BMB
reports

Predicting tissue-specific expressions based on sequence characteristics

Hyojung Paik[1,2], *Taewoo Ryu*[3], *Hyoung-Sam Heo*[1], *Seung-Won Seo*[1,4], *Doheon Lee*[2,*] *& Cheol-Goo Hur*[1,4,*]

[1]Plant Systems Engineering Center, KRIBB, Daejeon, [2]Department of Bio and Brain Engineering, KAIST, Daejeon, Korea, [3]Red Sea Laboratory for Integrative Systems Biology, Computational Biosciences Research Center, Division of Chemical & Life Sciences and Engineering, King Abdullah University of Science and Technology (KAUST), Kingdom of Saudi Arabia, [4]University of Science and Technology (UST), Daejeon, Korea

In multicellular organisms, including humans, understanding expression specificity at the tissue level is essential for interpreting protein function, such as tissue differentiation. We developed a prediction approach via generated sequence features from overrepresented patterns in housekeeping (HK) and tissue-specific (TS) genes to classify TS expression in humans. Using TS domains and transcriptional factor binding sites (TFBSs), sequence characteristics were used as indices of expressed tissues in a Random Forest algorithm by scoring exclusive patterns considering the biological intuition; TFBSs regulate gene expression, and the domains reflect the functional specificity of a TS gene. Our proposed approach displayed better performance than previous attempts and was validated using computational and experimental methods. [BMB reports 2011; 44(4): 250-255]

INTRODUCTION

Tissue-specific (TS) genes are exclusively expressed only in one or a few tissue types (1), in which the encoded proteins play functional roles specific to their expressing tissues. Conversely, housekeeping (HK) genes are constantly expressed across a broad range of tissues (1) and are responsible for maintaining vital functions, such as the cell cycle and energy metabolism (2). Tissue-specific (TS) genes and proteins contribute to biological phenomena, which occur in specific tissues including tissue development, aging symptoms (3), and metabolic disease (4). Precisely identifying the expressed tissues of a gene is fundamental for understanding related protein function due to the tissue-dependency of protein interactions (5) and metabolic pathways (6). Unraveling these biological networks has attracted great interest, whereas little is known about when and where each of these interactions occurs. Recent studies have highlighted the importance of TS/HK interactions in biological networks (5), so identifying TS/HK gene expression in humans is necessary to develop a detailed understanding of the biological mechanisms.

Several experimental methods have been used to identify expression specificity, including microarrays and next-generation sequencing methods (7). In terms of statistical methods, Audic's test (8) applies the statistical assumption that the observation frequency of expressed sequence tags (ESTs) in a putative TS gene is consistent with the tissue ratio of ESTs in an entire population. Based on this assumption, our previously reported method of tissue-specific alternative splicing (TISA) used Audic's test to assess gene expression tissue specificity in a public dataset (9). This statistical method allows TS genes to be partially predicted without the experimental efforts described above, but the usefulness of Audic's approach is limited by the biased observation of transcripts from various tissues. Moreover, these data-driven approaches are inappropriate to meet the hottest issue of gene expression studies; identifying sequence features that contribute to the gene expression landscape (10, 11).

HK and TS genes have several distinct features, including their CpG island frequency, gene length, gene ontology (GO) terms, evolutionary rates , and regulatory modules (e.g., TFBSs) (12-16). Several of these features have previously been used to classify HK and TS genes computationally. For example, regulatory modules, such as *cis*-acting regulatory elements, have been used to predict TS genes (17). Other sequence features have been used to identify HK genes, such as exon number and gene length (18). These promising results show that the identification and machine learning of HK/TS gene sequence features may shed light on the underlying features that define expression specificity.

