and global context scores were set at the minimum of the sum of errors of these two criteria (see Figure 2 and Figure 3 respectively). As can be seen from these two figures the relative contribution to the prediction power of the global context is higher then of the local context. The sum of the errors for local context is approximately two times higher then the sum of the errors of the global context. This means that by applying this complex criterion, in approximately 13% of the cases we may miss an identification of functional HNF4α binding sites (false negative rate of the method is 13%).

Analyzing ChIP-chip data for HNF4α sites

Being enabled with the HNF4α PWM, which was built on a representative set of 73 known functional HNF4α binding sites in mammalian genes, and having two new methods available (see above, local and global content for estimating the DNA context around functional HNF4α binding sites), we analyzed the ChIP-chip data for HNF4α reported by Odom et al. [11]. We interrogated two sets of sequences, i.e. "positives" - a set of 1,605 sequences that were reported as HNF4α-targeted genes in hepatocytes, and "negatives" - a set of 10,852 sequences that were reported not to be contacted by HNF4α in hepatocytes and in pancreatic islets. The average length of the sequences reported by Odom et al. was approximately 1Kb [11]. In each sequence of both sets, we computed the number of potential HNF4α binding sites (matrix score > 0.8), the sum of the scores for all sites, and the maximal score of the sites found in the sequence. Thereafter, we calculated the local context score (*d*) and the global context score (*v*) for each potential HNF4α binding site in these sequences and reported the maximal scores obtained in each sequence. We applied the complex recognition criterion (see above) to the sequences in these two sets. As a result, only 21% of the "positive" set (i.e. 375 sequences out of 1,605) passed the criterion. Indeed, 79% of the sequences were rejected, since they did not pass one or several requirements as defined above. In order to estimate the rate of false positives of our method, we applied it to the set of "negative" sequences. Our complex criterion rejected 97.4% of these sequences, giving us an overall estimate of 2.6% as false positive rate. Figure 4 depicts a plot of the "global" and "local" context scores, comparing distribution of the 375 sequences selected from the "positive" set versus distribution of all sequences in the "negative" set. Obviously, the selected sequences are characterized by highest scores of global and local context, whereas the majority of the "negative" sequences are characterized by the low values of these two scores. The list of the 375 sequences that passed our criterion are given in the additional file 2. Furthermore, Figure 5 summarizes the data obtained in the analysis of known HNF4α binding sites, as well as "positive" and " negative" sets of sequences derived from ChIP-chip experiments reported by Odom et al. [11]. These data clearly show that the majority of the sequences revealed in ChIP-experiments of Odom et al. [11] differ quite significantly in their local and global context from the sequences of known and experimentally confirmed HNF4α binding sites. We estimate that only 20% of these sequences fulfill our requirements to be considered as faithful functional HNF4α binding sites. Note, Odom et al. [11] assume a 16% false discovery rate in an identification of binding sites in their ChIP experiments. Application of our analysis to the Odom et al. data suggests about 80% of ChIP-chip identified targets not to meet the contextual requirements which characterize biologically functional sites and therefore may not be involved in HNF4α dependent regulation of gene transcription.

Linking HNF4α binding sites to gene expression

We further applied our computational method to data reported by Naiki et al. [7] and Lucas et al. [9]. Notably, these investigators carried out microarray experiments to identify genes whose expression differed upon targeted overexpression of HNF4α. From these studies a list of differentially expressed genes was obtained. Additionally, we compared the differentially expressed genes with findings reported by Odom et al. [11] who performed ChIP-chip experiments with HNF4α. We thus compared data from two different approaches, i.e. targeted overexpression of HNF4α and ChIP-chip data for the identification of novel HNF4α target genes. We then applied our computational approach (i.e. by use of the complex recognition criteria described above) to interrogate the data sets. The results are presented in Table 3. Only a small fraction of identified genes could be compared directly, i.e. 75 and 70 differentially expressed genes (= Up + Dn) and 150 genes whose expression did not change (= NC). As can be seen from the data given in Table 3, our computational method and the ChIP-chip data are similar when correlated with the gene expression data of HNF4α targeted genes (see table footer 2), i.e. approximately 18-20% of differentially expressed genes were similarly identified by the ChIP-chip and our computational method based on the data of 145 differently expressed genes. Indeed, several genes targeted by HNF4α are identified by both methods (e.g. 5 genes: ACADVL, RBKS, SLC35D1, ATP7B, MGST2 out of 70 genes of the data set of Lucas et al. [9]).

At the same time, our computational method for identifying HNF4α gene targets is less error-prone, i.e. 2.7% false results based on the computational method versus 13.3% false results determined for the ChIP-chip method (based solely on gene expression data from Lucas et al. [9]) (Table 3, last row). It is of considerable importance that the simple use of a single HNF4 PWM and by ignoring local and global sequence context the prediction of HNF4α target genes becomes false positive error prone, i.e. 19% as shown in Table 3.

Search for HNF4α functional sites amongst all known human gene promoters

We applied the method developed for an identification of putative HNF4α gene targets to the full set of promoters of human genes annotated in TRANSPro™ database rel. 2.1 (containing 15,455 promoters). First of all, we scanned promoters in the region from –500 to +100 around the start of transcription for matches of the HNF4 weight matrix with the matrix score $q > 0.8$ accompanied by local context score $d > 0.48$. We identified 3,009 promoters which had at least one site passing both these criteria. Next, we chose the highest scoring match of the HNF4 matrix in each of the promoters and retrieved 500 bp-flanking regions around the match. We applied the complex criterion (see above) to obtain a set of sequences, which led to the prediction of 375 target promoters; among them were 121 promoter of genes encoding transcription factors and other components of the signaling system in the cell. These genes attracted our attention for experimental verification by EMSA assays as reported herein. The full list of the predicted target promoters is given in the additional file 2.

EMSA confirmation

Supershift experiments with probes for established recognition sites for HNF4α, i.e. promoter regions derived from HNF1α, AAT, APOB, AGT, APOC3, CYP2D6, TF, ALDH2, APOC2 and PCK1, resulted in binding of HNF4α (see Figure 6A). This exemplifies selectivity and sensitivity of the EMSA assay to validate HNF4α binding sites for 10 arbitrarily chosen but known targets of HNF4α. From the list of 375 predicted HNF4α target genes (see above) we have selected further 10 novel HNF4α binding sites for experimental confirmation that are characterized by the high PWM score and scores of local and global context, and that have not been reported in the study of Odom et al. Note, EMSA revealed binding of HNF4α to NCOA2, TFF2, CHEK1, CD63, SH3Gl2, RND2, ESRRBL2 and DDB1, whereas supershift experiments did not confirm HNF4α binding to NEUROG3 and IL6 (see Figure 6B), thus providing an estimate of 80% for the sensitivity of our computational method for de novo prediction of HNF4α binding sites. A summary of the biological function of these newly identified HNF4α target genes is given in Table 4.

In addition, we wished to verify HNF4α binding sites predicted by ChIP-chip experiments [11]. Note that nearly 80% of proposed HNF4α binding sites were rejected by our computational method, which combines analysis of HNF4 matrices with local and global context of the sequences. Here, we selected 10 genes. These genes were reported by Odom et al. [11] to be targeted by HNF4α in hepatocytes, but our computational method characterized them by extremely low scores of the HNF4 weight matrix as well as low scores of local and global context (all four tests of the complex criteria set by us failed to identify these genes as HNF4α targets). Therefore, these 10 potential sites were analyzed for HNF4α binding. Strikingly, none of these sites, i.e. in promoters of genes: NPAS2, GPHN, PPP1R3C, AKR1C3, CFL2, MDM2, CLCN3, CBX3, AZI2 and C14orf119, were bound by HNF4α, as evidenced by supershift experiments (see Figure 6C).

Discussion

De novo computational identification of genes targeted by various transcription factors is a challenging task especially in genomes of high eukaryotic organisms which are characterized by extremely large gene regulatory regions. Indeed, binding of transcription factors to their cognate sites on DNA is a complex process that requires presence of a specific short sequence pattern in DNA, commonly described by a position weight matrix (PWM). Furthermore, the specific local sequence context in the vicinity of the binding site is required to provide favorable conditions for DNA confirmation and DNA flexibility (e.g. binding sites for the HNF1 factor require a significantly different DNA melting parameters of the surrounding region [22]). In addition, local structures such as short repeats and palindromes are often observed and, as discussed before, are needed to enable an optimal environment for homo- and heterodimerization of transcription factors [4]. A particularly important role of the “global context” of TF binding sites in determining cooperative binding of factors with other transcription factors to their neighboring DNA sites is broadly recognized [1]. A broad collection of experimentally proven facts on cooperative binding of two and more transcription factors to so called composite regulatory elements with synergistic effects on regulation of gene expression is provided by the TRANSCompel® database [4]. Among them there are several known examples of nuclear receptors to be involved in such composite elements (e.g. glucocorticoid receptor (GR), androgen receptor (AR) and others). But there are no bioinformatics tools available so far that would allow a systematic analysis of the combinatorial sequence context of genomic binding sites.

In general, there is a definitive need to develop novel computational approaches to improve description of the DNA patterns required for TF binding. Ellrott and co-authors applied a Markov chain model to identify HNF4α binding sites in order to improve recognition accuracy of the DNA binding pattern [23]. They have demonstrated that the approach performs better than PWMs alone, but this approach does not consider any local context on the flanks of sites that indeed play a crucial role in promoter activation and DNA binding *in vivo*.

Recently, the local context in the form of short repeats has been successfully implemented to improve recognition of binding sites for nuclear receptors [24, 25]. Extending the previously published approach [24] towards application of hidden Markov models (HMMs) Sandelin and Wasserman [25] modeled various known constellations of direct, inverted and everted repeats for different sites of nuclear receptors and were able to improve general precision of the recognition. This approach looks very promising, though it lacks any capability to classify predicted sites in order to identify which particular transcription factor from the large family of nuclear receptors is able to bind to the predicted sites. In addition, we here showed binding sites for such nuclear receptors as HNF4α to be highly enriched by various different repeat structures, which does not completely fit with the existing paradigm about a canonical structure of HNF4 sites as DR1 repeat. This makes it extremely difficult to judge factor recognition based on an oversimplified model of repeat structure of the sites.

We, therefore, developed a novel approach for recognition of functional HNF4α binding sites by analyzing the “local” and “global” context of targeted genes. The method is based on the assumption that sequence context which surrounds TF binding sites in DNA is very important for both the process of TF binding to the site and, most importantly, for providing specificity of the transcription factor in the regulation of gene expression – by either activation or repression of the gene in particular cellular situations. The sequence context of the TF binding sites actually makes them functional: in the absence of the proper context, the possible binding of TF to a particular site on DNA can be impaired or made functionally neutral (which means that the factors are bound to the DNA, but do not influence expression of the gene; such sites are, therefore, non-functional).

In the current work, we performed a thorough analysis of the local nucleotide context on the flanks of known functional HNF4α sites, as well as in the whole local region occupied by the sites. We improved our earlier published approach of analyzing the local context [2], which is based on a SiteVideo method [14], and introduced new types of contextual features that modeled various repeated structures in the sequences on the flanks of the sites. Interestingly, the revealed short oligonucleotide features and repeats can be classified into three categories: 1) Oligonucleotides like ANGD and MDDR and repeats like AV – VS and VS – YA and $(RBNH)^2$ that fit to different parts of the HNF4 consensus sequence and appeared to be overrepresented in a rather wide area around the center of the binding sites. We can interpret such features as a signature of overrepresentation of HNF4-site-like patterns in the local area surrounding the functional HNF4 site, which may play a role in increasing the probability of HNF4α factors to bind to this site. 2) Oligonucleotides like CDDM and repeats like $(DNCD)^2$ are overrepresented in a quite small area corresponding to the central positions of the sites. Such features correspond to the central HNF4 site pattern, but they reveal some contextual features of the functional HNF4 sites which can not be described by the PWM matrix model, e.g. correlation between neighboring nucleotides that can not be captured in full by the mononucleotide weight matrix. 3) “Negative” features, that reveal oligonucleotides are underrepresented at functional binding sites if compared with the background sequences. Such negative features can be “local” as in the case of the repeats $(BNDK)^2$,$(NBHV)^2$ and $(NVYB)^2$, which again describes some mutual nucleotide correlations that can not be captured by PWMs, or rather distributed as BR – NT, which can be interpreted as an “echo” of some physical-chemical properties of DNA that may interfere with the binding or functioning of the transcription factors. Notably, the length of the oligonucleotides tested by our method is restricted by 4 letters of the extended code, mainly, because of high computational complexity of the calculations. Yet, this oligoncleotide length seems quite optimal for revealing statistically significant features of DNA sequences.

