Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W403–W407***
doi:10.1093/nar/gki466

Identifying synonymous regulatory elements in vertebrate genomes

Ivan Ovcharenko* and Marcelo A. Nobrega[1]

Energy, Environment, Biology, and Institutional Computing, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA and [1]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

Received March 2, 2005; Revised March 23, 2005; Accepted April 13, 2005

ABSTRACT

Synonymous gene regulation, defined by regulatory elements driving shared temporal and/or spatial aspects of gene expression, is most probably predicated on genomic elements that contain similar modules of certain transcription factor binding sites (TFBS). We have developed a method to scan vertebrate genomes for evolutionary conserved modules of TFBS in a predefined configuration, and created a tool, named SynoR that identifies synonymous regulatory elements (SREs) in vertebrate genomes. SynoR performs *de novo* identification of SREs utilizing known patterns of TFBS in active regulatory elements (REs) as seeds for genome scans. Layers of multiple-species conservation allow the use of differential phylogenetic sequence conservation filters in search of SREs and the results are displayed such as to provide an extensive annotation of the genes containing the detected REs. Gene Ontology categories are utilized to further functionally classify the identified genes, and integrated GNF Expression Atlas 2 data allow the cataloging of tissue-specificities of the predicted SREs. SynoR is publicly available at http://synor.dcode.org.

INTRODUCTION

The complex patterns of gene expression in vertebrates arise from the combinatorial interaction of multiple transcription factors with target *cis*- regulatory units consisting of modules of transcription factor binding sites (TFBS). Although in simpler organisms, such as yeast, bacteria and viruses, REs are usually associated with the promoters of their target genes, in more complex organisms, especially vertebrates, REs modulate promoter activity and are often positioned remotely from the genes they regulate—sometimes being as far away as a megabase from the transcriptional start site of a gene (1). Therefore, the general architectural features of complex gene regulatory networks, consisting of multiple distant REs distributed over long distances, upstream and downstream of a gene, make their identification challenging. Comparative genomics was shown to be a powerful tool in facilitating the genomic search for REs (2), but despite the progress in identifying REs, the location of the majority of vertebrate REs remains unknown, owing partially to our lack of understanding about what are the fundamental components of REs, and whether their organizational rules (if any) can be used as signatures for the genome-wide identification. Toward this end, it has recently been shown that in invertebrates searching for TFBS clustered in defined configuration allows for the identification of REs with a predefined function (3,4). Several tools have been created to identify specific modules of TFBS in promoters of co-expressed genes, including Toucan (5) and Crème 2.0 (6). Recently, those observations were expanded to genome scans of vertebrates using defined TFBS motifs as seeds (7,8).

Our goal was to expand on these observations, develop and test a new strategy to carry out genome-wide scans for REs using evolutionarily conserved TFBS (cTFBS) motifs. Known TFBS structures of REs, defined as a cluster of TFBS and their defined spatial order and distribution, were used as seeds to search for novel REs that can dramatically differ from the original REs at the sequence level, but are synonymous in function—synonymous REs (SREs). We created a publicly available tool, SynoR (http://synor.dcode.org), which provides the users with the ability to extend the knowledge derived from a single gene's regulatory genomic structure to the whole genome and identify novel genes with synonymous regulation.

METHODS

Genome-wide annotation of cTFBS

The ECR Browser tool generates whole-genome *blastz*-based alignments of vertebrate and invertebrate genomes (9). To generate a dataset of cTFBS for SynoR scans, we have established an automated annotation of evolutionarily cTFBS based

*To whom correspondence should be addressed. Tel: +1 925 422 5035; Fax: +1 925 422 2099; Email: ovcharenko1@llnl.gov

Table 1. cTFBS in alignments of the human genome (hg17) to the mouse (mm5), chicken (galGal2), frog (xenTro1) and fugu (fr1) genomes (assembly indexes from the UCSC genome browser)

Organism	Mouse	Chicken	Frog	Fugu
No. of cTFBS	13 069 048	1 945 164	859 769	402 784

on the ECR Browser alignments. This was created by using the rVista 2.0 tool (http://rvista.dcode.org) (10) with 'optimized-for-function' position weight matrix thresholds (11). The automated ECR Browser/rVista 2.0 annotation gradually expands the list of available genome alignments with cTFBS for the subsequent SynoR processing. Table 1 summarizes the number of cTFBS in the human genome as compared with different species.

