Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W475–W484***
DOI: 10.1093/nar/gkh353

CONFAC: automated application of comparative genomic promoter analysis to DNA microarray datasets

Suresh Karanam[1,2] **and Carlos S. Moreno**[2,*]

[1]Program in Bioinformatics, School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA and [2]Department of Pathology and Laboratory Medicine and Winship Cancer Institute, Emory University School of Medicine, Atlanta, GA 30322, USA

Received January 29, 2004; Revised and Accepted February 23, 2004

ABSTRACT

The advent of DNA microarray technology and the sequencing of multiple vertebrate genomes has provided a unique opportunity for the integration of comparative genomics with high-throughput gene expression analysis. Here we describe the conserved transcription factor binding site (CONFAC) software that enables the high-throughput identification of conserved transcription factor binding sites (TFBSs) in the regulatory regions of hundreds of genes at a time (http://morenolab.whitehead.emory.edu/cgi-bin/confac/login.pl). The CONFAC software compares non-coding regulatory sequences between human and mouse genomes to enable identification of conserved TFBSs that are significantly enriched in promoters of gene clusters from microarray analyses compared to sets of unchanging control genes using a Mann–Whitney *U*-test. Analysis of random gene sets demonstrated that using our approach, over 98% of TFBSs had false positive rates below 5%. As a proof-of-principle, we have validated the CONFAC software using gene sets from four separate microarray studies and identified TFBSs known to be functionally important for regulation of each of the four gene sets.

INTRODUCTION

Even though the binding site specificities of many transcription factors have been experimentally defined, these transcription factor binding sites (TFBSs) are short (6–12 bp) degenerate motifs that are very common throughout genomic sequences. The specificities of transcription factors are typically represented as position weight matrices (PWMs) that are found in the TRANSFAC database. The MATCH and MatInspector software packages (1,2) use these PWMs to identify sequence matches, but because TFBSs are so common, the vast majority of detected TFBSs are not functionally important. One approach that greatly decreases the false positive rate for detection of functionally relevant TFBSs is the use of ‘phylogenetic footprinting’ or comparative genomics (3–10). These methods are based on the hypothesis that non-coding genomic sequences that are functionally important for gene expression will be more highly conserved during evolution than unimportant sequences.

The use of high-density DNA microarrays to identify sets of genes with similar expression patterns is rapidly becoming a widespread approach for understanding biological processes. Typically, microarray data is analyzed by hierarchical clustering, self-organizing maps, *K*-means clustering or principle component analysis. Most of these approaches readily identify clusters of tens to hundreds of genes that demonstrate similar expression patterns. One logical systematic approach to study a cluster of genes with similar expression profiles is to analyze the promoter sequences for each member of the gene clusters and attempt to identify transcription factors that might be crucial for regulating their expression.

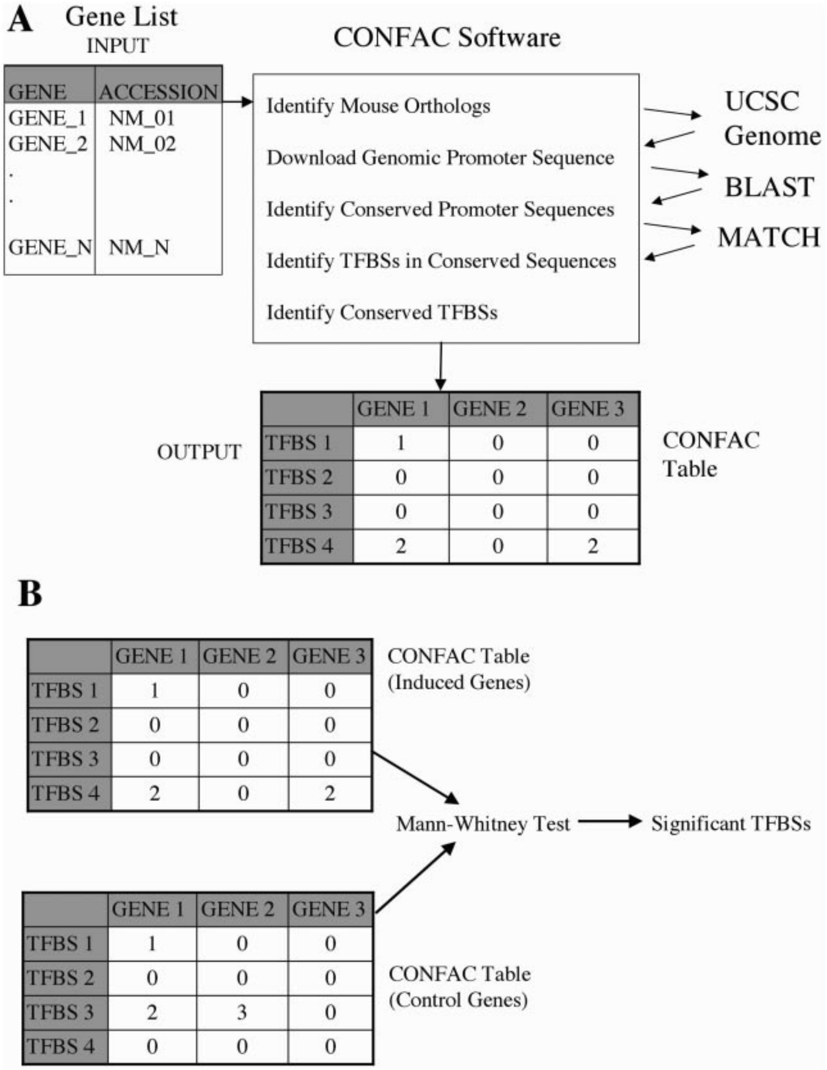
Here we describe the conserved transcription factor binding site (CONFAC) software that enables the high-throughput identification of conserved TFBSs in the regulatory regions of hundreds of genes at a time (http://morenolab.whitehead.emory.edu/cgi-bin/confac/login.pl). Our novel approach allows identification of TFBSs that are significantly more common in promoters of a group of genes of interest from microarray analyses than in a set of unchanging control genes. Although other tools such as *cis*-regulatory module explorer (CREME) (11) and TOUCAN (12,13) can analyze multiple genes, they do not allow for direct comparisons against user-defined

[*]To whom correspondence should be addressed at: Whitehead Biomedical Research Building, Room 105A, 615 Michael St., Atlanta, GA 30322, USA. Tel: +1 404 712 2809; Fax: +1 404 727 8538; Email: cmoreno@emory.edu

control gene lists. Moreover, in this study we have validated the CONFAC software using gene sets from four separate microarray studies and identified TFBSs known to be functionally important for regulation of each of the four gene sets.

