*4488–4497* *Nucleic Acids Research, 2008, Vol. 36, No. 13* *Published online 7 July 2008*
*doi:10.1093/nar/gkn407*

# MOPAT: a graph-based method to predict recurrent *cis*-regulatory modules from known motifs

**Jianfei Hu[1,2], Haiyan Hu[2,3] and Xiaoman Li[1,2,*]**

[1]Division of Biostatistics, [2]School of Informatics and [3]Center for Computational Biology and Bioinformatics, School of Medicine, Indiana University, 410 West 10th Street, Indianapolis, IN 46202, USA

Received January 26, 2008; Revised June 1, 2008; Accepted June 10, 2008

## ABSTRACT

**The identification of *cis*-regulatory modules (CRMs) can greatly advance our understanding of eukaryotic regulatory mechanism. Current methods to predict CRMs from known motifs either depend on multiple alignments or can only deal with a small number of known motifs provided by users. These methods are problematic when binding sites are not well aligned in multiple alignments or when the number of input known motifs is large. We thus developed a new CRM identification method MOPAT (motif pair tree), which identifies CRMs through the identification of motif modules, groups of motifs co-ccurring in multiple CRMs. It can identify 'orthologous' CRMs without multiple alignments. It can also find CRMs given a large number of known motifs. We have applied this method to mouse developmental genes, and have evaluated the predicted CRMs and motif modules by microarray expression data and known interacting motif pairs. We show that the expression profiles of the genes containing CRMs of the same motif module correlate significantly better than those of a random set of genes do. We also show that the known interacting motif pairs are significantly included in our predictions. Compared with several current methods, our method shows better performance in identifying meaningful CRMs.**

## INTRODUCTION

Identifying *cis*-regulatory modules (CRMs) is an important problem in this postgenomic era. CRMs are short DNA regions of a few hundred base pairs that contain multiple transcription factor-binding sites (TFBSs). It is estimated that there are five-to-ten times as many CRMs in a genomes as there are genes (1). In high eukaryotes, CRMs instead of individual TFBSs often determine the spatial temporal expression patterns of neighboring genes. Therefore, identification of the CRMs is important not only for the understanding of gene transcriptional regulation but also for the annotation of high eukaryotic genomes.

However, to identify CRMs in high eukaryotes is challenging. The difficulty lies in the following two facts. First, the possible residing regions of the CRMs in one gene can be as long as thousands of base pairs or even hundreds of thousands of base pairs. Second, the TFBSs are in general 6–14 bp long and there is some degeneracy at almost every position of the TFBSs of a transcription factor (TF). Thus, if we scan a DNA sequence even with one known motif, we will obtain many false motif hits; if we scan a sequence with a large number of known motifs, we will find motif hits at nearly every position of the sequence, which is full of false ones.

In the past several years, a number of computational approaches for CRM identification have been developed (2–12). Of them, those with high degrees of success are based on known motifs, represented by position weight matrices (PWMs). There are two types of methods based on known PWMs. The first type predicts CRMs from a small set of known PWMs that are expected to form a motif module (6,7,13–16). Here, a motif module is a group of motifs with instances co-occurring in many CRMs. The second type (9) depends on multiple sequence alignments to predict CRMs from a large set of PWMs, such as all motifs deposited in Transfac database (17,18). In practice, it is often difficult to provide the motifs in a motif module since most motif modules are unknown. On the other hand, TFBSs are so short that the counterpart TFBSs in orthologous sequences are often not aligned in multiple sequence alignments. Thus, many CRMs can be missed by the current computational CRM identification methods based on known PWMs.

Here, we developed a new method, motif pair tree (MOPAT) that identifies CRMs from known PWMs. Our method can handle a large number of input motifs. It does not rely on the multiple sequence alignments either. The major difference between our method and the

*To whom correspondence should be addressed. Tel: +1 317 278 7273; Fax: +1 317 278 9217; Email: shawnli@iupui.edu

published methods is that we predict CRMs through the identification of motif modules, while many current CRM prediction methods predict CRMs in one region independently from other regions. The identification of a motif module, the recurrent combination of motifs shared by many CRMs, greatly improves the accuracy of CRM predictions. By applying this method to mouse developmental genes, we found many motif modules that significantly overlap with known interacting motifs. We also found that the expression profiles of genes containing CRMs of the same motif modules significantly correlate better than those of random genes do. Compared with several available CRM identification methods based on known motifs, our method shows better performance in identifying motif modules. The software based on our method can be freely downloaded from the following link http://evolution.compbio.iupui.edu/li/page/software.

## MATERIALS AND METHODS

### Transfac motifs and upstream sequences of mouse developmental genes

All 522 vertebrate PWMs from Transfac 9.2 (17,18) was extracted for the CRM analysis. Pseudo counts are introduced to regularize these PWMs, as is described subsequently.

The mouse developmental genes are obtained from three sources: (i) those annotated with GO:0032502 (developmental process); (ii) those annotated by the offspring of GO:0032502 and (iii) those with their orthologs in other 10 species (human, rat, dog, chicken, frog, fugu, zebrafish, nematode, sea squirt and fly) annotated with GO:0032502 or the offspring of GO:0032502. The 5-kb long noncoding sequences around the transcription start sites (TSS) of the mouse developmental genes are extracted from Ensembl website (http://www.ensembl.org/biomart/, release 46) by BioMart software (19) according to the Ensembl gene IDs. If the annotated gene start codon is within 2.5 kb of the TSS, we only use the sequence from the −2 kb position to the start codon. The repeat sequences in these 5-kb long sequences are masked by RepeatMasker software (http://www.repeatmasker.org). The reason that we apply our method to mouse is that there are many mouse developmental microarray data available for the validation of our predicted CRMs and motif modules.

### Score of a DNA segment given a PWM

$$\text{Score (a segment of length } d) = \sum_{i=1}^{d} \log \frac{f(b,i)}{f(b)}$$

where $f(b)$ is the average frequency of nucleotide $b$ in the above mouse developmental gene upstream sequences, $f(b,i)$ is the frequency of nucleotide $b$ at position $i$ of the motif PWM under consideration, $d$ is the width of the motif. A pseudo count 0.375 has been added to all the computation of frequency, as was used by Claverie and Audie (20).

### Identification of candidate motif hits in each sequence

We scan the nonrepetitive regions of each sequence to identify hits of each of the known motifs, using the above defined log-likelihood ratio score. We first compute the nucleotide distribution in the above developmental gene sequences, and produce a 100-kb long random sequence with the same nucleotide distribution. We then compute the score of every motif at each position of the random sequence to obtain the score distribution of the motif. Finally, for every motif, we take the 99.99% quartile of the score distribution of this motif as the score cutoff of this motif.

