BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 6 2005, pages 803–810
doi:10.1093/bioinformatics/bti038

Databases and ontologies

Expression-based monitoring of transcription factor activity: the TELiS database

Steve W. Cole[1,4,5,6,*], Weihong Yan[2], Zoran Galic[1,4], Jesusa Arevalo[1] and Jerome A. Zack[1,3,4,5,6]

[1]UCLA Department of Medicine, [2]Department of Chemistry and Biochemistry, [3]Department of Microbiology, Immunology, and Molecular Genetics, [4]The UCLA AIDS Institute, [5]The Jonsson Comprehensive Cancer Center, and [6]The UCLA Molecular Biology Institute, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095–1678

Received on June 19, 2004; revised on August 31, 2004; accepted on September 7, 2004
Advance Access publication September 16, 2004

ABSTRACT

Motivation: In microarray studies it is often of interest to identify upstream transcription control pathways mediating observed changes in gene expression. The Transcription Element Listening System (TELiS) combines sequence-based analysis of gene regulatory regions with statistical prevalence analyses to identify transcription-factor binding motifs (TFBMs) that are over-represented among the promoters of up- or down-regulated genes. Efficiency is maximized by decomposing the problem into two steps: (1) a priori compilation of prevalence matrices specifying the number of putative binding sites for a variety of transcription factors in promoters from all genes assayed by a given microarray, and (2) real-time statistical analysis of pre-compiled prevalence matrices to identify TFBMs that are over- or under-represented in promoters of differentially expressed genes. The interlocking JAVA applications namely, PromoterScan and PromoterStats carry out these tasks, and together constitute the TELiS database for reverse inference of transcription factor activity.

Results: In two validation studies, TELiS accurately detected *in vivo* activation of NF-κB and the Type I interferon system by HIV-1 infection and pharmacologic activation of the glucocorticoid receptor in peripheral blood mononuclear cells. The population-based statistical inference underlying TELiS out-performed conventional statistical tests in analytic sensitivity, with parametric studies demonstrating accurate identification of transcription factor activity from as few as 20 differentially expressed genes. TELiS thus provides a simple, rapid and sensitive tool for identifying transcription control pathways mediating observed gene expression dynamics.

Availability: http://www.telis.ucla.edu

Contact: coles@ucla.edu

INTRODUCTION

It is now possible to monitor the transcriptional activity of an entire genome using massively parallel measurement technologies such as DNA microarrays or Serial Analysis of Gene Expression (SAGE) (Schena *et al*., 1995; Velculescu *et al*., 1995; Lockhart *et al*., 1996). Once global changes in gene expression have been defined, it is often of interest to identify the transcription control pathways mediating those dynamics. This article presents a sensitive and efficient computational strategy for monitoring the activity of multiple biological signaling pathways via their impact on the expression of genes bearing known transcription factor binding motifs (TFBMs) in their upstream regulatory regions.

Biological signals are transduced through a variety of receptor-mediated signaling pathways that converge on a small set of biochemical reactions modulating gene expression (Carey and Smale, 2000). Chief among these is a system of transcription factors that bind to DNA in a sequence-specific manner and recruit generic transcriptional machinery to a gene's core promoter (Mitchell and Tjian, 1989; Pabo and Sauer, 1992; Smale, 2001). Each transcription factor binds to a characteristic DNA motif such as GGGGCGGGG for *Sp1* or TGACGTCA for *CREB* (Letovsky and Dynan, 1989; Hill and Treisman, 1995). This basic relationship between nucleotide sequence and transcription factor binding permits inferences about which signaling pathways are likely to modulate a gene's expression based on the sequence of its promoter (Wingender *et al*., 1996). The cascading relationships among extracellular events, receptor-mediated signal transduction, transcription factor activation and genome regulation constitute a directed information flow by which cells adapt to environmental conditions. The advent of genome-wide expression monitoring provides an opportunity to reverse this causal sequence and infer upstream signaling dynamics from changes in global gene expression. At the most immediate level, it should be possible to identify the specific transcription factors mediating observed changes in gene expression based on the prevalence of their characteristic TFBMs in the promoters of co-regulated genes. An extensive body of research linking transcription factor activation to upstream second messenger systems and extracellular ligand/receptor networks (Hill and Treisman, 1995) also provides an opportunity for more distal inferences about the extracellular conditions modulating cellular activity [e.g. ambient proinflammatory cytokines activating nuclear factor- κB (NF-κB)] (Ghosh *et al*., 1998). Both the proximal inference of transcription factor activity and the distal inference of signal transduction depend on the ability to detect TFBMs that are over- or under-represented in the promoters of co-regulated genes relative to the genome as a whole. Genome-wide promoter analyses have recently been used to identify novel TFBMs (Roth *et al*., 1998; Spellman *et al*., 1998; van Helden *et al*., 1998; Hertz and Stormo,

*To whom correspondence should be addressed.

1999; Wagner, 1999; Wolfsberg *et al*., 1999; Bussemaker *et al*., 2000, 2001; Holmes and Bruno, 2000; Chiang *et al*., 2001; Liu *et al*., 2001; Ohler and Niemann, 2001). We focus on the distinct problem of surveying known TFBMs to identify the specific factors driving observed changes in gene expression.

Reverse inference of transcription factor activity would seem to be a straightforward problem, but several difficulties have hampered its widespread utilization. One obstacle is the intensive computation required to retrieve and scan large numbers of promoters for sequence homology. This effectively restricts reverse-inference analyses to users with strong bioinformatic skills and computational resources. A second difficulty is the lack of an efficient analytic framework for evaluating the statistical significance of variations in TFBM prevalence. Development of a valid statistical approach has been complicated by the combinatorial nature of gene regulation and the poor signal-to-noise characteristics of genome-wide expression assays. Most genes are regulated through the coordinated actions of multiple transcription factors, so the presence of a single TFBM in a gene's promoter does not guarantee that it will be expressed even if its cognate transcription factor is activated (Mitchell and Tjian, 1989; Wagner, 1999; Carey and Smale, 2000; Holmes and Bruno, 2000; Chiang *et al*., 2001). Conversely, the absence of a TFBM for a given factor does not ensure the absence of regulation because many transcriptional dynamics are mediated indirectly by secondary waves of transcription factor activity (e.g. factor A induces the expression of factor B, and factor B subsequently activates promoters bearing no consensus binding site for factor A). As a result, the presence of a TFBM is only loosely linked to the array of genes regulated by an active transcription factor. This problem is compounded by the fact that current analyses can severely underestimate the number of genes showing true differential expression (Cole *et al*., 2003). All these dynamics effectively contaminate the group of 'unregulated control' promoters with genes that should actually be assigned to the 'differentially expressed' subset, and vice versa. Such cross-contamination is known as the 'errors in variables' problem in the statistical literature, and it can profoundly degrade analytic accuracy (Miller, 1986). As a result, it is risky to rely on the results of reverse-inference analyses unless they can be shown to perform accurately in validation studies.

