CANCER GENOMICS & PROTEOMICS *4*: 301-308 (2007)

Review

Interpreting Microarray Data: Towards the Complete Bioinformatics Toolkit for Cancer

MICHAEL L. ROBERTS and STAVROS D. KOTTARIDIS

Regulon A.E., Afxentiou 7, Athens 11525, Greece

Abstract. *Functional genomics has been applied in the study of human malignancies since the inception of this field nearly a decade ago. Microarray analysis has been specifically used in an attempt to reclassify carcinomas at the molecular level, to aid in diagnosis/prognosis and to predict how various types of tumour respond to different therapeutic agents. Bioinformatics is now at the forefront of the post-genomics era and is providing a number of tools with which to mine the large datasets produced by genome-wide analysis. Of particular importance is the emergence of techniques that give the ability to reveal the transcription regulatory networks that are active in the cell in response to environmental stimuli or disease states. Deciphering the transcription networks that function in malignant cells not only will provide the knowledge to understand how carcinomas progress, but would also allow the construction of useful therapeutic tools for their effective treatment. In this review the recent advances that have been made in functional genomics that allow microarray data to be more fully interpreted and reveal the transcription networks that have gone awry in transformed cells are described.*

The application of microarray analysis, in the study of human malignancies, has undergone a paradigm shift in recent years so that datasets are now being used to elucidate the complex transcriptional programs that are active in human cancer. The use of integrative bioinformatics that allows this type of interpretation is a relatively new phenomenon and most studies to date have been carried out using microarray data from yeast. However, a number of studies have been recently published that have used meta-analysis of cancer datasets in an effort to identify transcriptional programs that are specifically active in cancer. In this review the recent advances that have been made in the development of bioinformatics resources that are used to analyse microarray data will be summarised, paying particular attention to algorithms that function to delineate cellular transcription networks. The studies that have been conducted, thus far, that have used these new tools to analyse the cellular networks that have gone awry during the development of cancer will be specifically discussed.

Correspondence to: Michael L. Roberts, Director of Gene Therapy, Regulon A.E., Afxentiou 7, Athens 17455, Greece. Tel: +30 210 9886502, Fax: +30 210 9858453, e-mail: michael@regulon.org

Key Words: Microarray data, bioinformatics, genomics, functional genomics, cancer, review.

Transcription Factor Databases

For nearly two decades scientists have been compiling databases that catalogue the *trans*-factors and *cis*-elements that are responsible for gene regulation (see Table I) (1). This has resulted in the emergence of useful tools, such as **TRANSCompel** (2), **TRANSFAC** (3), **ABS** (4), **JASPAR** (5) and **HTPSELEX** (6) that index transcription factors and their target sequences based on experimental data, and **TRED** (7), which utilizes both experimental and automated data. Databases of known transcription factor binding sites can be used to detect the presence of protein-recognition elements in a given promoter, but only when the binding site of the relevant DNA-binding protein and its tolerance to mismatches *in vivo* is already known. Because this knowledge is currently limited to a small subset of transcription factors, much effort has been applied to the discovery of regulatory motifs by comparative analysis of the DNA sequences of promoters. By finding conserved regions between multiple promoters, motifs can be identified with no prior knowledge of transcription factor binding sites. A number of models have emerged that achieve this by searching for statistically overrepresented conserved sequences flanking a gene's coding region. These algorithms, *e.g.* **YMF** (8, 9) and **SCORE** (10) function by aligning multiple untranslated regions from the entire genome and identifying sequences that are significantly overrepresented in comparison to what is expected by random. At present these tools are mainly applied in the study of lower eukaryotes where the genome is