We assume that in addition to the local context, the global context of the TF binding sites in the regulatory regions of genes dictates whether these sites are functional. The global context, which we model by specific combinations of binding sites of various TFs, provides some sort of “scaffold” on DNA to enable cooperative or antagonistic interactions between TFs. These multiple and complex interactions, if correctly organized in space and time, give rise to the regulatory function of the transcription factor binding sites under investigation. It is clear by now that binding of a single TF to its cognate site on DNA alone does not guarantee the proper functional activity of the targeted gene. More interaction with other TFs in the transcription complex and in the enhanceosome are necessary to acquire the full regulatory functionality.

Specifically, known functional combinations of TF binding sites were used before in a number of promoter analysis approaches, e.g. for an identification of muscle-specific promoters [26, 27], promoters of liver-enriched genes [28], of yeast genes [29], of immune-specific genes [30-32], promoters of genes regulated during cell cycle [33] or antibacterial defense responses [34, 35]. A number of approaches identifying composite motifs were created: BioProspector [36], Co-Bind [37], MITRA [38], dyad search [39]. These programs help to discover "*ab initio*" new regulatory sites for yet unknown transcription factors. Another set of methods have been developed to discover composite modules by utilizing information on potential binding sites for known transcription factors (stochastic methods: ClusterScan [40] and TOUCAN system [41] and probabilistic methods: [42]). We combined these two approaches by computing "local" context as an exhaustive "*ab initio*" composite motif discovery method with the "global" context - the powerful composite module discovery method based on application of a genetic algorithm.

Furthermore, we wish to point out that our method considers several alternative PWMs for the calculation of the global context. The use of such alternatives PWMs for constructing the composite module (see Method section) enabled more reliable predictions. Particularly, in cases with small training sets but with data derived from multiple rounds of computations the use of different matrices is meaningful. These computations are far from statistical saturation but new sites may eventually add a certain bias and potentially drive the new PWM matrix away from the functional binding site sequence context. In the case of HNF4α we included both half-site and full site matrices. Still the full length PWMs are able to capture some subtle differences in the spacing sequences between the "repeats". There are recent reports confirming our old observation that often such "gap" sequences between two repeats of the nuclear receptor sites are actually more conserved between species then the repeats themselves. Therefore, the combined usage of fixed length PWMs together with the distributed oligonucleotides on the flanks provides a more robust method for site detection then each method separately.

In order to compare findings of our algorithm with the best existing methods, we performed an independent run of the TOUCAN software [41] on the set of HNF4α sites. Notably, TOUCAN is similar to our method and is based on a genetic algorithm. It identified a combination of 14 PWMs for different transcription factors including two matrices for HNF4α factors (V$HNF4_01 and V$HNF4_01_B) and matrices for other factors (VUSF_01, VOCT1_02, V$SP1_Q6, V$PPARG_03 and some others) Interestingly, except HNF4 matrices, there were no further correlations with matrices selected by our method (see Table 2). This difference can be explained by the ability of our method to identify pairs of matrices and also by the ability to optimize the cut-off values, which is not possible in TOUCAN. Another advantage of our method is the possibility to include information on tissue specificity of the factors into consideration, through extensive use of factor expression annotation in TRANSFAC® database. We further compared our method with the NUBIscan algorithm [24]**.** Our approach combines many different features of the most conserved part of the sites as well as various features of local and global context, whereas the NUBIscan algorithm relies again solely on the "repeated" structure of the nuclear receptor sites, which is indeed a very profound property of these sites but not the only one. And, similarly to the later published Wasserman approach [25], the NUBIscan algorithm lacks capability to classify predicted sites in order to identify which particular transcription factor from the large family of nuclear receptors is able to bind to the predicted sites. Notably, our method was designed specifically for the recognition of HNF4 sites. Nonetheless, our strategy to define the local and global sequence context is a highly generalized method and can be applied to any transcription factor.

An additional point of consideration is the "regulatory potential score" as introduced in the work of Elnitski and co-workers [43] which refers to the 5-way multi-species alignment introduced into the UCSC Genome Browser. In our study we did not restrict our selves to conserved sites only based on multispecies homology of regulatory regions (phylogenetic footprinting). Although this concept is quite popular for the selection of evolutionary conserved regulatory sites, the method suffers from low sensitivity because frequently functional TF binding sites are missed mainly due to the very complicated evolutional history of mammalian regulatory sequences, which hardly can be modeled by the simple divergent concept – that is the basic concept of phyligenetic footprinting.

Furthermore, promoter regions are characterized by very specific average base composition as well as composition of di-nucleotides. It is known, that, whereas the overall genomic sequences are highly depleted by CG di-nucleotide, promoters are often located near high concentration of CG, in or near, so called, CpG – islands. We therefore compared the nucleotide and dinucleotide composition of the promoter sequences of the known HNF4α sites and sequences that were used as probes on the ChIP-chip experiment of Odom et al. [11], since these were the sequences to which our method was applied to. We did not find any significant difference in nucleotide and di-nucleotide composition of these two sets of sequences, which is not a surprise since the probes of the chip were designed using known genomic promoters. Consequently, since training and test sequences are very similar in their context, the CpG bias, if any, of the HNF4 containing promoters of the training set can not bias the results of analysis of the sequences from the ChIP-chip experiment.

Additionally, many authors attribute a certain functional role of the CpG islands in promoters. These islands can be some sort of centers of regulated DNA methylation, which can effectively contribute to the hepotocyte specific gene regulation, by providing HNF4alpha binding sites necessary functional context. Absence of such CpG "context" in the vicinity of HNF4alpha binding sites may potentially render them functionally neutral. Therefore, some CG dinucleotide – like features, which were included in the "local context" (e.g. elevated frequency of the oligonucleotide CDDM, where D=(A/T/G) and M=(A/C)) may reflect this "CpG" bias of functionally active promoter sequences and therefore help to identify the functionally active HNF4alpha binding sites.

We further studied the influence of the distance from HNF4α sites to transcription start site (TSS). As shown in additional file 1, the location of known HNF4 sites is variable in promoter sequences and may range from a position close to the TSS to up to + 10kb and –11kb. Under such high variability, it is improbable that the method can "memorize" the position of TSS during training since sequences in the training set are not aligned in relation to TSS.

We then applied the developed methods to analyze data derived from ChIP-chip experiments reported by Odom et al. [11] for HNF4α. This study is based on chromatin immunoprecipitation combined with DNA-DNA hybridization on a microarray containing 13,000 human promoter sequences. In the study of Odom et al. [11], the number of HNF4α-targeted promoters was unexpectedly high, i.e. 1,575 potential HNF4α target genes in hepatocytes were identified, corresponding to 42% of the genes occupied by RNA-polymerase II. Only 48 (= 3%) of 1,575 putative HNF4α targets, however, were verified in separate gene-specific ChIP experiments. Additionally, HNF4α DNA binding was not distinguished from protein-protein interactions, as *in vitro* binding was not analyzed. We applied our algorithm to the proposed HNF4α gene targets and found merely 20% of them to obey the complex computational criteria (presence of the appropriate local and global context) that can predict the functional activity of these binding sites. We further stratified our approach by comparing HNF4α functional sites identified by us with independent gene expression data. This comparison shows that our computational approach is versatile and predicts expressed genes directly targeted by the HNF4α transcription factor with similar sensitivity as ChiP experiments. In strong contrast, the false discovery rate of the computational method is almost 5 times lower than for the ChIP-chip method. This confirms our suspicion that many of the HNF4α binding sites predicted by Odom et al. [11] are functionally neutral, whereas the developed computational method is able to recognize the functionally active HNF4α binding sites based on verification of the local and global context of these sites.

Furthermore, in a recent paper of Gupta et al. [44], regulation of gene expression was studied in pancreatic cells of a HNF4α conditional knockout model. Expression analysis identified 133 genes as HNF4α regulated. Regulated genes could be compared with the promoter array data of Odom et al. [11]. Surprisingly, the overlap between differentially expressed genes and those bound by HNF4α is rather small. In other words, of 133 genes whose expression was dependent on HNF4α only 13 have been identified by the location analysis reported by Odom et al. [11]. Likewise, of 587 promoters occupied by HNF4α in the study of Odom et al. [11]. 574 showed no significant change in gene expression (see ref. 44). Therefore, 86% of Odom et al. [11] proposed HNF4α targeted genes did not differ in gene expression in the absence of HNF4α. These estimates agree well with our computational approach where only 20% of the Odom et al. [11] proposed target genes could be computationally confirmed.

In the most recent study [45] by the same investigators and through application of an improved ChIP-chip assay more then 4000 HNF4α target genes were identified. By comparing the list of genes identified in the ChIP-chip assay with the list of genes expressed in liver the authors determined combinatorial co-occupancy of binding sites of different factors in promoters of HNF4α target genes. Furthermore, this feature correlated well with expression of these genes in hepatocytes. This agrees well with our findings and confirms the utility of our method in defining the local and global context for specific combinations of different TF binding sites in the vicinity of functionally active HNF4alpha binding sites of promoters of genes whose expression is regulated by HNF4α. Notably, the combination of PWMs identified by the genetic algorithm (see Table 2) captured two transcription factors, i.e. FOX, CREB. Strikingly, these factors which were identified independently by Odom et al. [45] in an analysis of TF binding sites which co-accured with HNF4α sites (Table 2, matrices marked by the star).

In a further study of Odom et al. [46] the authors showed that two-thirds of the binding sites identified by ChIP-chip experiments are not conserved between human and mouse.

Taking into account the quite conservative liver expression patterns of genes between these two species we can conclude that far not all HNF4alpha binding sites identified by ChIP-chip method are directly contributing to the regulation of gene expression

In order to experimentally validate our predictions we selected two sets of promoters. The first set contained 10 *ab initio* and therefore novel HNF4α recognition sites predicted by the computational complex recognition criteria described above. Strikingly, 8 out of 10 binding sites could be confirmed as HNF4α binding in electromobility supershift experiments, i.e. NCOA2, TFF2, CHEK1, CD63, SH3Gl2, RND2, ESRRBL2 and DDB1 (see also Figure 6B). In addition, we studied another set of 10 promoters that were reported by Odom et al. [11] as targets for HNF4α, but our computational method rejected them because of extremely low scores of the HNF4 weight matrix as well as low scores of local and global context. None of these sites, i.e. NPAS2, GPHN, PPP1R3C, AKR1C3, CFL2, MDM2, CLCN3, CBX3, AZI2 and C14orf119, did in fact bind to HNF4α, as shown by electromobility supershift assays (see Figure 6C). These findings suggest a high error rate concerning the proposed targets by Odom et al. [11].

Finally, another computational approach [47] has been applied to analyze the same set of HNF4α (as well as HNF1 and HNF6) ChIP-chip data that was the focus of our current study. Indeed, Smith and co-authors demonstrated that an application of combinations of motifs allowed for improvements in the prediction of the genomic location of TF binding sites. In contrast to our approach, however, they performed a blind motif discovery instead of using the existing TF weight matrices. To the best of our knowledge, this makes the algorithm very complicated and increases the risk to miss important TF combinations that are characteristic for the functionally active regulatory sites.

Several further improvements of our algorithm can be considered in future. Among the most important, we should consider the possibility to take into account sequence conservation in the non-coding regulatory regions of genes between different species, e.g. human and mouse. It was demonstrated in recent studies [48 - 51] that sequence conservation can be a good indication of the functional importance of the region. Indeed, such regions can bear functional TF binding sites. Despite being quite useful, such considerations should be taken with care since regulatory regions are characterized by a high level of convergent evolution which can provide non-divergent means of forming a functional context of a TF binding site.