RESULTS AND METHODS

The CONFAC software runs in the Linux operating system, using cgi scripts written in the Perl programming language, and accepts lists of genes via a web-browser interface (http://morenolab.whitehead.emory.edu/cgi-bin/confac/login.pl). The user inputs a tab-delimited text file containing a unique identifier for the gene name in the first column and a GenBank accession number or RefSeq ID in the second column (Figure 1A). The CONFAC software then automatically identifies orthologous murine genes by accessing ortholog lookup tables obtained from the UCSC and ENSEMBL genome databases. The use of lookup tables is more conservative than using protein BLAT searches against the mouse genome, since it prevents the inadvertent comparison with non-expressed pseudogenes or hypothetical genes, and provides the user with well-curated sets of orthologs. While this approach does not identify orthologs for all genes submitted, a test run identified 350 murine orthologs upon submission of a set of 450 randomly selected human RefSeq IDs, corresponding to a hit rate of 78%. Once ortholog pairs are identified, 3 kb of genomic sequence 5′ of the transcriptional start site and up to 20 kb of the first intron are downloaded from the UCSC assemblies of the human and mouse genomes for each gene. Downloaded human and mouse genomic sequences from orthologous gene pairs are then compared by pairwise BLAST, and only significantly conserved (e-value < 0.001) sequences are analyzed for TFBSs via an automated interface with the MATCH software (1). The user has the option of defining core and matrix similarities used by MATCH when


A
Gene List
INPUT
CONFAC Software
Identify Mouse Orthologs
Download Genomic Promoter Sequence
Identify Conserved Promoter Sequences
Identify TFBSs in Conserved Sequences
Identify Conserved TFBSs
UCSC Genome
BLAST
MATCH
OUTPUT
CONFAC Table
B
CONFAC Table (Induced Genes)
Mann-Whitney Test
Significant TFBSs
CONFAC Table (Control Genes)

GENE	ACCESSION
GENE_1	NM_01
GENE_2	NM_02
.	
.	
GENE_N	NM_N

	GENE 1	GENE 2	GENE 3
TFBS 1	1	0	0
TFBS 2	0	0	0
TFBS 3	0	0	0
TFBS 4	2	0	2

	GENE 1	GENE 2	GENE 3
TFBS 1	1	0	0
TFBS 2	0	0	0
TFBS 3	0	0	0
TFBS 4	2	0	2

	GENE 1	GENE 2	GENE 3
TFBS 1	1	0	0
TFBS 2	0	0	0
TFBS 3	2	3	0
TFBS 4	0	0	0

Figure 1. (**A**) Schematic of data flow in CONFAC software. The user input is a tab-delimited list of genes of interest. The CONFAC software interfaces with the human and mouse genomes, local pairwise BLAST and local MATCH software to identify TFBSs that are conserved between human and mouse promoter regions. The output is a table of TFBS occurrences for each gene that has at least one conserved TFBS. (**B**) Identification of significantly enriched TFBSs. Two CONFAC output tables for affected and control gene sets are submitted to a Mann–Whitney U-test to identify sites that are significantly overrepresented in the affected gene list compared to controls.

submitting the initial gene list for analysis. All MATCH analyses in this study were performed with a core similarity of 0.95 and a matrix similarity of 0.85 using the vertebrate set of PWMs.

The CONFAC software compares the MATCH output from the conserved human and mouse genomic sequences and identifies those TFBSs that are present in both the human and mouse sequences within a 25 bp sliding window. The output of the CONFAC software is a table in which each column represents a given gene, and each row represents a transcription factor with a PWM in the TRANSFAC database. A schematic of the CONFAC software data flow and a sample output are shown in Figure 1. Each element of the table is the number of occurrences of any given TFBS in the human and mouse conserved promoter sequences for a submitted gene. Consolidated TFBS frequencies used in these studies were calculated by summing all PWM hits for the same transcription factor, so that for example, a sequence with two E2F_01 hits, three E2F_02 hits and one E2F_03 hit would have a consolidated total of six E2F hits. In addition to tabular output, graphical output of the conserved regulatory sequences and conserved TFBSs are also generated. The user has the option to download the entire genomic sequences analyzed for each gene set, conserved genomic sequences only, all TFBSs present in the conserved human genomic sequences, or only TFBSs that are conserved between the human and mouse genomic sequences. In a benchmark test of the speed of the CONFAC software, lists of 100 genes found 73–82 orthologous pairs and conserved TFBSs were identified for 59–65 gene pairs in under 8 min. This high level of automation and speed enables the completion of analyses in minutes that previously would take days or weeks if performed one gene at a time.