We developed the Transcription Element Listening system (TELiS) as a database-driven solution to the problems outlined above. This article describes the analytic strategy underlying TELiS and reports validation studies showing that it can accurately detect transcription factor activation under well-defined experimental conditions and amidst noisy *in vivo* pathology. We also present data on the comparative speed and accuracy of TELiS, with particular emphasis on alternative statistical strategies and methods for optimizing analytic sensitivity. Results show that population-based statistical inference can be coupled with genome-wide assessment of TFBM prevalence to provide accurate 'reverse inference' of transcription factor activity.

SYSTEM AND METHODS

The two major obstacles impeding reverse-inference analyses and addressed by TELiS. (1) To speed the inference process and make it available to biologists with no bioinformatic background, the most computationally intensive aspects of the problem are 'pre-solved' by generating a set of TFBM prevalence matrices. Each matrix records the number of putative binding sites for an array of transcription factors in promoters from all genes represented on a commonly used microarray (e.g. Affymetrix HuGene-FL, U95A, U133A, Mu11K, U74A, Mouse 430, U34A and Rat 230 high density oligonucleotide arrays). Such prevalence matrices can take weeks to compile depending upon the size of a promoter (bases analyzed), the number of genes analyzed (~20 000–30 000) and the number of TFBMs analyzed. However, once a matrix is available, it takes only seconds for analytic procedures to identify TFBMs that are over- or under-represented in promoters of an arbitrary set of differentially expressed genes. (2) To avoid the inferential difficulties associated with 'errors in variables', the statistical analysis is approached as a single-sample inference problem with known population parameters. In conventional statistical analyses such as the *t*-test, errors in variables lead to inaccurate estimates of the true sampling variability of TFBM prevalence in the population of all promoters (Miller, 1986). This undermines the accuracy of *p*-values testing differential representation because the standard error of that difference is estimated as a function of the inferred population sampling variance (Miller, 1986). However, a single-sample *z*-test does not require any sample-based inferences about TFBM variability because that parameter is already known at the population level (i.e. the mean and standard deviation of the number of TFBMs in each promoter is pre-compiled for all genes assayed by a given microarray). As a result, a population-based approach could potentially detect perturbations in TFBM prevalence with greater accuracy than conventional sample-based approaches such as the *t*-test.

Validation studies

To evaluate the performance of TELiS, we analyzed two datasets in which specific signaling pathways were known to be activated. The first test involved focal stimulation of the glucocorticoid receptor by exogenous hydrocortisone (cortisol). Peripheral blood mononuclear cells (2×10^7) were isolated by Ficoll density gradient centrifugation and cultured overnight in 10 ml of RPMI supplemented with 100 U/ml penicillin, 100 μg/ml streptomycin, 10% autologous donor serum, and either 1 μM hydrocortisone (Sigma) or an equivalent volume of medium. Twelve-hours later, total RNA was harvested (Qiagen RNEasy, Valencia CA), treated with DNAse (Qiagen), converted to fluorescent cRNA and hybridized to Affymetrix U133A high-density oligonucleotide arrays in the UCLA Gene Expression Core according to the manufacturer's protocol (Affymetrix, Santa Clara CA). Scanned images were analyzed for up-regulated genes using Affymetrix Microarray Suite v5 software with default analysis parameters (paired comparison of each donor's hydrocortisone-treated cells with parallel untreated cells). This experiment was repeated for three independent donors, and genes significantly up-regulated in all three replicates were subject to reverse-inference analysis by TELiS. Controlled pharmacologic stimulation ensured that a single transcription factor was initially activated.

In addition to the highly controlled glucocorticoid model, we also examined the capacity of TELiS to detect inflammatory signaling in a noisy *in vivo* pathology model. HIV-1 infection activates multiple proinflammatory transcription factors in lymphoid cells, including interferon response factors and NF-κB (Roulston *et al*., 1995; Corbeil *et al*., 2001; Keir *et al*., 2002; Miller *et al*., 2003; Yonezawa *et al*., 2003). To determine whether TELiS could identify such activation *in vivo*, we assessed differential gene expression in human fetal thymocytes after they had been stably engrafted in a SCID mouse host for 6 weeks and infected for 3 more weeks with the NL4-3 strain of HIV-1 or a mock infected control (Aldrovandi *et al*., 1993; Cole *et al*., 2003). Differentially expressed genes were identified as described above using Affymetrix HuGene-FL high-density oligonucleotide arrays and Microarray Suite v5 paired analysis of HIV versus mock-infected cells from the same thymic tissue donor (default analysis parameters). Genes identified as significantly increased in each of the two replicate experiments were subject to reverse-inference analysis by TELiS.