Defining TFBS modules as seeds for the genome scans

Transcription Factor (TF) molecules interact with each other and bind to DNA to establish a gene transcription signal. The number of different TFs in a module, the number of TFBS, the spatial constraints, the order of TFBS and relative strands of TFBS differ for different regulatory pathways. SynoR requires user input describing a TFBS module structure to initiate a genome scan. In practical terms, three tiers of information on TFBS modules might be available: (i) a list of TFs known to participate in a particular regulatory pathway, (ii) a set of spatial constraints separating different TFBS and (iii) the order and orientation of individual TFBS in a module. While the TF content is essential for the genome scans, the other two tiers of information effectively refine the module signature and are provided as optional features. SynoR is limited in selection of TFBS to the list of TFBS available from the TFANSFAC Professional database (12), which is utilized by the rVista 2.0 tool in genome scans.

Identification of statistically enriched Gene Ontology (GO) categories

To predict the putative biological function of the identified elements, SynoR performs a GO category enrichment analysis for the genes that either flank or contain the non-coding subset of these elements. The tool employs binomial distribution analysis as an approximation for the hypergeometric distribution to accomplish this analysis (applicable owing to the number of genes in each GO category being significantly smaller than the number of genes in the genome). The statistical analysis depicts GO categories that contain significantly more genes than would be expected just by chance given the number of identified genes with GO annotation and the distribution of all the genes in the genome among different GO categories. At the final step of the analysis SynoR utilizes Holm's sequential Bonferroni correction (13) to correct for the multiple testing. Significantly enriched GO categories (as quantified by the P-value < 0.05) are reported to the user.

Establishing tissue-specificities of identified genes

To predict tissue specificity (if any) of the identified genes with non-coding elements, SynoR analyzes the GNF Expression Atlas 2 data (14). It performs a two-step clustering analysis of tissue specificity in expression of the genes. First, the clustering of the data into groups of co-expressed genes is performed using the Cluster 3.0 tool (15) and the results are visualized in a micro-array expression profile style plot (Figure 2). At the second step, SynoR identifies a set of tissues, in which the genes are either significantly overexpressed or suppressed. In order to do so, the tool calculates the difference between the number of overexpressed and the number of suppressed genes for each tissue i, δ_j. An estimate for an average difference $\bar{\delta}$ and a corresponding standard deviation σ_δ are calculated using the distribution of δ_j across all the tissues. That allows defining a z_i-value describing deviation in the observed difference in the number of overexpressed and suppressed genes versus the expectation for a given tissue:

$$z_i = \frac{\delta_i - \bar{\delta}}{\sigma_\delta}.$$

The expression in tissues with an absolute z-value >2.0 is reported as significantly increased/decreased, and in tissues with an absolute z-value >1.0 as changed.

In the search for tissue-specificities, performed by the Cluster 3.0 tool, SynoR eliminates absolute differences in expression in between different genes from the analysis. In order to do so, expression pattern of each gene across different tissues is normalized by dividing expression score in a particular tissue by the highest expression score in all the tissues. This effectively brings the average expression of highly expressed genes and the genes with a low level of expression to the same level and strongly highlights the differences in gene expression across different tissues. Also, GNF Atlas2 expression patterns in cancer cell lines and cell lines without profound tissue-specificity are excluded from the analysis to provide sampling of co-expression in normal tissues; thus, providing a link between a predicted SRE and normal tissue specificity.

RESULTS

Design and features of the SynoR tool

SynoR utilizes pre-computed annotations of cTFBS in vertebrate genomes (as obtained through multi-species genome alignments) adopted from the ECR Browser (9) (http://ecrbrowser.dcode.org). It scans the genome distribution of cTFBS in search for modules of TFBS in defined spatial configurations that match the seed profile defined by the user (Figure 1). SynoR overlaps the identified TFBS modules with gene annotation ['UCSC known genes' (16)] to categorize them into promoter elements, UTRs, introns, intergenic elements and coding exons. The ratio of newly identified elements overlapping with coding exons is expected to be small in comparisons of evolutionary distant species (such as humans and fish or humans and chickens), serving as an immediate quantifier of the specificity of predictions. The online results page also includes the multi-species conservation analysis of all the identified modules. Genes bracketing the identified noncoding elements or including them are selected for a further two-step analysis of shared activity. First, the Gene Ontology (GO) (17) categories for each gene, reflecting their biological function, are defined. Enrichment in GO categories that match the known functional activity of the seed RE allows us to evaluate