Some motifs may have the tendency to occur together, merely due to the similarity of their PWMs. To deal with it, many software predicting CRMs (7,14,15,21,22) or predicting interacting motif pairs (23–25) require that the motifs do not overlap with each other. Similarly, we require that the start positions of any two motif hits must be separated by at least 4 bp, which is also used by Sharan *et al.* (28). We sorted the motif hits of candidate motifs by their start positions and deleted the overlapped motif hits based on following rule: when the start positions of two motif hits are <4 bp, the motif hit with the lower score will be discarded.

### Identification of CRMs and motif modules in a motif pair tree

We define the following parameters in our method. $K_{\min}$ and $K_{\max}$ are the minimum number and the maximum number of motifs in a motif module, respectively. The $w$ is the allowed maximal length of a CRM. The $g$ is the required minimal number of genes that contain the instances of a motif module.

Our method carries out the following three steps to identify CRMs and motif modules: First, we extract the motif pair information from the motif hits by hashing. The motif pair information includes which two motifs have motif hits co-occurring within a $w$-bp window and how many sequences contain the instances of such co-occurring motif pairs. For every such motif pair, we also store the list of the gene names where the corresponding sequences contain instances of both motifs in the motif pair. For example, in seq1, assuming pos1 is the start position of the motif $m_6$ hit and pos2 is the end position of the motif $m_0$ hit, $m_6$ and $m_0$ form a motif pair since pos2–pos1 is less than $w$. Therefore, we add the gene name seq1 to the gene list of $m_6m_0$. Please note that seq1 is added to the gene list of $m_6m_0$ just one time no matter how many times $m_6m_0$ occurs in seq1 (Figure 1b). After all sequences are analyzed, we scan the hash table to delete the motif pairs that occur in less $g$ genes and modify the motif degree accordingly. For every motif in the remaining motif pairs, we then store the motif degree, which is the number of the distinct motifs that form motif pair with this motif under consideration. For example, in Figure 1a, the motif degree for motif m4 is one, because motif pair m4m3 occur in no less than $g = 2$ genes. In the following, motifs always mean those in the motif pairs that occur in at least $g$ genes.

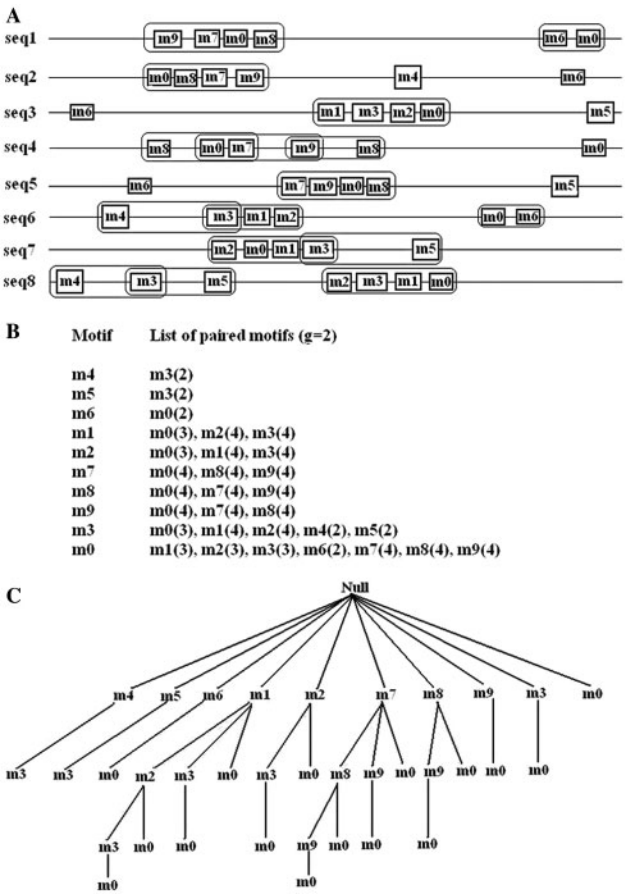


**Figure 1.** Construction of motif pair tree. (**A**) Motif hits of ten motifs in eight sequences. The motifs can overlap with each other as long as their start positions are separated by at least 4 bp. The motifs in the same box are all paired with each other. (**B**) Motifs and their paired motif list. The number in the parenthesis is the number of the genes that contain instances of the motif pair. (**C**) Motif pair tree. Each node in motif pair tree represents a motif. Each path in the motif pair tree represents a potential motif module.

Second, we construct a motif pair tree and output potential motif modules. We define the motif degree of a node in the motif pair tree as the motif degree of the motif represented by the node. We first take null as the root node of the tree. We then construct a node for every motif as a child node of the root node. Here and in the following, the child nodes of a node are always sorted from left to right with the increment of the motif degree. For every leaf node in the current tree, we then take every other node with higher motif degree as its child nodes, if the motif with higher motif degree and the motif represented by the leaf node under consideration form motif pairs at the first step. We repeat this procedure until no child node can be added to any current leaf nodes. For instance, the tree in Figure 1c is the motif pair tree for the motif hits in Figure 1a. It is obvious that any path in the motif pair tree starting from the root is a potential motif module. Each potential motif module is represented by and only by one path of the motif pair tree. In practice, we will construct the motif pair tree in a depth-first format. We will also add a gene list to each node when we construct a path of the motif pair tree. The gene list of a child node is a subset of the gene list of its parent nodes. Taking the path null–$m_1$–$m_2$–$m_3$–$m_0$ in Figure 1c as an example, the gene list of the node $m_0$ is the set of genes that contain instances of motif pairs $m_1m_2$, $m_1m_3$, $m_1m_0$, $m_2m_3$, $m_2m_0$, and $m_3m_0$. With the gene list for each node, we can also stop to extend a path to include a node if the number of the genes in its gene list is less than $g$, which increases the efficiency of our method. In this way, we only need to store one branch of the tree at one time, which enables our method to handle a large number of known motifs and the input gene noncoding sequences.

Third, we check whether the potential motif modules really have instances in at least $g$ genes. For each motif module, we check the genes in its gene list one by one to see whether the motif module can be found in the genes. For each gene, we scan sequence to see whether any $w$-bp window contains hits of all motifs of the potential motif module. The gene is claimed to contain instances of the motif module if such a $w$-bp window exists. A motif module is output if the number of genes containing instances of this motif module is no less than $g$. The $w$-bp windows that contain instances of this motif module in these genes are the CRMs of this motif module.

## Statistical evaluation

We use *Poisson clump heuristic* to compute the $P$-value of a motif module, with the assumption that each motif occurs independently according to a Poisson process (26). Independent motif occurrence is assumed by many CRM identification methods to analyze the significance of a motif cluster or CRM (13,27,28).