Another direction for further improvements is considering known protein-protein interaction data between different transcription factors. Such data are partially available in databases as such TRANSFAC®, TRANSPATH® and BIND. Known interactions between transcription factors can help to find proper combinations of neighboring binding sites for these factors.

A further step in improving our method will be the use of PWMs for factors whose expression is tissue specific, as indicated in TRANSFAC. This will greatly improve the predictive power of the method. To achieve that, more extensive annotation of expression information of transcription factors is needed and will be the task of the future. One possibility for obtaining this information resides in ChIP-chip experiments in conjunction with gene expression data. This will help to identify transcription factor activity in a given cellular environment.

Taking gene expression data into account will significantly help to determine the global and local sequence context and therefore functional transcription factor binding sites. Recently, we applied the algorithm described in this paper to analyze promoters of differentially expressed genes [2, 15, 51]. Such an integrative approach is now available in the software system *ExPlain* for a mechanistic interpretation of gene expression changes in various physiological and pathological situations of eukaryotic cells [52].

Conclusions

We report a new approach based on machine learning techniques for the *de novo* identification of novel HNF4α binding sites. The genetic algorithms developed by us significantly improved data analysis of various experimental sources. The here described method can be applied to any transcription factor and enables computational prediction of genome-wide functional transcription factor binding sites. By applying our method, interactions between different transcription factors can be taken into account. This provides clues for the mechanisms responsible for promoter activation and even for antagonistic binding of transcription factors, as it is known for HNF4α and Coup-TF who successfully compete for the same binding site but differ in activity under various biological conditions. Indeed, while both factors can bind to the same sequence, the individual local and global sequence context determines the actual binding activity and may therefore provide an estimate of transcription factor activity in particular cellular or physiological situations.

Material and methods

Databases

In this study, databases provided by BIOBASE GmbH were used, e.g, TRANSFAC® , which is a database on gene regulation [53]. It collects data on transcription factors and their binding sites in promoters and enhancers of eukaryotic genes as well as a library of PWMs. This work was done with TRANSFAC release 9.4. Additionally, to retrieve promoters of human genes we used *TRANSPro™ release 2.1* [54], which is based on genomic sequence from Ensembl release v35, Nov 2005. Final verification of the composite modules was done with the help of TRANSCompel™ database [4].

HNF4α binding sites in genome

In this work, we significantly updated the collection of known genomic HNF4α sites in TRANSFAC®. Additional file 1 lists the collected sites with information about the target genes, positions in the promoters of the genes, and the site sequence.

First of all, we compiled all known HNF4 binding sites from the literature and extended them upstream (28bp) and downstream (34bp). This is set as Y-local. After that, we prepared the background sequences. This is set as N-local. After that, we split set Y-local into two parts: “training set” and “test set” (sites included into the “training set” are indicated in the additional file 1 by stars). And we split the N-local into two parts: “training background set” and “test background set”. The training of the method was done by comparison of training set versus training background set. The testing of the method and building of the histogram in the Figure 2 was done on test set versus test background set – on two sets that were not used in the training. This procedure of preparing of four sets is the best possible statistical procedure for training and testing of the recognition methods.

Positional weight matrix (= PWM) for HNF4α binding sites

Based on the collection of HNF4α sites we constructed a position weight matrix (acc. Number: M01031) and two "half-site" matrices (acc. numbers: M01032, M01033) and deposited them in the TRANSFAC database (see Table 5). Construction of PWMs was done according the general outline described in [51] and as detailed in the protocol of TRANSFAC matrix construction (see TRANSFAC documentation). The half-site matrices were created by manual splitting of each site into two parts and were used independently for the alignment. Together with preexisting HNF4α matrices in TRANSFAC (acc. numbers: M00762, M00764, and M00967), the new matrices were used to search for HNF4α binding sites in genomic sequences. For this basic search we employed the MATCH™ algorithm calculating scores for the matches by applying the so-called information vector [55]. This algorithm is implemented in the ExPlain™ software system. This software was also used for analysis of the flanking regions of HNF4α sites to search for other TF binding sites from the most up-to-date library of matrices derived from the TRANSFAC® Professional database. The cut-offs for the matrices were set to minFN to maximize the sensitivity of the site prediction (false negative rate of 10%).

Machine learning techniques for building methods of identification of genomic sites

In order to identify novel and functional HNF4α binding sites in the human genome, we first analyzed flanking regions of DNA binding sites for this factor and determined what kind of additional contextual rules appeared to be molecular descriptors for these sites. We used two machine learning techniques for revealing such rules and applied them for building methods for recognition of functional genomic HNF4α sites. The techniques used here are similar to the methods applied for recognition of AhR sites [2]. The main ideas of the techniques as well as their recent improvements are reported here.

Defining local context. Revealing short sequence motifs with the help of an exhaustive feature selection algorithm.

Initially, we interrogated short sequence motifs with the help of an *exhaustive feature selection* algorithm and searched for cliques in the feature correlation graph. Specifically, we analyzed flanking regions of HNF4α bindings sites, 28 bp upstream and 33 bp downstream, and applied a modification of the search algorithm that was recently developed [14, 33]. It should be noted that the algorithm searches for a specific composition of over- and underrepresented short oligonucleotides (= features of local context) in the flanking regions of HNF4α sites and uses them for construction of a site recognition method. In the first step, the algorithm - through an exhaustive search - selects a set of such features of the local context. In the second step, it creates a graph of correlations between the features, selects non-redundant combinations of them through identification of cliques in the graph and builds the site recognition method using a final set of features of local context.

The algorithm compares two sets of sequences of equal length L: a training set $Y_{\mathbf{Local}}$ consisting of the functional HNF4α sites including their flanking regions (see additional file 1), and a background set of sequences $N_{\mathbf{Local}}$. The $N_{\mathbf{Local}}$ set is made by running the TRANSFAC® accompanying tool Match^{TM} [55] in the set of human intergenic regions (located at least 1 Mb from any known gene) using the HNF4α position weight matrix (acc. number M00967 with the score cut-off value $q_{cut\text{-}off} = 0.8$. This cut-off which guarantees recognition of all known sites collected in additional file 1). Then, we randomly selected 100 matches together with their – 28bp and +33 bp flanks and placed them into the background set $N_{\mathbf{Local}}$. Therefore, the set $N_{\mathbf{Local}}$ consists of sequences that contain a central motif fitting to the HNF4α matrix, however, because of its position in the genome on such large distances from any known gene and also since the threshold was so low in the motif matching, a randomly chosen subset is likely to contain mostly false positives. By comparison of the sets $Y_{\mathbf{Local}}$ and $N_{\mathbf{Local}}$ we could reveal contextual features that characterize the sequence environment (local context) of functional HNF4α sites. In addition, such comparison allowed us to reveal features of the core motif at the HNF4α binding site that are not captured by the positional weight matrix method alone (e.g. correlation between positions of the site).

We extended the approach described in Kel et al. [2] and consider now 3 types of contextual features *(φ)*: 1) frequency of occurrence of short motifs $\lambda = (a_1 a_2 .. a_k)$ ($a \in \{A,T,G,C,W,S,R,Y,M,K,B,V,H,D,N\}$.(we use the following one-letter code for different combinations of alternative nucleotides: W-(A/T, (read A or T)); R-(A/G); M-(A/C); K-(T/G); Y-(T/C); S-(G/C); B-(T/G/C);V-(A/G/C); H-(A/T/C); D-(A/T/G); N-(A/T/G/C)) of the length $k \leq 4$ in a window $w = [t_1,t_2]$ ($0 < t_2 < t_1 < L-k+1$); 2) frequency in the same window of dinucleotide pairs: $(\lambda \rightarrow \delta)$, where $\lambda = (a_1 a_2)$ and $\delta = (b_1 b_2)$ with a distance between them varying from r_{min} to r_{max}; 3) frequency of four-nucleotide repeats $\lambda^2 = (\lambda \rightarrow \lambda)$, where $\lambda = (a_1 a_2 a_3 a_4)$ with the varying distance r_{min} to r_{max}.

In our previous work [2, 14, 33], we described the statistical criteria which are based on utility theory that permits an identification of single motifs λ and the windows *w* that are characterized by a significant difference of their frequencies $f(\lambda,w,S)$ in the sequences S from the sets Y and N. Here, we extend this algorithm to an identification of significant pairs of dinucleotides and four-nucleotide repeats. Found contextual features are then used for creating a context analyzer that is able to perform an additional filtering of the potential sites as revealed by the weight matrix method.

The context analyzer is developed in two steps. On the first step, we perform feature selection. For that, we analyze correlation between all contextual features found by the statistical criteria described above and choose a limited set of features which are characterized by the lowest level of mutual correlations using the means of an algorithm revealing maximal cliques in a weighted graph. The contextual features selected at the previous step (φ_1, φ_2,..., φ_m) (10-20 relatively independent features) are used for construction of a linear classification function discriminating sets $Y_{\mathbf{Local}}$ and $N_{\mathbf{Local}}$. So, for every sequence X we calculate the score of context:

$$d = \beta + \sum_{i=0}^{m} \alpha_i \times f(\varphi_i, w_i, X) \qquad (1)$$

where α_i and β are the coefficients of the discriminating function. This coefficients are obtained through least-square method of estimating linear regression.

In order to validate the obtained context analyzer we applied it to the control sets: $Y_{\text{Local-Control}}$ and $N_{\text{Local-Control}}$, that do not contain sequences used at the training steps. The set $Y_{\text{Local-Control}}$ consists of new HNF4α sites annotated most recently in TRANSFAC® and $N_{\text{Local-Control}}$ was constructed through the same procedure as described above but using also randomly chosen but different human intergenic regions compared to those used to construct N_{Local}.

It should be mentioned here, that repeating the same training procedure with a new random negative set may result in different set of features for the context analyzer. This happened due to the procedure of feature selection, which does not necessarily select the full set of important features but gives the most discriminative sub-subset of "representative" mutually uncorrelated features. Different independent runs of the algorithm can result in different subsets of features, although, often, the most discriminative features do not change, and several other features, being literally different, nevertheless represent quite similar oligonucleotides in similar position windows. Such different sets of contextual features usually achieve similar recognition power (similar level of sensitivity and specificity) and can characterize different "sub-populations" of the training set of the sites by looking at it from different "angle" of different background sequences. Under the condition of rather limited training set of real sites, and in order to avoid overfitting, we can not increase the number of selected features too much. So, for further analysis we take the set of mutually uncorrelated features obtained in the first run of the *exhaustive feature selection* algorithm.

Defining the global context: identification of composite modules in HNF4α-targeted promoters using a genetic algorithm.

Composite regulatory modules in the promoter regions flanking functional HNF4α binding sites were identified by using the recently developed software tool CMA [56]

Potential TF binding sites in the flanking regions were identified by Match^{TM} [55] that uses a library of about 500 PWMs for vertebrate transcription factors (TRANSFAC® release 9.4).

CMA was applied for analyzing combinations of TF binding sites (composite modules, CMs) in promoters of differentially expressed genes. The definition of a CM is now significantly improved compared to the previous application [2]. It is defined as a set of individual PWMs and pairs of PWMs that are characteristic for co-regulated promoters. CMs are characterized by the following parameters: *K,* the number of individual PWMs in the module, *R*, the number of pairs of PWMs, cut-off value $q_{cut-off}^{(k)}$, relative impact values $\phi^{(k)}$, maximal number of best matches $\kappa^{(k)}$ that were assigned to every weight matrix *k* (*k = 1,K*), as well as cut-off value $q_{cut-off}^{(r)}$, relative impact values $\phi^{(r)}$ and maximal distances $d_{\max}^{(r)}$ and maximal number of best matches $\kappa^{(r)}$ that were assigned to every matrix pair *r* (*r = 1,R*) in the CM. A composite module score (CM score) is calculated for all sequences *X* according to the following equation:

$$v(X) = \sum_{k=1,K_i} \varphi^{(k)} \times \sum_{j=1}^{\kappa^{(k)}} q_j^{(k)}(x) + \sum_{r=1,R_i} \varphi^{(r)} \times \sum_{i=1}^{\kappa^{(r)}} (q_{1,i}^{(r)}(x) + q_{2,i}^{(r)}(x)) \qquad (2)$$

where $q_j^{(k)}(x)$ is the score of *j*-th match of the *k*-th PWM and $q_j^{(k)}(x) > q_{cut-off}^{(k)}$; and $q_{1,i}^{(r)}(x)$ and $q_{2,i}^{(r)}(x)$ are scores of two sites in a pair *r* and $q_{1,i}^{(r)}(x), q_{2,i}^{(r)}(x) > q_{cut-off}^{(r)}$ and the distance between these sites in the pair: $d_{\min}^{(r)} < d^{(r)} < d_{\max}^{(r)}$. Normalization is then applied as in Waleev et al. [15].