ALGORITHM AND IMPLEMENTATION

Transcription element listening system consists of four interacting components outlined in Figure 1. A JAVA application called PromoterScan assesses the incidence of TFBMs in promoters for all

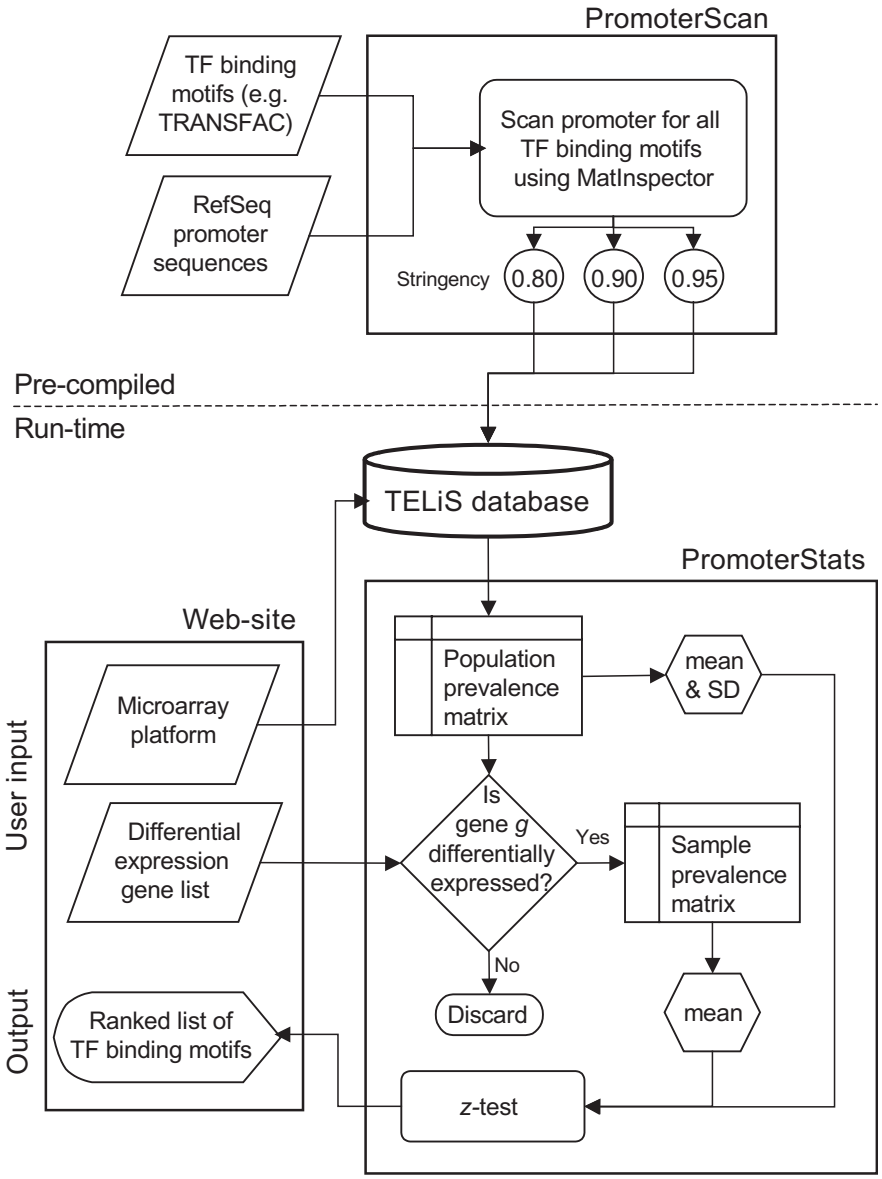

Fig. 1. Structure of TELiS. Four interlocking components provide rapid identification of TFBMs that are over- or under-represented in promoters of differentially expressed genes. PromoterScan establishes a set of sampling frames corresponding to specific microarray platforms. Promoters for each gene are scanned for an array of TFBMs from the TRANSFAC database, and the number of sites in each promoter is stored in the TELiS database as a population prevalence matrix (P promoters $\times$ T TFBMs). In response to a user request, the TELiS website passes a list of differentially expressed genes and the microarray platform used to detect them to the JAVA servlet, PromoterStats. PromoterStats retrieves the appropriate population prevalence matrix and generates a sample prevalence matrix containing TFBM frequencies for the subset of differentially expressed genes. For each TFBM, representation in the differentially expressing promoters is compared to the sampling frame as a whole by z-test (or a binomial test for binary present/absent data). Test statistics, p-values and prevalence information are returned to the user via the web interface to identify transcription factors that drive observed expression dynamics.

genes in a genome and stores the resulting 'population prevalence matrix' in the TELiS database. Users interact with a World Wide Web interface (http://www.telis.ucla.edu) to supply a list of differentially expressed genes and specify the sampling frame in which they were identified (e.g. the microarray used). Based on that input, a JAVA servlet called PromoterStats retrieves TFBM prevalence data from the TELiS database and computes statistical summaries of over- or under-representation in regulated promoters relative to the basal prevalence of TFBMs across the entire sampling frame. Results are ranked according to statistical significance and returned to the user in real time. The web page also has a utility for downloading raw data on TFBM prevalence in promoters of specified genes. Details are provided below.

PromoterScan and the TELiS database PromoterScan retrieves promoter sequences for all transcripts in a specified genome, scans each promoter for a fixed array of TFBMs and stores the number of sites identified in the TELiS database. Nucleotide sequences come

from the NCBI RefSeq database (Pruitt and Maglott, 2001) (human transcripts as of September 2003, and mouse and rat transcripts as of December 2003), and promoters are defined as nucleotide sequences spanning positions −300 to +0, −600 to +0 or −1000 to +200 relative to the RefSeq transcription start site (TSS). Each promoter is scanned with 192 nucleotide position matrices representing all vertebrate TFBMs in the anonymous FTP release of TRANSFAC v3.2 (V$ matrices) (Wingender *et al.*, 1996). Scans utilize the TRANSFAC MatInspector algorithm (Quandt *et al.*, 1995) at mat_sim stringencies of 0.80, 0.90 and 0.95. Prevalence values for transcripts with multiple putative start sites are averaged to provide a single value for each gene. Results are stored in MySQL 4.0 as separate prevalence tables for each combination of species (human, mouse, rat), promoter size (300, 600, 1200 bases), and scan stringency (0.80, 0.90, 0.95), with HGNC Gene Symbols serving as unique keys. The database also contains parallel tables from PromoterScan analyses treating TFBM incidence as a binary variable (present versus not present in each promoter).