Let $N$ be the total number of motifs, $G$ be the total number of sequences, $L$ be the average length of sequence and $\lambda_k$ be the rate parameter of the Poisson process for the motif k. Suppose $m_1, m_2, \ldots, m_n$ is a motif module with instances co-occurring in $g$ sequences. For each position of a sequence, the probability that $m_i$ occurs is $\lambda_{mi}$. For a window covering this position and containing $m_i$, the probability that instances of other $n-1$ motifs occur at least one time is $\prod_{1\le j\le n, j\ne i}(1-e^{-w\lambda_{mj}})$. Therefore, the probability that this motif module occurs in this sequence is bound by,

$$P_c = 2L \times \sum_{i=1}^{n} \lambda_{m_i} \prod_{1\le j\le n, j\ne i} (1 - e^{-w\lambda_{mj}})$$

The probability that this motif module occur in at least $g$ sequences is,

$$P_{gc} = 1 - \sum_{k=0}^{g-1} C_G^k P_c^k (1-P_c)^{(G-k)}$$

$N$ motifs can produce $C_N^n$ different motif clusters containing $n$ distinct motifs. So if we take 0.05 as the significance level, after Bonferroni correction for multiple comparisons, the probability that we can find a motif module containing $n$ distinct motifs and occurring in no less than $g$ sequences, $P_{gc}$, should be smaller than

$$0.05/C_N^n$$

## RESULTS

We developed a method and software, MOPAT, to search for CRMs and motif modules in DNA sequences. MOPAT includes two parts. The first part, written by C++, takes a set of sequences in FASTA format and a set of motifs in count-matrix format as input to predict the hits of motifs in the input sequences. The second part, written by Perl, takes the set of motif hits as input to find CRMs and motif modules.

Given a set of sequences, a set of motifs and a motif probability cutoff, the number of predicted motif modules and the running time of MOPAT are mainly influenced by two factors: the window size cutoff $w$ and the gene number cutoff $g$. The larger the $w$ and the smaller the $g$, the more motif modules we can find and the more time required to run the program. Our program also allows the user to set the minimum and maximum number of motifs in a motif module. The constraint by these two parameters can further improve the speed of program. When using our program, the user can set a loose parameter set (such as $w = 200$, $g = 20$) and spend more time to identify as many motif modules as possible; or they can set a stringent parameter set (such as $w = 100$, $g = 20$) and spend less time to identify only the motif modules with high significance.

We have applied MOPAT to the 5-kb long sequences around the TSS of the 5530 mouse developmental genes by using the following parameters: $P = 0.0001$, $w = 200$, $g = 20$, $K_{min} = 3$ and $K_{max} = 8$. We use 3 as the minimum number of motifs allowed in a motif module, since motif modules composing of only two motifs occur very frequently. We use window size as 200 bp and gene number as 20 to ensure most predicted motif modules are significant according to our analysis above. With these parameters, we have predicted 144 490 motif modules, which cover 494 motifs and 5492 genes (Supplementary data). Taking 0.05 as the Bonferroni corrected $P$-value cutoff, we divide the predicted motif modules into two parts according to the significance. The first part, Result I, contains all the motif modules with Bonferroni corrected $P < 0.05$. The second part, Result II, contains the rest motif modules. Result I contains 33 361 motif modules, which cover 489 distinct motifs and 5486 distinct genes. Result II contains 111 129 motif modules, which cover 471 motifs and 5491 genes. Our analyses in the following are based on the two groups of predictions. Note that many motif modules in Result II are also significant although their Bonferroni corrected $P$-values are $>0.05$.

### Validate predicted CRMs and motif modules by expression data

We attempt to validate our predicted motif modules by using microarray expression data here. This validation strategy has been used in the previous study to support the predicted motif modules in upstream 1-kb regions by other group (28). The basic assumption of this validation strategy is that the genes containing the instances of the same motif module should have more similar expression patterns. That is, the CRMs with the same combination of the motifs will control similar temporal spatial expression patterns (29,30).

Three microarray datasets have been downloaded from GEO database (http://www.ncbi.nlm.nih.gov/geo/): mouse epididymis development (GDS2202), mouse ovary development (GDS2203) and mouse cochlear nucleus postnatal development (GDS2144). We have evaluated our results by using the three microarray datasets respectively, and the performance of our method is similar for different microarray datasets. Therefore, we will use GDS2202 as an example to show the quality of our predicted motif modules.

As the previous study (28), we first calculate the Pearson's correlation coefficient to measure the similarity of the gene expression. Note that some genes may have multiple expression measurements, and there are multiple similarities between each pair of these genes. In this case, we take the maximal one as the expression similarity of the gene pair. In the following, we justify the significance of the predicted motif modules from two different aspects.

In the first way, we first extract all gene pairs in GDS2202 and compute the expression similarity of gene pairs, and draw the histogram of these similarities. We call this distribution the background distribution. The background distribution is like a normal distribution, with 0.14 as the mean (Figure 2a). The nonzero mean shows that the dataset is not completely random and some genes coexpressed with each other. We next extract the gene pairs containing the CRMs of the same motif modules in our predicted results and calculate the expression similarities of gene pairs. We first calculate the similarities of expression profiles of gene pairs in every gene set containing instances of same motif modules and then plot the histogram of these similarities from all gene sets (Figures 2b and 3c). It can be seen from Figure 2b and c, relative to the background, the distributions of the expression similarity of the genes containing instances of the same motif modules are apparently skew toward to the high correlation end. The mean expression similarity of the genes containing instances of the same motif modules from Result I is 0.31, higher than that from Result II, 0.27. The mean similarities from Result I and Result II are both significantly higher than that from the background distribution, 0.14 ($P < 0.0001$ by $t$-test). Note that 39.7 and 14.5% of the mean similarities of the expression of the genes containing instances of our predicted motif modules in Result I and Result II are higher than that of the typical example in ref. (28), 0.341, respectively, although Sharan *et al.* make use of the prior biological knowledge and the sequence conservation information to identify motif modules.

In the second way, we first construct two groups of random gene sets from GDS2202, Random I and Random II, such that Random I and Random II have the same distribution of the size of the gene sets in Result I and Result II, respectively. For instance, the percentage of gene sets containing 25 genes are 4.4% in both Result I and Random I. Random I includes 166 805 ($33\,361 \times 5$) gene sets, and Random II includes 555 645 ($111\,129 \times 5$) gene sets, five times as many gene sets as that in Result I and Result II, respectively. We next