Here, we generated sequence features using weighted indexing and stochastic methods and conjugated the random forest (RF) algorithm (19) to establish a multiclass classifier capable of distinguishing HK genes and of classifying TS genes to

*Corresponding author. Doheon Lee, Tel: 82-42-350-4316; Fax: 82-42-350-8680; E-mail: dhlee@kaist.ac.kr, Cheol-Goo Hur, Tel: 82-42-879-8560; Fax: 82-42-869-8569; hurlee@kribb.re.kr
DOI 10.5483/BMBRep.2011.44.4.250

Received 29 December 2010, Accepted 20 January 2011

Keywords: Domain, Housekeeping, Random forest, Tissue-specific, Transcription factor binding site

their expressing tissues in humans. We called this computational pipeline the multiclass classification of tissue-specific genes (McTis). The classifier comprises two parts: a first step in which HK and TS genes are distinguished, and a second step in which the TS genes are assigned to their target tissues using a multiclass classification. To validate the performance of McTis, we used the reverse transcription-polymerase chain reaction (RT-PCR) to compare the McTis results experimentally with those from an independent set derived from TS gene related database, The Catalog of Tissue-Specific Regulatory Motifs (TCat) (20) and tissue-specific alternative splicing analogs (TISA) (9) data.

RESULTS

Dataset construction

To ensure the selection of appropriate human HK genes, we included only genes identified as HK genes in at least two of the three reports utilized (16, 21, 22). We prepared 2,024 candidate TS genes from three other studies (23-25). We experimentally confirmed the tissue specificity of each candidate TS gene using RT-PCR to minimize the possibility for false positives trainings with the TS genes (see Methods). Our final TS dataset consisted of 143 genes expressed as follows: 14 in brain, 15 in kidney, 70 in liver, 15 in lung, 15 in testis, and 14 in the gastrointestinal tract (GI) (stomach, small intestine, and colon) (Supplemental Table 1). These 279 genes, comprising 136 HK and 143 TS genes, were finally prepared for further analysis. To prevent confusion with the set derived from TCat and TISA, we refer to our set as "the dataset" and the other set as "the independent set".

First step: classification of HK genes using sequence characteristics

McTis consists of two different classifiers (Supplemental Fig. 1). The first separates HK and TS genes and the second classifies TS genes according to the target tissues. In the first step, gene length and the number of upstream CpG islands were analyzed. Although previous studies have considered various HK gene patterns, a statistical analysis suggested that gene length and CpG count were significant features ($P < 0.05$). According to these results and consistent with previous studies (12, 16), gene length and the number of CpG islands were used to separate the HK genes in the first round of classification. The error rate of the first classifier in McTis was estimated to be 29.5% using out-of-bag (OOB) data.

Second step: multiclass classification of TS genes using generated sequence indices

For the multiclass classification of the TS genes in the second step, we developed indexing features for tissue-specific TFBSs and domains, which are related to the regulation of tissue-specific expression and its functional roles, respectively (15, 17, 26). As expected, we observed that the tissue-specific domains and TFBSs of the TS genes in our dataset were compatible with the characteristic functions and expression specificities of these genes and their encoded proteins. For example, a gene with a GABA-B receptor domain showed brain-specific function, and a gene with the TFBS of *Olf1* suggested brain-specific regulation of neuronal expression with brain-specific expression (Supplemental Table 2). Therefore, we were able to identify the tissue-specific expression and functional characteristics of the i^{th} gene using a developed domain and TFBS indices, thereby training the input vectors of the dataset for the multiclass classification of TS genes.

The functional relationships between the expressing tissues and the identified domains of the TS genes were supported by calculating Jaccard's index (27), a measure of functional similarity. As depicted in Table 1, we defined the target-expressing tissues as being functionally related if the comparison had a P value less than the Bonferroni corrected P value (8E-03). Jaccard's index for kidney-specific and liver-specific genes was 0.09, with a significant P value (2E-04), which was consistent with the functional relatedness of the liver and kidney (28-30). Similarly, the brain- and testis-specific gene domains reflected the developmental interdependence of the brain and testis (31). Conversely, the nonsignificant P value for Jaccard's index of the GI tract and testis (2E-2) was consistent with their relative lack of functional relatedness.

After training these sequence characteristics with the McTis system, the classification errors for the tested target tissues were 7% for the brain, 7% for the GI tract, 13% for the kidney, 0% for the liver, 13% for the lung, and 6% for the testis. Consequently, the average error rate was estimated to be 4.9% over the six target tissues. All error rates were estimated with OOB data obtained from 500 iterations of random tree growing.