Table I. *Databases employed in the identification of promoter sequences.*

Resource	Description	Citation
DBTSS	Database of transcriptional start sites	Suzuki *et al.*, 2001 (11, 12)
TRAFAC	Conserved *cis*-element search tool	Jegga *et al.*, 2002 (35)
TRANSCompel	Database of composite regulatory elements	Kel-Margoulis *et al.*, 2002 (2)
TRANSFAC	Eukaryotic transcription factor database	Matys *et al.*, 2003 (3)
Phylofoot	Tools for phylogenetic footprinting purposes	Lenhard *et al.*, 2003 (31)
CORG	Multi-species DNA comparison and annotation	Dieterich *et al.*, 2003 (30)
CONSITE	Explores trans-factor binding sites from two species	Lenhard *et al.*, 2003 (31)
CONFAC	Conserved transcription factor binding site finder	Karanam and Moreno, 2004 (32)
CisMols	Identifies *cis*-regulatory modules from inputed data	Jegga *et al.*, 200534)
TRED	Catalogue of transcription regulatory elements	Zhao *et al.*, 2005 (7)
Oncomine	Repository and analysis of cancer microarray data	Rhodes *et al.*, 2005 (61)
ABS	Database of regulatory elements	Blanco *et al.*, 2006 (4)
JASPAR	Database of regulatory elements	Sandelin *et al.*, 2004 (5)
HTPSELEX	Database of composite regulatory elements	Jagannathan *et al.*, 2006 (6)
PReMod	Database of transcriptional regulatory modules	Blanchette *et al.*, 2006 (18, 19)
CisView	Browser of regulatory motifs and regions	Sharov *et al.*, 2006 (20)
BEARR	Batch extraction algorithm for microarray data	Vega *et al.*, 2004 (21)
VISTA	Align and compare sequences from multiple species	Dubchak and Ryaboy, 2006 (22)
PromAn	Promoter analysis by integrating databases	Lardenois *et al.*, 2006 (23)
CRSD	Composite regulatory signature database	Liu *et al.*, 2006 (24)
MPromDb	Portal for genome-wide promoter analysis	Sun *et al.*, 2006 (60)

less complex and regulatory elements are easier to identify, thus extending these algorithms to the human genome has proved to be somewhat more difficult. In order to clarify this issue a number of groups have shown that it is possible to mine the genome of higher eukaryotes by searching for conserved regulatory elements adjacent to transcription start site motifs such as TATA and CAAT boxes, *e.g.* as catalogued in the **DBTSS** resource (11, 12), or by searching for putative *cis*-elements in the CpG rich regions that are present in promoter sequences, and in higher proportions (13). Alternatively, with the co-emergence of microarray technology and the complete sequence of the human genome, it is now possible to search for potential transcription factor binding sites by comparing the upstream non-coding regions of multiple genes that show similar expression profiles under certain conditions. Gene sets for comparative analysis can be chosen based on clustering, such as hierarchical and k-means (14), from a simple expression ratio (15) or functional analysis of gene products (16). This provides scientists with the opportunity to identify promoter elements that are responsive to certain environmental conditions, those that play a key role in mediating the differentiation of certain tissues or those that may be particularly active in mediating pathological phenotypes. Databases of predicted transcription factors also exist. **DBD** is one such example which functions by analysing the structure of target proteins and searching for the presence of protein domains that are indicative of DNA binding (17).

Recently, transcription factor databases have been compiled with the aim of utilising them to unravel regulatory networks active in response to diverse environmental stimuli. The **PReMod** database describes more than 100,000 computational predicted transcriptional regulatory modules within the human genome. These modules represent the regulatory potential for 229 transcription factors families and are the first genome-wide/transcription factor-wide collection of predicted regulatory modules for the human genome (18, 19). **CisView** is a browser of regulatory motifs and regions in the genome (20). Information on transcription factor binding sites can be viewed in the context of sequence conservation, neighbouring genes and their structure, GO annotations, protein domains, DNA repeats and CpG islands. It can then be used to define gene regulatory modules by searching for genes with specific combinations of binding sites and by identifying binding sites over represented in a given set of gene promoters and/or enhancers. **BEARR** is a batch extraction algorithm that allows users to analyse the *cis*-regulatory regions of hundreds or even thousands of differentially expressed genes identified in microarray studies (21). **VISTA** is a comprehensive suite of programs that allows users to align and compare sequences from multiple species, search for transcription factor binding sites with comparative sequence analysis, compare sequences with whole genome assemblies and to analyze multiple DNA sequence alignments from different species to unravel

phylogenic relationships (22). **PromAn** is a modular web-based tool that facilitates promoter analysis by integrating a variety of databases, methods and programs. It combines prediction programs and experimental databases to locate transcription start sites and promoter regions within a large genomic input sequence (23). Transcription factor binding sites (TFBSs) can then be predicted using several public databases and user-defined motifs. The composite regulatory signature database (**CRSD**) is another tool that can be applied when investigating complex regulatory networks using data from microarray analyses (24). **CRSD** has the additional ability to integrate data from both microRNA and transcription factor regulatory signatures. The program utilises human UniGene, mature microRNAs, putative promoter, TRANSFAC, pathway and gene ontology databases in order to predict potential gene regulation networks. The Mammalian Promoter Database (**MPromDb**) is an integrated database for gene promoters with experimentally supported annotation of transcription start sites (TSS), *cis*-regulatory elements, CpG islands and ChIP-chip experimental results (60). The current version of MPromDb contains information on 27,945 genes, 6,394 transcription factor binding sites and 1,771 transcription factors with links to PubMed and GenBank references. Target promoters of 5 transcription factors identified by chromatin immunoprecipitation microarray (ChIP-chip) assays are also integrated into the database. **MPromDb** serves as a portal for genome-wide promoter analysis of data generated by ChIP-chip experimental studies.