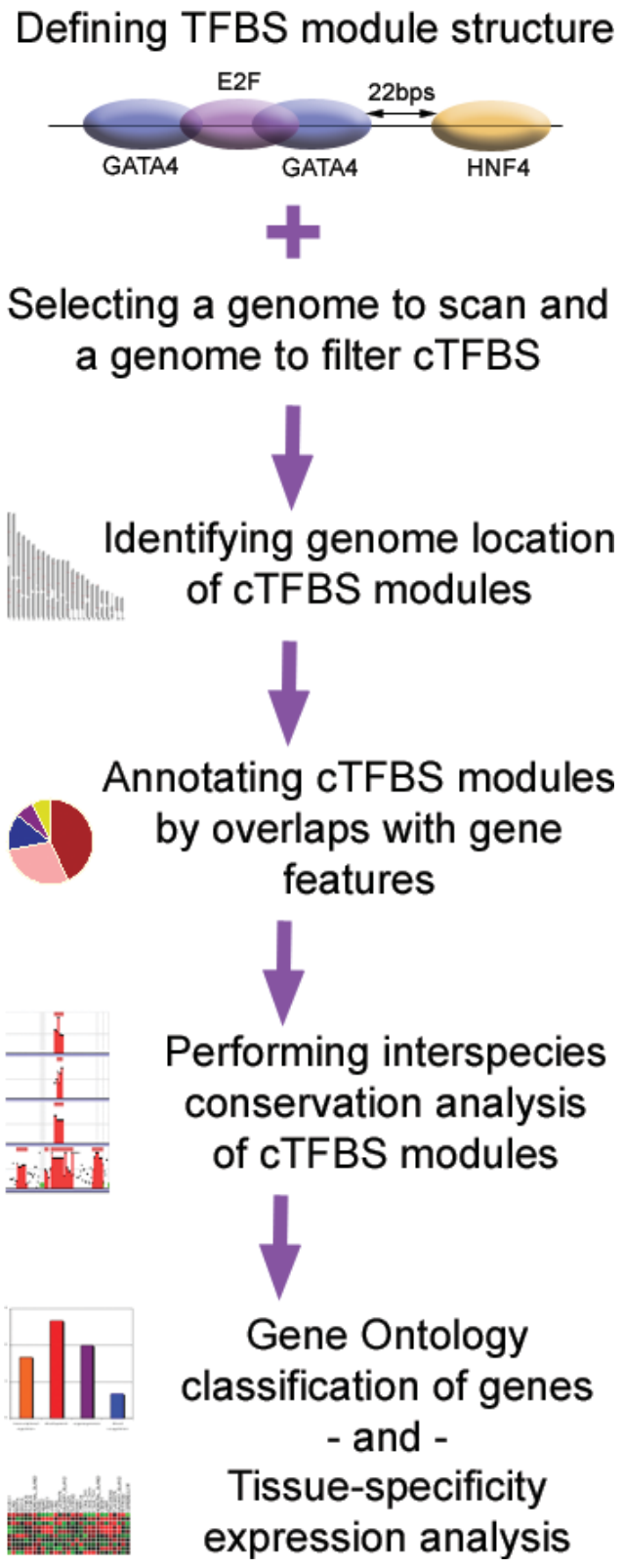

Figure 1. The schematic profile of SynoR genome scans and data analysis.

the sensitivity of genome scans. Next, the analysis using the GNF Expression Atlas 2 (14) is performed to define the tissue specificity of the genes. Comparative analysis of the expression of identified genes versus the average expression of all the genes in the genome highlights a subset of tissues, in which the genes bracketing the identified SREs are preferentially expressed.

A priori knowledge of TFs or TFBS modules involved in a particular biochemical process [such as neuronal development (18), heart formation (19) and muscle development (20)] is helpful in determining the seed signatures used in the SynoR scans. Studies that generate additional positional sequence information for active, multiple TFBS (10,21) can be effectively used to establish the configuration of spatial constraints and TFBS ordering, thus increasing the specificity of a SynoR search.

