



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Identifying Synonymous Regulatory Elements in Vertebrate Genomes

I. Ovcharenko, M. A. Nobrega

February 7, 2005

Nucleic Acids Research

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

Identifying Synonymous Regulatory Elements in Vertebrate Genomes.

Ivan Ovcharenko^{1,*} and Marcelo Nobrega²

1 Energy, Environment, Biology, and Institutional Computing, Lawrence Livermore National Laboratory, Livermore, California 94550, USA

2 Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

* Phone: 1 (925) 422 5035; Fax: 1 (925) 422 2099; Email: ovcharenko1@llnl.gov

ABSTRACT.

Synonymous gene regulation, defined as driving shared temporal and/or spatial expression of groups of genes, is likely 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 identify 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 the search of SREs and the results are displayed as to provide an extensive annotation of genes containing 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. We illustrate how this new tool can be used to establish a linkage between human diseases and noncoding genomic content. SynoR is publicly available at <http://synor.dcode.org>.

INTRODUCTION

The complex patterns of gene regulation in vertebrates arise from the combination of the regionalized expression of multiple transcription factors and their interactions with target cis-regulatory units consisting of modules of TFBSs. The outcome of these interactions, either the activation or repression of genes, as well as the structural nature of the REs have served as the basis for their classification in various categories, such as enhancers, repressors, silencers, insulators, and locus control regions. 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 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 (1,2). Therefore, the general architectural features of complex gene regulatory networks, consisting of multiple distant REs distributed over long distances up and downstream of a gene, makes their identification challenging. Comparative genomics was shown to be instrumental in facilitating the genomic search for REs (3-5). For example, comparisons of phylogenetically distant species proved especially effective in detecting REs associated with certain categories of genes, such as developmental genes involved in embryogenesis (1,2,6). Despite this progress in identification of 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 of regulatory elements that drive the expression of multiple genes in similar, synonymous patterns (SREs). Toward this end, it has been shown recently that, in invertebrates, searching for TFBSs clustered in defined motifs allow for the identification of SREs (7-9). Recently, those observations were expanded to vertebrates, with studies illustrating that genome scans searching for SREs using defined TFBS motifs as seeds are feasible (10,11).

Our goal was to expand on these observations, developing and testing a strategy to carry out genome-wide scans searching for SREs using evolutionarily conserved TFBS motifs. Known TFBS structures of REs, defined as a cluster of TFBS and their defined spatial order and distribution, were used as seeds to scan genomes in search for novel REs that might differ from the seed element at the sequence level, but are synonymous in function – SREs. We created a publicly available tool, SynoR (<http://synor.dcode.org>), which provides the users with the ability to extend the identification of single gene regulatory genomic structures to the whole genome scale and to identify novel genes with synonymous regulation. Illustrative examples are provided as how *de novo* observations obtained with the use of SynoR can be used in prioritizing conserved elements for studies of human diseases.

RESULTS

Design and features of the SynoR tool.