Phylogenetic footprinting, or comparative genomics, is now being applied to identify novel promoter elements by comparing the evolutionarily conserved untranslated elements proximal to known genes from a variety of organisms (25). The availability of genome sequences between species has notably advanced comparative genomics and the understanding of evolutionary biology in general. The neutral theory of molecular evolution provides a framework for the identification of DNA sequences in genomes of different species. Its central hypothesis is that the vast majority of mutations in the genome are neutral with respect to the fitness of an organism. Whilst deleterious mutations are rapidly removed by selection, neutral mutations persist and follow a stochastic process of genetic drift through a population. Therefore, non-neutral DNA sequences (functional DNA sequences) must be conserved during evolution, whereas neutral mutations are acquired. Initial studies sufficiently demonstrated that the human genome could be adequately compared to the genomes of other organisms allowing for the efficient identification of homologous regions in functional DNA sequences (26-28). Subsequently, a number of bioinformatic tools have emerged that operate by comparing non-coding regulatory sequences between the genomes of various organisms to enable the identification of conserved transcription factor binding sites that are significantly enriched in the promoters of candidate genes or from clusters identified by microarray analysis. Examples of these software suites include **TRAFAC** (29), **CORG** (30), **CONSITE** (31), **CONFAC** (32), **VAMP** (33) and **CisMols Analyser** (34). Typically these tools work by aligning the upstream sequences of target genes between species thus identifying conserved regions that could potentially function as *cis*-regulatory elements and have consequently been applied in the elucidation of transcription regulatory networks in a variety of models. In one study relating to cancer, **TRAFAC** was used to illustrate a mechanism whereby the oncogene c-myc regulates the enzymes involved in glycolysis and was also able to reveal a noncanonical E box that mediates expression of some of these enzymes (35). Myc overexpression has been suggested to aberrantly enhance tumour glycolysis even in the presence of oxygen, a phenomenon designated as the Warburg effect. The authors had previously shown, by gene expression profiling, that myc increases the expression of specific glycolytic enzyme genes, but were unable to demonstrate that these increases were a direct effect of myc. Moreover, by using traditional experimental approaches such as *in vitro* reporter- and electrophoresis mobility shift- assays, they were able to identify a number of myc-sensitive glycolytic genes. However, these methodologies still did not provide evidence that Myc directly activated the transcription of these genes. **TRAFAC** was therefore used to identify potential myc binding sites in the promoter regions of the glycolytic genes. The biological significance of these sequences was then confirmed experimentally by the chromatin immuno-preciptiation assay. By adopting this approach the authors were able to apply phylogenetic footprinting to determine the architecture of the myc target glycolytic gene network and further dissect the molecular basis of myc-induced altered glucose metabolism.

The emergence of bioinformatics tools that function to identify specific transcriptional elements in the sequenced human genome is transforming the application of functional genomics. Until recently the interpretation of data from microarray analysis has been limited to the identification of genes whose function may be important in a single pathway or response. How this related to global changes in the cellular phenotype had been largely ignored, as the necessary tools to examine this simply did not exist. With the advancement of bioinformatics we are now in a position to utilize all the data that is obtained from large-scale gene expression analysis and combine it with knowledge of the completed sequence of the human genome and with transcription factor, gene ontology and molecular function databases, thereby more fully utilizing the large datasets that are generated by global gene expression studies.