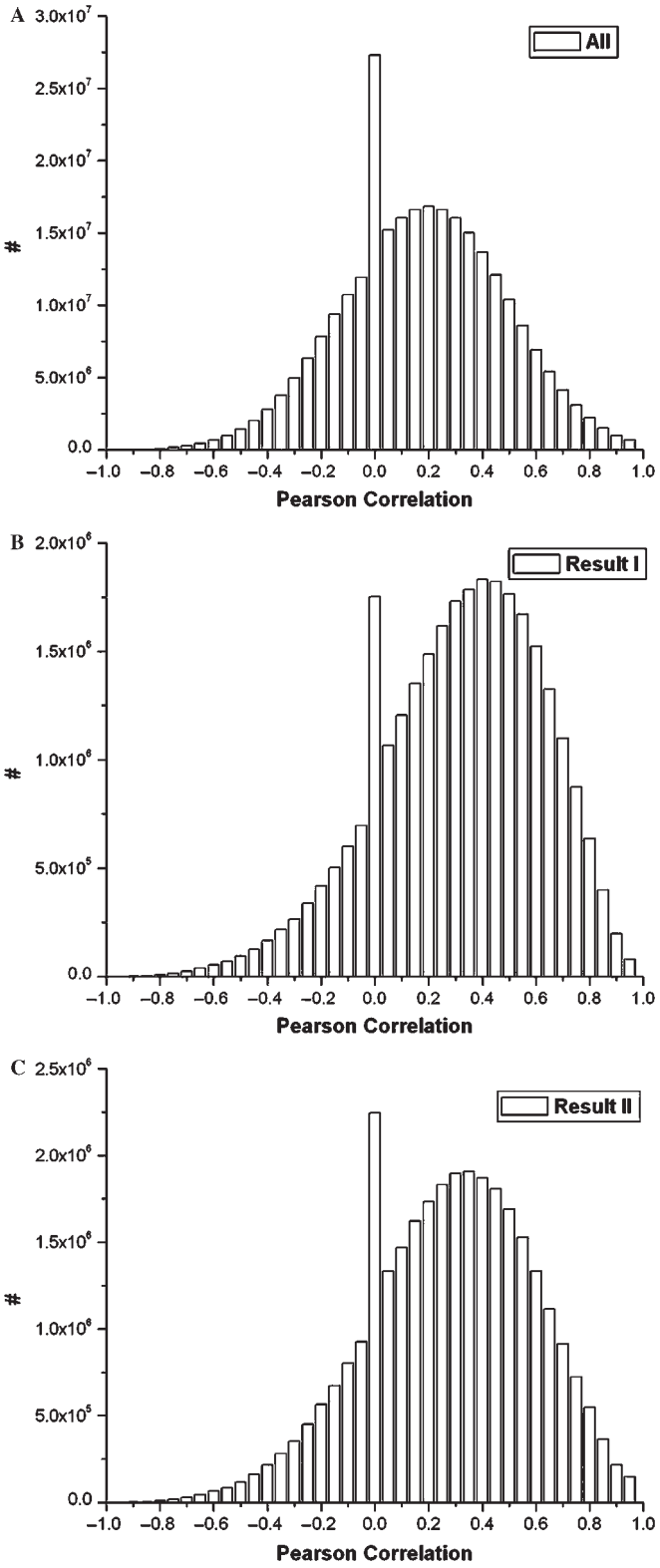


**Figure 2.** Histogram of the expression similarities of gene pairs. (**A**) Expression similarity of gene pairs in GDS2202, average = 0.14. (**B** and **C**) Expression similarity of gene pairs containing instances of the same motif modules in Results I and II, mean = 0.31 and 0.27, respectively.

calculate the median gene pair expression similarity in each gene set in Random I and Random II, and draw the histogram of the median similarities, respectively (Figure 3a and b). For Random I, only 5% of median similarities are higher than 0.263, and for Random II, only 5% of median similarities are higher than 0.277. We thus take 0.263 and 0.277 as threshold to define the significant motif modules in Result I and Result II, respectively. Under such significance levels, 31 125 of 33 361 motif modules (93%) of Result I are significant (Figure 3c), and 76 641 of 1 11 129 motif modules (69%) of Result II are significant (Figure 3d).

### Validate predicted CRMs and motif modules by experimentally verified composite regulatory elements in Transfac

Composite regulatory elements or composite elements (CEs) contain two closely located TFBSs of two different TFs and represent minimal functional units that provide combinatorial transcriptional regulation. Both factor–DNA and factor–factor interactions contribute to the function of CE (31,32). In the past decades, many CEs have been experimentally verified and deposited in public databases. These CE can be used to verify the predicted motif modules.

In Transfac database, CE is represented as interacting factors. If some TFs interact with certain factor TF0, then we can find these TFs from TF0's interacting factors in Transfac database. We have thus extracted the TF interacting information for each TF from the Transfac database and translated them into motif pairs according to the map between TF and PWMs. We call the motif pair as CE motif pair if the two factors that bind the two motifs interact with each other. In total, 2515 vertebrate CE motif pairs are found in Transfac database, after the motif pairs consisting of the same motif PWM are removed.

Without considering the order of the motifs, the predicted 144 490 motif modules above contain 27 546 distinct motif pairs, of which 625 are CE motif pairs. In detail, 33 361 motif modules in Result I contain 15 263 distinct motif pairs, of which 440 are CE motif pairs; 111 129 motif modules in Result II contain 22 479 distinct motif pairs, of which 481 are CE motif pairs. Our results only contain a small part of CE motif pair, which may be due to the following facts. First, we only analyzed mouse genes here, while the CE motif pairs are from many vertebrates. Second, we only analyzed developmental genes in mouse, which account for <20% of the total mouse genes. Third, not all the CE motif pairs in Transfac database require the physical binding of the TFs with DNA in order to interact. On the other hand, although not all CE motif pairs are included in our predictions, our predicted motif modules still significantly overlap with the CE motif pairs. The significance of the overlap between the CE motif pairs and our predicted result can be tested by a hypergeometric distribution:

$$P(N, M, n, m) = 1 - \sum_{t=0}^{m-1} \frac{C_M^t \, C_{N-M}^{n-t}}{C_N^M}.$$

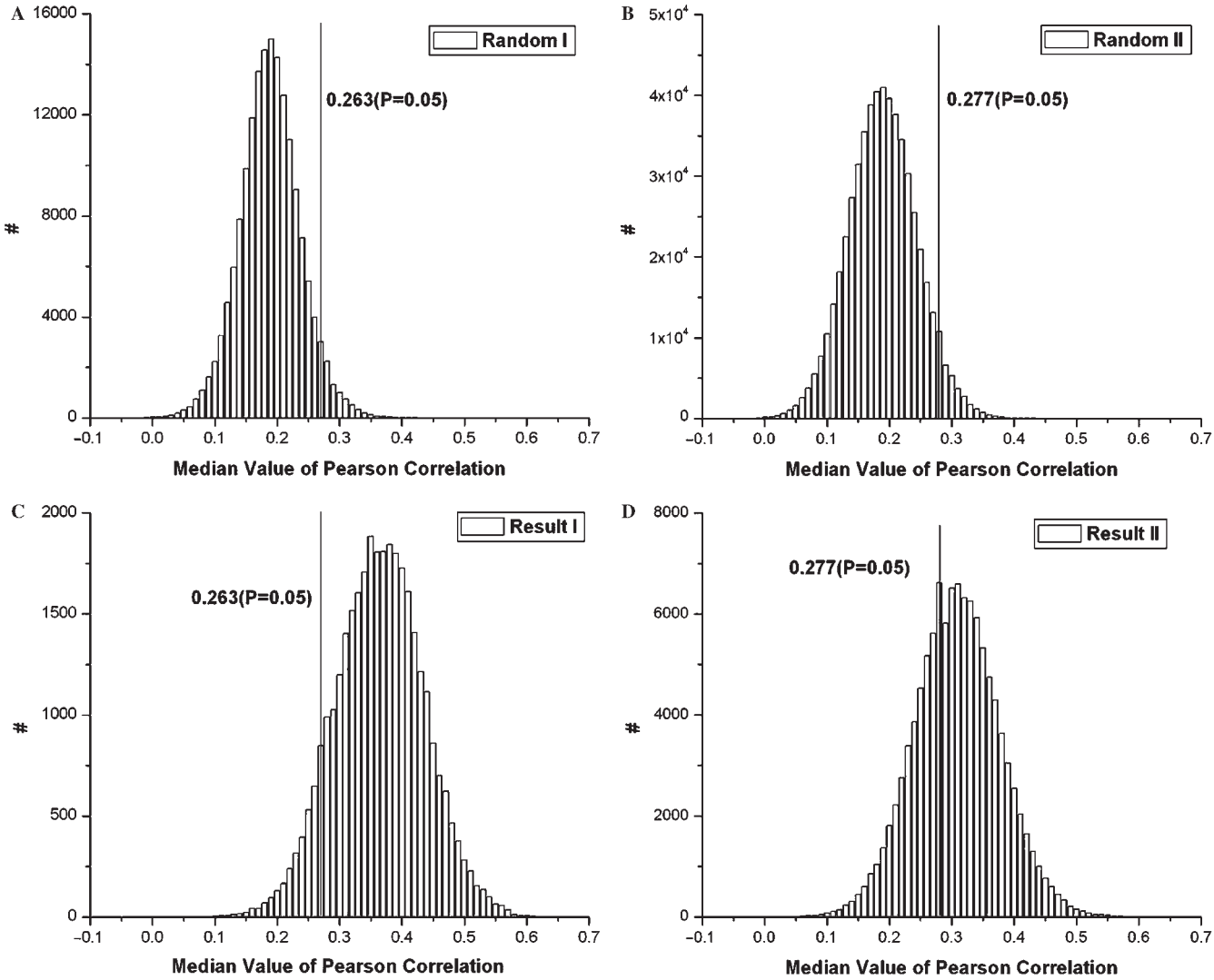


**Figure 3.** Histogram of the median expression similarity of gene pairs in gene sets. (**A**) 166 805 random gene sets generate from the same distribution of the size of the gene sets in Result I. The 95% quartile of this histogram is 0.263. (**B**) 555 645 random gene sets generate from the same distribution of the size of the gene sets in Result II. The 95% quartile of this histogram is 0.277. (**C**) 33 361 target gene sets of the predicted motif modules in Result I. Ninety-three percent of the target genes have a median gene pair expression similarity larger than 0.263. (**D**) 111 129 target gene sets from the predicted motif modules in Result II. Sixty-nine percent of the target genes have a median gene pair expression similarity larger than 0.277.

Here, $N = 522 \times (522 - 1)/2 = 135\,981$ is the total number of distinct motif pairs constituted by 522 vertebrate motifs; $M = 2515$ is the number of CE motif pairs deposited in Transfac database; $n$ is the number of distinct motif pairs included in our predicted motif modules; $m$ is the number of CE motif pairs included in our predicted motif modules. The $P$-value of observing so many CE motif pairs in all 144 490 motif modules, 33 361 motif modules in Result I and 111 129 motif modules in Result II is $P_A$ (135 981, 2515, 27 546, 625) $= 2.29 \times 10^{-9}$, $P_I$ (135 981, 2515, 15 263, 440) $= 1.32 \times 10^{-21}$ and $P_{II}$ (135 981, 2514, 22 479, 481) $= 5.00 \times 10^{-5}$, respectively (Table 1).

**Table 1.** Significance of proportion of CE motif pairs in predicted motif modules

| Result | $N$ | $M$ | $n$ | $m$ | $P$ |
|---|---|---|---|---|---|
| I | 135 981 | 2515 | 15 263 | 440 | 1.32E-21 |
| II | 135 981 | 2515 | 22 479 | 481 | 5E-5 |
| I + II | 135 981 | 2515 | 27 546 | 625 | 2.29E-09 |

See text for the meaning of $N$, $M$, $n$ and $m$.

### An example of the predicted motif modules

To further show that the predicted CRMs and motif modules make sense, here we give an example of the predicted CRMs of a motif module. This motif module comprises three motifs, M00380, M00423 and M00724, which corresponds to the TF Pax4, Fox-j2 and hepatocyte nuclear factor-3alpha (Hnf-3α), respectively (Figure 4).

The three TFs all involve in the metabolism process. Pax4 is a well-known diabetes-linked TF, and plays an important role in diabetes-related metabolism process (33–37). Fox-j2 belongs to the forkhead TF family, a large gene family known for their key function in