SynoR utilizes pre-computed annotations of conserved TFBS (cTFBS) in vertebrate genomes (as obtained through multi-species genome alignments) adopted from the ECR Browser (12,13) (<http://ecrbrowser.dcode.org>). It scans the genome distribution of cTFBS in a search for types of TFBSs in defined spatial configurations that match the seed profile defined by the user (Figure 1). Upon the localization of the TFBS modules in the genome, SynoR overlaps their coordinates with genomic annotation (known genes from the UCSC Genome Browser (14)) to categorize the identified modules into promoter elements, UTRs, introns, intergenic elements, and coding exons. The ratio of newly identified elements overlapping with coding exons is expected to be negligibly small, serving as an immediate quantifier of the specificity of prediction (although sometimes a sizeable ratio may reflect the detection of genes with duplicated protein domains being recognized as binding sites). The online results page also offers the conservation analysis of all the identified modules, obtained from the genome alignment and comparisons with different species. Genes bracketing the identified noncoding elements or contained within them are selected for a further two-step analysis aimed at determining the synonymous functional nature (if any) of the identified TFBS modules. As a first step, the Gene Ontology (GO) (15,16) categories for each gene, reflecting their biological function, are defined. An enrichment in GO categories that match the known functional activity of the seed RE provides with an effective verification of the sensitivity of the genome scans. In a later step, the analysis using the GNF Expression Atlas 2 (17) is performed to define the tissue-specificity of the genes in the vicinity of the identified SREs. 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 transcription factors (TF) or TFBS modules involved in a particular biochemical process is helpful in determining the seed signatures used in the SynoR scans. Numerous studies have already reported on the existence of specific combinatorial TFBS in modules responsible for the regulation of various biological processes including neuronal development (18-20), heart formation (21-23), oncogenesis (24-26), muscle development (27), etc. All these pre-defined patterns of TFBS organization can be utilized to define the minimal TFBS content of a seed RE and initiate a genome scan searching for SREs of that element. Studies that generate additional sequence positional information on active multiple TFBS (such as (28) or (29), for example) can be effectively used to establish the configuration of spatial constraints and TFBS ordering, thus increasing the specificity of a SynoR search. Nevertheless, the program also accommodates user-defined seeds, allowing for the investigation of new modular patterns of TFBSs that may be enriched in the vicinity of genes with synonymous regulation.

SynoR application for biological discoveries.

The *PAX6* gene is a member of the PAX family of transcription factors that are crucial during early development, especially in the specification of eyes and developing central nervous system. Recently, it was demonstrated that aspects of *PAX6* regulation is achieved partially by autoregulation, through an intronic element deeply conserved in vertebrates, including mammals and fish (30). *PAX6* protein binds to its binding site

within a conserved element (CE2) in *PAX6* intron 7 resulting in the upregulation in the expression of *PAX6* (30). Using SynoR, we searched for an enriched module of conserved *PAX6* sites to identify other putative elements with an activity synonymous to the CE2 element of *PAX6*. A defined parameter was set requiring the presence of a cluster of at least 3 *PAX6* TFBS, conserved in human and chicken, and with each TFBS no further apart from each other than 40 bp. Only three such modules were identified in the human genome, with only one displaying deep sequence conservation in vertebrates (humans, mice, chicken, frog, and fish), a pattern strongly reminiscent of that of the *PAX6* CE2 element. Remarkably, this deeply conserved module is embedded in an intron of *PAX5*, another member of the PAX family of transcription factors. Further analysis of the 196kb *PAX5* locus conservation identified 159 human/mouse ECRs that could potentially represent REs of this gene. The SRE identified in the 5th intron of *PAX5* overlaps with one of these 159 ECRs, 832bps long and 88% conserved between human and mouse. This ECR that has been conserved throughout the vertebrate lineage, including chicken, frog and fugu is the only ECR in this locus conserved between humans and fish. The module of 3 *PAX6* TFBS is located in the middle of this intronic ECR and the three TFBS are well conserved in human, mouse, chicken and frog lineages, with one of these three sites also conserved in fugu. These data suggest that the SynoR genome scan is sufficient for the identification of a critical *PAX5* regulatory element responsible for establishing proper gene regulation by the *PAX6* protein through a cis-regulatory unit synonymous to the autoregulatory element in *PAX6*. From a biological perspective, it is very important to understand the mechanisms of *PAX5* regulation as the strategies to downregulate *PAX5* expression in donor pro-B cells could be used to restore

T-cell development in patients with various immunodeficiencies, ranging from inherited syndromes to AIDS (31).. Our results suggest that the direct targeting of this genomic element or indirect targeting of the *PAX5* expression through decreasing the *PAX6* protein concentration might provide the means to achieve the sought after, clinically relevant, downregulation of *PAX5* expression.