PromoterStats and the TELiS website The TELiS website and its associated servlets are housed on an Apache Tomcat 4.0 server with JDBC connections to the TELiS database. For reverse-inference analyses, the website collects a list of D differentially expressed genes and an indication of the sampling frame in which those changes were measured (the specific microarray used). This information is passed to PromoterStats, which then generates two data matrices; a 'population prevalence matrix' specifying the number of binding sites for each TFBM in promoters of all P genes in the sampling frame, and a 'sample prevalence matrix' indicating the number of TFBMs in promoters of the D differentially expressed genes. The mean prevalence of each TFBM in the sample prevalence matrix is computed and tested for over- or under-representation relative to the population mean prevalence using a single-sample z-test (Kanji, 1999). For TFBM t, the test statistic z_t is defined as:

$$z_t = (\bar{x}_t - \mu_t) D^{1/2} / \sigma_{xt}, \quad (1)$$

where $\bar{x}t$ is the mean number of detected binding sites for transcription factor t among the D promoters in the sample prevalence matrix, μ_t is the mean number of sites for factor t among the P promoters in the population prevalence matrix, and $\sigma_{\bar{x}t}$ is the standard deviation of the number of sites for factor t in the population prevalence matrix. Positive values of z indicate over-representation of TFBM t in promoters of differentially expressed genes and negative values indicate under-representation (a possible inhibitory effect). Each z-value generates a two-tailed p-value gauging statistical significance. Binary (present/not) data are analyzed in a standard binomial test, with p-values derived from the probability of observing S_t or more positive promoters in a sample of D Bernoulli trials, each of which has a probability of positive outcome equal to the prevalence of TFBM t in the sampling frame as a whole (Kanji, 1999). Statistically significant under-representation is assessed by the probability of observing S_t or fewer positive promoters, with S indicating the number of promoters in the differentially expressing subset that contain at least one instance of TFBM t. Population[1] prevalence matrices are derived from genes listed in microarray manufacturers' annotation files (e.g. HG-U133A_annot.csv at http://www.affymetrix.com/analysis/download_center.affx). It could be argued that the most appropriate reference population for analysis is the set of genes found to be expressed in the experimental samples, rather than the entire population of transcripts assayed by the microarray. The web interface includes a form for submitting both a list of genes constituting the sampling frame and a differentially expressed subset: http://www.med.ucla.edu:8080/telis/TELiSDifferentialExpressionCustomSamplingFrame.htm. A servlet is also provided to analyze TFBM frequencies relative to a Poisson-distributed population with intensity parameter $= D\mu_t$: http://www.med.ucla.edu:8080/telis/TELiSDifferentialExpressionPoisson.htm. However, the z-test is recommended instead because population TFBM frequencies do not generally follow a Poisson distribution (detailed in the section titled, 'Performance relative to alternative approaches').

In addition to individual statistical results for each TFBM, PromoterStats also estimates the false discovery rate (FDR) (Benjamini and Hochberg, 1995) across the entire set of significant results. The FDR gives the fraction of significant results that are likely due to chance and is estimated as,

$$FDR_p = T\Phi_p / n_p, \quad (2)$$

where T is the number of TFBMs surveyed, Φ_p is the expected proportion of false positive errors at a specified significance level (e.g. $p < 0.01$) and n_p is the number of observed tests significant at that level. Φ_p is estimated by Monte Carlo analysis of significant results in 100 000 randomly sampled gene lists of size D drawn from the same sampling frame. In addition to FDR estimates for the default significance threshold of $p < 0.01$, p-value thresholds that control the FDR at 10, 20, 30 and 40%, are derived from regression analysis of the relationship between estimated FDR and p-values ranging between 0.03 and 0.0001.

[1]The appropriate population is not the set of all human genes, but the set of all transcripts that could possibly be observed to change in a given experiment (e.g. all genes on the microarray used). This is a significant distinction because genes in the sampling frame are not necessarily representative of the genome as a whole in terms of TFBM prevalence. For example, the genes represented on the Affymetrix U133A GeneChip have approximately half as many glucocorticoid response elements in their promoters (mean = 0.083 per promoter, SD = 2.88) as do the entire set of sequenced human genes (mean = 0.151, SD = 3.88; difference $p = 0.019$ by a single-sample z-test). Inappropriate use of the genome-wide sampling frame could thus prevent detection of glucocorticoid signaling in a GeneChip experiment even if glucocorticoid response elements were 2-fold over-represented among promoters of differentially expressed genes.

It could be argued that the most appropriate population is the subset of assayed transcripts that are expressed in one or more of the experimental samples. However, the use of an 'expressed gene' sampling frame consistently weakened the detection of known signal transduction activity in both the glucocorticoid and HIV validation studies. All target TFBMs were still detected at statistically significant levels, but the over-representation ratios and p-values were noticeably attenuated. The sensitivity loss stemmed mainly from inaccuracies in the definition of the expressed population that resulted from negative biases in transcript 'present' calls by Affymetrix Microarray Suite v5. It is unclear whether similar problems might exist for other low-level expression analyses, but conservatism suggests using 'microarray population' sampling frames to avoid such biases. However, the option of restricting the sampling frame remains available at: http://www.med.ucla.edu:8080/telis/TELiSDifferentialExpressionCustomSamplingFrame.htm.

Table 1. Performance of alternative statistical analyses in detecting over-representation of TFBMs

Experiment (TFBM matrix)	Parameter[a]	Statistical test		
		2-sample *t*-test[b]	1-sample *t*-test	*z*-test
Glucocorticoid (V$GRE_C)	Mean difference	0.0045	0.0044	0.0044
	SD	0.0736/0.0301[c]	0.0736	0.0320
	SE of difference	0.0038	0.0038	0.0017
	Test statistic	1.18	1.15	2.65
	p-value	0.240	0.251	0.008
HIV (V$NFKB_C)	Mean difference	0.0257	0.0252	0.0252
	SD	0.1909/.1082[c]	0.1909	0.1108
	SE of difference	0.0166	0.0166	0.0196
	Test statistic	1.55	1.52	2.62
	p-value	0.124	0.131	0.009
HIV (V$IRF2_01)	Mean difference	0.0177	0.0173	0.0173
	SD	0.1490/.0718[c]	0.1490	0.0743
	SE of difference	0.0130	0.0129	0.0064
	Test statistic	1.36	1.34	2.69
	p-value	0.173	0.182	0.007

[a]Mean difference = mean number of TFBMs in promoters of up-regulated genes − unregulated genes (2-sample test) or − population mean prevalence (1-sample tests); SD = estimated population standard deviation in number of TFBMs per promoter; SE of difference = estimated standard error of mean difference, test statistic = *z*-value or *t*-value; *p*-value = two-tailed *p*-value associated with test statistic.