So, if $\nu(X)$ is higher than a predefined threshold $\nu_{cut-off}$, the program reports a match of the composite module to the sequence.

The CMA program is based on a genetic algorithm. It takes as input two sets of sequences (set under study Y_{Global} and the background set N_{Global}) and a set of PWMs for transcription factors. The program optimizes parameters *K* and *R,* the set of matrices selected, their number, their cut-offs, the relative impact, and the maximum number of best matches. We defined the fitness function of the genetic algorithm as a weighted sum of several statistical parameters characterizing the difference between two distributions - the distributions of the CM scores (ν(X)) in the two sets of promoters, as described in Kel et al. [56]. Calculating the fitness function allows to assess the usability of the obtained solutions for classification of individual sequences. The output of the program is the best discriminative CM with the optimized parameters.

Molecular biology experiments

To confirm predicted functional HNF4α binding sites, we performed electromobility supershift assays with nuclear extracts of Caco-2 cells. Note that this cell line is well characterized for its abundant expression of HNF4α, as reported elsewhere [57].

Isolation of nuclear extracts

Nuclear extracts from Caco-2 cells were isolated by a modified method of Dignam et al. [58]. Eleven days after seeding, cells were washed twice with ice-cold phosphate buffered saline (PBS), scraped into microcentrifuge tubes and centrifuged for 5 min at 2000 x g, 4 °C. Cell pellets were resuspended in hypotonic buffer (10 mM Tris, pH 7.4, 2 mM $MgCl_2$, 140 mM NaCl, 1 mM DTT, 4 mM Pefabloc, 1% Aprotinin, 40 mM ß-glycerophosphate, 1 mM sodiumorthovanadate and 0.5% TX100) at 4 °C for 10 min (300 µl for 1 x 10^7 cells), transferred onto one volume of 50% sucrose in hypotonic buffer and centrifuged at 14000 x g and 4 °C for 10 min. Nuclei were resuspended in Dignam C buffer (20 mM Hepes, pH 7.9, 25% glycerol, 420 mM NaCl, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 1 mM DTT, 4 mM Pefabloc, 1% Aprotinin, 40 mM ß-glycerophosphate, 1 mM sodiumorthovanadate, 30 µl for 1 x 10^7 cells) and gently shaken at 4 °C for 30 min. Nuclear debris was removed by centrifugation at 14000 x g at 4 °C for 10 min. The extracts were aliquoted and stored at –70 °C.

Electrophoretic Mobility Shift Assay (EMSA)

The oligonucleotides were purchased from MWG Biotech (Ebersberg/Muenchen, Germany); for sequence information see Table 6. The central five nucleotides are highlighted. 2.5 µg nuclear extract and 10^5 cpm (0.027 ng) ^{32}P-labeled oligonucleotides were incubated in binding buffer consisting of 25 mM Hepes, pH 7.6, 5 mM $MgCl_2$, 34 mM KCl, 2 mM DTT, 2 mM Pefabloc, 2% Aprotinin, 40 ng of poly (dI-dC)/µl and 100 ng of bovine serum albumin/µl. Oligonucleotides and nuclear proteins were incubated for 20 minutes on ice. Free DNA and DNA-protein complexes were resolved on a 6% polyacrylamide gel. Supershift experiments were done with an HNF4α-specific antibody (sc-6556x, Santa Cruz Biotechnology, Heidelberg, Germany). Gels were blotted to Whatman 3 MM paper, dried under vacuum, exposed to imaging screens for autoradiography and analyzed using a phosphor imaging system (Molecular Imager FX pro plus; Bio-Rad Laboratories GmbH, Muenchen, Germany).

Abbreviations

AR, Androgen receptor; ChIP, Chromatin immunoprecipitation; CM, Composite module; CRM, Composite regulatory module; CMA, Composite module analyst; DR1, Direct repeat 1; ENCODE, Encyclopedia of DNA elements; GR, Glucocorticoid receptor; HNF4α, hepatic nuclear factor 4α; HMM, Hidden Markov model; PBS, phosphate buffered saline; PWM, Positional weight matrices; TF, Transcription factor; TFBS, Transcription factor binding site; TSS, Transcription start site.

Authors' contributions

AK & JB are responsible for the conceptional design of the study. AK developed the genetic algorithms, the selection schemes and analyzed the results. Computational and simulation analysis were done by AK, VM and RZ. Authors VM and RZ helped to code the algorithms. Authors MN & JB are responsible for the molecular biology studies. AK & JB drafted the manuscript and are responsible for its final content. All authors critically evaluated the data. All authors read and approved the final manuscript.

Additional data files

The following additional data are available with the online version of this paper. Additional file 1 informs on TRANSFAC database annotated HNF4alpha binding sites. Additional data file 2 is a table listing 375 sequences that passed our criterion.

Acknowledgements

Parts of the work were funded by a grant from the German Federal Ministry of Education and Research (BMBF) together with BioRegioN GmbH “BioProfil”, grant no. 0313092; EU grants “TRANSISTOR” (MRTN-CT-2004-512285), “EuroDia” (LSHM-CT-2006-518153), VALAPODYN (LSHG-CT-2006-037277) and Net2Drug (LSHB-CT-2007-037590).

The authors confirm that there is no conflict of interest.

References

1. Merika M, Thanos D: **Enhanceosomes.** *Curr Opin Genet Dev* 2001*,* **11:**205-208.

2. Kel A, Reymann S, Matys V, Nettesheim P, Wingender E, Borlak J: **A novel computational approach for the prediction of networked transcription factors of aryl hydrocarbon-receptor-regulated genes.** *Mol Pharmacol* 2004, **66:**1557-1572.

3. Jin VX, Rabinovich A, Squazzo SL, Green R, Farnham PJ: **A computational genomics approach to identify cis-regulatory modules from chromatin immunoprecipitation microarray data--a case study using E2F1.** *Genome Res* 2006, **16:**1585-1595.

4. Kel-Margoulis O, Kel AE, Reuter I, Deineko IV, Wingender E: **TRANSCompel: a database on composite regulatory elements in eukaryotic genes***.* *Nucleic Acids Res* 2002, **30:**332-334.

5. Schrem H, Klempnauer J, Borlak J: **Liver enriched transcription factors in liver function and development. Part II: The C/EBPs and DBP in cell cycle control, carcinogenesis, circadian gene regulation, liver regeneration, apoptosis, and liver-specific gene regulation.** *Pharmacol Rev* 2004, **56:**291-330.

6. Schrem H, Klempnauer J, Borlak J: **Liver enriched transcription factors in liver function and development. Part I: The hepatocyte nuclear factor network and liver-specific gene expression.** *Pharmacol Rev* 2002, **54:**129–158.

7. Naiki T, Nagaki M, Shidoji Y, Kojima H, Imose M, Kato T, Ohishi N, Yagi K, Moriwaki H: **Analysis of gene expression profile induced by hepatocyte nuclear factor 4alpha in hepatoma cells using an oligonucleotide microarray.** *J Biol Chem* 2002, **277:**14011-14019.

8. Thomas H, Senkel S, Erdmann S, Arndt T, Turan G, Klein-Hitpass L, Ryffel GU: **Pattern of genes influenced by conditional expression of the transcription factors HNF6, HNF4alpha and HNF1beta in a pancreatic beta-cell line.** *Nucleic Acids Res* 2004, **32:**e150.

9. Lucas B, Grigo K, Erdmann S, Lausen J, Klein-Hitpass L, Ryffel GU: **HNF4alpha reduces proliferation of kidney cells and affects genes deregulated in renal cell carcinoma.** *Oncogene* 2005, **24:**6418-6431.

10. Battle MA, Konopka G, Parviz F, Gaggl AL, Yang C, Sladek FM, Duncan SA: **Hepatocyte nuclear factor 4alpha orchestrates expression of cell adhesion proteins during the epithelial transformation of the developing liver.** *Proc Natl Acad Sci USA* 2006, **103:**8419-8424.

11. Odom DT, Zizlsperger N, Gordon DB, Bell GW, Rinaldi NJ, Murray HL, Volkert TL, Schreiber J, Rolfe PA, Gifford DK, Fraenkel E, Bell GI, Young RA: **Control of Pancreas and Liver Gene Expression by HNF Transcription Factors.** *Science* 2004, **303:**1378-1381.

12. Rada-Iglesias A, Wallerman O, Koch C, Ameur A, Enroth S, Clelland G, Wester K, Wilcox S, Dovey OM, Ellis PD, Wraight VL, James K, Andrews R, Langford C, Dhami P, Carter N, Vetrie D, Ponten F, Komorowski J, Dunham I, Wadelius C: **Binding sites for metabolic disease related transcription factors inferred at base pair resolution by chromatin immunoprecipitation and genomic microarrays.** *Hum Mol Genet* 2005, **14:**3435-3447.

13. Nishiyama C, Hi R, Osada S, Osumi T: **Functional interactions between nuclear receptors recognizing a common sequence element, the direct repeat motif spaced by one nucleotide (DR-1).** *J Biochem* 1998, **123:**1174-1179.

14. Kel AE, Ponomarenko MP, Likhachev EA, Orlov YL, Ischenko IV, Milanesi L, Kolchanov NA: **SITEVIDEO: A computer System for Functional site Analysis and Recognition. Investigation of the human splice sites.** *Comput Applic Biosci* 1993, **9:**617-627.

15. Waleev T, Shtokalo D, Konovalova T, Voss N, Cheremushkin E, Stegmaier P, Kel-Margoulis O, Wingender E, Kel A: **Composite Module Analyst: Identification of transcription factor binding site combinations using genetic algorithm.** *Nucleic Acids Res* 2006, **34:**W541-W545.

16. Lee SK, Choi HS, Song MR, Lee MO, Lee JW: **Estrogen receptor, a common interaction partner for a subset of nuclear receptors.** *Mol Endocrinol* 1998, **12:**1184-1192.

17. Louet JF, Hayhurst G, Gonzalez FJ, Girard J, Decaux JF: **The coactivator PGC-1 is involved in the regulation of the liver carnitine palmitoyltransferase I gene expression by cAMP in combination with HNF4 alpha and cAMP-response element-binding protein (CREB).** *J Biol Chem* 2002, **277:**37991-38000.

18. Nitsch D, Boshart M, Schuetz G: **Activation of the tyrosine aminotransferase gene is dependent on synergy between liver-specific and hormone-responsive elements.** *Proc Natl Acad Sci USA* 1993, **90:**5479-5483.

19. Galson DL, Tsuchiya T, Tendler DS, Huang LE, Ren Y, Ogura T, Bunn HF: **The orphan receptor hepatic nuclear factor 4 functions as a transcriptional activator for tissue-specific and hypoxia-specific erythropoietin gene expression and is antagonized by EAR3/COUP-TF1.** *Mol Cell Biol* 1995, **15:**2135-2144.

20. Liu Y, Yang N, Teng CT: **COUP-TF acts as a competitive repressor for estrogen receptor-mediated activation of the mouse lactoferrin gene.** *Mol Cell Biol* 1993, **13:**1836-1846.

21. Nakajima K, Kusafuka T, Takeda T, Fujitani Y, Nakae K, Hirano T: **Identification of a novel interleukin-6 response element containing an Ets-binding site and a CRE-like site in the junB promoter.** *Mol Cell Biol* 1993, **13:**3027-3041.

22. Ponomarenko JV, Ponomarenko MP, Frolov AS, Vorobyev DG, Overton GC, Kolchanov NA: **Conformational and physicochemical DNA features specific for transcription factor binding sites.** *Bioinformatics* 1999, **15:**654-668.