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 (17). These data provide immediate, indicative 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 that SRE motif, identified by SynoR. To assess the applicability of SynoR tissue-specificity analysis of predicted SREs, we scanned the human genome for the combinatorial module of two cTFBS, *SRF* (serum response factor) and *SPI*. Multiple lines of evidence support the notion that these TFs cooperatively participate in orchestrating gene expression in heart and vascular tissues (32-36). We applied SynoR to predict targets of synergistic *SRF/SPI* gene regulation in the human genome using as a seed motif the presence of these TFBS separated by less than 40 bps and being conserved in human and mouse. One hundred fourteen noncoding modules were identified in this scan, twenty-three (20%) of which overlapping with promoter regions. Taking into account the density of human/mouse ECRs in the human genome (37), the probability of such a ratio of elements being in

promoters by chance is less than 10^{-5} , suggestive of 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 3). 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 of identified elements and search patterns. Together, these data support the notion that using this pair of cTFBS as seeds for a genome-wide scan successfully identifies SREs likely responsible for the shared pattern of expression of their corresponding genes. Further biochemical studies are required to assess the in vivo functional activity of these elements and investigate their possible role in cardio-vascular diseases.

Other SynoR features

Further functions and features associated with SynoR include GO classification of genes bracketing the identified noncoding modules, multi-species evolutionary conservation analysis of identified modules, and categorization of modules based on gene annotation (as promoter, UTR, intronic, intergenic or coding elements). To illustrate the applicability of the latter function, we scanned the human genome for a module of 3 NRSF human/mouse cTFBS. *NRSF* (neuron-restrictive silencer factor) plays a key role in neuronal differentiation (38) and mediation of transcriptional repression of neuron-specific genes in non-neuronal cells (39). Ten noncoding modules were identified, of

which 3 within promoters , 4 in introns, and 4 in intergenic intervals. One of the 3 promoters corresponds to that of *Barhl1*, a gene associated with neuronal migration (37), in an expression resembling that of the NRSF regulatory pathway. The remaining 2 promoters identified in this scan correspond to uncharacterized genes, and these results raise, thus, the possibility that these genes represent new members in the NRSF pathway.

METHODS

Genome-wide annotation of conserved TFBS.

The ECR Browser tool (<http://ecrbrowser.dcode.org>) generates whole genome *blastz*-based alignments of vertebrate and invertebrate genomes (12). It currently operates with the genomes of the human, mouse, rat, chicken, dog, frog, 3 fish (tetraodon, zebrafish, and fugu), and 6 *Drosophila* species. To generate a dataset of conserved TFBS for SynoR scans, we have established an automated annotation of evolutionary conserved TFBS based on the ECR Browser alignments. This was created by applying the rVista 2.0 tool (<http://rvista.dcode.org>) (29) with “optimized-for-function” position weight matrix (PWM) thresholds (40) to different pairwise genome alignments. Currently, the annotation of the conserved TFBS is available for the human genome in alignments with several other genomes including mouse, chicken, frog, and fugu as well as for the mouse genome in the alignment with the chicken genome. The automated ECR Browser/rVista 2.0 annotation processes other available genomes, gradually expanding the list of genome alignments with the conserved TFBS. Table 1 summarizes the number of conserved TFBS in the human genome as compared to different species. SynoR scans through these conserved TFBS to identify specific TFBS modules.

Defining TFBS modules as seeds for the genome scans.

DNA footprint of a regulatory TFBS module is a two-dimensional projection of the three-dimensional complex of TF molecules interacting with each other and with the chromosome 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: (1) a list of TFs known to participate in a particular regulatory pathway, (2) a set of spatial constraints separating different TFBS, and (3) 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. Previous studies on regulatory pathways and signaling networks might be sufficient to identify key players of a particular biochemical process and consequently to define the TF content. Previous characterization of a locus using tools similar to rVista 2.0 or multiTF (40) can establish a detailed structure of regulatory modules and relative order of TFBS. SynoR is limited in selection of TFBS to the list of TFBS available from the TFANSFAC database (41), which is utilized by the rVista 2.0 tool in genome scans.