Experimental validation

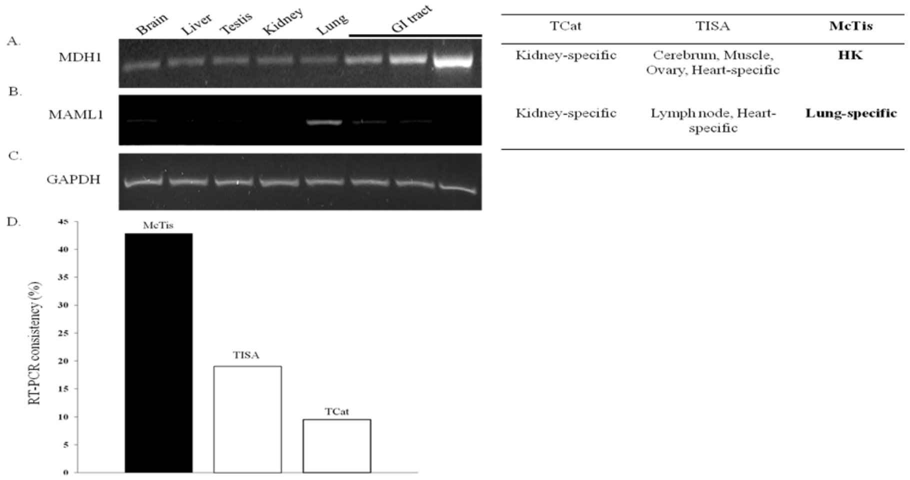
Because the newly developed indices were generated by a dataset that was subsequently tested using the indices, we next sought to validate our method using an independent set. Although relatively few TS gene datasets are available, we regarded a set of TS genes previously predicted by TCat as a reliable independent set of TS genes (20). An independent set consisting of 15 kidney-specific and six liver-specific genes was prepared using the TCat voting score. We also used our previous TISA study (9) to predict TS genes from the same test set and to compare the results obtained from the McTis and Audic's test. However, it must be noted that the independent set came from studies using TCat, which infers TS genes using microarray analysis, dbEST information, and GO terms and TISA, which utilize Audic's test. We compared our direct experimental results for the independent set with those obtained from TCat, TISA, and McTis, and with direct experimental confirmations from RT-PCR.

Our RT-PCR analyses were consistent with the McTis predictions for these genes (Fig. 1). In contrast, both genes were

Table 1. Jaccard's index P values for domain similarities

	Brain	GI tract	Kidney	Liver	Lung	Testis
Brain	0 (58)*	2E-02 (1)*	3E-02 (3)*	2E-02 (3)*	3E-02 (2)*	5E-04 (11)*
GI tract		0 (115)*	4E-02 (5)*	2E-02 (17)*	8E-03 (8)*	2E-02 (1)*
Kidney			0 (82)*	2E-04 (29)*	2E-02 (2)*	2E-02 (1)*
Liver				0 (254)*	2E-02 (3)*	2E-02 (1)*
Lung					0 (65)*	2E-02 (5)*
Testis						0 (67)*

The values shown in bold indicate significant domain similarities based on the Bonferroni corrected P value (<8E-03). *Number of common domains.



TCat	TISA	McTis
Kidney-specific	Cerebrum, Muscle, Ovary, Heart-specific	**HK**
Kidney-specific	Lymph node, Heart-specific	**Lung-specific**

Fig. 1. Comparison of RT-PCR, multiclass classification of tissue-specific genes (McTis), tissue-specific alternative splicing (TISA), and The Catalog of Tissue-Specific Regulatory Motifs (TCat) predictions. (A, B) The RT-PCR results for the prediction validations from McTis, TISA, and TCat. (C) The loading control for the RT-PCRs. (D) Overall consistency between the RT-PCR results and predictions from McTis, TISA, and TCat.

classified incorrectly by TCat and TISA. In the first step of McTis, 11 of 21 genes were classified as HK genes, and seven of these predictions were compatible with the RT-PCR results. The remaining 10 genes (putative TS genes) were further classified in the second step. Two of these 10 genes were classified correctly according to the RT-PCR results, for a total consistency rate of 42% (9 of 21). TCat and TISA correctly identified only one and four of the 21 genes, respectively, producing consistency rates of 9% and 19%, respectively, indicating that McTis was approximately 33% and 23% more accurate than TCat and TISA, respectively (Fig. 1D). Overall, our experimental results indicate that the RF training of sequence features (gene length and CpG pattern) and our new sequence features (the TFBS and domain indices) distinguish HK genes and classify TS genes into multiple classes with greater precision than previous methods.