23. Ellrott K, Yang C, Sladek FM, Jiang T: **Identifying transcription factor binding sites through Markov chain optimization.** *Bioinformatics* 2002, **18**(Suppl 2):S100-S109.

24. Podvinec M, Kaufmann MR, Handschin C, Meyer UA: **NUBIScan, an in Silico Approach for Prediction of Nuclear Receptor Response Elements.** *Mol Endocrinol* 2002, **16:**1269-1279.

25. Sandelin A, Wasserman WW: **Prediction of nuclear hormone receptor response elements.** *Mol Endocrinol* 2005, **19:**595-606.

26. Frech K, Quandt K, Werner T: **Muscle actin genes: a first step towards computational classification of tissue specific promoters.** *In Silico Biol* 1998, **1:**29-38.

27. Wasserman WW, Fickett JW: **Identification of regulatory regions which confer muscle-specific gene expression.** *J Mol Biol* 1998, **278:**167-181.

28. Tronche F, Ringeisen F, Blumenfeld M, Yaniv M, Pontoglio M: **Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome.** *J Mol Biol* 1997, **266:**231-245.

29. Brazma A, Vilo J, Ukkonen E: **Finding Transcription Factor Binding Site Combinations in Yeast Genome.** In *Computer Science and Biology. Proc German Conference on Bioinformatics GCB: 22-24 September 1997; Martinsried, Germany.* Edited by Frishman D, Mewes HW; 1997:57-59.

30. Boehlk S, Fessele S, Mojaat A, Miyamoto NG, Werner T, Nelson EL, Schlondorff D, Nelson PJ: **ATF and Jun transcription factors, acting through an Ets/CRE promoter module, mediate lipopolysaccharide inducibility of the chemokine RANTES in monocytic Mono Mac 6 cells.** *Eur J Immunol* 2000, **30:**1102-1112.

31. Fessele S, Boehlk S, Mojaat A, Miyamoto NG, Werner T, Nelson EL, Schlondorff D, Nelson PJ: **Molecular and in silico characterization of a promoter module and C/EBP element that mediate LPS-induced RANTES/CCL5 expression in monocytic cells.** *FASEB J* 2001, **15:**577-579.

32. Kel A, Kel-Margoulis O, Babenko V, Wingender E: **Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells.** *J Mol Biol* 1999, **288**:353-376.

33. Kel AE, Kel-Margoulis OV, Farnham PJ, Bartley SM, Wingender E, Zhang MQ: **Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors**. *J Mol Biol* 2001, **309:**99-120.

34. Shelest E, Wingender E: **Construction of predictive promoter models on the example of antibacterial response of human epithelial cells.** *Theor Biol Med Model* 2005, **2:**2.

35. Segal E, Sharan R: **A discriminative model for identifying spatial cis-regulatory modules.** *J Comput Biol* 2005, **12:**822-834.

36. Liu X, Brutlag DL, Liu JS: **BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes.** *Pac Symp Biocomput* 2001, **6**:127-138.

37. Guha Thakurta D, Stormo GD: **Identifying target sites for cooperatively binding factors.** *Bioinformatics* 2001, **17:**608-621.

38. Eskin E, Pevzner PA: **Finding composite regulatory patterns in DNA sequences.** *Bioinformatics* 2002, **18**(Suppl 1):S354-S363.

39. van Helden J, Rios AF, Collado-Vides J: **Discovering regulatory elements in non-coding sequences by analysis of spaced dyads.** *Nucleic Acids Res* 2000, **28:**1808-1818.

40. Kel-Margoulis O, Ivanova TG, Wingender E, Kel AE: **Automatic annotation of genomic regulatory sequences by searching for composite clusters.** *Pac Symp Biocomput* 2002, **7:**187-198.

41. Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B: **Computational detection of cis-regulatory modules.** *Bioinformatics* 2003, **19**(Suppl 2):ii5-ii14.

42. Sinha S, van Nimwegen E, Siggia ED: **A probabilistic method to detect regulatory modules.** *Bioinformatics* 2003, **19**(Suppl 1):i292–i301.

43. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W: **Aligning multiple genomic sequences with the threaded blockset aligner.** *Genome Res* 2004, **14:**708–715.

44. Gupta RK, Gao N, Gorski RK, White P, Hardy OT, Rafiq K, Brestelli JE, Chen G, Stoeckert CJ Jr, Kaestner KH: **Expansion of adult beta-cell mass in response to increased metabolic demand is dependent on HNF-4alpha.** *Genes and Development* 2007, **21:**755-769.

45. Odom DT, Dowell RD, Jacobsen ES, Nekludova L, Rolfe PA, Danford TW, Gifford DK, Fraenkel E, Bell GI, Young RA: **Core transcriptional regulatory circuitry in human hepatocytes.** *Mol Syst Biol* 2006, **2:**2006.0017.

46. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E. **Tissue-specific transcriptional regulation has diverged significantly between human and mouse.** *Nat Genet* 2007, **39:**730-732.

47. Smith AD, Sumazin P, Das D, Zhang MQ: **Mining ChIP-chip data for transcription factor and cofactor binding sites.** *Bioinformatics* 2005, **21**(Suppl 1):i403-i412.

48. Cheremushkin E, Kel A: **Whole genome human/mouse phylogenetic footprinting of potential transcription regulatory signals.** *Pac Symp Biocomput* 2003, **8**:291-302.

49. Liu Y, Liu XS, Wei L, Altman RB, Batzoglou S: **Eukaryotic regulatory element conservation analysis and identification using comparative genomics.** *Genome Res* 2004, **14:**451-458.

50. Sauer T, Shelest E, Wingender E: **Evaluating phylogenetic footprinting for human-rodent comparisons.** *Bioinformatics* 2006, **22:**430-437.

51. Moehle C, Ackermann N, Langmann T, Aslanidis C, Kel A, Kel-Margoulis O, Schmitz-Madry A, Zahn A, Stremmel W, Schmitz G: **Aberrant intestinal expression and allelic variants of mucin genes associated with inflammatory bowel disease.** *J Mol Med* 2006, **84:**1055-1066.

52. Wingender E, Crass T, Hogan JD, Kel AE, Kel-Margoulis OV, Potapov AP: **Integrative content-driven concepts for bioinformatics "beyond the cell".** *J Biosci* 2007, **32:**169-180.

53. Matys V, Kel-Margoulis O, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel A, Wingender E: **TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes.** *Nucleic Acids Res* 2006, **34:**D108-D110.

54. Chen X, Wu JM, Hornischer K, Kel A, Wingender E: TiProD: **The Tissue-specific Promoter Database.** *Nucleic Acids Res* 2006, **34:**D104-D107.

55. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E: **MATCH: A tool for searching transcription factor binding sites in DNA sequences.** *Nucleic Acids Res* 2003, **31:**3576-3579.

56. Kel A, Konovalova T, Waleev T, Cheremushkin E, Kel-Margoulis O, Wingender E: **Composite Module Analyst: a fitness-based tool for identification of transcription factor binding site combinations.** *Bioinformatics* 2006, **22:**1190-1197.

57. Niehof M, Borlak J: **RSK4 and PAK5 are novel candidate genes in diabetic rat kidney and brain.** *Mol Pharmacol* 2005, **67:**604-611.

58. Dignam JD, Lebovitz RM, Roeder RG: **Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei.** *Nucleic Acids Res* 1983, **11:**1475-1489.

Figure legends

Figure 1. Repeats in the structure of HNF4 binding sites (from TRANSFAC®). a) examples of multiple repeats forming canonical DR1 as well as DR2, inverted (IR) and "everted" (ER) repeats. The centrally located arrows, marked as DR1 or DR2, indicate the repeat with the maximal score (sum of the scores of single elements); b) statistic of the repeats of different types (direct repeats - DR0-4, everted repeats – ER0-4 and inverted repeats IR0-4) in the structure of HNF4 sites. Dark bars show the observed number of repeats found in the structure of 73 sequences of known HNF4 binding sites (listed in additional file 1) considering one repeat with the maximal score per sequence. Grey bars show the total number of repeats found in this set of HNF4 sites.

Figure 2. Two histograms of the distributions of the **score of context** in the –28bp/+33bp flanks of real HNF4α sites (test set, see additional file 1) *versus* –28bp/+33bp flanks of PWM matches (PWM score > 0.8) in random genomic positions. Mean values of the two distributions are 0.5849 and 0.348 respectively. On the axis: x-axis – score of context; left y-axis – number of PWM matches in random genomic positions with the corresponding score of context; right y-axis – number of real HNF4α sites with the corresponding score of context.

Figure 3. Two histograms of distributions of the **CM (composite module) score** in the +/-500bp flanks of HNF4α sites (solid bars) *versus* +/-500bp flanks of PWM matches (PWM score > 0.8) in random genomic positions (empty bars). The average *CM* score for real HNF4α sites equals 0.499, whereas for PWM matches in random genomic positions (in the set $N_{\mathbf{Global}}$) equals 0.050 (ratio =9.98 t-test p-value =1.4896*10^{-26}).

Figure 4. Plot of the distribution of global and local context in the 375 sequences (red squares) selected from the "positive" set of ChIP-chip results reported by Odom et al. [11] versus all 10852 sequences from the "negative" (not binding) (H13K_noHNF4) set (green dots) reported for the same experiment. The selected sequences are characterized by highest scores of global and local context whereas the majority of the "negative" sequences are characterized by the low values of these two scores. With the vertical and horizontal lines we show two thresholds chosen for the global context score (0.28) and local context score (0.18).

Figure 5. Percentages of sequences passing the complex recognition criteria in the set of known HNF4 binding sites (TRANSFAC® HNF4 sites), in the set of "positive" sequences based on ChiP-chip experiments reported by Odom et al. [11] for hepatocytes and in the set of "negative" sequences described by Odom et al. [11]. From the last set we estimate that the percentage of false results of our method is about 2.6%.

Figure 6. EMSA confirmation experiments.

A. EMSA with established HNF4α recognition sites. Electrophoretic mobility shift experiment with 2.5 µg Caco-2 cell nuclear extracts and oligonucleotides corresponding to promoter regions derived from HFN1, AAT, APOB, AGT, APOC3, CYP2D6, TF, ALDH2, APOC2 and PCK1 as ^{32}P labeled probes. For supershift analysis an antibody directed against HNF4α was added (+). **B.** EMSA with predicted novel HNF4α recognition sites. Electrophoretic mobility shift experiment with 2.5 µg Caco-2 cell nuclear extracts and oligonucleotides corresponding to promoter regions derived from NCOA2, TFF2, CHEK1, CD63, SH3GL2, RND2, ESRRBL2, DDB1, NEUROG3 and IL6 as ^{32}P labeled probes. For supershift analysis an antibody directed against HNF4α was added (+). **C.** EMSA with potential recognition sites from putative HNF4α targets reported by Odom et al. [11]. Electrophoretic mobility shift experiment with 2.5 µg Caco-2 cell nuclear extracts and oligonucleotides corresponding to promoter regions derived from AZI2, CFL2, GPHN, C14orf119, PPP1R3C, AKR1C3, NPAS2, MDM2, CLCN3 and CBX3 as ^{32}P labeled probes. For supershift analysis an antibody directed against HNF4α was added (+).