[b]Standard deviation of TFBM prevalence in regulated promoters significantly exceeded that of unregulated promoters in all cases ($p < 0.001$ by Levene's test). All 2-sample *t*-tests therefore use the Welch formula for unequal variances (Miller, 1986).

[c]Standard deviation of TFBM prevalence in the group of regulated and unregulated promoters, respectively. Standard error of difference for 2-sample *t*-test is a sample-size weighted function of both SDs (Miller, 1986).

Validation

To assess the accuracy of reverse inference by TELiS, we analyzed data from a controlled experimental system involving pharmacologic stimulation of the glucocorticoid receptor in peripheral blood mononuclear cells. Cells were cultured for 12 h in the presence of 1 μM hydrocortisone or vehicle control and mRNA expression was surveyed by Affymetrix U133A high-density oligonucleotide arrays. Among 22 215 assayed transcripts, 304 showed consistent up-regulation across three replicate experiments. TELiS revealed significant over-representation of glucocorticoid response elements among promoters of up-regulated genes (V$GRE_C: 6.7-fold increase relative to unregulated genes, $z = 3.12$, $p = 0.0018$). Ten other TFBMs were also identified as over-represented, and eight of those corresponded to transcription factors known to interact with or be regulated by glucocorticoid receptors (Oct family members, AP1 family members, SRE, Elk1, CDP and E2F) (Karagianni and Tsawdaroglou, 1994; Rhee *et al*., 1994; Pearce *et al*., 1998; Prefontaine *et al*., 1998; Miyazaki *et al*., 2000; Zhu and Dudley, 2002). In the context of the total 192 TFBMs surveyed, these results give a specificity >90% and a positive predictive value of 82%. Thus, TELiS can accurately detect focal transcription factor activation under carefully controlled conditions, even in a background of low basal gene expression (cells were not stimulated by any mitogens).

To evaluate performance in a noisier *in vivo* environment, we tested the ability of TELiS to detect activation of proinflammatory transcription factors during HIV-1 infection of human thymocytes in a thy/liv SCID-hu mouse model. Three weeks following inoculation of implanted human thymic tissue with HIV-1 or a vehicle control, viral pathology was documented by PCR detection of HIV-1 provirus and depletion of CD4+/CD8+thymocytes relative to mock-infected cells (data not shown). mRNA was harvested in parallel and assayed using Affymetrix HuGene-FL high-density oligonucleotide arrays. Of the 7070 assayed transcripts, 105 showed significant up-regulation in each of the two replicates. Among the promoters of those genes, TELiS identified a substantial over-representation of binding sites for interferon response factor 1 (V$IRF1_01: 8.4-fold increase, $z = 6.44$, $p < 10^{-10}$) and interferon response factor 2 (V$IRF2_01: 5.5-fold increase, $z = 3.21$, $p = 0.0013$), as well as the consensus interferon-stimulated response element (V$ISRE_01: 18.2-fold increase, $z = 10.07$, $p < 10^{-10}$) and two matrices defining consensus NF-κB response elements (V$NFKAPPAB_01: 2.3-fold increase, $z = 2.73$, $p = 0.0064$; V$NFKB_Q6: 4.5-fold increase, $z = 4.35$, $p = 0.0000134$). Previous experimental studies have shown that each of these signaling pathways is in fact activated during HIV-1 infection (Roulston *et al*., 1995; Corbeil *et al*., 2001; Keir *et al*., 2002; Miller *et al*., 2003; Yonezawa *et al*., 2003), and these motifs constituted four of the top five hits identified in the over-representation analysis. In contrast, TELiS failed to indicate significant over-representation of TFBMs for transcription factors known not to be induced by HIV-1 infection (e.g. Oct1, V$OCT1_Q6: 0.930-fold change, $z = -0.11$, $p = 0.918$; Sp1, V$SP1_Q6: 0.776-fold change, $z = -1.33$, $p = 0.185$). Among 192 TFBMs surveyed, 184 were found not to be significantly over-represented, for a specificity exceeding 90%. Thus, even amidst noisy *in vivo* pathology, TELiS can accurately detect physiologically relevant transcription factor activity.

Performance relative to alternative approaches

To evaluate the statistical approach underlying TELiS, we compared results of its population-based z-test [Equation (1)], with the findings produced by a single-sample or two-sample t-test (the latter equivalent to a one-way analysis of variance) (Miller, 1986). The single-sample t-test is similar to the z-test in comparing the prevalence of TFBMs in up-regulated promoters with that of the sampling frame as a whole, but the t-test treats the population sampling variance as an unknown quantity that must be inferred from the sample data (i.e. from the D differentially expressing promoters rather than the P promoters in the sampling frame). As shown in Table 1, the single-sample t-test yielded considerably larger p-values than the z-test, and was therefore unable to identify glucocorticoid signaling in the glucocorticoid stimulation study or IRF and NF-κB activation in the HIV study. A 2-sample t-test comparing TFBM prevalence in up-regulated versus unregulated promoters also failed to detect each of those signals. The t-test's poor sensitivity stemmed from the fact that the sampling variability of TFBM prevalence in the D differentially expressing promoters was not representative of that in the population as a whole (in Table 1, compare SD and SE values for 1- and 2-sample t-tests with the corresponding value for the z-test). Sample standard deviations over-estimated their population values by 2–4-fold, leading to inflated standard errors and loss of sensitivity (i.e. increased p-values).