DISCUSSION

We established a prediction scheme to multiclassify TS expressions via developed scoring functions derived from sequence features (the TFBS and domain indices). The classification process, McTis, was configured with a stochastic analysis of sequence features and indexing approaches and consisted of two steps: 1) distinguishing HK genes and 2) the multiclass classification of TS genes. The error rates for McTis were estimated to be 29% and 4.9% for the first and second steps, respectively. The RT-PCR results showed that the McTis predictions were more accurate than other methods (33% and 19% improvement comparing the TCat and TISA results).

Although identifying overrepresented TFBSs is still a challenging problem (32), successful prediction (a 4.9% error rate) was achieved using the indexing McTis method. Future studies should assess TS expression and develop a more advanced algorithm to identify TFBSs to refine our indexing approach considering overfitting issues from sample observations. McTis indexing approaches classify TS genes based on an unbiased characterization of domains/TFBSs in a non-parametric manner and the trained target expressing tissues. Therefore, our concept of patterning and weighting indexing approaches suggests hopeful results to identify tissue-specific regulatory patterns (TFBS), including brain and kidney-specific TFBS (33) as depicted Table 2.

The indices identified by McTis were created from observed (not predicted) TS genes. Therefore, McTis is not limited by the Audic's test drawback, which is a potential discrepancy between statistical assumptions and observed data. Moreover the features are sequence derived, so McTis does not suffer from experimental limitations. In conclusion, we have reported a framework for the multiclass classification of TS expression sequence characteristic indices. To the best of our knowledge, this is the first study to attempt such a multiclass classification of TS expressions using sequence characteristics. This technique may prove useful for the high-throughput classification of TS genes and to uncover the contribution of regulatory elements and functional domains.

MATERIALS AND METHODS

Dataset preparation

The HK and TS gene datasets were prepared using six previous studies (16, 21-25). The tissue specificities of the TS genes in the dataset were confirmed by RT-PCR to minimize errors in the training data. Details of the RT-PCR validation and criteria for selecting HK and TS genes are presented in Supplemental Data (Supplemental Fig. 2 and Supplemental Tables 3, 4 and 5).

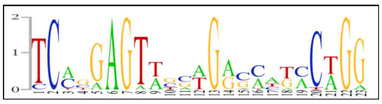
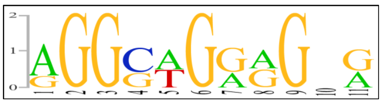
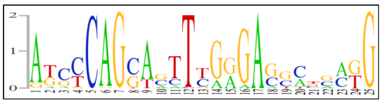
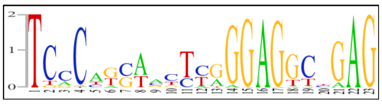
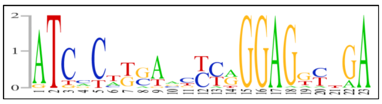
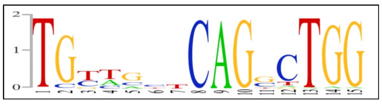
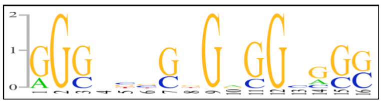
Exploring the sequence features for McTis training

In the first step of McTis, putative HK and TS genes were distinguished based on gene length and the number of CpG islands in the upstream region (−5,000 bp). CpGproD (34) was used to predict the CpG islands. In the second step of McTis, the TS genes were further classified based on their domain and TFBS indices. InterProScan (35) and AlignACE (36) were used to analyze the domains and TFBSs, respectively, and the results were transformed into domain and TFBS indices.

Random forest (RF)

McTis was implemented using the RF algorithm. Briefly, the RF algorithm (19) produces many classification trees from randomly selected datasets and populates the resulting forest with these trees. Because the test set error is estimated internally with Out-Of-Bag (OOB) data during tree drawing, we considered that the OOB error rates represented the accuracy of the classifier without cross-validation.