Tables

Table 1. Oligonucleotides and short repeats found in the local context of genomic HNF4α sites. Mode: (I) – search for oligonucleotides, (II) – dinucleotide pairs, (III) – four-nucleotide repeats; wind_from, wind_to – sequence window in which the motif was found (HNF4α site is located between positions 29 and 42, flanks are 28 and 33 bp long respectively); rmin, rmax – distances between dinucleotide pairs and repeats; alpha – coefficients in the linear function (1); avrfreq_Y – the average frequency of the oligonucleotides in the corresponding window among sequences of the real sites and ,avrfreq_N – background sequences. The values of standard error is given in the parenthesis.

		wind_	**wind**					
Olig	**Mode**	**from**	**_to**	**rmin**	**rmax**	**alpha**	**avrfreq_Y**	**avrfreq_N**
MDDR	(I)	22	66	0	0	0.003082	13.433 (3.665)	11.051 (4.782)
ANGB	(I)	20	38	0	0	0.016132	5.358 (2.529)	3.582 (2.845)
CDDM	(I)	36	38	0	0	0.020372	4.346 (2.332)	3.212 (1.732)
AV – VS	(II)	1	34	33	33	0.008246	6.694 (3.663)	4.893 (3.088)
MD – DB	(II)	20	70	20	25	-0.003212	16.941 (3.078)	14.711 (3.344)
BR – NT	(II)	33	37	9	18	-0.003942	4.802 (5.460)	7.702 (5.144)
VS – YA	(II)	1	34	11	11	0.0103	4.237 (1.742)	3.121 (2.640)
VB – HA	(II)	1	34	5	5	0.008647	9.028 (2.985)	6.517 (3.741)
HM – GN	(II)	40	50	2	4	0.006468	7.783 (4.335)	4.672 (3.764)
$(RBNH)^2$	(III)	20	51	5	12	0.030376	7.259 (1.778)	5.961 (2.168)
$(MVKN)^2$	(III)	20	51	7	13	0.015979	3.123 (1.413)	2.388 (1.155)
$(BNDK)^2$	(III)	32	32	7	7	-0.002214	0.000 (0.000)	14.343 (28.652)
$(DNCD)^2$	(III)	28	42	7	7	0.068635	4.176 (2.797)	1.051 (2.196)
$(NBHV)^2$	(III)	26	26	7	7	-0.001045	0.000 (0.000)	13.626 (28.102)
$(NVYB)^2$	(III)	29	29	7	7	-0.001696	0.000 (0.100000)	12.909 (27.523)
							beta = -0.325584	

Table 2. Matrices and matrix pairs of the "global context" selected by CMA program. Matrix_ID(1) and Matrix_ID(2) are the TRANSFAC identifiers of the selected single matrix (or the first matrix in a pair and the second matrix in the pair, respectively). The other headings of the table correspond to the parameters of the composite module score (see equation (2) in the Methods section). The first six lines of the table represent the single matrices selected by the algorithm to represent the "global context", the other lines represent the pairs of matrices selected by the CMA program. The corresponding values of the parameters (in columns: cut-offs, κ and ϕ) are optimized by the CMA algorithm. (*) – matrices corresponding to TFs whose binding site combinatorial co-occupancy was found in Odom et al. [45] for promoters of liver-expressed genes.

Matrix_ID(1)	TFs(1)	cut-off(1)	Matrix_ID(2)	TFs(2)	cut-off(2)	dmin	dmax	κ	φ
V\$MAZ_Q6	**MAZ**	0.89						4	0.020763
V\$ER_Q6	**ER-α**	0.913						4	0.047177
V\$HEB_Q6	**HTF4**	0.969						4	0.078905
V\$HNF4_Q6_01(*)	**HNF-4α, HNF-4α2, HNF-4γ**	0.976						4	0.210340
V\$HEN1_02	**HEN1**	0.854						4	0.099368
V\$CREB_Q2(*)	**CRE-BP2,CREM, ATF-1,2,3,4,6**	0.888						4	0.086618
V\$HNF4_Q6_01(*)	**HNF-4α, HNF-4α2, HNF-4γ**	0.8325	**V\$EFC_Q6**	**RFX1 (EF-C)**	0.6825	8	100	2	0.043344
V\$COUP_01	**COUP-TF1, COUP-TF2,**	0.8005	**V\$KROX_Q6**	**Egr-1,2,3,4**	0.8315	8	100	2	0.053285
V\$PEBP_Q6	**PEBP2α/AML 1,3; PEBP2β**	0.84	**V\$TEL2_Q6**	**Tel-2a,b,c**	0.878	8	100	2	0.214469
V\$ELK1_01	**Elk-1**	0.785	**V\$WHN_B**	**FOXN1**	0.948	8	100	2	0.111909
V\$CMYB_01	**c-Myb, B-Myb**	0.86	**V\$KROX_Q6**	**Egr-1,2,3,4**	0.841	8	100	2	0.100922
V\$FOXO1_02(*)	**FOXO1,2,4, FOXJ3**	0.8715	**V\$FXR_Q3**	**FXRα/RXR**	0.8135	8	500	2	0.100184
V\$HNF4_Q6_01(*)	**HNF-4α, HNF-4α2, HNF-4γ**	0.8065	**V\$HNF4_01**(*)	**HNF-4α, HNF-4α2, HNF-4γ**	0.8705	8	200	2	0.080381
V\$XBP1_01	**XBP-1**	0.8845	**V\$FOXO1_02**(*)	**FOXO1,2,4, FOXJ3**	0.8715	8	200	2	0.112402
Intercept									-0.098626

Bold text indicates transcription factors identified by the CMA.

Table 3. Comparison of gene lists between HNF4α expression data, ChIP-chip data, and computational prediction of target promoters.

	Gene sets in ChIP-chip experiment		HNF4 targets identified by PWM V$HNF4_Q6_1 (cut-off=0.9)	HNF4 targets identified by local+global context
	Positive[2]	Negative[3]		
Gupta et al. [44] Up+Dn (133)	13 (9.8%)	ND	66 (49.6%)	41 (30.8%)
Naiki et al. [7] Up + Dn (75)[1]	17 (22.7%)	32 (42.7%)	15 (20%)	14 (18.7%)
Lucas et al. [9] Up + Dn (70)[1]	13 (18.6%)	39 (55.7%)	17 (24.3%)	13 (18.6%)
Lucas et al. [9] NC (150)[1]	20 (13.3%)	99 (66%)	29 (19.3%)	4 (2.7%)

[1] The number of genes whose expression was Up or Down by more than two fold; NC – genes with no change of expression.

[2] The number of differentially expressed genes with HNF4α binding sites as identified by ChIP-chip experiments.

[3] The number of differentially expressed genes with no HNF4α binding as determined by ChIP-chip experiments.

Table 4. Biological functions of novel predicted HNF4α gene targets.

Gene symbol	Gene name	Biological function
CD63	CD63 antigen (melanoma 1 antigen)	localization plasma membrane endocytosis
CHEK1	CHK1 checkpoint homolog	cell cycle negative regulation of cell proliferation DNA damage response
ESRRBL1	estrogen-related receptor beta like 1	induction of neuronal apoptosis
DDB1	damage-specific DNA binding protein 1, 127kDa	DNA repair
NCOA2	nuclear receptor coactivator 2	regulation of transcription signal transduction
RND2	Rho family GTPase 2	signal transduction protein transport dendrite development
SH3GL2	SH3-domain GRB2-like 2	central nervous system development signal transduction endocytosis
TFF2	trefoil factor 2	defense response digestion

Table 5. Positional weight matrix for HNF4α sites (TRANSFAC accession number M01031, identifier V$HNF4_Q6_01).

A	**19**	**5**	**11**	**5**	**2**	**52**	46	**49**	**0**	**3**	**3**	**1**	**46**	17
C	**8**	**2**	**3**	**16**	**48**	**1**	2	**0**	**0**	**1**	**19**	**47**	**2**	15
G	**21**	**46**	**30**	**20**	**2**	**0**	10	**9**	**58**	**26**	**7**	**2**	**6**	14
T	**10**	**5**	**14**	**17**	**6**	**5**	0	**0**	**0**	**28**	**29**	**8**	**4**	12
Consensus*	**N**	**G**	**G**	**N**	**C**	**A**	A	**A**	**G**	**K**	**Y**	**C**	**A**	N

* Two “half-matrices” (M01032, V$HNF4_Q6_02 and M01033, V$HNF4_Q6_03) corresponding to the DR1 repeat are given in bold and refer to the frequency of a nucleotide of the matrix.

Table 6. Shift-probe sequences.

Gene symbol[1]	Gene name[2]	Oligo-name[3]	Location (rel.TSS) [4]	Score[5]	Sequence[6]
TCF1/HNF1	hepatic nuclear factor 1	HNF1	-265	0.988	AAGGCTGAAGTC**CAAAG**TTCAGTCCCTTC
APOB	Apolipoprotein B	APOB	-86	0.905	GGAAAGGTC**CAAAG**GGCGCCTTG
SERPINA1/AAT	alpha-1-antitrypsin	GS21	-134	0.865	CAACAGGGG**CTAAG**TCCACTGGC
AGT	angiotensinogen	GS47	-429	0.905	TGCAGAGGG**CAGAG**GGCAGGGGA
APOC3	Apolipoprotein C3	GS104	-93	0.995	GGCGCTGGG**CAAAG**GTCACCT GC
CYP2D6	cytochrome P450, family 2, subfamily D, polypeptide 6	GS105	-69	0.989	AGCAGAGGG**CAAAG**GCCATCATC
TF	Transferring	GS106	-76	0.817	ACGGGAGGT**CAAAG**ATTGCGCCC
ALDH2	aldehyde dehydrogenase 2 family (mitochondrial)	GS107	-332	0.817	CATTGGGGT**CAAAG**GCACACATT
APOC2	apolipoprotein C2	GS108	-159	0.916	TGTCTAGGC**CAAAG**TCCTGGCCA
PCK1	phosphoenolpyruvate carboxykinase 1 (soluble)	GS109	-455	0.923	GGTCACAGT**CAAAG**TTCATGGGA
NCOA2	nuclear receptor coactivator 2	GS110	-485	0.981	ATGGGAGGG**CAAAG**GGCAATGCC
TFF2	trefoil factor 2	GS111	-495	0.978	AAGATGGGA**CAAAG**GGCATCGTG
CHEK1	CHK1 checkpoint homolog	GS112	+5	0.976	AGTGGTGGG**CAAAG**GACAGTCCG
CD63	CD63 antigen (melanoma 1 antigen)	GS113	-182	0.967	CTGCAGGAG**CAAAG**GACAGAAGT
SH3GL2	SH3-domain GRB2-like 2	GS114	-393	0.964	CGCCAGGCT**CAAAG**GGCAGGAGG
RND2	Rho family GTPase 2	GS115		0.923	AGGGCAGGT**CAGAG**TTCAAGCGA
ESRRBL1	estrogen-related receptor beta like 1	GS116	+63	0.91	CAGAACGGA**CAGAG**TCCAGCGTG

Gene symbol[1]	Gene name[2]	Oligo-name[3]	Location (rel.TSS) [4]	Score[5]	Sequence[6]
DDB1	damage-specific DNA binding protein 1, 127kDa	GS117	-295	0.909	GGGGAAGGG**CAAAG**GGCGCGGAA
NEUROG3	neurogenin 3	GS118	-225	0.896	GATTCCGGA**CAAAG**GGCCGGGGT
IL6	interleukin-6	GS119	-149	0.889	ACTAGGGGG**AAAAG**TGCAGCTTA
AZI2	5-azacytidine induced 2	GS120	-217	0.793	GGACCCCCC**AAAAG**GACACTGAG
CFL2	cofilin 2 (muscle)	GS121	-676	0.792	CGAGGCGAG**AAAAG**CCCCCCGCA
GPHN	gephyrin	GS122	+733	0.79	GACTGAGAG**GAAAG**GATAGCACA
C14orf119	Chromosome 14 open reading frame 119	GS123	-610	0.786	CAAGCGGCT**CAAAG**GGGTGAGGA
PPP1R3C	protein phosphatase 1, regulatory (inhibitor) subunit 3C	GS124	-142	0.772	CGAGACGTG**CAGAG**AGCTATCTG
AKR1C3	aldo-keto reductase family 1, member C3 (3-alpha hydroxysteroid dehydrogenase, type II)	GS125	-481	0.763	GAAAATGTA**AAAAG**GCAAATATT
NPAS2	neuronal PAS domain protein 2	GS126	-395	0.759	GAGCCGGCC**CAGAG**GAGAGGCAA
SAG	S-arrestin	GS127	-106	0.754	CCTGGGAGA**CAGAG**CAAGACTCC
CLCN3	chloride channel 3	GS128	-377	0.753	AGCGTCACG**CAGAG**TTCGGATCC
CBX3	chromobox homolog 3 (HP1 gamma homolog, Drosophila)	GS129	-525	0.748	GCGGAAGGC**TAGAG**TCCTGCTAG

[1], [2], the gene symbol and gene name of the gene where the corresponding HNF4 site was analyzed.