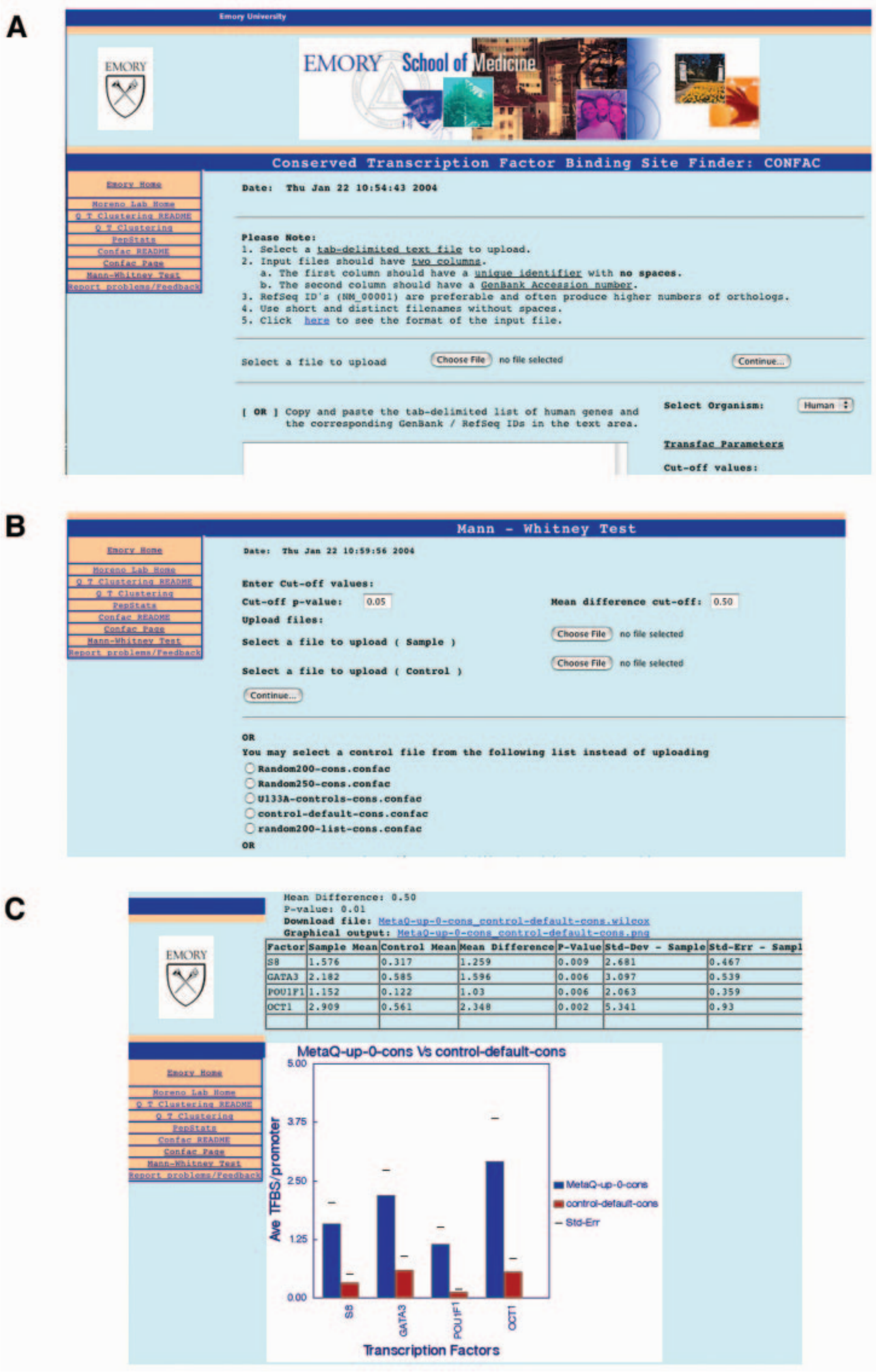
A critical and novel component in our method for the identification of functionally significant TFBSs is the statistical comparison of identified TFBSs between two gene sets (Figure 1B). Typically, we identify a control set of genes that show little or no variation in a microarray experiment, and a set of experimental genes that are increased or decreased in expression after some perturbation. Users have the option to upload their own control gene sets, or to use several available control lists, including normalization control lists from Affymetrix U133A arrays, our default set that shows no change in cancer cells, and sets of randomly selected genes. To allow the user to perform statistical comparisons of identified TFBSs between two gene sets, the CONFAC software uses the *R* statistical programming environment and the RSPerl module that enables integration of *R* functions with Perl cgi scripts. The CONFAC software identifies TFBSs that are significantly more common in the experimental gene set than in the control gene set using a non-parametric Wilcoxon Rank Sum test equivalent to a Mann–Whitney *U*-test. While the CONFAC data can have many ties (0, 1, 2 TFBS), the Mann–Whitney test (in which the null hypothesis is that the medians do not differ against a two-sided alternative) adjusts for ties. The key assumption of the Mann–Whitney test is that the underlying distributions being compared must be continuous, but need not be symmetric. However, others (14) have investigated the use of the Mann–Whitney test in comparing discrete distributions and have found only small losses in power when applying the test to grouped data or data with a limited number of values when appropriate provision for ties has been addressed. A list of TFBSs that are significantly more common in the experimental gene set are then returned to the user along with the corresponding *P*-value. The user can define both the *P*-value cutoff as well as a mean-difference cutoff, which sets a minimum threshold for the differences in the average TFBS frequencies between the two groups. Screenshots of the user interface are shown in Figure 2 for uploading gene lists (Figure 2A), submitting CONFAC results for Mann–Whitney tests (Figure 2B) and the output table of significant TFBSs (Figure 2C). In addition to tabular statistical results, the Mann–Whitney page generates bar graphs of the average conserved TFBS frequencies for the sample and control gene sets.

False positive analysis of CONFAC software

To investigate the false positive rate of conserved TFBS detection, we generated 25 random sets of 100 RefSeq IDs and identified conserved TFBSs in these random gene sets. We then performed all 300 possible pairwise comparisons between these 25 random gene sets using the Mann–Whitney *U*-test with a mean-difference cutoff of 0.5 and a *P*-value of 0.05. Only three TFBSs, (GATA3, ZTA and NKX25) had false positive rates exceeding the predicted *P*-value of 5% (Table 1, column 1). Since the CONFAC software looks for over 200 TFBSs, at least 197/200 or 98.5% of TFBSs had false positive rates below 5%. We then went on and generated an additional 25 random sets of 100 RefSeq IDs and compared all 50 random gene sets to random sets of 200 and 250 control genes that are available as control comparison files on the CONFAC Mann–Whitney test page. As can be seen in Table 1, only one TFBS (NKX25) exceeded 5% false positive rate for the set of 200 random control genes, and only one TFBS (E2F) exceeded 5% for the set of 250 random control genes.

We next compared the 50 random gene sets to our default control gene set of 41 genes that were expressed but exhibited very little change between normal and tumor samples. This control gene set includes ribosomal proteins, subunits of protein phosphatases and actin-associated proteins, among others (Supplementary Table 1). Because the genes in the default control set were not randomly selected, these genes had a higher false positive rate compared to random gene sets. With a *P*-value of 0.05 and a mean-difference cutoff of 0.5, this control gene set produced 18 TFBSs that exceeded a false positive rate of 5% when compared to the 50 random gene sets (not shown). However, using a more stringent *P*-value cutoff of 0.01, only five TFBSs (CEBPDELTA, NKX25, LPOLYA, POU1F and S8 HOX) had false positive rates that exceeded 5%. For the validation analyses used in this study, we used a *P*-value cutoff of 0.01 and a mean-difference cutoff of 0.5 unless otherwise noted. Nevertheless, many of the TFBSs that were significantly different between random gene sets and this default control gene set at the less stringent *P*-value of 0.05 were homeobox and forkhead box (FOX) binding sites. These data suggest that comparison of non-changing genes in a microarray experiment against multiple random datasets can also potentially provide a measure of what TFBSs are important in changes seen in microarray data.

This effect was even more pronounced in the set of Affymetrix U133 control genes. With the mean-difference cutoff of 0.5 and a *P*-value of 0.05, the random gene sets produced



Conserved Transcription Factor Binding Site Finder: CONFAC

Date: Thu Jan 22 10:54:43 2004

Please Note:
1. Select a tab-delimited text file to upload.
2. Input files should have two columns.
 a. The first column should have a unique identifier with **no spaces**.
 b. The second column should have a GenBank Accession number.
3. RefSeq ID's (NM_00001) are preferable and often produce higher numbers of orthologs.
4. Use short and distinct filenames without spaces.
5. Click here to see the format of the input file.