SREs associated with synergistic activation of gene expression in cardiac myocytes

The GNF Expression Atlas 2 summarizes the expression patterns of human, mouse and rat genes in several selected tissues using whole-genome microarray experiments (14). These data provide evidence of tissue specificity of genes bracketing predicted SREs. If a particular SRE is associated with a gene expressed in a set of defined tissues, for example, these tissues should also correspond to the expression pattern of the candidate genes sharing the SRE motif identified by SynoR. To assess the applicability of SynoR's tissue specificity analysis of predicted SREs, we scanned the human genome for combinatorial modules of two cTFBS, *SRF* (serum response factor) and *SP1*. Multiple lines of evidence support the notion that these TFs cooperatively participate in orchestrating gene expression in the heart and the vascular tissues (22,23). We applied SynoR to predict targets of synergistic *SRF*/*SP1* gene regulation in the human genome using as a seed motif the presence of these TFBS separated by <40 bp. Human–mouse conservation threshold was utilized. A total of 114 non-coding modules were identified in this scan, 23 (20%) of which overlapped with the promoter regions. Taking into account the density of human/mouse ECRs in the human genome (24), the probability of this high ratio of elements associated with promoters by chance is $<10^{-5}$, suggesting an enrichment in functional SREs identified in this scan. Expression analysis of the genes that either contain or flank the identified SREs presented a very distinct tissue specificity of these genes. Sixty-four percent of them (88 out of 138) are specifically expressed in cardiac myocytes while others are expressed in smooth muscle, heart and other tissues (Figure 2). This general observation is in agreement with the experimental data on expression of the studied TFs supporting the notion that GNF Expression Atlas 2 data integrated with SynoR predictions may provide an effective and straightforward annotation of tissue specificity for the identified elements and the search patterns. Together, these data support the idea that using this pair of cTFBS as a seed for genome-wide scan successfully identifies SREs, which are most probably responsible for the shared pattern of expression of their corresponding genes. Further studies are required to assess the *in vivo* functional activity of these elements and to investigate their possible roles in cardiovascular diseases.

Features categorization

To illustrate the application of SynoR's categorization of identified elements based on gene annotation, we scanned