Frequency data are often analyzed under the assumption of a Poisson distribution (Santner and Duffy, 1989), so we compared the performance of a single-sample Poisson analysis with that of the z-test. In both the glucocorticoid and HIV studies, single-sample Poisson tests identified more TFBMs as being significantly over-represented. This difference appears to stem from increased false positive error by the Poisson analysis rather than increased sensitivity. In Monte Carlo studies carried out to estimate FDR Φ_p values, Poisson analyses consistently yielded Type I error rates exceeding the nominal p-value (e.g. Table 2). These errors stemmed from the fact that observed TFBM frequency distributions showed greater variance than assumed by the Poisson distribution (variance $= \mu_t$), leading Poisson analyses to underestimate the true sampling variability. For example, in the frequency data analyzed in Table 2, 98% of TFBMs showed a population variance greater than μ_t, with 96% showing significant over-dispersion at $p < 0.01$ [Fisher's χ^2-test of Poisson fit (Santner and Duffy, 1989)]. Similar results emerged for all combinations of sample size, promoter length and scan stringency. In contrast, the z-test accurately controlled Type I errors in all Monte Carlo studies (Table 2) because it utilizes the empirically correct variance. The z-test is therefore recommended as the primary test of TFBM differential representation. However, for those who wish to use a Poisson-based analysis, the statistical output includes a comparison of the empirical and assumed variance to users can assess the Poisson approach for a particular TFBM.

Optimizing analytic sensitivity

The analyses reported above were based on default settings for TELiS: analysis of 300 bases upstream of the TSS using a MatInspector stringency of 0.90. These defaults were derived from a set of parametric studies examining the effects of alternative stringencies (0.80, 0.90, 0.95) and promoter lengths (300 600 or 1200 bases adjacent to the transcription start site). As summarized in Figure 2A, analyses of short promoter sequences (300 bases) with moderate stringency (0.90) generally provided optimal signal detection.

Table 2. False positive error rates for single-sample z-test and Poisson test[a]

Sample size	Nominal p-value					
	0.01		0.001		0.0001	
	z-test	Poisson	z-test	Poisson	z-test	Poisson
10	0.0145	0.0450	0.0041	0.0210	0.0021	0.0122
30	0.0130	0.0488	0.0030	0.0221	0.0014	0.0127
100	0.0111	0.0510	0.0022	0.0229	0.0008	0.0130
300	0.0099	0.0514	0.0015	0.0224	0.0004	0.0125
1000	0.0072	0.0469	0.0007	0.0201	0.0001	0.0111
3000	0.0029	0.0333	0.0001	0.0134	0.0000	0.0074

[a]Table entries give the fraction of 192 TFBMs identified as significantly over- or under-represented at each nominal p-value, averaged over 100 000 random samples (without replacement) of the size indicated in column 1. Displayed results are based on data from low-stringency (0.80) scanning of 1200-base human promoter sequences. Similar results emerged in analyses of data from other species, scan stringencies and promoter lengths.

Analyses using longer sequences or lower stringency produced poorer signal-to-noise ratios due to increased non-specific detection events. High-stringency analyses (0.95) produced enhanced signal-to-noise ratios, but sometimes yielded no results at all with short promoter sequences (e.g. IRF1 Figure 2B). These results suggest that long promoter sequences (1200 bases) should be utilized when high-stringency analyses are required. Power analyses (Figure 2C) showed that sample sizes of $D > 20$ differentially expressed genes were generally sufficient to yield statistically significant results.

DISCUSSION

Transcription Element Listening System combines real-time data on transcriptional dynamics with a stored database of genomic promoter characteristics to identify transcription factors driving global changes in gene expression. The validation studies reported above show that this approach can successfully detect transcription factor activation in both well-defined experimental systems and complex *in vivo* pathology. The core bioinformatic resource supporting these inferences is the TELiS database—a collection of sampling frames that store information on the prevalence of each TFBM in the promoters of all genes assayed by a given microarray. These sampling frames reduce solution times by pre-solving the most computationally intensive aspect of the reverse inference problem—scanning large nucleotide sequences for multiple TFBMs. They also establish the conceptual population required for the most sensitive statistical analysis—a z-test. The only input required from the user is a list of differentially expressed genes and the microarray platform used to find them. Given this combination of simple input, rapid results and sensitive detection, the TELiS database search tool should considerably increase the use of reverse-inference analyses to define the transcription control pathways driving gene expression dynamics.

A key advance for reverse inference is the development of an efficient statistical framework for detecting TFBMs that are over-represented among co-regulated genes. Conventional inferential statistics such as the t-test fail in this task because they attempt to estimate the sampling variability in the population of all genes from the sampling variability in the subset of differentially expressed genes. However, the prevalence of TFBMs varies by 2–4-fold more among activated promoters than it does in the population as a whole,