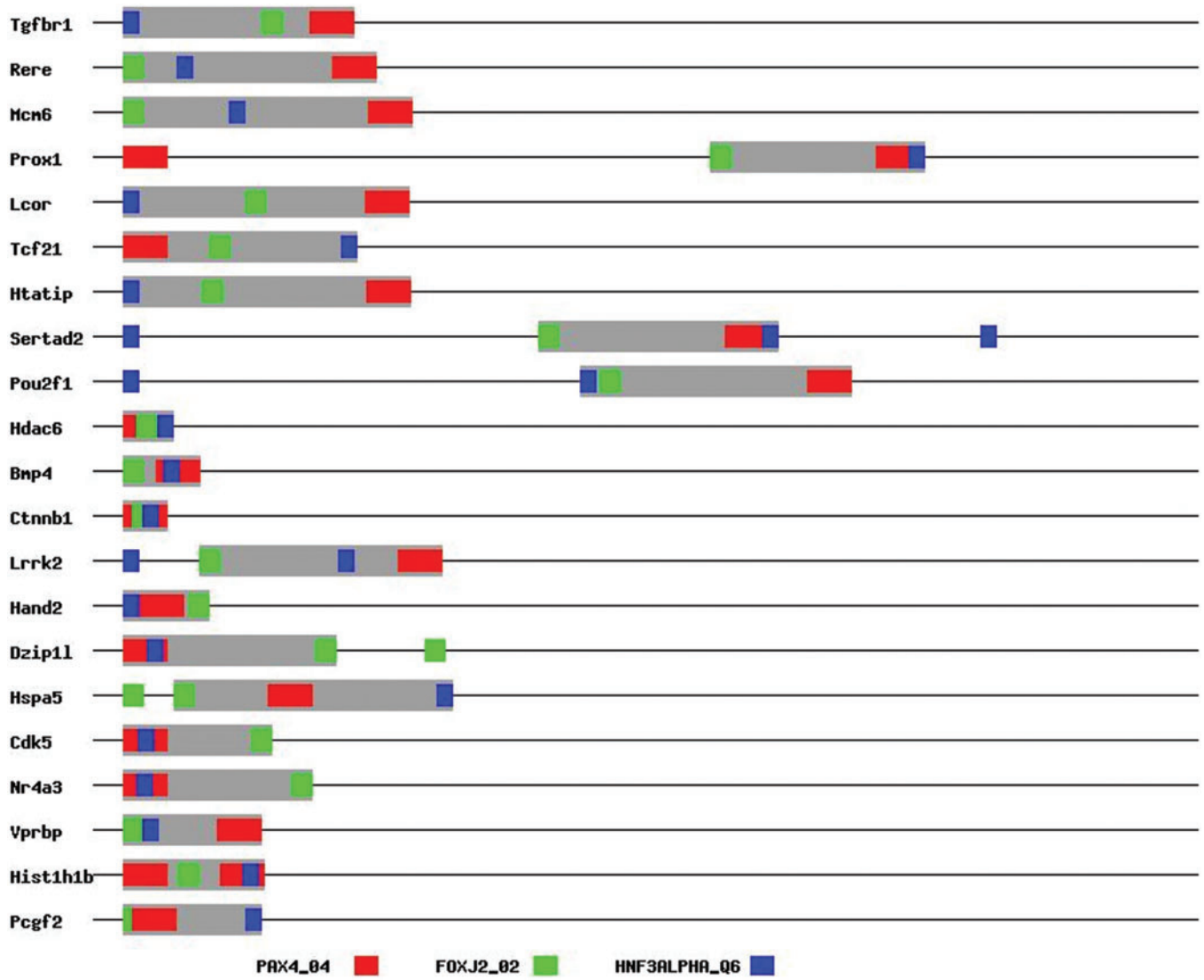


**Figure 4.** One example of the predicted recurrent motif modules.

development and metabolism (38). It has been reported that the deregulation of the forkhead family genes can lead to congenital disorders, diabetes mellitus or carcinogenesis (39). Hnf3 (also known as Fox-a2) is also a member of the forkhead TF family, and functions in diabetes-related metabolism (40,41).

There are 21 genes containing instances of this motif module. Our GO analysis demonstrates 19 of them relate to cellular metabolism (GO:0044237). Further analysis indicates that a majority of them relate to diabetes. Ten of them are annotated to be related to diabetes in past literature: ctnnb1 (42,43), tgfbr1 (44), prox1 (45), bmp4 (46), mcm6 (47), pou2f1 (48), hspa5 (49), Nr4a3 (50), cdk5 (51,52) and hdac6 (53). Five of the rest, lcor, pcgf2, htatip, dzip1l and rere locate in a region that has been implicated in susceptibility to type 1 diabetes (54). Lrrk2, hand2, vprbp, tcf21, sertad2 and hist1h1b, which are not well annotated, may represent the unknown genes related to diabetes as well.

### Compare our method with other methods

We compared MOPAT with three popularly used methods, Cbust (http://zlab.bu.edu/cluster-buster/cbust.html), Compel (http://compel.bionet.nsc.ru/FunSite/CompelPatternSearch.html) and Dire (http://dire.dcode.org/). These three methods can be classified into two groups according to the input data. Cbust and Compel, belong to the first group. Cbust takes a set of sequences and a set of candidate motifs as input to predict CRMs. Compel takes a sequence set as input, then it uses the CE motif pair information deposited in TRANScompel database to predict CRMs. Dire belongs to the second group that takes a list of gene names as input.

To compare our method with the methods in the first group, we generate random sequences with implanted CRMs. We use random sequences rather than real sequences because it is difficult to know what motif modules and CRMs may be contained in real sequences. We first generate 10 motif modules, with the motifs randomly

selected from the 522 vertebrate motifs and the number of motifs in a motif module varying from 3 to 8. We then generate a sequence set for each motif module, and randomly inserted instances of each motif in the motif module into a 200-bp window in each sequence. The sequence sets are produced completely randomly with the same nucleotide frequency as the developmental sequences we used. The number of the sequences for one motif module varies between 15 and 30, and the length of each sequence is 1 kb. We then mix the sequences for every motif module together as our test dataset. Totally, our test dataset includes implanted instances of 51 distinct motifs and 202 1-kb long sequences (Table 2). Taking the 202 sequences and 522 vertebrate motifs as input, we then compare our method with other methods. The default parameters are used for all methods.

MOPAT predicted 43 motif modules and 633 CRMs, which totally include 61 distinct motifs (42 of them are true positives). The number of distinct motifs in CRMs is four on average, with three as the minimum and eight as the maximum (Table 3). Cbust predicted 190 CRMs, which totally include 515 distinct motifs and 51 of them are true positive. The number of distinct motifs in CRMs is 94 on average, with 17 as the minimum and 249 as the maximum. Thus, the result of Cbust includes many false positives. For Compel, 61 distinct motifs are predicted and 10 of them are true. The number of distinct motifs in a CRM is 8.5 on average, with 5 as the minimum and 13 as the maximum.

We further judge whether MOPAT can predict true motif modules and CRMs according to two criteria. First, if the genes contain CRMs of a predicted motif module are included in the group of genes containing CRMs of one implanted motif module, we call the predicted motif module a true motif module. Second, if a CRM contains exactly the same instances compared with those in an implanted CRM, we call the predicted CRM a true CRM. For MOPAT, all predicted motif modules satisfy the first criterion and 109 predicted CRMs satisfy the second criterion (Table 3). Furthermore, for the second criterion, if we allow to delete or to add at most one instance to the predicted CRMs, 401 predicted CRMs satisfy the second criteria. For Cbust and Compel, none of predicted motif modules or CRMs satisfies the first or the second criterion even we allow one mismatch. Note that there are also false positives in MOPAT results. This is because there are always a few new motifs occurring frequently in random sequences, given a large number of motifs input.

Given that Dire just accepts a list of input gene names and does not accept the input of sequence sets and motif sets, we have to try another way to compare it with MOPAT. We mixed the 21 diabetes-metabolism related genes in our example with other 200 mouse developmental genes, and take these 221 gene names as input of Dire to see what shared motifs can be found in the 21 genes output from Dire. We found that no motif is shared by the 21 genes according to Dire. Even with the 21 diabetes metabolism-related genes as input, Dire cannot identify any motif shared by the 21 genes, which may be because of the requirement that corresponding motif instances must be aligned. Note that MOPAT identifies a motif module shared by these 21 genes from the 5530 developmental genes input.

**Table 2.** Ten groups of randomly generated CRMs and genes

| Group | No. of motifs | No. of genes |
|---|---|---|
| 1 | 3 | 16 |
| 2 | 3 | 24 |
| 3 | 4 | 15 |
| 4 | 4 | 29 |
| 5 | 5 | 26 |
| 6 | 5 | 17 |
| 7 | 6 | 25 |
| 8 | 6 | 17 |
| 9 | 7 | 17 |
| 10 | 8 | 16 |
| Total | 51 | 202 |

## DISCUSSION

We have developed a method to identify CRMs based on a large set of known motifs deposited in Transfac. We show that our predicted motif modules significantly overlap with the known interacting motif pairs. We also show that the expression profiles of the genes containing CRMs of the same motif modules correlate significantly better than the genes in random gene sets do, which suggests many predicted CRMs could be used to predict the gene expression patterns.