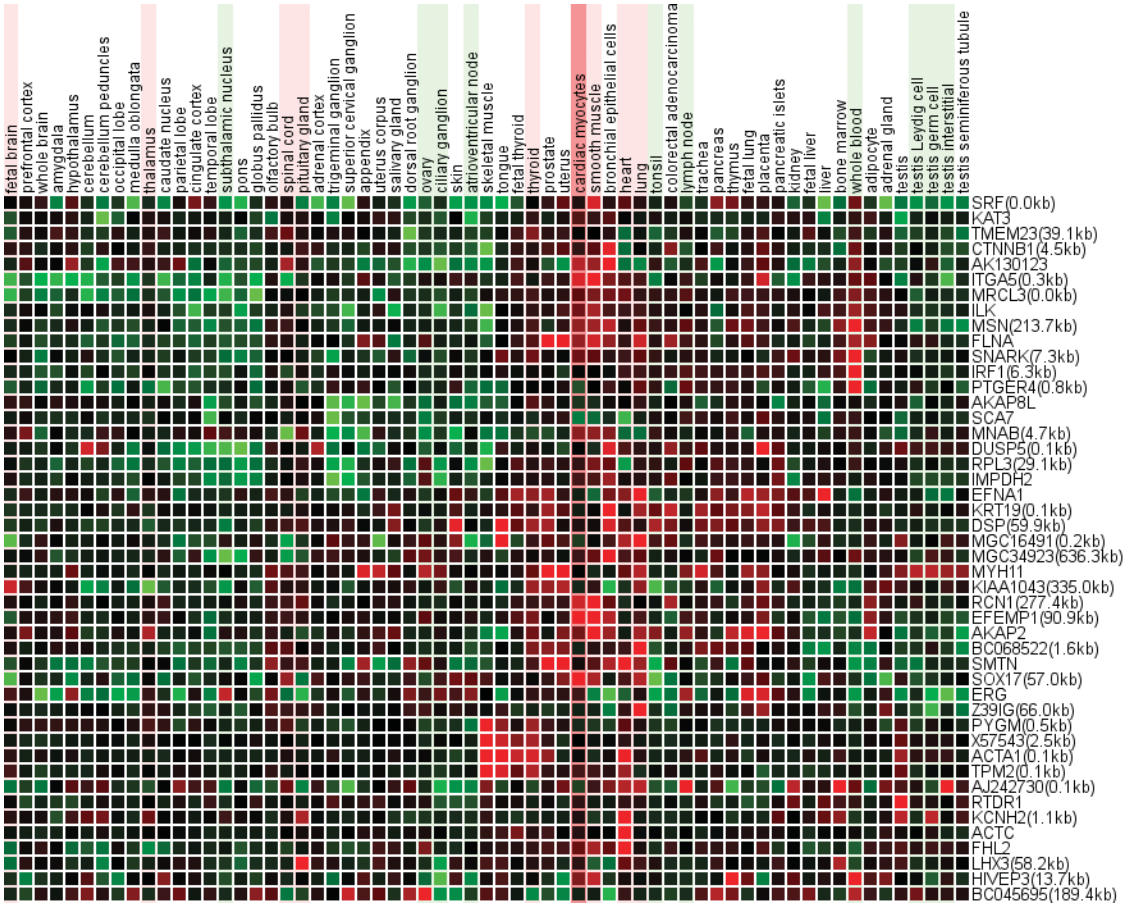

Figure 2. GNF Expression Atlas2 analysis for genes identified in the *SRF*/*SP1* SynoR scan of the human genome. A subset of 46 genes including the *SRF* gene is presented. Cardiac myocytes with significant overexpression identified by solid red background. Light red and light green backgrounds correspond to the over-expressed and suppressed tissue categories. Different columns correspond to different tissues listed on top and different rows correspond to the identified genes listed on the right. The number in parentheses following gene name provides a distance between an element and the gene in case of intergenic elements.

the human genome for a module of three NRSF human/mouse cTFBS. NRSF (neuron-restrictive silencer factor) plays a key role in neuronal differentiation (25) and mediation of transcriptional repression of neuron-specific genes in non-neuronal cells (26). Ten noncoding modules were identified, of which three were within promoters, four in introns, and four in intergenic intervals. One of the three promoters corresponds to that of *Barhl1*, a gene associated with neuronal migration (27), in an expression resembling that of the NRSF regulatory pathway. The remaining two promoters identified in this scan correspond to uncharacterized genes, and these results raise, thus, the possibility that these genes represent new members of the NRSF pathway.

DISCUSSION

The identification of non-coding sequences conserved among vertebrates has served as the most important pillar leading to the identification of functional gene REs in the human genome (1,2). Here we introduce a new tool based on cTFBS, named SynoR that performs genome scans for REs with shared biological activity. The fundamental inference behind the conceptualization of SynoR is that regulatory elements with similar function (SREs) operate under similar organizational principles, the modular distribution of a defined set of TFBS. This principle has been previously validated in several eukaryotes including yeast, worm and flies, and recent evidence suggests that the SREs can also be identified in humans (7). Our results support this notion, and SynoR was created to represent a publicly available tool for the search of SREs allowing for a broad range of user-defined options guiding the search of SREs in multiple regulatory pathways. SynoR is endowed with multiple checkpoint mechanisms to define the precise functional annotation of the identified elements, which include multispecies evolutionary conservation analysis, GO functional characterization and GNF Expression Atlas 2 analysis of tissue specificity of the genes bracketing the identified SREs. The results provided by SynoR allow for the immediate comparison between the functions of the genes in the vicinity of the identified SREs, ensuring the reliability of SynoR genome scans, while presenting the use with a categorized set of identified elements with distinct functions and evolutionary traits.

In summary, we present a strategy to identify SREs in eukaryotic genomes, and describe the design of a new tool, SynoR aiding in the identification of noncoding sequences that are most likely to correspond to regulatory elements, which can be tested in the laboratory.

ACKNOWLEDGEMENTS

The work was performed under the auspices of the United States Department of Energy by the University of California, Lawrence Livermore National Laboratory Contract No. W-7405-Eng-48 and was supported by DOE SCW0345 grant. Funding to pay the Open Access publication charges for this article was provided by DOE SCW0345 grant.

Conflict of interest statement. None declared.