Select a file to upload — no file selected

[OR] Copy and paste the tab-delimited list of human genes and the corresponding GenBank / RefSeq IDs in the text area.

Select Organism: Human

Transfac Parameters

Cut-off values:

Mann - Whitney Test

Date: Thu Jan 22 10:59:56 2004

Enter Cut-off values:
Cut-off p-value: 0.05
Mean difference cut-off: 0.50
Upload files:
Select a file to upload (Sample) — no file selected
Select a file to upload (Control) — no file selected

OR
You may select a control file from the following list instead of uploading
○ Random200-cons.confac
○ Random250-cons.confac
○ U133A-controls-cons.confac
○ control-default-cons.confac
○ random200-list-cons.confac
OR

Mean Difference: 0.50
P-value: 0.01
Download file: MetaQ-up-0-cons_control-default-cons.wilcox
Graphical output: MetaQ-up-0-cons_control-default-cons.png

Factor	Sample Mean	Control Mean	Mean Difference	P-Value	Std-Dev - Sample	Std-Err - Sampl
S8	1.576	0.317	1.259	0.009	2.681	0.467
GATA3	2.182	0.585	1.596	0.006	3.097	0.539
POU1F1	1.152	0.122	1.03	0.006	2.063	0.359
OCT1	2.909	0.561	2.348	0.002	5.341	0.93

Figure 2. (**A**) A screenshot of the CONFAC user interface for uploading gene lists. The user can specify core and matrix similarities, and sets of PWMs. (**B**) A screenshot of the user interface for the Mann–Whitney test for statistical significance. The user can upload their own control datasets or choose from several default control sets. The user also specifies the *P*-value and mean-difference cutoffs for the analysis. (**C**) A screenshot of the output of the Mann–Whitney test, which lists significant TFBSs, the average frequencies for both sets, the mean difference and the *P*-values.

32 TFBSs with a false positive rate exceeding 5% and 12 of those exceeded 20% (not shown). Using a more stringent mean-difference cutoff of 1.0 and a *P*-value of 0.001, the number of TFBSs with false positive rates exceeding 5% was reduced to two (HFH3 and HNF3ALPHA). These results are also a direct result of the fact that the genes in the Affymetrix normalization control set are also non-random and were chosen because they show very little variation among a wide

Table 1. Summary of false positive TFBS analyses

TFBS	False positive in random versus random comparisons (%)	False positive in random versus random 200 (%)	False positive in random versus random 250 (%)	False positive in random versus control-default (%)	False positive in random versus Affymetrix U133 controls (%)
AP1	2	0	0	0	4
CAAT	1.7	2	0	2	0
CEBPDELTA	4.7	0	0	8	2
E2F	0.3	0	6	0	0
ER	2.7	4	0	0	0
GATA3	9.7	2	0	2	0
HFH3	3	0	0	2	8
HNF3ALPHA	3	0	0	0	18
HNF3B	1.3	0	0	0	4
IK1	3.3	2	0	2	2
LPOLYA	0	0	0	10	0
NKX25	6.7	8	0	6	0
OCT_1	5	0	0	4	0
POU1F1	1	0	0	8	0
S8	1.7	2	0	8	0
P-value cutoff	0.05	0.05	0.05	0.01	0.001
Mean difference	0.5	0.5	0.5	0.5	1.0

The first column is the result of 300 pairwise comparisons of 25 random sets of 100 genes. The last four columns are the result of 50 comparisons of 50 random gene sets against single sets of 200 random genes, 250 random genes, 41 genes that show little change between tumor and normal samples and 97 normalization control genes for Affymetrix U133 GeneChip arrays.

Table 2. Summary of FKHR target classes and CONFAC analysis

FKHR target class	Total genes analyzed	Orthologous pairs with conserved TFBS	Percent genes with conserved TFBS (%)	Type of FKHR target	Dependence on DNA binding domain
Class I	32	20	63	Activation	Dependent
Class IIA	20	13	65	Activation	Independent
Class IIB	37	26	70	Activation	Independent and stronger in DBD mutant
Class III	27	24	89	Repression	Independent

variety of biological conditions. The results in Table 1 show that the promoters of these normalization control genes are significantly lacking in AP1, HFH3, HNF3ALPHA and HNF3B sites. Thus, while the use of non-randomly selected control gene sets can increase the sensitivity of detection of significant TFBSs, it is important for the user to be aware of the effects that this can introduce and either adjust the *P*-value and mean-difference cutoffs, or be cautious of results with respect to specific TFBSs that have high false positive rates.

Validation of the CONFAC software with forkhead overexpression microarray data

To validate the CONFAC approach using publicly available microarray data, we analyzed the results from Ramaswamy *et al.* (15), in which the authors identified three classes of genes that respond to overexpression of the forkhead-family transcription factor FKHR (or FOXO1). The first class of genes was induced by FKHR in a manner that was dependent on the ability of FKHR to bind to DNA, and thus one would predict that FOX sites and insulin response sites (IRS), to which FKHR has been shown to bind (15), would be enriched in the promoters of this Class I gene set (see Table 2). The second class of genes was induced by FKHR overexpression in a manner that was independent of the ability of FKHR to bind to DNA, suggesting that FKHR might activate the expression of this gene set by protein–protein interactions with other transcription factors involved in regulation of these genes. The third class of genes was repressed by FKHR overexpression, also in a manner independent of its DNA-binding domain (DBD).

We analyzed the promoters of FKHR Class I, Class IIA, Class IIB and Class III target genes for conserved TFBSs using CONFAC. The default control set of 41 constitutively expressed genes that exhibit very little variation in comparisons of normal and tumor tissues was used for the control gene set. The number of genes analyzed, the number of orthologous pairs with conserved TFBSs and the nature of each class are summarized in Table 2. A detailed list of the members of each gene class is given in Supplementary Table 2. Analysis of these four experimental and one control gene sets using the CONFAC software found that for those genes in Class I that require an intact FKHR DBD, there was a statistically significant increase in nine TFBSs (Figure 3A). Four of the significant TFBSs were FOX family sites and a fifth corresponds to the IRS to which FKHR has been shown to bind (15). The complete CONFAC tables for the significant TFBSs shown in Figure 3A are given in Supplementary Table 3. In the FKHR Class IIA and Class IIB target genes, which were activated by FKHR in a DBD-independent manner, no significant TFBSs were significantly enriched. Thus, the CONFAC software identified significant conserved TFBSs for FKHR-activated genes that require FKHR recognition of its