Our method has utilized the observation that instances of the motifs in the same motif module co-occur in

**Table 3.** Comparison between MOPAT and others

| Method | No. of motif inserted | No. of motif candidate | No. of CRMs predicted (true) | No. of motifs predicted (true) | No. of motifs in CRM |
|---|---|---|---|---|---|
| MOPAT | 51 | 522 | 633 (109[a], 401[b]) | 60 (42) | 4 (3[c], 8[d]) |
| Cbust | 51 | 522 | 190 (0,0) | 515 (51) | 94 (17, 249) |
| Compel | 51 | – | 202 (0,0) | 61 (10) | 8.5 (5, 13) |

[a]The number of CRMs that match the implanted CRMs perfectly.
[b]The number of CRMs that match the implanted CRMs with one mismatch.
[c]Minimum number of motifs in a CRM.
[d]Maximum number of motifs in CRM.
See text for the details.

many CRMs. This observation helps us to differentiate 'true' CRMs from the false ones, especially when the number of candidate motifs is very large, which can be shown from the comparisons between our method and others that do not utilize this observation. This observation also helps us find gene sets with similar expression pattern, which can be seen from the much higher correlations of gene expression of the gene pairs containing instances of a motif module.

Our method provides an option to identify CRMs without multiple sequence alignments. Many current CRM identification methods based on known motifs depend on sequence alignment. They first align noncoding sequences of orthologous genes to find conserved regions, then search motif hits or CRMs in the conserved regions. In most cases, it is valid. However, in some cases, it does not work well. There are reports showing that many functional elements in the human genome are seemingly unconstrained across mammalian evolution (55). Therefore, a reasonable way should be like this, it can utilize ortholog information without multiple alignment, thus can find conserved CRMs that cannot be aligned well in multiple alignments. Our method provides such an option to utilize the ortholog information. When extracting motif pair information (step one), we can require the motif pairs must occur in some or all its ortholog genes, which enable the identification of the conserved CRMs that are shuffled during the genome rearrangements.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

This project is supported by a NHGRI grant R01HG004359 and a Showalter Trust award. The computation is implemented on the IU supercomputers. Funding to pay the Open Access publication charges for this article was provided by NHGRI grant R01HG004359 and Showalter Trust award.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Davidson,E.H. (2006) *The regulatory genomes: Gene regulatory networks in development and evolution*, *1st edn*. Academic Press, Oxford.
2. Berman,B.P., Nibu,Y., Pfeiffer,B.D., Tomancak,P., Celniker,S.E., Levine,M., Rubin,G.M. and Eisen,M.B. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc. Natl Acad. Sci. USA*, **99**, 757–762.
3. Halfon,M.S., Grad,Y., Church,G.M. and Michelson,A.M. (2002) Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res.*, **12**, 1019–1028.
4. Rebeiz,M., Reeves,N.L. and Posakony,J.W. (2002) SCORE: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation. *Proc. Natl Acad. Sci. USA*, **99**, 9888–9893.
5. Boffelli,D., McAuliffe,J., Ovcharenko,D., Lewis,K.D., Ovcharenko,I., Pachter,L. and Rubin,E.M. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, **299**, 1391–1394.
6. Frith,M.C., Li,M.C. and Weng,Z. (2003) Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.*, **31**, 3666–3668.
7. Alkema,W.B., Johansson,O., Lagergren,J. and Wasserman,W.W. (2004) MSCAN: identification of functional clusters of transcription factor binding sites. *Nucleic Acids Res.*, **32**, W195–W198.
8. Sinha,S., Schroeder,M.D., Unnerstall,U., Gaul,U. and Siggia,E.D. (2004) Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. *BMC Bioinform.*, **5**, 129.
9. Blanchette,M., Bataille,A.R., Chen,X., Poitras,C., Laganiere,J., Lefebvre,C., Deblois,G., Giguere,V., Ferretti,V., Bergeron,D. *et al.* (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.*, **16**, 656–668.
10. Hallikas,O., Palin,K., Sinjushina,N., Rautiainen,R., Partanen,J., Ukkonen,E. and Taipale,J. (2006) Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell*, **124**, 47–59.
11. Gupta,M. and Liu,J.S. (2005) De novo cis-regulatory module elicitation for eukaryotic genomes. *Proc. Natl Acad. Sci. USA*, **102**, 7079–7084.
12. Zhou,Q. and Wong,W.H. (2004) CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl Acad. Sci. USA*, **101**, 12114–12119.
13. Frith,M.C., Spouge,J.L., Hansen,U. and Weng,Z. (2002) Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res.*, **30**, 3214–3224.
14. Bailey,T.L. and Noble,W.S. (2003) Searching for statistically significant regulatory modules. *Bioinformatics*, **19(Suppl 2)**, ii16–ii25.
15. Johansson,O., Alkema,W., Wasserman,W.W. and Lagergren,J. (2003) Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics*, **19(Suppl 1)**, i169–i176.
16. Sinha,S., van Nimwegen,E. and Siggia,E.D. (2003) A probabilistic method to detect regulatory modules. *Bioinformatics*, **19(Suppl 1)**, i292–i301.
17. Wingender,E., Dietze,P., Karas,H. and Knuppel,R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.
18. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
19. Durinck,S., Moreau,Y., Kasprzyk,A., Davis,S., De Moor,B., Brazma,A. and Huber,W. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.
20. Claverie,J.M. and Audic,S. (1996) The statistical significance of nucleotide position-weight matrix matches. *Comput. Appl. Biosci.*, **12**, 431–439.
21. Donaldson,I.J. and Gottgens,B. (2007) CoMoDis: composite motif discovery in mammalian genomes. *Nucleic Acids Res.*, **35**, e1.
22. Kel,A., Konovalova,T., Waleev,T., Cheremushkin,E., Kel-Margoulis,O. and Wingender,E. (2006) Composite Module Analyst: a fitness-based tool for identification of transcription factor binding site combinations. *Bioinformatics*, **22**, 1190–1197.
23. Makeev,V.J., Lifanov,A.P., Nazina,A.G. and Papatsenko,D.A. (2003) Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. *Nucleic Acids Res.*, **31**, 6016–6026.
24. Yu,X., Lin,J., Masuda,T., Esumi,N., Zack,D.J. and Qian,J. (2006) Genome-wide prediction and characterization of interactions between transcription factors in Saccharomyces cerevisiae. *Nucleic Acids Res.*, **34**, 917–927.
25. Yu,X., Lin,J., Zack,D.J. and Qian,J. (2006) Computational analysis of tissue-specific combinatorial gene regulation: predicting