REFERENCES

1. Nobrega,M.A., Ovcharenko,I., Afzal,V. and Rubin,E.M. (2003) Scanning human gene deserts for long-range enhancers. *Science*, **302**, 413.
2. Loots,G.G., Locksley,R.M., Blankespoor,C.M., Wang,Z.E., Miller,W., Rubin,E.M. and Frazer,K.A. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, **288**, 136–140.
3. Markstein,M., Zinzen,R., Markstein,P., Yee,K.P., Erives,A., Stathopoulos,A. and Levine,M. (2004) A regulatory code for neurogenic gene expression in the Drosophila embryo. *Development*, **131**, 2387–2394.
4. Erives,A. and Levine,M. (2004) Coordinate enhancers share common organizational features in the Drosophila genome. *Proc. Natl Acad. Sci. USA*, **101**, 3851–3856.
5. Aerts,S., Thijs,G., Coessens,B., Staes,M., Moreau,Y. and De Moor,B. (2003) Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res.*, **31**, 1753–1764.
6. Sharan,R., Ben-Hur,A., Loots,G.G. and Ovcharenko,I. (2004) CREME: *cis*-regulatory module explorer for the human genome. *Nucleic Acids Res.*, **32**, W253–W256.
7. Donaldson,I.J., Chapman,M., Kinston,S., Landry,J.R., Knezevic,K., Piltz,S., Buckley,N., Green,A.R. and Gottgens,B. (2005) Genome-wide identification of *cis*-regulatory sequences controlling blood and endothelial development. *Hum. Mol. Genet.*, **14**, 595–601.
8. Thompson,W., Palumbo,M.J., Wasserman,W.W., Liu,J.S. and Lawrence,C.E. (2004) Decoding human regulatory circuits. *Genome Res.*, **14**, 1967–1974.
9. Ovcharenko,I., Nobrega,M.A., Loots,G.G. and Stubbs,L. (2004) ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res.*, **32**, W280–W286.
10. Loots,G.G. and Ovcharenko,I. (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res.*, **32**, W217–W221.
11. Ovcharenko,I., Loots,G.G., Giardine,B.M., Hou,M., Ma,J., Hardison,R.C., Stubbs,L. and Miller,W. (2005) Mulan: multiple-sequence local alignment and visualization for studying function and evolution. *Genome Res.*, **15**, 184–194.
12. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Pruss,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.
13. Holm,S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, **6**, 65–70.
14. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al*. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
15. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
16. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al*. (2003) The UCSC genome browser database. *Nucleic Acids Res.*, **31**, 51–54.
17. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
18. Beaujean,D., Rosenbaum,C., Muller,H.W., Willemsen,J.J., Lenders,J. and Bornstein,S.R. (2003) Combinatorial code of growth factors and neuropeptides define neuroendocrine differentiation in PC12 cells. *Exp. Neurol.*, **184**, 348–358.
19. Bruneau,B.G., Nemer,G., Schmitt,J.P., Charron,F., Robitaille,L., Caron,S., Conner,D.A., Gessler,M., Nemer,M., Seidman,C.E. *et al*. (2001) A murine model of Holt-Oram syndrome defines roles of the T-box transcription factor Tbx5 in cardiogenesis and disease. *Cell*, **106**, 709–721.
20. Jensen,A.M. (2004) Potential roles for BMP and Pax genes in the development of iris smooth muscle. *Dev. Dyn.*, **232**, 385–392.
21. Lien,C.L., McAnally,J., Richardson,J.A. and Olson,E.N. (2002) Cardiac-specific activity of an Nkx2-5 enhancer requires an evolutionarily conserved Smad binding site. *Dev. Biol.*, **244**, 257–266.
22. Miano,J.M., Ramanan,N., Georger,M.A., de Mesy Bentley,K.L., Emerson,R.L., Balza,R.O.,Jr, Xiao,Q., Weiler,H., Ginty,D.D. and Misra,R.P. (2004) Restricted inactivation of serum response factor to the cardiovascular system. *Proc. Natl Acad. Sci. USA*, **101**, 17132–17137.
23. Jimenez,S.K., Sheikh,F., Jin,Y., Detillieux,K.A., Dhaliwal,J., Kardami,E. and Cattini,P.A. (2004) Transcriptional regulation of FGF-2 gene expression in cardiac myocytes. *Cardiovasc. Res.*, **62**, 548–557.
24. Ovcharenko,I., Stubbs,L. and Loots,G.G. (2004) Interpreting mammalian evolution using Fugu genome comparisons. *Genomics*, **84**, 890–895.
25. Su,X., Kameoka,S., Lentz,S. and Majumder,S. (2004) Activation of REST/NRSF target genes in neural stem cells is sufficient to cause neuronal differentiation. *Mol. Cell. Biol.*, **24**, 8018–8025.
26. Murai,K., Naruse,Y., Shaul,Y., Agata,Y. and Mori,N. (2004) Direct interaction of NRSF with TBP: chromatin reorganization and core promoter repression for neuron-specific gene transcription. *Nucleic Acids Res.*, **32**, 3180–3189.
27. Bulfone,A., Menguzzato,E., Broccoli,V., Marchitiello,A., Gattuso,C., Mariani,M., Consalez,G.G., Martinez,S., Ballabio,A. and Banfi,S. (2000) *Barhl1*, a gene belonging to a new subfamily of mammalian homeobox genes, is expressed in migrating neurons of the CNS. *Hum. Mol. Genet.*, **9**, 1443–1452.