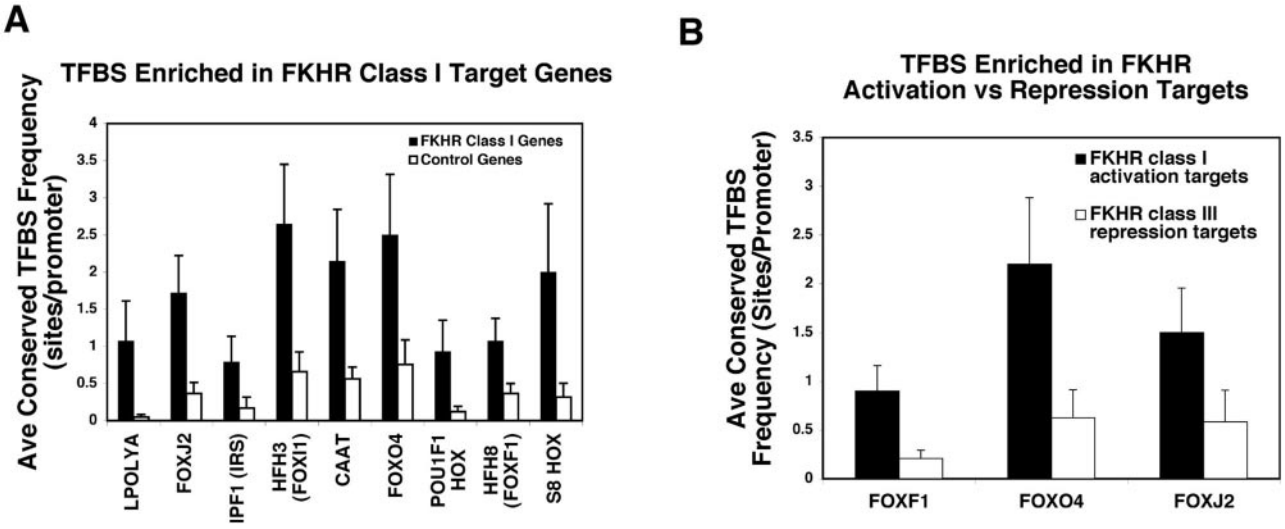


Figure 3. (**A**) The average frequency of TFBSs that are significantly enriched in 20 FKHR Class I target genes relative to 41 control genes are graphed for both Class I target genes and control genes. Four FOX sites and the FKHR-responsive IRS site were significantly overrepresented in this DBD-dependent gene set. Error bars represent the standard error for each conserved TFBS in this and all subsequent figures. (**B**) FOX sites were significantly more frequent in promoters of 20 FKHR Class I activation target genes than in promoters of 24 FKHR Class III repression target genes.

DNA binding site, but not for genes that were activated by FKHR independently of its ability to bind to DNA.

Finally, we analyzed 27 FKHR Class III repression target genes that are repressed independently of FKHR DNA binding, and analyzed the conserved TFBSs in 24 orthologous gene pairs. No TFBSs were significantly overrepresented in this gene set relative to control genes. However, when the Class III repression target genes were compared to the Class I activation target genes, the only significantly different TFBSs were FOX family sites (Figure 3B). These data are consistent with repression of these genes via a mechanism that does not require DNA binding by FKHR.

Validation of the CONFAC software with NF-κB target gene microarray data

It is well established that the transcription factor NF-κB activates transcription of target genes in response to signal transduction pathways activated by the cytokine, tumor necrosis factor-α (TNF-α) (16). We next analyzed a set of 29 genes found by DNA microarray analysis (17) to be upregulated 2-fold within 1 h of TNF-α treatment to determine whether these genes would be enriched for NF-κB binding sites (Supplementary Table 4). For this gene set, conserved TFBSs were identified in 21 orthologous gene pairs and 21 TFBSs were significantly increased relative to control gene sets, using a *P*-value threshold of 0.01 (Supplementary Table 5). Of those 21 TFBSs, the most significant 7 TFBSs had a $P < 0.001$, and are shown relative to control genes in Figure 4A. Of the seven TFBSs with a $P < 0.001$, five were NF-κB sites. The other highly significant TFBSs were C/EBPδ and ETS2, suggesting that C/EBP and ETS factors may play an important role in the NF-κB response to TNF signaling.

In another microarray study of NF-κB target genes in Hodgkin and Reed–Steinberg (HRS) tumor cells (18), the authors identified several known NF-κB targets and verified a number of novel ones. They also showed that NF-κB induced overexpression and activation of STAT5a in HRS cells. We analyzed a set of 26 NF-κB target genes from this study and identified conserved TFBSs in 20 orthologous pairs (Supplementary Table 6). A total of eight significant TFBSs enriched in these 20 genes relative to control genes using a *P*-value of 0.05 and a mean-difference cutoff of 0.25 are shown in Figure 4B. Not only were 50% of the significant TFBSs NF-κB sites, but also STAT and STAT5b sites were significantly enriched in this gene set, suggesting that besides activating STAT5a in HRS cells, NF-κB cooperates with activated STAT5 to activate target gene expression. While the IK1 site produced false positive hits when the random gene sets were compared to this control gene set (Table 1), the STAT sites showed no false positives even with a *P*-value of 0.05 and a mean difference of 0.25. The genes with both conserved NF-κB and STAT sites included ICAM-1, lymphotoxin-α, granulocyte–macrophage colony-stimulating factor (GM-CSF) and the antiapoptotic genes Bcl-x and immediate early response 3 (IEX-1).

CONFAC analysis of Cyclin D target genes identifies C/EBPδ TFBS

Recent studies have shown that C/EBP is critical for activation of genes regulated by Cyclin D1 (19). We analyzed a set of 43 genes from this study that were highly correlated with Cyclin D1 expression in a large set of tumor tissue samples (Supplementary Table 7) and found conserved TFBSs in 31 human–mouse ortholog pairs. CONFAC analysis and comparison with our default set of 41 control genes identified C/EBPδ as the only significantly enriched TFBS in the promoters of this gene set, again validating the CONFAC approach. An average of 1.5 conserved C/EBP sites were detected in the promoters of the 31 genes that were very highly correlated with cyclin D1 expression, while only 0.5 C/EBP sites were found on average in the promoters of the control gene set (*P*-value = 0.002).