Table 2. Tissue-specific transcriptional factor binding sites

Expressed tissue	Pattern*	TRANSFAC	
		Description†	Score‡
Liver		· HNF-1 (M01011) · TCat presented liver- specific TFBS	2.2
Kidney		· Zic3 · M00450 · GL1 protein related · Related to renal morphogenesis (33)	1.5
Brain		· M00138 · Brain octamer-protein related TFBS	2.3
Brain		· Olf-1 · M00261 · DNA binding site for direct neuronal express	2.2
Lung		· STAT (M00224) · Unknown lung-specific TFBS	2.1
GI tract		· Muscle initiator (M00321) · Unknown GI tract- specific TFBS	1.8
Testis		· Spz1 (M00446) · Novel bHLH-Zip protein, testis-specific expressions	1.5

*Presented by WebLogo (http://weblogo.berkeley.edu) with the out file of AlignACE, †Annotation of TRANSFAC; Official symbol of TFBS, ID of TRANSFAC, ‡MatCompare scoring results.

Indexing feature generation

In the second step of McTis, we trained the input vector of the i^{th} gene, g_i. We assumed that the j^{th} tissue-specific domains and TFBSs have distinct characteristics for regulating tissue-specific expression and functional protein roles. Let C_j represent the j^{th} expressing tissue, and let g_i denote one of the target tissues among C_j. Let the following hold:

$$g_i \leftarrow (X_i)$$

$$\textit{where } X_i \text{ is the input vector of } g_i;$$

$$\textit{for } C_j,\ 1 \leq j \leq \textit{total number of target tissues; and}$$

$$\textit{for } g_i,\ 1 \leq i \leq \textit{total number of genes.}$$

To generate the j^{th} tissue-specific domains and TFBSs, we analyzed the j^{th} tissue-specific domains and TFBSs from our dataset using InterproScan and AlignACE, as described above. If we let D_j and T_j represent the j^{th} tissue-specific domains and TFBSs, respectively, then:

$$D_j = \{d_{j1}, d_{j2} \cdots d_{jp}\}, \text{ for } j \textit{ where } 1 \leq p \leq \textit{total number of } j^{th} \textit{ tissue-specific domains; and}$$

$$T_j = \{t_{j1}, t_{j2} \cdots t_{jq}\}, \text{ for } j \textit{ where } 1 \leq q \leq \textit{total number of } j^{th} \textit{ tissue-specific TFBSs.}$$

We used the MAP score obtained from the AlignACE analysis to identify t_{jq}. The higher MAP score of AlignACE presents significance of predicted motif. In this study, t_{jq} yielded MAP scores larger than the average MAP score for the j^{th} tissue-specific TFBSs. However, d_{jp} showed a heterogeneous distribution throughout the j^{th} tissue-specific genes. Therefore, we weighted d_{jp} with its frequency of observation. If we let k_{jp} be the number of d_{jp} genes observed among the j^{th} tissue-specific genes and i_{jp} be the copy number of d_{jp} within g_i, then the domain index of j^{th} tissue specificity can be expressed as:

$$D_{ij} = \sum i_{jp} \cdot k_{jp} \cdot d'_{jp} \qquad [1]$$

$$d'_{jp} = \begin{cases} 1, \textit{ if } d_{jp} \text{ is assigned to } g_i \\ 0, \textit{ otherwise} \end{cases}$$

Similarly, the TFBS index was configured with weighting factors according to the following equation:

$$T_{ij} = \sum i_{jq} \cdot k_{jq} \cdot t'_{jp} \qquad [2]$$

$$t'_{jp} = \begin{cases} 1, \textit{ if } t_{jp} \text{ is assigned to } g_i \\ 0, \textit{ otherwise} \end{cases}$$

When the total number of target tissues is j, X_i consists of the j-dimensional vector of the domain (D_{ij}) and TFBS (T_{ij}) indices, respectively; that is,

$$x_i = ((D_{i1}, D_{i2} \cdots . D_{ij}), (T_{i1}, T_{i2} \cdots T_{ij})).$$

Analysis of generated sequence indices

A position frequency matrix T_j was generated using the AlignACE output to analyze the j^{th} tissue-specific TFBS. The tissue-specific motifs in the upstream regions were compared with known TFBSs from TRANSFAC using MatCompare (17, 37) and were represented using WebLogo (http://weblogo.berkeley.edu). The functional similarity between the target tissues was analyzed by domain comparisons. Jaccard's index is a statistic used to compare the similarities of sample sets (27). We defined the sample sets as the domains of the target tissue-specific genes. If we let J_{ab} be Jaccard's index for expressing tissues *a* and *b*, then:

$$J_{ab} = \frac{N_{ab}}{N_a + N_b - N_{ab}} \qquad [4]$$

where N_a and N_b represent the number of distinct domains of tissues *a* and *b*, respectively, and N_{ab} denotes the number of domains commonly assigned to both tissues. In this case, the value of Jaccard's index increases as the domain similarity between the TS gene sets increases. A background model was constructed with random permutations of the NCBI RefSeq database of human proteins to evaluate Jaccard's index statistically.