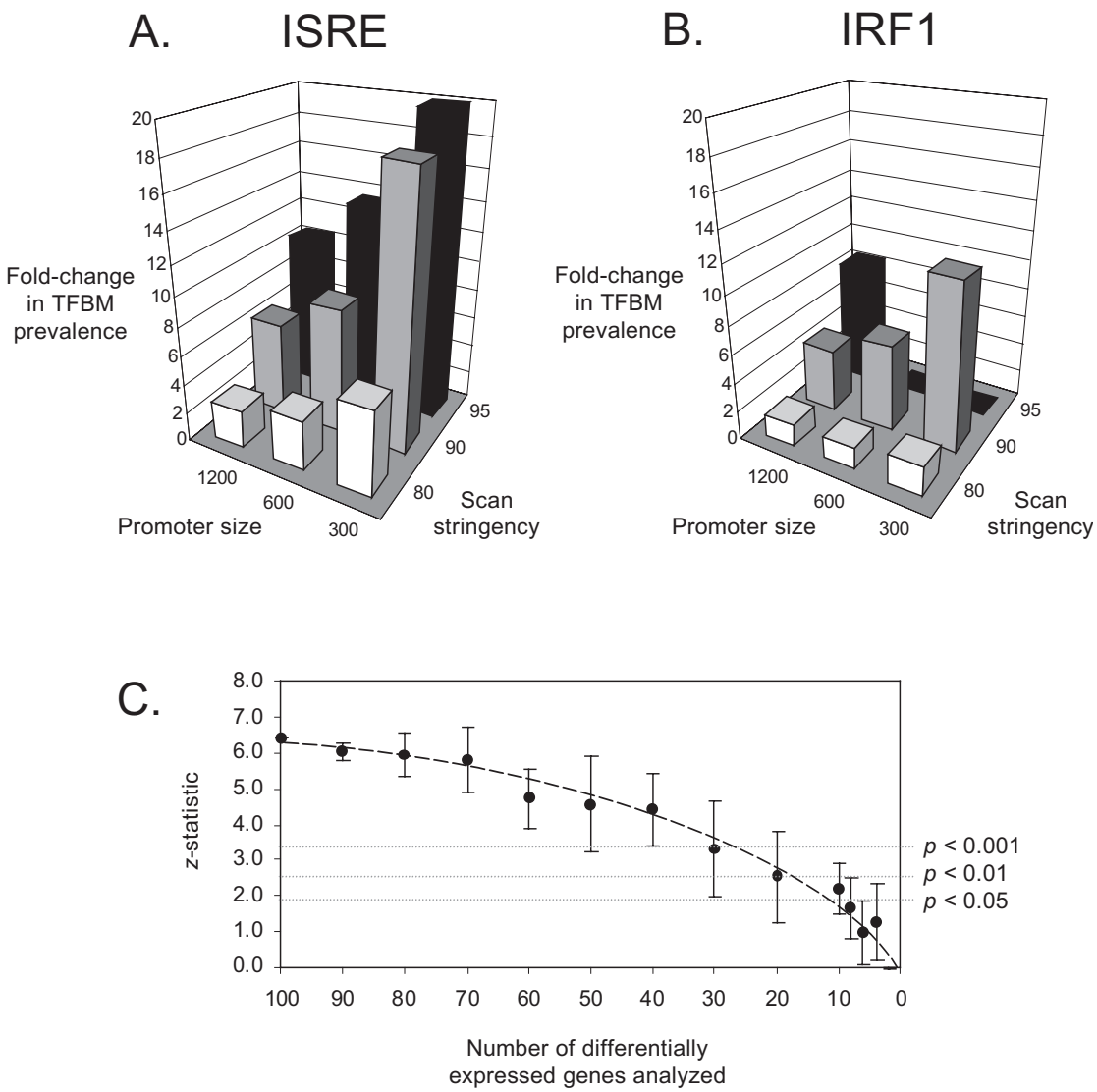

Fig. 2. Optimizing detection sensitivity. Promoter size and scan stringency were parametrically varied to identify optimal settings for detection of interferon responsive elements in promoters of 105 genes up-regulated in HIV-infected thymocytes. (**A**) For the low-stringency ISRE motif (matrix V$ISRE_01), signal-to-noise ratios consistently increased as stringency was elevated and promoter size decreased. (**B**) For the high stringency IRF1 matrix (V$IRF1_01), signal-to-noise ratios also increased as promoter sizes were reduced in low- and intermediate-stringency analyses (mat_sim = 0.80 and 0.90). However, high stringency analyses (0.95) failed to identify any IRF1 motifs in all but the longest promoter sequences (1200 bases). Similar results emerged from analyses of glucocorticoid response (V$GRE_C) in hydrocortisone-stimulated leukocytes (data not shown). (**C**) To define the number of genes required to detect over-represented TFBMs, random samples of varying size were drawn from the set of all genes over-expressed in the HIV study and analyzed using default parameters (300 bases, stringency 0.90). The resulting empirical power curve (defined by the mean ± standard error of resulting z-test statistics) indicated that at least 20 genes were required to yield consistently significant results. Similar results emerged from analyses of other inflammation-related motifs in the HIV study and glucocorticoid response elements in the hydrocortisone study (data not shown).

resulting in inflated p-values and failure to detect over-represented motifs even when they are known to exist. Population-based z-tests are generally more sensitive, but typically not feasible because the population mean and standard deviation are unknown (Miller, 1986). Fortunately, the TELiS database provides exactly the population parameters required to support a z-test because it is based on an exhaustive census of promoters. As a result, this key data resource fundamentally transforms the analytic approach to yield qualitative improvements in signal detection.

In object-oriented programming, a 'listener' passively monitors an ongoing process and activates itself in response to a predefined condition. TELiS represents each transcription factor as a listener that scans induced promoters for variations in the incidence of its signature TFBM. Listeners 'call out' their statistical confidence in their own differential representation, and a system-level referee aggregates those calls into a set of inferences about signaling pathways driving observed changes in gene expression. The validation studies reported here show that this approach can detect transcription factor activation

in cases where other statistical approaches fail. However, the statistical component of the analysis depends crucially on access to a full census of promoter sequences. TELiS is currently implemented for human, mouse and rat genomes assayed by Affymetrix GeneChips, and it can easily be extended to other genomes and assay systems such as SAGE or proteomic arrays by development of appropriate sampling frames. Some difficult problems remain to be addressed, including combinatorial effects of multiple transcription factors (Wagner, 1999; Carey and Smale, 2000; Holmes and Bruno, 2000; Chiang *et al*., 2001; Michelson, 2002) and more refined mapping of promoter elements. However, the development of a simple, fast and sensitive system for monitoring upstream transcription control pathways will help advance gene expression analysis beyond a descriptive mode to provide a causal understanding of genome dynamics.

ACKNOWLEDGEMENTS

This research was supported by the University of California Universitywide AIDS Research Program (K99-LA-030, CC02-LA-001), the Norman Cousins Center at UCLA, the James L. Pendelton Charitable Trust and the National Institute of Allergy and Infectious Diseases (AI33259, AI49135, A152737). We gratefully acknowledge the assistance of Greg Baran and Boris Sorkin in web deployment.

REFERENCES

Aldrovandi,G.M., Feuer,G., Gao,L., Kristeva,M., Chen,I.S.V., Jamieson,B. and Zach,J.A. (1993) HIV-1 infection of the SCID-hu mouse: an animal model for virus pathogenesis. *Nature*, **363**, 732–736.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B*, **57**, 289–300.

Bussemaker,H.J., Li,H. and Siggia,E.D. (2000) Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc. Natl Acad. Sci. USA*, **97**, 10096–100100.

Bussemaker,H.J., Li,H. and Siggia,E.D. (2001) Regulatory element detection using correlation with expression. *Nat. Genet.*, **27**, 167–171.