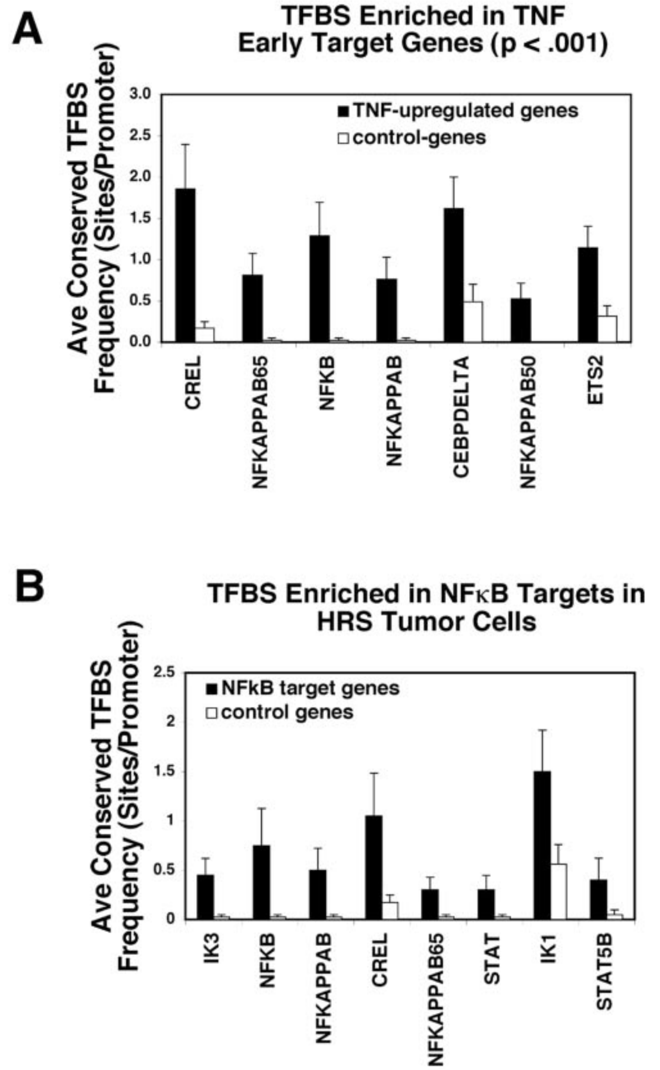


Figure 4. (**A**) The average frequency of TFBSs that are highly significantly enriched ($P < 0.001$) in 21 TNF-inducible genes relative to 41 control genes is shown. Five of the most significant ($P < 0.001$) TFBSs were NF-κB sites. (**B**) The five most significantly enriched TFBSs in promoters of 20 NF-κB target genes from HRS tumor cells were NF-κB sites. STAT sites were also significantly enriched relative to control genes using P-value <0.05 and mean difference >0.25.

CONFAC analysis of genes upregulated in prostate cancer

We next wanted to apply the CONFAC analysis to a set of genes for which the expected TFBSs were unknown but that would be of clinical interest. A number of studies have applied microarray analysis to the study of genes affected in prostate cancer (20–25). One meta-analysis of several prostate cancer microarray studies identified sets of genes that were upregulated in multiple microarray studies, and assigned a false-discovery rate, or q-value for each gene (23). We selected the 46 genes with the most significant q-values ($q < 0.1$) for CONFAC analysis, and found conserved TFBSs in 33 ortholog pairs (Supplementary Table 8). Four TFBSs were significantly enriched in the promoters of these 33 genes that are overexpressed in prostate cancers and three of these four sites were homeobox family sites (Figure 5). Interestingly, when these genes were analyzed with a less stringent P-value cutoff of 0.05, hypoxia-inducible factor 1 (HIF-1), FOXA1, FOXN1 and growth factor-independent1 (GFI1) sites were also significantly enriched in genes that are overexpressed in prostate cancer (not shown). In addition, the average TFBS frequency of FOXA1 and FOXN1 sites was greater than two sites/promoter, suggesting that many of these prostate cancer specific genes could be activated by the AKT–FKHR pathway in cooperation with homeobox, and GATA3 factors. Consistent with this hypothesis, both functional and physical interactions have been demonstrated between FKHR and HOXA5 (26), and between FOXA2 and GATA4 (27). Furthermore, the enrichment for HIF1 sites is consistent with an established role of HIF1-α in prostate cancer and interactions between HIF1-α and the PI3K/AKT pathway in human prostate cancer cells (28–30).

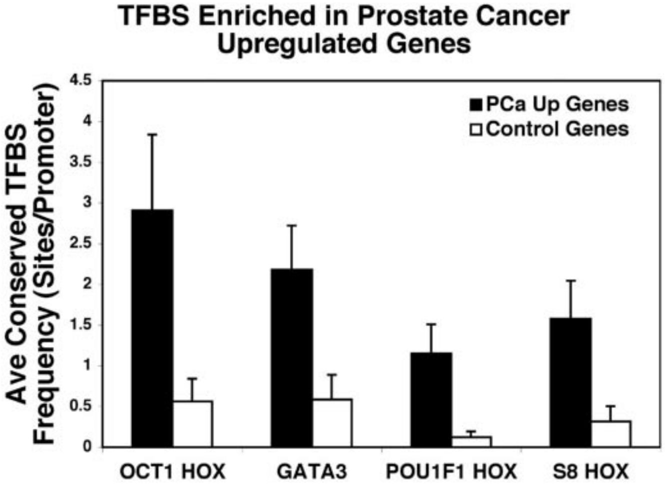


Figure 5. The average frequency of TFBSs that are significantly enriched in 33 genes strongly upregulated in prostate cancer (23) relative to 41 control genes is shown. Three of the four significant TFBSs were homeobox family sites.

DISCUSSION

Here we have described a rapid, novel and practical approach for identification of transcription factors that may regulate clusters of genes from DNA microarray experiments. We have also validated our approach with data from four separate studies (15,17–19). Our approach for integrating high-throughput comparative genomics with TFBS analysis and statistical comparisons can be applied to virtually any microarray experiment. For example, cells that have been treated with a drug, cytokine or growth factor can be analyzed by microarray expression profiling, and the sets of induced genes analyzed using the CONFAC software to identify the transcription factors that act downstream of signal transduction pathways. Such analyses could also provide insights into mechanisms of drug actions. Moreover, it may be possible to identify co-factors that interact with a given transcription factor to regulate various clusters of genes if sets of TFBSs are detected together. The CONFAC approach is not necessarily limited to sets of microarray data, since the input to the software is only a list of genes. Thus, sets of genes that are thought to be co-regulated can be analyzed even without access to microarray evidence. Ultimately, this approach can (and will) be applied to the entire genome.

There are several important assumptions that the user should be aware of that underlie the application of the CONFAC software to identify important TFBSs in clustered microarray gene lists. First, there is the assumption that sets of similarly expressed genes are regulated by common transcription factors. In some cases, sets of genes may be activated or repressed via a handful of distinct mechanisms, and this will decrease the sensitivity of the CONFAC approach. The more tightly clustered in expression pattern a gene list is, the more likely they will be regulated by common transcription factors, and the more likely the CONFAC analysis will provide useful insights into molecular mechanisms that affect expression of gene clusters.