Identification of statistically enriched GO categories.

To predict the putative biological function of the identified elements SynoR performs a stepwise GO classification of the host genes that either flank or contain the

noncoding subset of these elements. At the first step, the GO annotation is done by independently assigning a list of corresponding GO categories to each of these genes. Subsequently, the population counts are established for different GO categories – how many genes contain a particular GO category in their annotation. These GO category population counts are then compared to the population counts originating from all the genes in the base genome (with the limitation to the GO categories that have 10 or more population counts in the base genome). Finally, the program determines the GO categories that are statistically enriched in population counts as compared to what would be expected in a purely by chance manner.

In practice, first we count the expectation number of population counts for each i -th GO category as:

$$N_{ex}^i = N_{total}^i \cdot \frac{G_h}{G_{total}},$$

where G_h is the number of the host genes, G_{total} is the total number of genes in the base genome, and N_{total}^i is the total population count of the base genome corresponding to the i -th GO category (gene counts are given for nonoverlapping genes). Then, the population count of the host genes for the i -th GO category, N_h^i is compared to the N_{ex}^i to calculate the deviation z -value corresponding to the i -th GO category:

$$z_i = \frac{N_h^i - N_{ex}^i}{\sigma_i},$$

where the standard deviation σ_i is estimated using the Poisson distribution as $\sqrt{N_{ex}^i}$.

All the GO categories with the absolute z -value greater than 2.0 (that corresponds to the

less than 5% probability of the observation occurring simply by chance) are reported to the user as either statistically enriched or depleted (depending on the sign of the z -value).

Establishing tissue-specificities of identified genes.

To predict tissue-specificity (if any) of the identified genes with noncoding elements, SynoR analyzes the GNF Expression Atlas 2 (17) data that corresponds to these genes. 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 (42) and the results are visualized in a micro-array expression profile style profile (Figure 3). This allows a straightforward visual identification of subsets of genes co-expressed in particular tissues. 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 over 2.0 is reported as significantly increased/decreased, and in tissues with an absolute z -value over 1.0 as changed. It effectively allows prediction of tissue-specificity of the identified elements.

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 largest 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.

DISCUSSION

The identification of noncoding sequences conserved among vertebrates has served as the most important pillar leading to the identification of functional gene regulatory elements in the human genome (1-3,30). Nevertheless, the sheer degree of sequence conservation among mammals, associated with the time consuming nature of the functional assays designed to test these sequences preclude the ability to test all of these conserved noncoding elements. Moreover, recent data suggest that at least a sizeable fraction of these conserved sequences may not represent regulatory elements or be amenable for testing in the our current laboratory setting. Therefore, it has become essential to devise strategies aiming at the prioritization of a subset of conserved sequences for functional testing. The design and implementation of SynoR, a tool that allows for the identification of regulatory sequences with shared function represent an

important ancillary strategy to identify the conserved genomic elements that are most likely to be functional, and testable in various transcription assays.

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 lower eukaryotes such as yeast, worm and flies, and recently evidence suggested that SREs may also be identifiable in humans. Our results support this notion, and SynoR represents a publicly available tool for the search of SREs with a broad range of options in adopting the search to different regulatory pathways. SynoR is equipped with multiple mechanisms of functional annotation of identified elements, which include multi-species evolutionary conservation analysis, GO functional characterization, and GNF Expression Atlas 2 analysis of tissue specificity of the identified genes. These mechanisms allow quantifying the reliability of SynoR genome scans and allows dissecting the set of identified elements into subcategories 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 non-coding sequences that are most likely to correspond to regulatory elements, that 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.

WEB SITES

<http://synor.dcode.org/> - *SynoR* genome miner for synonymous regulation,

<http://rvista.dcode.org/> - *rVista* 2.0; identification of conserved TFBS in pairwise alignments,

<http://ecrbrowser.dcode.org/> - *ECR Browser*; alignment of multiple genomes, genome-wide annotation of conserved TFBS.