Carey,M. and Smale,S.T. (2000) *Transcriptional Regulation in Eukaryotes: concepts, Strategies, and Techniques*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Chiang,D.Y., Brown,P.O. and Eisen,M.B. (2001) Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles. *Bioinformatics*, **17**(Suppl. 1), S49–S55.

Cole,S.W., Galic,Z. and Zack,J.A. (2003) Controlling false-negative errors in microarray differential expression analysis: a PRIM approach. *Bioinformatics*, **19**, 1808–1816.

Corbeil,J., Sheeter,D., Genini,D., Rought,S., Leoni,L. Du,P., Ferguson,M., Masys,D.R., Welsh,J.B., Fink,J.L. *et al*. (2001) Temporal gene regulation during HIV-1 infection of human CD4+T cells. *Genome. Res.*, **11**, 1198–11204.

Ghosh,S., May,M.J. and Kopp,E.B. (1998) NF-kappa B and Rel proteins: evolutionarily conserved mediators of immune response. *Ann. Rev. Immunol.*, **16**, 225–260.

Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

Hill,C.S. and Treisman,R. (1995) Transcriptional regulation by extracellular signals: mechanisms and specificity. *Cell*, **80**, 199–211.

Holmes,I. and Bruno,W.J. (2000) Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Proc. Int. Conf. Intell Syst. Mol. Biol.*, **8**, 202–210.

Kanji,G.K. (1999) *100 Statistical Tests*. Sage, London.

Karagianni,N. and Tsawdaroglou,N. (1994) The c-fos serum response element (SRE) confers negative response to glucocorticoids. *Oncogene*, **9**, 2327–2334.

Keir,M.E., Stoddart,C.A., Linquist-Stepps,V., Moreno,M.E. and McCune,J.M. (2002) IFN-alpha secretion by type 2 predendritic cells up-regulates MHC class I in the HIV-1-infected thymus. *J. Immunol.*, **168**, 325–331.

Letovsky,J. and Dynan,W.S. (1989) Measurement of the binding of transcription factor Sp1 to a single GC box recognition sequence. *Nucleic Acids Res.*, **17**, 2639–2653.

Liu,X., Brutlag,D.L. and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, 127–138.

Lockhart,D., Dong,H., Byre,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittman,M., Wang,C., Kobayashi,M., Horton,H. and Brown,E.L. (1996) Expression monitoring by high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.

Michelson,A.M. (2002) Deciphering genetic regulatory codes: a challenge for functional genomics. *Proc. Natl Acad. Sci. USA*, **99**, 546–548.

Miller,E.D., Smith,J.A., Lichtinger,M., Wang,L. and Su,L. (2003) Activation of the signal transducer and activator of transcription 1 signaling pathway in thymocytes from HIV-1-infected human thymus. *Aids*, **17**, 1269–1277.

Miller,R.G. (1986) *Beyond ANOVA: Basics of Applied Statistics*. Wiley, New York.

Mitchell,P.J. and Tjian,R. (1989) Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*, **245**, 371–378.

Miyazaki,Y., Tsukazaki,T., Hirota,Y., Yonekura,A., Osaki,M., Shindo,H. and Yamashita,S. (2000) Dexamethasone inhibition of TGF beta-induced cell growth and type II collagen mRNA expression through ERK-integrated AP-1 activity in cultured rat articular chondrocytes. *Osteoarthritis Cartilage*, **8**, 378–385.

Ohler,U. and Niemann,H. (2001) Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet.*, **17**, 56–60.

Pabo,C.O. and Sauer,R.T. (1992) Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.*, **61**, 1053–1095.

Pearce,D., Matsui,W., Miner,J.N. and Yamamoto,K.R. (1998) Glucocorticoid receptor transcriptional activity determined by spacing of receptor and nonreceptor DNA sites. *J. Biol. Chem.*, **273**, 30081–30085.

Prefontaine,G.G., Lemieux,M.E., Giffin,W., Schild-Poulter,C., Pope,L., LaCasse,E., Walker,P. and Hache,R.J. (1998) Recruitment of octamer transcription factors to DNA by glucocorticoid receptor. *Mol. Cell Biol.*, **18**, 3416–3430.

Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucl. Acids Res.*, **29**, 137–140.

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.

Rhee,K., Ma,T. and Thompson,E.A. (1994) The macromolecular state of the transcription factor E2F and glucocorticoid regulation of c-myc transcription. *J. Biol. Chem.*, **269**, 17035–17042.

Roth,F.P., Hughes,J.D., and Estep,P.W. and Church,G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.

Roulston,A., Lin,R., Beauparlant,P., Wainberg,M.A. and Hiscott,J. (1995) Regulation of human immunodeficiency virus type 1 and cytokine gene expression in myeloid cells by NF-kappa B/Rel transcription factors. *Microbiol. Rev.*, **59**, 481–505.

Santner,T.J. and Duffy,D.E. (1989) *The Statistical Analysis of Discrete Data*. Springer-Verlag, New York.

Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene expression profiles in complementary DNA microarrays. *Science*, **270**, 467–470.

Smale,S.T. (2001) Core promoters: active contributors to combinatorial gene regulation. *Genes Dev.*, **15**, 2503–2508.

Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Bolstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

van Helden,J., Andre,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281**, 827–842.

Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.

Wagner,A. (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, **15**, 776–784.

Wingender,E., Dietze,P., Karas,H. and Knüppel,R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.

Wolfsberg,T.G., Gabrielian,A.E., Campbell,M.J., Cho,R.J., Spouge,J.L. and Landsman,D. (1999) Candidate regulatory sequence elements for cell cycle-dependent transcription in *Saccharomyces cerevisiae*. *Genome Res.*, **9**, 775–792.

Yonezawa,A., Morita,R., Takaori-Kondo,A., Kadowaki,N., Kitawaki,T., Hori,T. and Uchiyama,T. (2003) Natural alpha interferon-producing cells respond to human immunodeficiency virus type 1 with alpha interferon production and maturation into dendritic cells. *J. Virol.*, **77**, 3777–3784.

Zhu,Q. and Dudley,J.P. (2002) CDP binding to multiple sites in the mouse mammary tumor virus long terminal repeat suppresses basal and glucocorticoid-induced transcription. *J. Virol.*, **76**, 2168–2179.