Comparative Functional Genomics in Cancer

Cancer is a complex biological phenomenon that is thought to arise out of a multi-step process of genetic and epigenetic alterations in the cellular DNA, ultimately resulting in the transformation of the cell and its uncontrolled growth, division and migration. Identifying the aberrant molecular pathways that mediate cellular transformation has been a major challenge in understanding how malignancy develops. The advent of functional genomics has given scientists the prospect of examining global changes in gene expression, thus providing molecular phenotypes that could potentially help in establishing more effective techniques of diagnosis and prognosis in a variety of carcinomas (36-38). Utilising microarrays to decipher the molecular events that result in tumour progression has proved to be a more difficult task, particularly since microarray data only provides a snapshot into a cell's transcriptome at a specific point in time. Moreover, many carcinomas contain multiple genetic alterations, making it difficult to ascribe specific changes in gene expression profiles to particular alterations in the genome of a transformed cell. Despite this some effort has been made to examine the effects of oncogenes on global gene expression (39, 40-42). However, these studies have not made full use of all the data available from the resultant gene expression profile, as researchers have tended to sift through the differentially expressed genes choosing to focus on those whose function is known to them, ignoring large volumes of data in the process and thus introducing experimenter bias into the analysis. In the past few years it has become apparent that microarray data can have wider applications in the study of cancer, particularly with the advent of comparative genomic microarray analysis (43, 44). In this type of analysis, gene expression data can be mapped to chromosomes, revealing potential sites of chromosomal aberrations, such as amplifications or deletions, which may predominate in particular types of cancer. There is also now a growing trend for researchers to analyse microarray data in terms of 'gene modules' instead of the presentation of differentially regulated gene lists (45). By grouping genes into functionally related modules it is possible to identify subtle changes in gene expression that may be biologically (if not statistically significantly) important, to more easily interpret molecular pathways that mediate a particular response and to compare many different microarray experiments from different tumour types in an effort to uncover the commonalities and differences in multiple clinical conditions. Therefore, we are now moving into a new era of functional genomics, where the large datasets generated by the evaluation of global gene expression studies can be more fully interpreted by improvements in computational methods. It is important, in the study of cancer, that these improved bioinformatics tools be applied to this complex disease in an effort to unravel the molecular processes that mediate the malignant phenotype, so that ultimately improved targeted therapeutics can be effectively designed.

Identifying Transcription Networks in Cancer

Traditionally, microarray data have been computationally analysed by clustering algorithms, such as hierarchical, k-means or self-organising maps. It has been proposed that genes that behaved in a similar fashion, *i.e.* those that clustered together, could be related, in that they would be under the control of the same transcription factors. Reconstructing transcription networks based on clustering techniques has not met with much success as genes within a cluster may actually not be related to one another and genes that are part of the same network may fall into different clusters. Indeed, this was elegantly demonstrated in a recent study conducted by the group of Michael Hubank (46). In this study, p53 targets were identified by applying simple differential equations and a small p53-responsive training set of genes to predict the activity profile of p53 and infer genes within a microarray dataset that behaved in a similar fashion. The authors presented p53-responsive genes with different levels of confidence and validated their approach using siRNA. It was demonstrated that responsive genes fell into a number of different clusters when the microarray data was analysed by k-means clustering. Thus demonstrating the importance of using prior biological knowledge in the mathematical analysis of microarray data and suggesting that all microarray analyses of this type should be conducted taking into account the specific biological action of the regulator under study.

The regulation of gene expression in eukaryotes is highly complex and often occurs through the coordinated action of multiple transcription factors. The use of *trans*-factor combinations in the control of gene expression allows a cell to employ a relatively small number of transcription factors in the regulation of disparate biological processes. As discussed above, a number of tools have been developed that allow us to utilise microarray data to identify novel *cis*-regulatory elements. It is also possible to use this information to decipher the transcriptional networks that are active in cells under different environmental conditions. In an innovative study conducted in yeast, the importance of the combinatorial nature of transcriptional regulation was established by specifically examining clusters of up-regulated genes for the presence of combinations of specific *cis*-regulatory motifs (47). To identify motif combinations that control gene expression patterns, a database of known and putative regulatory motifs was established and all the genes in the *S. cerevisiae* genome containing each motif in their promoter were identified. Gene expression profiling of

genes whose promoters contained the particular motif or motif combination was then used to evaluate the effect of each motif on gene expression. For each motif or combination, a score was given, reflecting the similarity of the expression profiles of all the genes containing that motif, in several different conditions, including different stages of the cell cycle, sporulation, diauxic shift, heat and cold shock, and treatment with DTT, pheromone and DNA-damaging agents. Using this information motif synergy maps displaying the discovered motif associations were generated, thus providing a global view of the connections between regulators of the transcriptional networks within the cell under the different environmental conditions tested. The authors demonstrated that a relatively small number of motifs functioned as nodes, exerting their influence on a variety of different biological processes.