Second, it is assumed that important regulatory sequences will be evolutionarily conserved between mouse and human genomes. Several studies have shown this to be a valid assumption for 70–90% of regulatory sequences (3–10). Analysis of 28 muscle-specific genes found that 74/75 (98%) of experimentally defined TFBSs were located within regions that are conserved between human and mouse genomes (6). In a study of Bruton's tyrosine kinase (BTK), a highly conserved 3.5 kb section adjacent to the first exon was shown to confer lineage-specific expression to reporter constructs (3). In another study, 1Mb of orthologous human and mouse sequences containing a cytokine gene cluster were compared and 35/40 (88%) of experimentally defined AP-1 and NFAT sites were present in conserved, orthologous sequences (8). Moreover, the total number of AP-1 and NFAT sites that were identified was reduced by 95% when searching only the conserved sequences compared to the entire 1Mb segment. Yet another study of 14 gene pairs and 40 verified TFBSs found that the total number of sites detected was reduced by 85% while maintaining detection of 83% of verified sites (7). Thus, comparison of human and mouse genome sequences can greatly reduce the background noise of false positive TFBSs with only a small loss in the overall sensitivity for detection of functionally significant TFBSs. While other resources (rVISTA (8), TraFaC (31) and ConSite (7)) that allow for the alignment and TFBS analysis of orthologous sequences do exist, these resources are designed to be used a single gene at a time, so that analysis of 50 genes or more becomes impractical. Other tools such as (CREME) (11) and TOUCAN (12,13) have been developed for analysis of multiple genomic regulatory regions. CREME enables the identification of sets of TFBSs that co-occur in a significant manner, but it requires an input of at least 50 genes, and does not compare the data from the gene list of interest with a control gene list. Moreover, the promoter regions that are analyzed in CREME do not include more than one sequence conserved in human–mouse alignments, which is often fairly short (<200 bp), and it does not examine any intronic sequences. The TOUCAN software does compare data from genes of interest with background files that are mathematical representations of the noise in genomic sequences, but it does not allow direct comparisons with user-selected control sequences, and often identifies many more TFBSs than the Mann–Whitney tests built into the CONFAC software.

Third, it is assumed that the mouse and human orthologous genes will be similarly regulated transcriptionally, which should be true for most (but not all) ortholog pairs. Fourth, it is assumed that the important regulatory sequences lie in the proximal 3 kb of genomic sequence upstream of the transcription start site and in the first intron. While this assumption is clearly not true for all genes, it is for a large number of them. We are in the process of developing improvements to the CONFAC software that will allow the user to define the size of the genomic regions to be analyzed. Finally, the CONFAC software is currently limited in its sensitivity by the available PWMs that are present in the TRANSFAC 4.0 database. Thus, it cannot detect what it does not know to look for. Planned enhancements to the CONFAC software are to enable the user to upload their own PWMs, provide an alternative set of unique, non-redundant PWMs and interfaces with *de novo* detection algorithms to identify novel, conserved, DNA-binding motifs. In addition, the CONFAC software is currently designed for use only with human microarray datasets. Future plans will enable analysis of mouse and rat gene lists, as well. In addition, algorithms to identify sets of TFBSs that may function as control modules are under development.

Another critical component of the CONFAC software is the choice of which control gene list should be used to compare against sets of induced or repressed genes. The optimal choice for a control gene list is a set of genes that are expressed but show little or no variation in the same microarray experiment as the experimental gene list. If such a list is not available because the user is analyzing published data without access to the entire dataset, a good (and fairly sensitive) second choice is a set of normalization controls for Affymetrix U133A GeneChips or the default list of 41 genes used in this report that have exhibited very little change in numerous hybridizations. However, the user should apply the P-value and mean-difference cutoffs used in Table 1 with these datasets and be aware which sites are more likely to produce false positive hits with these control gene sets. A third, more conservative choice, is to use a large, randomly selected gene set. Randomly selected gene sets will usually decrease the sensitivity of the CONFAC approach in Mann–Whitney comparisons. However, if one detects highly significant TFBSs even against several randomly selected gene sets, the user can be quite confident of the significance of the results. For example, the NF-κB sites found in the TNF-induced target genes were highly significant ($P < 0.001$) against every control set tested.

Another issue is how to most appropriately deal with statistical multiple testing issues, since we are testing approximately 200 TFBSs for statistical significance. One approach would be to simply reduce the cutoff P-value. Another would be to perform resampling of the genes into 1000 randomized groups in a bootstrapping approach, and we are currently investigating the utility of this method. Nevertheless, we have found that an effective heuristic method for reducing the number of false positives while retaining likely important TFBSs is by use of the mean-difference cutoff, as shown in the microarray validation studies described in this report. This approach is similar to the use of fold change cutoffs in SAM analyses (32) of DNA microarray data, and performed quite well at reducing false positives while retaining true positives.

While the CONFAC software is a highly useful approach for microarray data analysis and hypothesis generation, it does not prove in any way the functional importance of any transcription factor in regulation of a gene cluster. Nevertheless, it can provide useful insights into potential mechanisms of gene regulation and interactions between transcription factors that can then be followed up experimentally *in vitro* and *in vivo*. We have validated the CONFAC approach using microarray datasets from four different studies, and in each case detected patterns of significant TFBSs that match molecular observations, demonstrating the utility of combining comparative genomics, gene clustering and statistical comparisons to identify functionally important TFBSs. Moreover, these approaches may be useful in future studies for identification of novel, uncharacterized TFBSs. As the number of validated TFBSs grows, so will the power of the CONFAC approach described here.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Drs Jeremy Boss, Kirk Easley, Andrew Neish and Paul Wade for critical reading of this manuscript. This research was supported in part by NIH grant K22- CA 96560 to C.S.M. and the Department of Pathology and Laboratory Medicine, Emory University School of Medicine. This work is the subject of a US provisional patent application.