[3] the internal name of the oligonucleotide used in the study

[4] location of the site in the promoter of the gene. Position of the 5' end of the site is given relative to the transcription start site (TSS)

[5] The score was computed using the HNF4 positional weight matrix constructed in the course of this study (see Table 2).

[6] sequence of the oligonucleotide used in the study. The central part of the oligonucleotide, which correspond to the most conserved core of the HNF4 motif is given in bold.

Figure 1

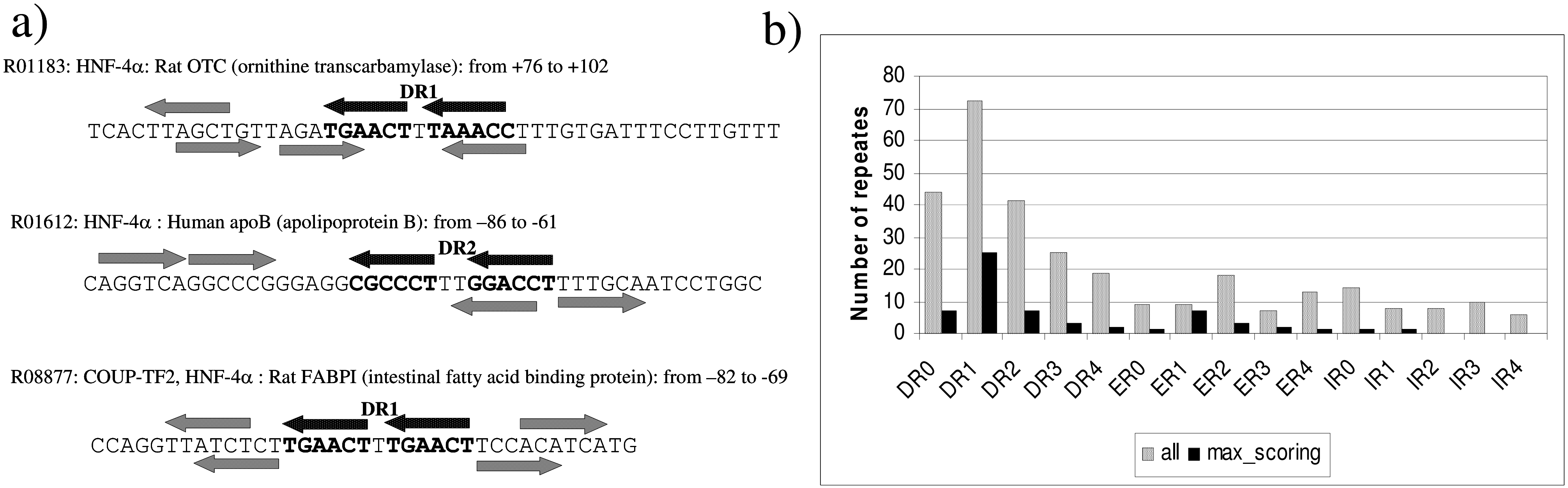
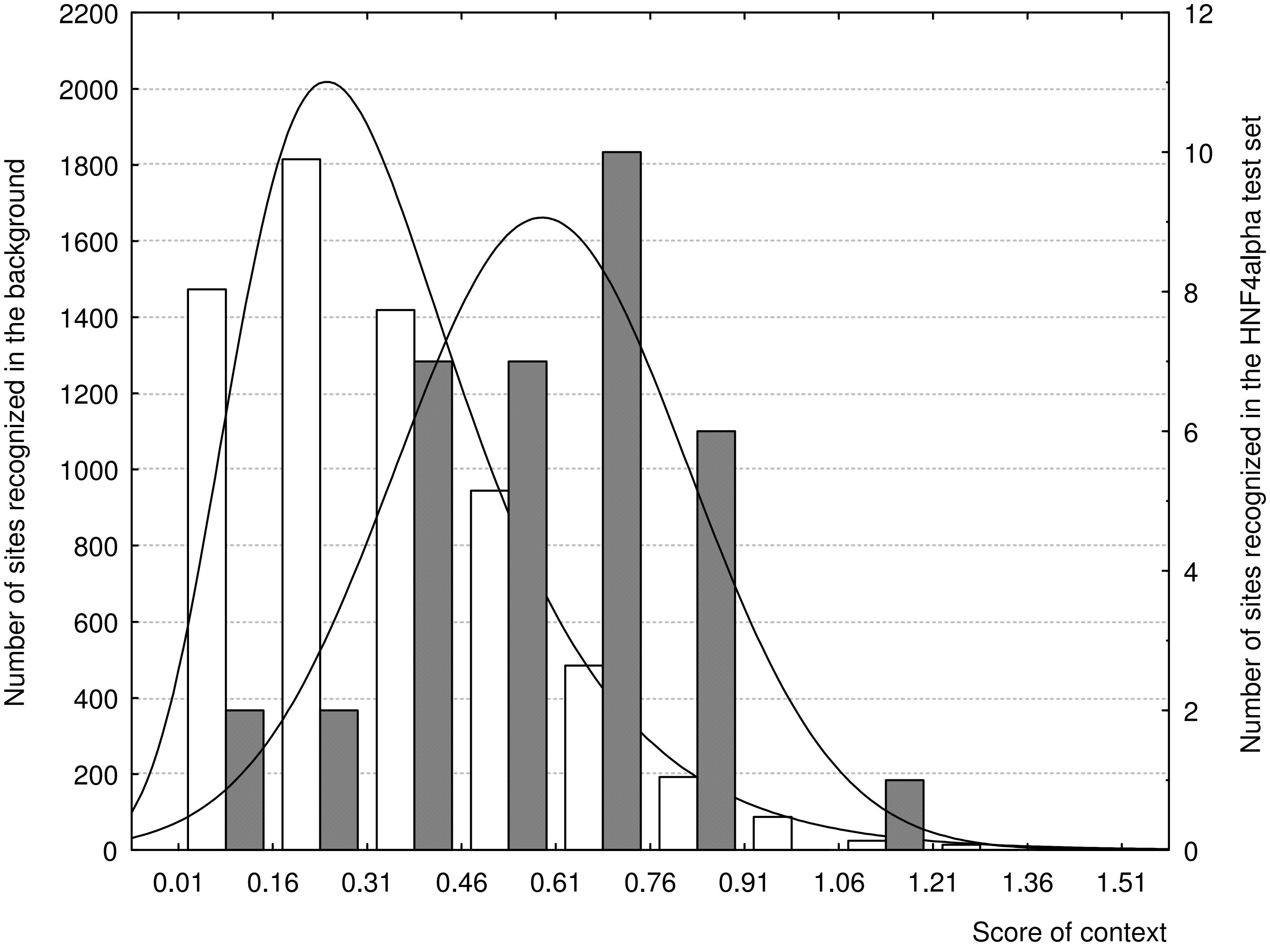
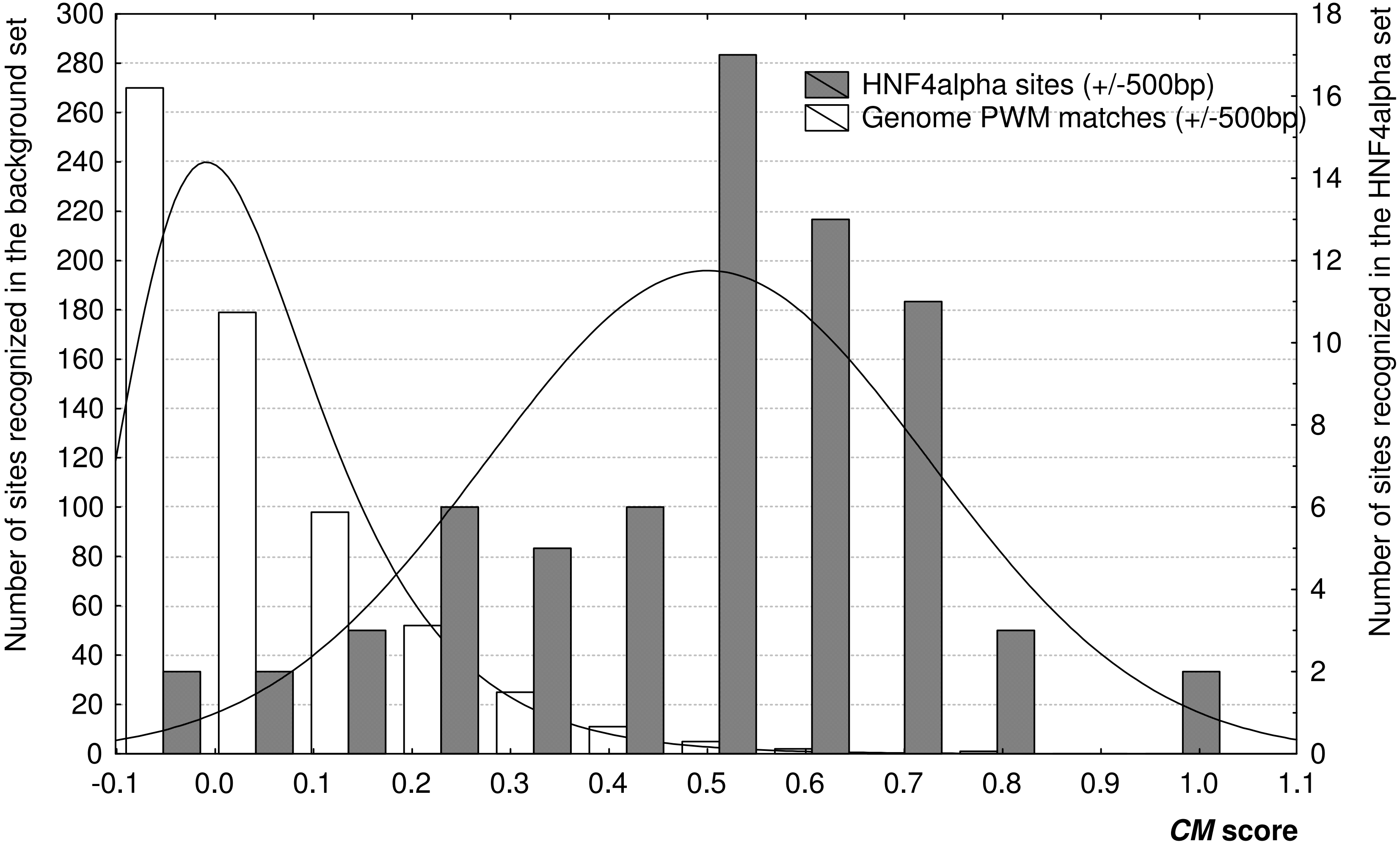

a)
R01183: HNF-4α: Rat OTC (ornithine transcarbamylase): from +76 to +102
DR1
TCACTTAGCTGTTAGATGAACTTTAAACCTTTGTGATTTCCTTGTTT
R01612: HNF-4α : Human apoB (apolipoprotein B): from –86 to -61
DR2
CAGGTCAGGCCCGGGAGGCGCCCTTTGGACCTTTTGCAATCCTGGC
R08877: COUP-TF2, HNF-4α : Rat FABPI (intestinal fatty acid binding protein): from –82 to -69
DR1
CCAGGTTATCTCTTGAACTTTGAACTTCCACATCATG
b)
Number of repeates
80
70
60
50
40
30
20
10
0
DR0
DR1
DR2
DR3
DR4
ER0
ER1
ER2
ER3
ER4
IR0
IR1
IR2
IR3
IR4
▨ all ■ max_scoring