TABLES

Table 1. Conserved TFBS 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 (14)).

Organism	mouse	chicken	frog	fugu
# of conserved TFBS	13,069,048	1,945,164	859,769	402,784

FIGURES

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

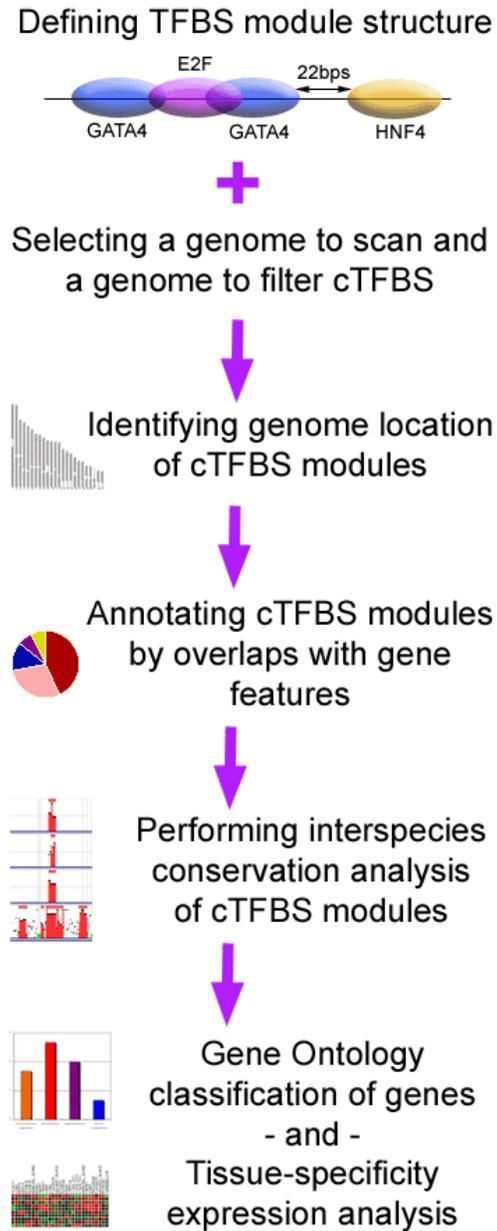


Figure 2. Human/mouse/chicken/frog/fugu conservation visualization of the PAX5 locus with a zoom into an ECR conserved in all the species (chr9:36,968,521-36,969,352; NCBI Build 35). A module of 3 conserved PAX6 binding sites is located in the middle of this ECR as depicted by yellow boxed in the zoom in panel. Alignments were obtained from the ECR Browser (<http://ecrbrowser.dcode.org>)