REFERENCES

1. Quandt,K., Frech,K., Karas,H., Wingender, E. and Werner,T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.
2. Kel,A.E., Gossling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.
3. Oeltjen,J.C., Malley,T.M., Muzny,D.M., Miller,W., Gibbs,R.A. and Belmont,J.W. (1997) Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res.*, **7**, 315–329.
4. Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278**, 167–181.
5. Fickett,J.W. and Wasserman,W.W. (2000) Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biotechnol.*, **11**, 19–24.
6. Wasserman,W.W., Palumbo,M., Thompson,W., Fickett,J.W. and Lawrence,C.E. (2000) Human–mouse genome comparisons to locate regulatory sites. *Nat. Genet.*, **26**, 225–228.
7. Lenhard,B., Sandelin,A., Mendoza,L., Engstrom,P., Jareborg,N. and Wasserman,W.W. (2003) Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.*, **2**, 13.
8. Loots,G.G., Ovcharenko,I., Pachter,L., Dubchak,I. and Rubin,E.M. (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.*, **12**, 832–839.
9. Hardison,R.C., Oeltjen,J. and Miller,W. (1997) Long human–mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.*, **7**, 959–966.
10. Flint,J., Tufarelli,C., Peden,J., Clark,K., Daniels,R.J., Hardison,R., Miller,W., Philipsen,S., Tan-Un,K.C., McMorrow,T. *et al.* (2001) Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the alpha globin cluster. *Hum. Mol. Genet.*, **10**, 371–382.
11. 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.
12. Aerts,S., Van Loo,P., Thijs,G., Moreau,Y. and De Moor,B. (2003) Computational detection of *cis*-regulatory modules. *Bioinformatics*, **19**(Suppl. 2), II5–II14.
13. Aerts,S., Thijs,G., Coessens,B., Staes,M., Moreau,Y. and De Moor,B. (2003) Toucan: deciphering the *cis*-regulatory logic of coregulated genes. *Nucleic Acids Res.*, **31**, 1753–1764.
14. McNeil,D.R. (1967) Efficiency loss due to grouping in distribution-free tests. *J. Am. Statist. Assoc.*, **62**, 954–965.
15. Ramaswamy,S., Nakamura,N., Sansal,I., Bergeron,L. and Sellers,W.R. (2002) A novel mechanism of gene regulation and tumor suppression by the transcription factor FKHR. *Cancer Cell*, **2**, 81–91.
16. Van Antwerp,D.J., Martin,S.J., Verma,I.M. and Green,D.R. (1998) Inhibition of TNF-induced apoptosis by NF-kappa B. *Trends Cell Biol.*, **8**, 107–111.
17. Zeng,H., Carlson,A.Q., Guo,Y., Yu,Y., Collier-Hyams,L.S., Madara,J.L., Gewirtz,A.T. and Neish,A.S. (2003) Flagellin is the major proinflammatory determinant of enteropathogenic Salmonella. *J. Immunol.*, **171**, 3668–3674.
18. Hinz,M., Lemke,P., Anagnostopoulos,I., Hacker,C., Krappmann,D., Mathas,S., Dorken,B., Zenke,M., Stein,H. and Scheidereit,C. (2002) Nuclear factor kappaB-dependent gene expression profiling of Hodgkin's disease tumor cells, pathogenetic significance, and link to constitutive signal transducer and activator of transcription 5a activity. *J. Exp. Med.*, **196**, 605–617.
19. Lamb,J., Ramaswamy,S., Ford,H.L., Contreras,B., Martinez,R.V., Kittrell,F.S., Zahnow,C.A., Patterson,N., Golub,T.R. and Ewen,M.E. (2003) A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. *Cell*, **114**, 323–334.
20. Dhanasekaran,S.M., Barrette,T.R., Ghosh,D., Shah,R., Varambally,S., Kurachi,K., Pienta,K.J., Rubin,M.A. and Chinnaiyan,A.M. (2001) Delineation of prognostic biomarkers in prostate cancer. *Nature*, **412**, 822–826.
21. Luo,J., Duggan,D.J., Chen,Y., Sauvageot,J., Ewing,C.M., Bittner,M.L., Trent,J.M. and Isaacs,W.B. (2001) Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res.*, **61**, 4683–4688.
22. Singh,D., Febbo,P.G., Ross,K., Jackson,D.G., Manola,J., Ladd,C., Tamayo,P., Renshaw,A.A., D'Amico,A.V., Richie,J.P. *et al.* (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.
23. Rhodes,D.R., Barrette,T.R., Rubin,M.A., Ghosh,D. and Chinnaiyan,A.M. (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res.*, **62**, 4427–4433.
24. Luo,J., Dunn,T., Ewing,C., Sauvageot,J., Chen,Y., Trent,J. and Isaacs,W. (2002) Gene expression signature of benign prostatic hyperplasia revealed by cDNA microarray analysis. *Prostate*, **51**, 189–200.
25. Ernst,T., Hergenhahn,M., Kenzelmann,M., Cohen,C.D., Bonrouhi,M., Weninger,A., Klaren,R., Grone,E.F., Wiesel,M., Gudemann,C. *et al.* (2002) Decrease and gain of gene expression are equally discriminatory markers for prostate carcinoma: a gene expression analysis on total and microdissected prostate tissue. *Am. J. Pathol.*, **160**, 2169–2180.
26. Foucher,I., Volovitch,M., Frain,M., Kim,J.J., Souberbielle,J.C., Gan,L., Unterman,T.G., Prochiantz,A. and Trembleau,A. (2002) Hoxa5 overexpression correlates with IGFBP1 upregulation and postnatal dwarfism: evidence for an interaction between Hoxa5 and Forkhead box transcription factors. *Development*, **129**, 4065–4074.
27. Denson,L.A., McClure,M.H., Bogue,C.W., Karpen,S.J. and Jacobs,H.C. (2000) HNF3beta and GATA-4 transactivate the liver-enriched homeobox gene, Hex. *Gene*, **246**, 311–320.
28. Zhong,H., Agani,F., Baccala,A.A., Laughner,E., Rioseco-Camacho,N., Isaacs,W.B., Simons,J.W. and Semenza,G.L. (1998) Increased expression of hypoxia inducible factor-1alpha in rat and human prostate cancer. *Cancer Res.*, **58**, 5280–5284.
29. Zhong,H., Chiles,K., Feldser,D., Laughner,E., Hanrahan,C., Georgescu,M.M., Simons,J.W. and Semenza,G.L. (2000) Modulation of hypoxia-inducible factor 1alpha expression by the epidermal growth factor/phosphatidylinositol 3-kinase/PTEN/AKT/FRAP pathway in

human prostate cancer cells: implications for tumor angiogenesis and therapeutics. *Cancer Res.*, **60**, 1541–1545.

30. Mabjeesh,N.J., Willard,M.T., Frederickson,C.E., Zhong,H. and Simons,J.W. (2003) Androgens stimulate hypoxia-inducible factor 1 activation via autocrine loop of tyrosine kinase receptor/phosphatidylinositol 3′-kinase/protein kinase B in prostate cancer cells. *Clin. Cancer Res.*, **9**, 2416–2425.
31. Jegga,A.G., Sherwood,S.P., Carman,J.W., Pinski,A.T., Phillips,J.L., Pestian,J.P. and Aronow,B.J. (2002) Detection and visualization of compositionally similar *cis*-regulatory element clusters in orthologous and coordinately controlled genes. *Genome Res.*, **12**, 1408–1417.
32. Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci., USA*, **98**, 5116–5121.