This type of co-ordinated data combination from the results of different microarray experiments, with a view to examining global regulation of transcription is termed meta-analysis and has been broadly applied in the study of cancer. The ability to compare gene expression profiles from multiple platforms has been made possible by examining the degree of statistical significance that a gene is differentially expressed relative to the control rather than looking at the expression level of the gene *per se* (48). Meta-analyses of cancer datasets have permitted the identification of gene modules, allowing for the reduction of complex cancer signatures to small numbers of activated transcription programs and even to the identification of a common transcription program that is active in most types of cancer (49). Indeed, it has even been suggested that all new microarray studies should incorporate meta-analysis into their experimental design to overcome the high numbers of false positives that the analysis of single datasets yield (50). Meta-analysis can also help to identify specific transcription factors whose deregulation plays a key role in tumour development. For instance, in one study, the importance of aberrant E2F activity in cancer was reaffirmed during a search for the regulatory programs linking transcription factors to the target genes found to be up-regulated in specific cancer types (51). It was shown that E2F target genes were disproportionately up-regulated in more than half of the gene expression profiles examined, which were obtained from a multitude of different cancer types. In addition to re-affirming an established hypothesis, meta-analyis can also be used to identify important underlying transcription programs active in specific carcinomas. In one of the pioneering meta-analysis studies, a synchronous network of transcriptional regulation in the polyamine and purine biosynthesis pathways was identified in prostate cancer using data from four different datasets, representing the two major microarray platforms (52). Therefore, it is now transpiring that integrative bioinformatics analyses have the potential to generate new hypotheses about cancer progression, allowing the identification of new targets for cancer therapy and moving us towards a complete understanding of the aberrant transcription programs active in cancer.

Meta-analyses have also been used in an effort to identify potential common transcription modules responsible for metastasis in diverse tumour types. In one study, a gene expression signature that distinguished primary from metastatic adenocarcinomas was identified by analysing primary *versus* metastatic tumours from a variety of origins, including lung, breast, prostate, colon, uterus and ovaries (53). A 128-gene signature (or reduced 17-gene signature) was then applied to gene expression profiles obtained from stage I/II lung, breast and prostate adenocarcinomas and was able to predict the metastatic potential of each tumour. Particularly, tumours predicted to contain the metastatic module were also associated with a poor clinical outcome. These results suggest that the metastatic potential is encoded within the bulk of a primary tumour, thus challenging the idea that metastases arise from rare cells within a primary tumour. Meta-analysis has also been applied to predict survival times in patients diagnosed with stage I/II non-small cell lung carcinoma (54). In this study, a 64-gene signature was identified from microarray analysis by correlating gene expression profiles to the corresponding two- and five-year survival times of the patients from which the tumours were originally derived. This signature was then used to effectively identify the risk factor for aggressive disease progression in newly diagnosed patients, and it was proposed that the gene signature could potentially be used in classifying which patients should receive more intense chemotherapeutic dosing regimes.

This ability to use gene expression data to identify gene modules, which mediate specific responses to environmental stimuli (or to a diseased state) and to correlate their regulation to the *cis*-regulatory elements present upstream of the genes in each module, has transformed the way in which microarray data are interpreted (55). For instance, by using the modular approach it is possible to examine whether particular gene modules are active in a variety of different carcinomas, or whether individual carcinomas require the function of unique gene modules. This has allowed us to look for transcriptional commonalities between different carcinomas, which should aid in the design of widely applicable anti-cancer therapeutic strategies. In one particular study, gene expression data from 1,975 microarrays, spanning 22 different carcinomas were used to identify gene modules that were activated or deactivated in specific types of cancer (56). Using this approach the authors found that a bone osteoblastic module was active in a number of carcinomas whose primary metastatic site was known to be the bone. Thus, a common

mechanism of bone metastasis between a variety of different carcinomas was identified, which could be targeted in the development of novel anticancer therapies. It is also possible to identify the higher-level regulator that controls the expression of the genes in each module (55). Examination of the upstream regulatory sequences of each gene in a module reveals the presence of common *cis*-regulatory elements that are known to be the target of the module's regulator. Therefore, by identifying specific regulatory proteins that control the activation of gene modules in different carcinomas, it should be possible to extrapolate the important *cis*-elements that mediate transcription in the transformed cell.

The growing number of publicly available microarray datasets has facilitated the meta-analysis phenomenon. Indeed, such a large amount of data is currently available that researchers interested in specific genes have begun to use it to identify all the conditions under which their gene is active before embarking on traditional wet bench experimentation. This allows the researcher to uncover novel pathways that their gene of interest may be involved in. Finnochario and colleagues used this approach to identify expression datasets containing information on genes regulated by the tumour suppressor genes p16 and pRb, and were able to uncover a strong correlation with genes regulated by the EWS/FLI fusion protein, the gene responsible for the development of Ewings sarcoma (57). The authors were then able to use this data to predict the molecular mechanisms underpinning the progression of Ewings sarcoma and pointed out the importance that altered control of the cell cycle plays in this process.