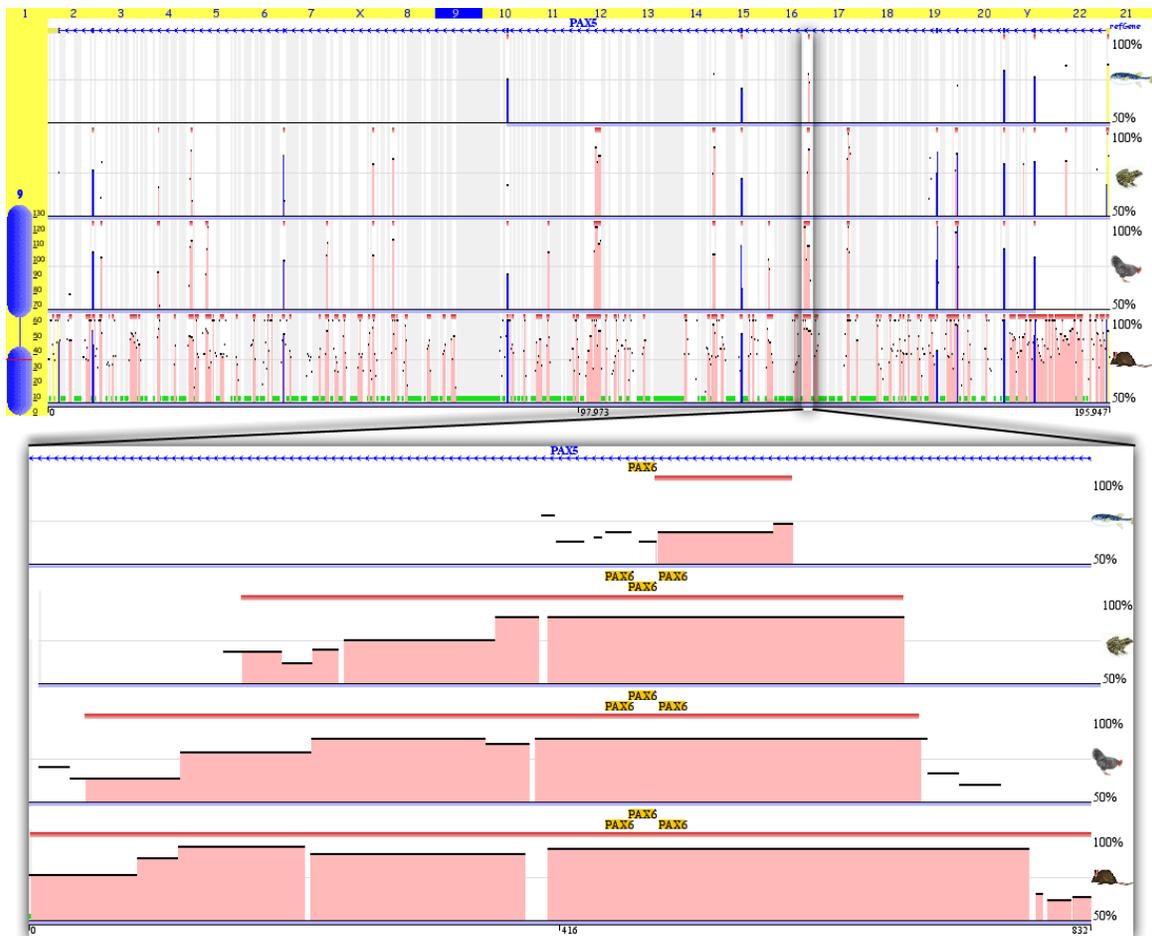
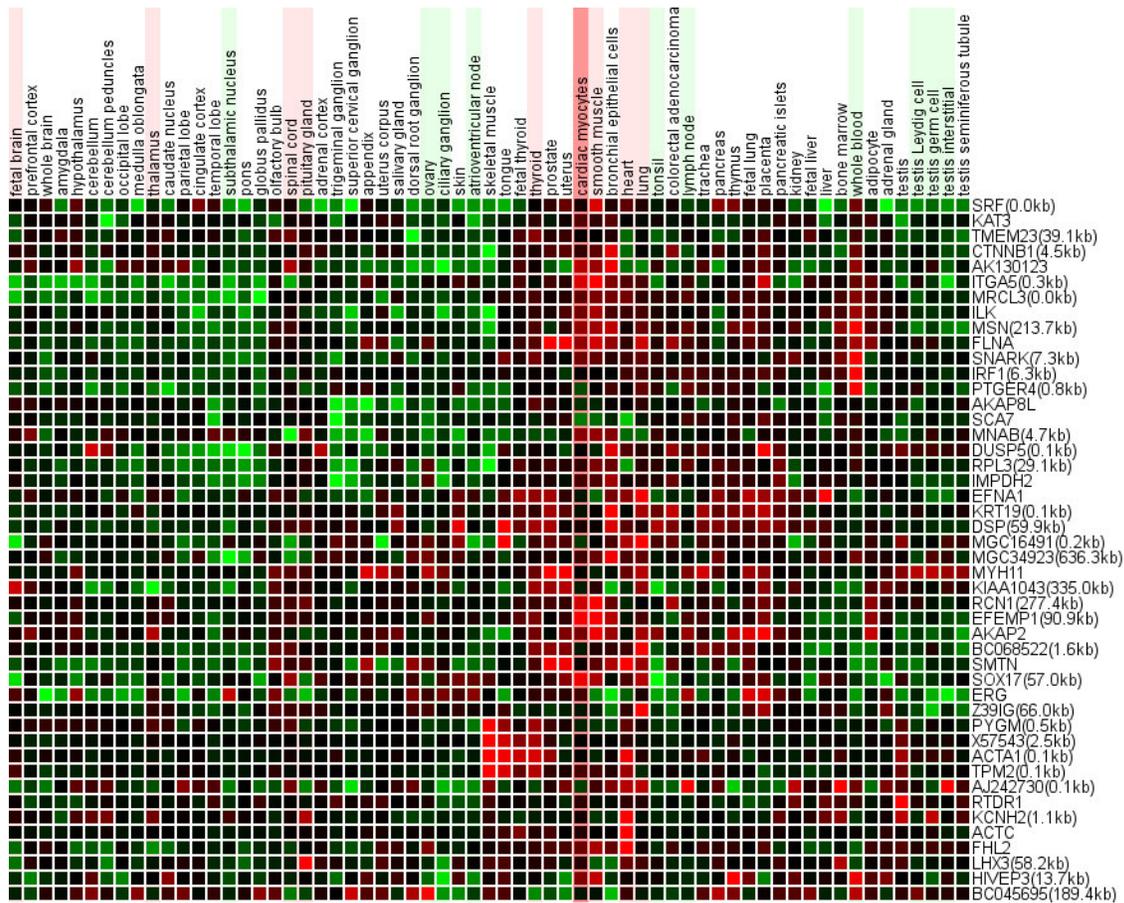