Acknowledgments

This work was supported by grants from an Omics-Based Integration Database for Cancer Interpretation (KGM1161012 and KGM0661113) funded by the Korea Research Institute of Bioscience and Biotechnology (KRIBB). DL was supported by the World Class University program (R32-2008-000-10218-0), the Korean Systems Biology Research Project (20100002164), and the Basic Research Laboratory grant (2009-0086964) of the Ministry of Education, Science and Technology through the National Research Foundation of Korea.

REFERENCES

1. Zhu, J., He, F., Hu, S. and Yu, J. (2008) On the nature of human housekeeping genes. *Trends Genet.* **24**, 481-484.
2. Butte, A. J., Dzau, V. J. and Glueck, S. B. (2001) Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues". *Physiol. Genomics* **7**, 95-96.
3. Andersson, T., Simonyte, K., Andrew, R., Strand, M., Buren, J., Walker, B. R., Mattsson, C. and Olsson, T. (2009) Tissue-specific increases in 11beta-hydroxysteroid dehydrogenase type 1 in normal weight postmenopausal women. *PLoS One* **4**, e8475.
4. Sopasakis, V. R., Liu, P., Suzuki, R., Kondo, T., Winnay, J., Tran, T. T., Asano, T., Smyth, G., Sajan, M. P., Farese, R. V., Kahn, C. R. and Zhao, J. J. (2010) Specific roles of the p110alpha isoform of phosphatidylinsositol 3-kinase in hepatic insulin signaling and metabolic regulation. *Cell Metab.* **11**, 220-230.
5. Bossi, A. and Lehner, B. (2009) Tissue specificity and the human protein interaction network. *Mol. Syst. Biol.* **5**, 260.
6. Shlomi, T., Cabili, M. N., Herrgard, M. J., Palsson, B. O. and Ruppin, E. (2008) Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.* **26**, 1003-1010.
7. Morozova, O. and Marra, M. A. (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**, 255-264.
8. Stephan, A. and Jean-Michel, C. (1997) The significance

of digital gene expression profiles. *Genome Res.* **7**, 986-995.