The continuing development of the human interactome is also contributing to a more effective interpretation of large-scale gene expression datasets. Interactomes comprise maps of proteins that are known to functionally interact with one another, and are generated using the results from yeast two-hybrid screens, by collecting information from the literature and from computational methodologies (58). Although the human interactome is far from complete it is still being effectively used to interpret microarray data with a view to identifying important networks responsible for tumour progression. For instance the importance of aberrantly active Raf in the development of multiple myeloma has been reiterated using this approach, where it was demonstrated that Raf is a key protein that interacts with a number of proteins found to be up-regulated in this cancer by microarray analysis (59). Furthermore, proteins mutated in cancer were found to contain a high ratio of domains with a high propensity for mediating protein interactions (58). It could be argued that the completion of the human interactome is a critical step in the full integration of genome-wide expression analysis. The ability to map gene (and protein) expression data to such maps would afford us the opportunity to directly visualise networks of diverse protein interactions induced under specific environmental and disease-specific conditions and allow the identification of key nodes that could act as targets for the design of new therapeutic agents.

Conclusion

The ability of molecular biologists to analyse genome-wide expression changes in response to environmental stimuli has provided scientists with an opportunity to unravel the gene regulatory processes that underpin complex biological responses. In the post-genomic era it is now the responsibility of the bioinformatician to provide biologists with the tools with which to accomplish this goal. In recent years a number of resources have been made available that for the first time give a realistic chance of truly understanding how complex biological regulatory networks dictate the cellular phenotype. A few studies have begun to explore new computational methods that allow cancer-specific transcriptional programs to be deciphered from genomic sequence data and microarray meta-analyses. Identification of these networks should afford us the opportunity to link the well-characterised aberrant upstream signalling events, associated with tumour progression, to the more complex transcriptional pathways that they control. This will ultimately lead to the development of more efficacious anticancer therapies that are rationally designed based on an intricate understanding of the transformed phenotype.

References

1 Wingender E: Compilation of transcription regulating proteins. Nucleic Acids Res *16*: 1879-1902, 1988.

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

3 Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S and Wingender E: TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res *31*: 374-378, 2003.

4 Blanco E, Farre D, Alba MM, Messeguer X and Guigo R: ABS: a database of annotated regulatory binding sites from orthologous promoters. Nucleic Acids Res *34*: D63-D67, 2006.

5 Sandelin A, Alkema W, Engstrom P, Wasserman WW and Lenhard B: JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res *32*: D91-D94, 2004.

6 Jagannathan V, Roulet E, Delorenzi M and Bucher P: HTPSELEX--a database of high-throughput SELEX libraries for transcription factor binding sites. Nucleic Acids Res *34*: D90-D94, 2006.

7 Zhao F, Xuan Z, Liu L and Zhang MQ: TRED: a transcriptional regulatory element database and a platform for *in silico* gene regulation studies. Nucleic Acids Res *33*: D103-D107, 2005.

8 Sinha S and Tompa M: YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. Nucleic Acids Res *31*: 3586-3588, 2003.

9 Sinha S and Tompa M: Discovery of novel transcription factor binding sites by statistical overrepresentation. Nucleic Acids Res *30*: 5549-5560, 2002.

10 Rebeiz M, Reeves NL and Posakony JW: SCORE: a computational approach to the identification of *cis*-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation. Proc Natl Acad Sci USA *99*: 9888-9893, 2002.

11 Suzuki Y, Yamashita R, Sugano S and Nakai K: DBTSS, database of transcriptional start sites: progress report 2004. Nucleic Acids Res *32*: D78-D81, 2004.

12 Suzuki Y, Yamashita R, Nakai K and Sugano S: DBTSS: database of human transcriptional start sites and full-length cDNAs. Nucleic Acids Res *30*: 328-331, 2002.

13 Davuluri RV, Grosse I and Zhang MQ: Computational identification of promoters and first exons in the human genome. Nat Genet *29*: 412-417, 2001.

14 Roth FP, Hughes JD, Estep PW and Church GM: Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol *16*: 939-945, 1998.

15 Bussemaker HJ, Li H and Siggia ED: Regulatory element detection using correlation with expression. Nat Genet *27*: 167-171, 2001.

16 Jensen LJ and Knudsen S: Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. Bioinformatics *16*: 326-333, 2000.

17 Kummerfeld SK and Teichmann SA: DBD: a transcription factor prediction database. Nucleic Acids Res *34*: D74-D81, 2006.