Figure 3. GNF Expression Atlas2 analysis for genes identified in the *SRF/SPI* SynoR scan of the human genome. A subset of 46 genes including the *SRF* gene is presented. Cardiac myocytes with significant overexpression identified by solid red background. Light red and light green backgrounds correspond to the overexpressed 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 parenthesis following gene name provides a distance between an element and the gene in case of intergenic elements.



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. Lettice, L.A., Heaney, S.J., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E. and de Graaff, E. (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet*, **12**, 1725-1735.
3. 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.
4. Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L. and Rubin, E.M. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, **299**, 1391-1394.
5. Ghanem, N., Jarinova, O., Amores, A., Long, Q., Hatch, G., Park, B.K., Rubenstein, J.L. and Ekker, M. (2003) Regulatory roles of conserved intergenic domains in vertebrate Dlx bigene clusters. *Genome Res*, **13**, 533-543.
6. Uchikawa, M., Takemoto, T., Kamachi, Y. and Kondoh, H. (2004) Efficient identification of regulatory sequences in the chicken genome by a powerful combination of embryo electroporation and genome comparison. *Mech Dev*, **121**, 1145-1158.
7. 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.
8. Erives, A. and Levine, M. (2004) Coordinate enhancers share common organizational features in the Drosophila genome. *Proc Natl Acad Sci U S A*, **101**, 3851-3856.
9. Senger, K., Armstrong, G.W., Rowell, W.J., Kwan, J.M., Markstein, M. and Levine, M. (2004) Immunity regulatory DNAs share common organizational features in Drosophila. *Mol Cell*, **13**, 19-32.
10. 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.
11. 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*.
12. 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**, W280-286.
13. Loots, G.G. and Ovcharenko, I. (2005, in press) Dcode.org anthology of comparative genomic tools. *Nucleic Acids Res*, **33**.
14. 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.