Figure 2



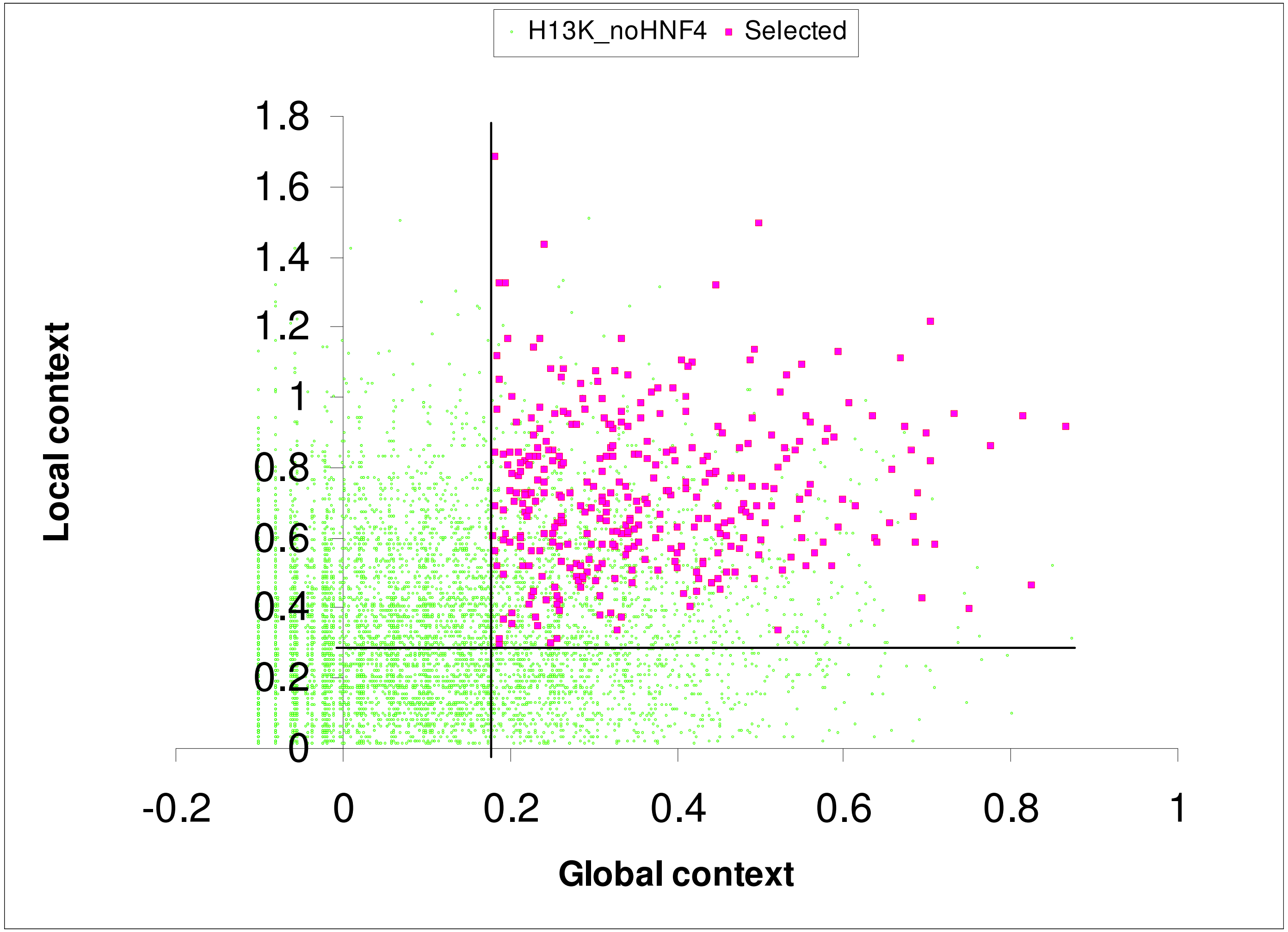
Number of sites recognized in the background
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0.01
0.16
0.31
0.46
0.61
0.76
0.91
1.06
1.21
1.36
1.51
Score of context
Number of sites recognized in the HNF4alpha test set
12
10
8
6
4
2
0

Figure 3



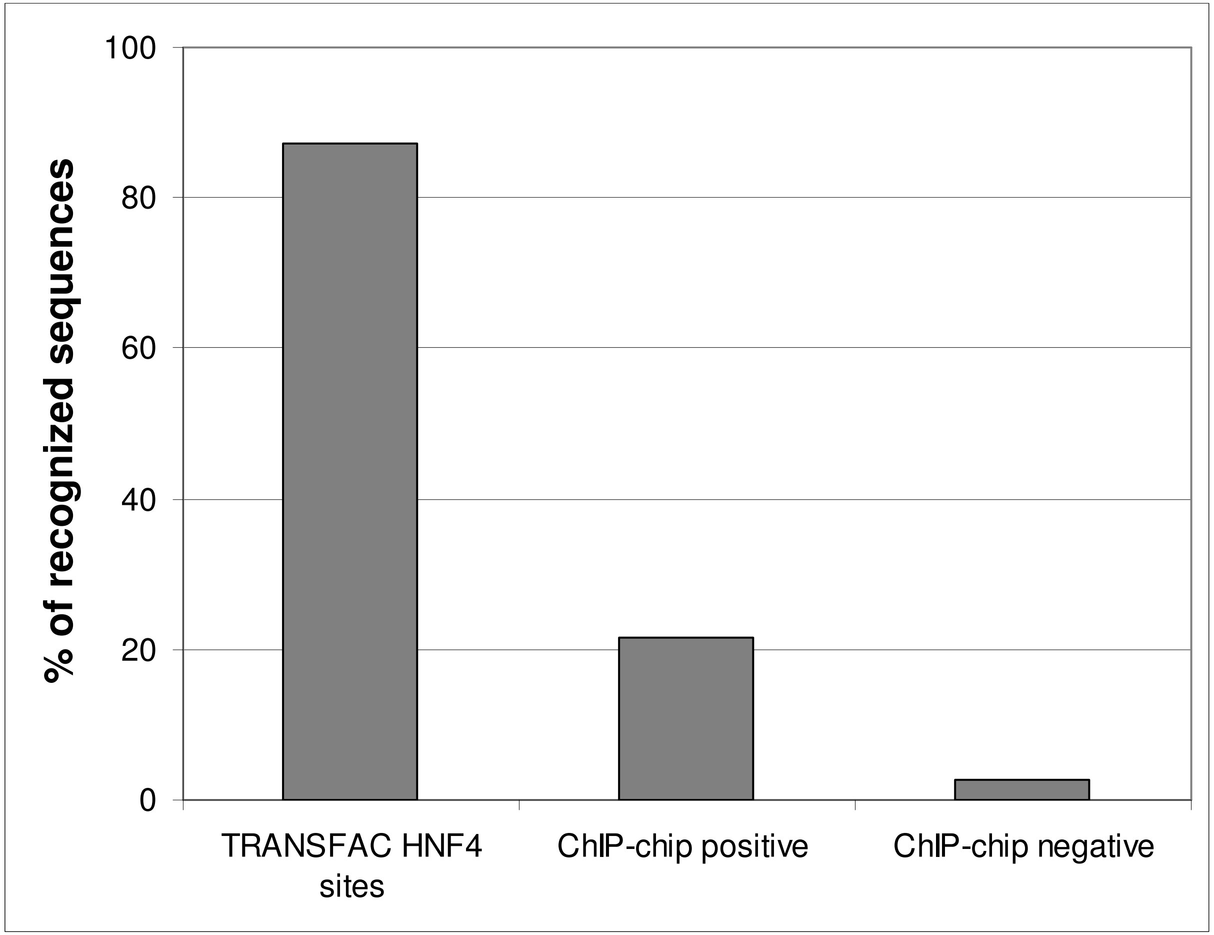
Number of sites recognized in the background set
300
280
260
240
220
200
180
160
140
120
100
80
60
40
20
0
Number of sites recognized in the HNF4alpha set
18
16
14
12
10
8
6
4
2
0
-0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
CM score
HNF4alpha sites (+/-500bp)
Genome PWM matches (+/-500bp)

Figure 4



H13K_noHNF4
Selected
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
0.2
0.4
0.6
0.8
1
Local context
Global context

Figure 5


% of recognized sequences
100
80
60
40
20
0
TRANSFAC HNF4 sites
ChIP-chip positive
ChIP-chip negative

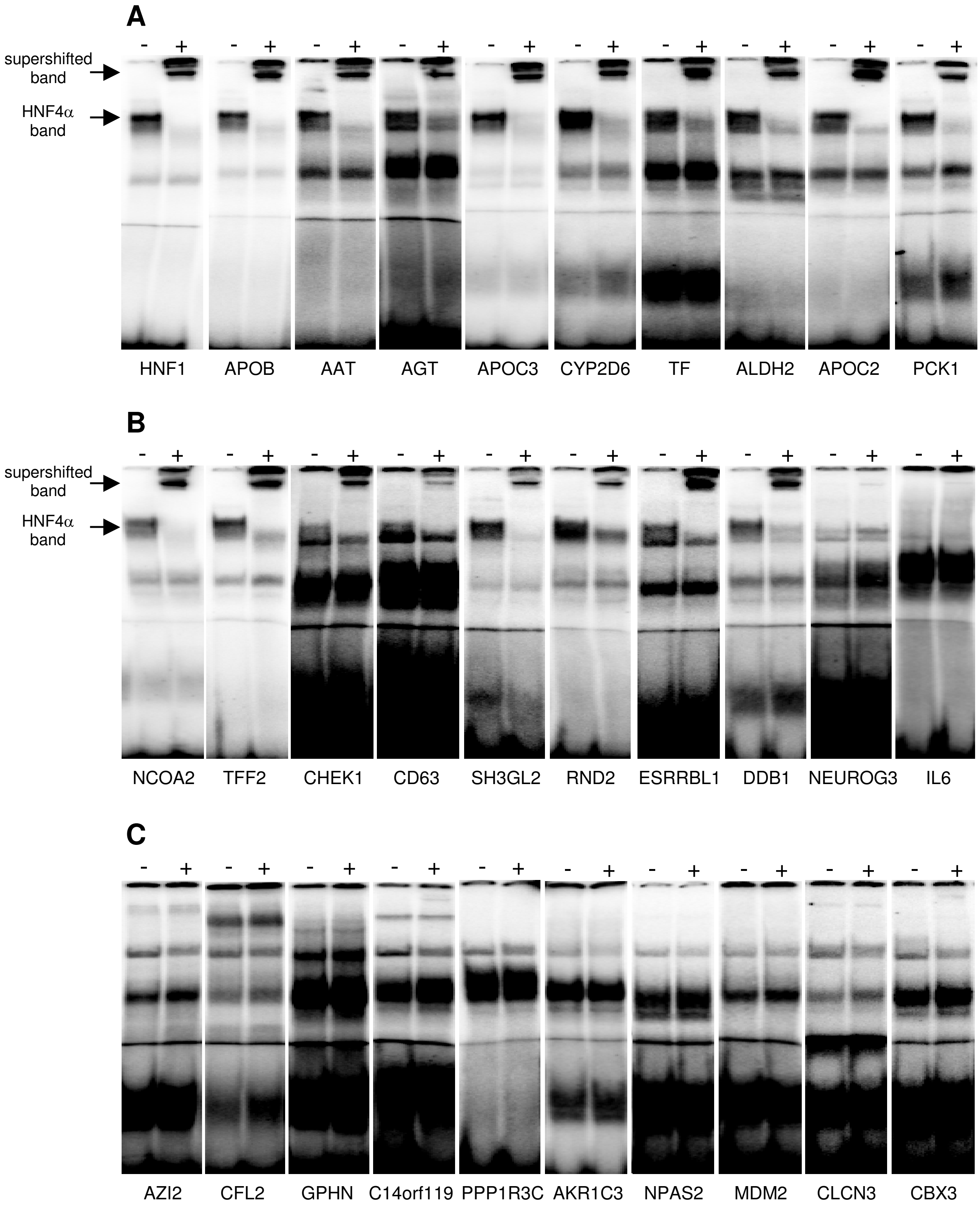

A
supershifted band
HNF4α band
- +
HNF1
APOB
AAT
AGT
APOC3
CYP2D6
TF
ALDH2
APOC2
PCK1
B
supershifted band
HNF4α band
- +
NCOA2
TFF2
CHEK1
CD63
SH3GL2
RND2
ESRRBL1
DDB1
NEUROG3
IL6
C
- +
AZI2
CFL2
GPHN
C14orf119
PPP1R3C
AKR1C3
NPAS2
MDM2
CLCN3
CBX3

Figure 6

Additional files provided with this submission:

Additional file 1: additional file 1.doc, 182K
http://genomebiology.com/imedia/1690046580183734/supp1.doc
Additional file 2: additional file 2.xls, 158K
http://genomebiology.com/imedia/2110420837183735/supp2.xls