18 Blanchette M, Bataille AR, Chen X, Poitras C, Laganiere J, Lefebvre C, Deblois G, Giguere V, Ferretti V, Bergeron D, Coulombe B and Robert F: Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res *16*: 656-668, 2006.

19 Ferretti V, Poitras C, Bergeron D, Coulombe B, Robert F and Blanchette M: PReMod: a database of genome-wide mammalian *cis*-regulatory module predictions. Nucleic Acids Res *35*: D122-D126, 2007.

20 Sharov AA, Dudekula DB and Ko MS: CisView: a browser and database of *cis*-regulatory modules predicted in the mouse genome. DNA Res *13*: 123-134, 2006.

21 Vega VB, Bangarusamy DK, Miller LD, Liu ET and Lin CY: BEARR: batch extraction and analysis of *cis*-regulatory regions. Nucleic Acids Res *32*: W257-W260, 2004.

22 Dubchak I and Ryaboy DV: VISTA family of computational tools for comparative analysis of DNA sequences and whole genomes. Methods Mol Biol *338*: 69-89, 2006.

23 Lardenois A, Chalmel F, Bianchetti L, Sahel JA, Leveillard T and Poch O: PromAn: an integrated knowledge-based web server dedicated to promoter analysis. Nucleic Acids Res *34*: W578-W583, 2006.

24 Liu CC, Lin CC, Chen WS, Chen HY, Chang PC, Chen JJ and Yang PC: CRSD: a comprehensive web server for composite regulatory signature discovery. Nucleic Acids Res *34*: W571-W577, 2006.

25 Ureta-Vidal A, Ettwiller L and Birney E: Comparative genomics: genome-wide analysis in metazoan eukaryotes. Nat Rev Genet *4*: 251-262, 2003.

26 Makalowski W and Boguski MS: Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc Natl Acad Sci USA *95*: 9407-9412, 1998.

27 Makalowski W, Zhang J and Boguski MS: Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. Genome Res *6*: 846-857, 1996.

28 Wheelan SJ, Boguski MS, Duret L and Makalowski W: Human and nematode orthologs – lessons from the analysis of 1800 human genes and the proteome of *Caenorhabditis elegans.* Gene *238*: 163-170, 1999.

29 Jegga AG, Sherwood SP, Carman JW, Pinski AT, Phillips JL, Pestian JP and Aronow BJ: Detection and visualization of compositionally similar *cis*-regulatory element clusters in orthologous and coordinately controlled genes. Genome Res *12*: 1408-1417, 2002.

30 Dieterich C, Wang H, Rateitschak K, Luz H and Vingron M: CORG: a database for comparative regulatory genomics. Nucleic Acids Res *31*: 55-57, 2003.

31 Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N and Wasserman WW: Identification of conserved regulatory elements by comparative genome analysis. J Biol *2*: 13, 2003.

32 Karanam S and Moreno CS: CONFAC: automated application of comparative genomic promoter analysis to DNA microarray datasets. Nucleic Acids Res *32*: W475-W484, 2004.

33 La Rosa P, Viara E, Hupe P, Pierron G, Liva S, Neuvial P, Brito I, Lair S, Servant N, Robine N, Manie E, Brennetot C, Janoueix-Lerosey I, Raynal V, Gruel N, Rouveirol C, Stransky N, Stern MH, Delattre O, Aurias A, Radvanyi F and Barillot E: VAMP: visualization and analysis of array-CGH, transcriptome and other molecular profiles. Bioinformatics *22*: 2066-2073, 2006.

34 Jegga AG, Gupta A, Gowrisankar S, Deshmukh MA, Connolly S, Finley K and Aronow BJ: CisMols Analyzer: identification of compositionally similar *cis*-element clusters in ortholog conserved regions of coordinately expressed genes. Nucleic Acids Res *33*: W408-W411, 2005.

35 Kim JW, Zeller KI, Wang Y, Jegga AG, Aronow BJ, O'Donnell KA and Dang CV: Evaluation of myc E-box phylogenetic footprints in glycolytic genes by chromatin immunoprecipitation assays. Mol Cell Biol *24*: 5923-5936, 2004.

36 Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO and Staudt LM: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature *403*: 503-511, 2000.

37 Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB and Hanash S: Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med *8*: 816-824, 2002.

38 van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, Van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH and Bernards R: A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med *347*: 1999-2009, 2002.

39 Schuldiner O and Benvenisty N: A DNA microarray screen for genes involved in c-MYC and N-MYC oncogenesis in human tumors. Oncogene *20*: 4984-4994, 2001.