interaction between transcription factors in human tissues. *Nucleic Acids Res.*, **34**, 4925–4936.
26. Waterman,M.S. (1995) edn). *Introduction to Computational Biology: Maps, Sequences and Genomes*, *1st edn*. Chapman & Hall/CRC, New York, NY.
27. Wagner,A. (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, **15**, 776–784.
28. Sharan,R., Ovcharenko,I., Ben-Hur,A. and Karp,R.M. (2003) CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics*, **19(Suppl 1)**, i283–i291.
29. Pilpel,Y., Sudarsanam,P. and Church,G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **29**, 153–159.
30. Bussemaker,H.J., Li,H. and Siggia,E.D. (2001) Regulatory element detection using correlation with expression. *Nat. Genet.*, **27**, 167–171.
31. Kel-Margoulis,O.V., Kel,A.E., Reuter,I., Deineko,I.V. and Wingender,E. (2002) TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res.*, **30**, 332–334.
32. Kel-Margoulis,O.V., Romashchenko,A.G., Kolchanov,N.A., Wingender,E. and Kel,A.E. (2000) COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation. *Nucleic Acids Res.*, **28**, 311–315.
33. Shimajiri,Y., Sanke,T., Furuta,H., Hanabusa,T., Nakagawa,T., Fujitani,Y., Kajimoto,Y., Takasu,N. and Nanjo,K. (2001) A missense mutation of Pax4 gene (R121W) is associated with type 2 diabetes in Japanese. *Diabetes*, **50**, 2864–2869.
34. Kanatsuka,A., Tokuyama,Y., Nozaki,O., Matsui,K. and Egashira,T. (2002) Beta-cell dysfunction in late-onset diabetic subjects carrying homozygous mutation in transcription factors NeuroD1 and Pax4. *Metabol. Clin. Exp.*, **51**, 1161–1165.
35. Brun,T., Duhamel,D.L., Hu He,K.H., Wollheim,C.B. and Gauthier,B.R. (2007) The transcription factor PAX4 acts as a survival gene in INS-1E insulinoma cells. *Oncogene*, **26**, 4261–4271.
36. Brun,T., Franklin,I., St-Onge,L., Biason-Lauber,A., Schoenle,E.J., Wollheim,C.B. and Gauthier,B.R. (2004) The diabetes-linked transcription factor PAX4 promotes {beta}-cell proliferation and survival in rat and human islets. *J. Cell Biol.*, **167**, 1123–1135.
37. Tokuyama,Y., Matsui,K., Ishizuka,T., Egashira,T. and Kanatsuka,A. (2006) The Arg121Trp variant in PAX4 gene is associated with beta-cell dysfunction in Japanese subjects with type 2 diabetes mellitus. *Metabol. Clin. Exp.*, **55**, 213–216.
38. Carlsson,P. and Mahlapuu,M. (2002) Forkhead transcription factors: key players in development and metabolism. *Dev. Biol.*, **250**, 1–23.
39. Katoh,M. and Katoh,M. (2004) Human FOX gene family (Review). *Int. J. Oncol.*, **25**, 1495–1500.
40. Kaestner,K.H. (2000) The hepatocyte nuclear factor 3 (HNF3 or FOXA) family in metabolism. *Trends Endocrinol. Metab.*, **11**, 281–285.
41. Wolfrum,C., Asilmaz,E., Luca,E., Friedman,J.M. and Stoffel,M. (2004) Foxa2 regulates lipid metabolism and ketogenesis in the liver during fasting and in diabetes. *Nature*, **432**, 1027–1032.
42. Smith,U. (2007) TCF7L2 and type 2 diabetes—we WNT to know. *Diabetologia*, **50**, 5–7.
43. Schlosshauer,P.W., Pirog,E.C., Levine,R.L. and Ellenson,L.H. (2000) Mutational analysis of the CTNNB1 and APC genes in uterine endometrioid carcinoma. *Mod. Pathol.*, **13**, 1066–1071.
44. McKnight,A.J., Savage,D.A., Patterson,C.C., Sadlier,D. and Maxwell,A.P. (2007) Resequencing of genes for transforming growth factor beta1 (TGFB1) type 1 and 2 receptors (TGFBR1, TGFBR2), and association analysis of variants with diabetic nephropathy. *BMC Med. Genet.*, **8**, 5.
45. Harvey,N.L., Srinivasan,R.S., Dillard,M.E., Johnson,N.C., Witte,M.H., Boyd,K., Sleeman,M.W. and Oliver,G. (2005) Lymphatic vascular defects promoted by Prox1 haploinsufficiency cause adult-onset obesity. *Nat. Genet.*, **37**, 1072–1081.
46. Goulley,J., Dahl,U., Baeza,N., Mishina,Y. and Edlund,H. (2007) BMP4-BMPR1A signaling in beta cells is required for and augments glucose-stimulated insulin secretion. *Cell Metabol.*, **5**, 207–219.
47. Pollard,J. Jr, Butte,A.J., Hoberman,S., Joshi,M., Levy,J. and Pappo,J. (2005) A computational model to define the molecular causes of type 2 diabetes mellitus. *Diabetes Technol. Ther.*, **7**, 323–336.
48. Lam,V.K.L., Ho,J.S.K.H., So,W.Y., Ma,R.C.W., Chan,J.C.N. and Ng,M.C.Y. (2006) Association of *POU2F1* genetic polymorphisms with type 2 diabetes in Hong Kong Chinese. *Program Nr: 1452 for the 2006 ASHG Annual Meeting,* Hong Kong, China.
49. Laybutt,D.R., Preston,A.M., Akerfeldt,M.C., Kench,J.G., Busch,A.K., Biankin,A.V. and Biden,T.J. (2007) Endoplasmic reticulum stress contributes to beta cell apoptosis in type 2 diabetes. *Diabetologia*, **50**, 752–763.
50. Fu,Y., Luo,L., Luo,N., Zhu,X. and Garvey,W.T. (2007) NR4A orphan nuclear receptors modulate insulin action and the glucose transport system: potential role in insulin resistance. *J. Biol. Chem.*, **282**, 31525–31533.
51. Frayling,T.M. (2007) Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nat. Rev.*, **8**, 657–662.
52. Ubeda,M., Rukstalis,J.M. and Habener,J.F. (2006) Inhibition of cyclin-dependent kinase 5 activity protects pancreatic beta cells from glucotoxicity. *J. Biol. Chem.*, **281**, 28858–28864.
53. Lee,H.B., Noh,H., Seo,J.Y., Yu,M.R. and Ha,H. (2007) Histone deacetylase inhibitors: a novel class of therapeutic agents in diabetic nephropathy. *Kidney Int.*, **72**, S61–S66.
54. Hulbert,E.M., Smink,L.J., Adlem,E.C., Allen,J.E., Burdick,D.B., Burren,O.S., Cassen,V.M., Cavnor,C.C., Dolman,G.E., Flamez,D. *et al.* (2007) T1DBase: integration and presentation of complex data for type 1 diabetes research. *Nucleic Acids Res.*, **35**, D742–D746.
55. Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigo,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T., Thurman,R.E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.