15. 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. *Nat Genet*, **25**, 25-29.
16. Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, **32 Database issue**, D258-261.
17. 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 U S A*, **101**, 6062-6067.
18. Miguel-Aliaga, I., Allan, D.W. and Thor, S. (2004) Independent roles of the dachshund and eyes absent genes in BMP signaling, axon pathfinding and neuronal specification. *Development*, **131**, 5837-5848.
19. 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.
20. Walshe, J. and Mason, I. (2003) Unique and combinatorial functions of Fgf3 and Fgf8 during zebrafish forebrain development. *Development*, **130**, 4337-4349.
21. Plageman, T.F., Jr. and Yutzey, K.E. (2004) Differential expression and function of Tbx5 and Tbx20 in cardiac development. *J Biol Chem*, **279**, 19026-19034.
22. 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.
23. Takeuchi, J.K., Ohgi, M., Koshiba-Takeuchi, K., Shiratori, H., Sakaki, I., Ogura, K., Saijoh, Y. and Ogura, T. (2003) Tbx5 specifies the left/right ventricles and ventricular septum position during cardiogenesis. *Development*, **130**, 5953-5964.
24. Elkon, R., Zeller, K.I., Linhart, C., Dang, C.V., Shamir, R. and Shiloh, Y. (2004) In silico identification of transcriptional regulators associated with c-Myc. *Nucleic Acids Res*, **32**, 4955-4961.
25. Troen, G., Nygaard, V., Jenssen, T.K., Ikonomou, I.M., Tierens, A., Matutes, E., Gruszka-Westwood, A., Catovsky, D., Myklebost, O., Lauritzsen, G. *et al.* (2004) Constitutive expression of the AP-1 transcription factors c-jun, junD, junB, and c-fos and the marginal zone B-cell transcription factor Notch2 in splenic marginal zone lymphoma. *J Mol Diagn*, **6**, 297-307.
26. Firlej, V., Bocquet, B., Desbiens, X., de Launoit, Y. and Chotteau-Lelievre, A. (2004) Pea3 transcription factor cooperates with USF-1 in regulation of the murine bax transcription without binding to an Ets binding site. *J Biol Chem*.
27. Jensen, A.M. (2004) Potential roles for BMP and Pax genes in the development of iris smooth muscle. *Dev Dyn*, **232**, 385-392.
28. 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.

29. Loots, G.G. and Ovcharenko, I. (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res*, **32**, W217-221.
30. Kleinjan, D.A., Seawright, A., Childs, A.J. and van Heyningen, V. (2004) Conserved elements in Pax6 intron 7 involved in (auto)regulation and alternative transcription. *Dev Biol*, **265**, 462-477.
31. Mikkola, I., Heavey, B., Horcher, M. and Busslinger, M. (2002) Reversion of B cell commitment upon loss of Pax5 expression. *Science*, **297**, 110-113.
32. Wang, S.X., Elder, P.K., Zheng, Y., Strauch, A.R. and Kelm, R.J., Jr. (2004) Cell cycle-mediated regulation of smooth muscle alpha -actin gene transcription in fibroblasts and vascular smooth muscle cells involves multiple adenovirus E1A-interacting cofactors. *J Biol Chem*.
33. Madsen, C.S., Regan, C.P. and Owens, G.K. (1997) Interaction of CArG elements and a GC-rich repressor element in transcriptional regulation of the smooth muscle myosin heavy chain gene in vascular smooth muscle cells. *J Biol Chem*, **272**, 29842-29851.
34. Barron, M.R., Belaguli, N.S., Trinh, M., Iyer, D., Lough, J.W., Parmacek, M.S., Bruneau, B.G. and Schwartz, R.J. (2004) Serum Response Factor, An Enriched Cardiac Mesoderm Obligatory Factor, Is A Downstream Gene Target For Tbx Genes. *J Biol Chem*.
35. 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 U S A*, **101**, 17132-17137.
36. 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.
37. Ovcharenko, I., Stubbs, L. and Loots, G.G. (2004) Interpreting mammalian evolution using Fugu genome comparisons. *Genomics*, **84**, 890-895.
38. 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.
39. 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.
40. 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.
41. 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.
42. 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 U S A*, **95**, 14863-14868.