40 Russo G, Claudio PP, Fu Y, Stiegler P, Yu Z, Macaluso M and Giordano A: pRB2/p130 target genes in non-small lung cancer cells identified by microarray analysis. Oncogene *22*: 6959-6969, 2003.

41 Roberts ML, Drosopoulos KG, Vasileiou I, Stricker M, Taoufik E, Maercker C, Guialis A, Alexis MN and Pintzas A: Microarray analysis of the differential transformation mediated by Kirsten and Harvey Ras oncogenes in a human colorectal adenocarcinoma cell line. Int J Cancer *118*: 616-627, 2006.

42 Frattini M, Ferrario C, Bressan P, Balestra D, De Cecco L, Mondellini P, Bongarzone I, Collini P, Gariboldi M, Pilotti S, Pierotti MA and Greco A: Alternative mutations of BRAF, RET and NTRK1 are associated with similar but distinct gene expression patterns in papillary thyroid cancer. Oncogene *23*: 7436-7440, 2004.

43 Louro ID, Bailey EC, Li X, South LS, McKie-Bell PR, Yoder BK, Huang CC, Johnson MR, Hill AE, Johnson RL and Ruppert JM: Comparative gene expression profile analysis of GLI and c-MYC in an epithelial model of malignant transformation. Cancer Res *62*: 5867-5873, 2002.

44 Crawley JJ and Furge KA: Identification of frequent cytogenetic aberrations in hepatocellular carcinoma using gene-expression microarray data. Genome Biol *3*: RESEARCH0075, 2002.

45 Segal E, Friedman N, Kaminski N, Regev A and Koller D: From signatures to models: understanding cancer using microarrays. Nat Genet *37 Suppl*: S38-S45, 2005.

46 Barenco M, Tomescu D, Brewer D, Callard R, Stark J and Hubank M: Ranked prediction of p53 targets using hidden variable dynamic modeling. Genome Biol *7*: R25, 2006.

47 Pilpel Y, Sudarsanam P and Church GM: Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet *29*: 153-159, 2001.

48 Rhodes DR and Chinnaiyan AM: Integrative analysis of the cancer transcriptome. Nat Genet *37 Suppl*: S31-S37, 2005.

49 Rhodes DR and Chinnaiyan AM: Integrative analysis of the cancer transcriptome. Nat Genet *37 Suppl*: S31-S37, 2005.

50 Choi JK, Choi JY, Kim DG, Choi DW, Kim BY, Lee KH, Yeom YI, Yoo HS, Yoo OJ and Kim S: Integrative analysis of multiple gene expression profiles applied to liver cancer study. FEBS Lett *565*: 93-100, 2004.

51 Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Barrette TR, Ghosh D and Chinnaiyan AM: Mining for regulatory programs in the cancer transcriptome. Nat Genet *37*: 579-583, 2005.

52 Rhodes DR, Barrette TR, Rubin MA, Ghosh D and Chinnaiyan AM: Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res *62*: 4427-4433, 2002.

53 Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES and Golub TR: Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci USA *98*: 15149-15154, 2001.

54 Lu Y, Lemon W, Liu PY, Yi Y, Morrison C, Yang P, Sun Z, Szoke J, Gerald WL, Watson M, Govindan R and You M: A gene expression signature predicts survival of patients with stage I non-small cell lung cancer. PLoS Med *3*: e467, 2006.

55 Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D and Friedman N: Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet *34*: 166-176, 2003.

56 Segal E, Friedman N, Koller D and Regev A: A module map showing conditional activity of expression modules in cancer. Nat Genet *36*: 1090-1098, 2004.

57 Finocchiaro G, Mancuso F and Muller H: Mining published lists of cancer related microarray experiments: identification of a gene expression signature having a critical role in cell-cycle control. BMC Bioinformatics *6(Suppl 4)*: S14-2005.

58 Jonsson PF and Bates PA: Global topological features of cancer proteins in the human interactome. Bioinformatics *22*: 2291-2297, 2006.

59 Rhodes DR and Chinnaiyan AM: Bioinformatics strategies for translating genome-wide expression analyses into clinically useful cancer markers. Ann NY Acad Sci *1020*: 32-40, 2004.

60 Sun H, Palaniswamy SK, Pohar TT, Jin VX, Huang THM and Davulur RV: MPromDb: an integrated resource for annotation and visualization of mammalian gene promoters and ChIP-chip experimental data. Nucleic Acids Res *34*: D98-D103, 2006.

61 Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A and Chinnaiyan AM: ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia *6*: 1-6, 2004.

Received March 8, 2007
Accepted March 23, 2007