9. Noh, S.-J., Lee, K., Paik, H. and Hur, C.-G. (2006) TISA: tissue-specific alternative splicing in human and mouse genes. *DNA Res.* **13**, 229-243.
10. Rao, A., Hero, A. O., 3rd, States, D. J. and Engel, J. D. (2007) Motif discovery in tissue-specific regulatory sequences using directed information. *EURASIP J. Bioinform. Syst. Biol.* **2007**, 13853.
11. Farre, D., Bellora, N., Mularoni, L., Messeguer, X. and Alba, M. M. (2007) Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biol.* **8**, R140.
12. Yamashita, R., Suzuki, Y., Sugano, S. and Nakai, K. (2005) Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. *Gene* **350**, 129-136.
13. Zhang, L. and Li, W.-H. (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol. Biol. Evol.* **21**, 236-269.
14. Zhixi, S., Yong, H. and Xun, G. (2007) Tissue-driven hypothesis with gene ontology (GO) analysis. *Annals of Biomedical Engineering* **35**, 1088-1094
15. Cohen-Gihon, I., Lancet, D. and Yanai, I. (2005) Modular genes with metazoan-specific domains have increased tissue specificity. *Trends Genet.* **21**, 210-213.
16. Eisenberg, E. and Levanon, E. Y. (2003) Human housekeeping genes are compact. *Trends Genet.* **19**, 362-365.
17. Smith, A. D., Sumazin, P., Xuan, Z. and Zhang, M. Q. (2006) DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *PNAS* **103**, 6275-6280.
18. Ferrari, L. D. and Aitken, S. (2006) Mining housekeeping genes with a Naive Bayes classifier. *BMC Genomics* **7**, 277.
19. Breiman, L. (2001) Random forests. *Machine Learning* **45**, 5-32.
20. Smith, A. D., Sumazin, P. and Zhang, M. Q. (2007) Tissue-specific regulatory elements in mammalian promoters. *Mol. Sys. Biol.* **3**, 73.
21. Hsiao, L.-L., Dangond, F., Yoshida, T., Hong, R., Clark, R. V., Haverty, P., Weng, Z., Mutter, G. L., Frosch, M. P., Donald, M. E. M., Milford, E. L., Crum, C. P., Bueno, R., Pratt, R. E., Mahadevappa, M., Warrington, J. A. and Stephanopoulos, G. (2001) A compendium of gene expression in normal human tissues. *Physiol. Genomics* **7**, 97-104.
22. Warrington, J. A., Nair, A., Mahadevappa, M. and Tsyganskaya, M. (2000) Comparision of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol. Genomics* **2**, 143-147.
23. Ge, X., Yamamoto, S., Tsutsumi, S., Midorikawa, Y., Ihara, S., Wang, S. M. and Aburatani, H. (2005) Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics.* **86**, 127-141.
24. Pao, S.-Y., Lin, W.-L. and Hwang, M.-J. (2006) In silico identification and comparative analysis of differnetially expressed genes in human and mouse tissues. *BMC Genomics* **21**, 86.
25. Shyamsundar, R., Kim, Y. H., Higgins, J. P., Montgomery, K., Jorden, M., Sethuraman, A., van de Rijin, M., Botstein, D., Brown, P. O. and Pollack, J. R. (2005) A DNA microarray survey of gene expression in normal human tissues. *Genome Biol.* **6**, R22.
26. Zadissa, A., McEwan, J. C. and Brown, C. M. (2007) Inference of transcriptional regulation using gene expression data from the bovine and human genomes. *BMC Genomics* **8**, 265.
27. Lin, K., Zhu, L. and Zhang, D.-Y. (2006) An initial strategy for comparing proteins at the domain architecture level. *Bioinformatics* **22**, 2081-2086.
28. Schwartz, D. T. (1967) The relation of cirrhosis of the liver to renal hypertension. A review of 639 autopsied cases. *Ann. Intern. Med.* **66**, 862-869.
29. Troup, G. M., Wagner, I. and Walford, R. L. (1966) Liver pigment, liver histidase, and renal lysozyme changes in relation to age in normal and irradiated Syrian hamsters. *Radiat. Res.* **29**, 489-498.
30. Gadano, A., Moreau, R., Heller, J., Chagneau, C., Vachiery, F., Trombino, C., Elman, A., Denie, C., Valla, D. and Lebrec, D. (1999) Relation between severity of liver disease and renal oxygen consumption in patients with cirrhosis. *Gut* **45**, 117-121.
31. Malorni, W., Barcellona, P. S., Campana, A. and De Martino, C. (1981) Relationship between brain cortex and testis maturation rate in two inbred strains of mice. *Ric. Clin. Lab.* **11**, 247-257.
32. Wang, Y., Zhang, X. S. and Xia, Y. (2009) Predicting eukaryotic transcriptional cooperativity by Bayesian network integration of genome-wide data. *Nucleic Acids Res.* **37**, 5943-5958.
33. Hu, M. C., Mo, R., Bhella, S., Wilson, C. W., Chuang, P. T., Hui, C. C. and Rosenblum, N. D. (2006) GLI3-dependent transcriptional repression of Gli1, Gli2 and kidney patterning genes disrupts renal morphogenesis. *Development* **133**, 569-578.
34. Ponger, L. and Mouchiroud, D. (2002) CpGProD: Identifying CpG islands associated with transciption start sites in large genomic mammalian sequences. *Bioinformatics* **18**, 631-633.
35. Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. and Lopez, R. (2005) InterProScan: Protein domains identifier. *Nucleic Acids Res.* **33**, 116-120.
36. Hughes, J. D., Estep, P. W., Tavazoie, S. and Church, G. M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae. *J. Mol. Biol.* **296**, 1205-1214.
37. Matys, V., Fricke, E., Geffers, R., Gössling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D.-U., Land, S., Lewicki-Potapov, B., Michael, H., Münch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. and Wingender, E. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31**, 374-378.