Available online at www.sciencedirect.com

SCIENCE DIRECT®

Genomics 87 (2006) 433 – 436

GENOMICS

www.elsevier.com/locate/ygeno

Short Communication

Different classes of tissue-specific genes show different levels of noncoding conservation

Sun Shim Choi, Eliot C. Bush, Bruce T. Lahn *

Howard Hughes Medical Institute, Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA

Received 21 June 2005; accepted 21 September 2005
Available online 21 November 2005

Abstract

We divide tissue-specific genes into two major classes: regulators, defined as genes participating in tissue-specific transcriptional regulation, and effectors, defined as genes involved in rendering the physiological properties of cells. We show that regulators tend to have significantly greater noncoding conservation than effectors. We further show that within the regulator class, tissue-specific transcription factors generally have the greatest noncoding conservation, whereas signal receptors generally have the least noncoding conservation. Using noncoding conservation as a proxy for the complexity of *cis*-regulatory DNA, we extrapolate that different classes of tissue-specific genes tend to have different levels of *cis*-regulatory complexity and that greater complexity can be found in genes involved in transcriptional regulation, especially transcription factors.

Keywords: Noncoding conservation; Cis-regulatory region; Gene expression; Regulatory network

The *cis*-regulatory regions of genes are the primary sites where information from the cell's regulatory network is integrated to impart gene expression patterns [1]. It is reasonable to suppose that genes with more complex expression patterns might require more complex *cis*-regulatory elements. Several recent studies argue that this may indeed be the case. Nelson and colleagues showed that genes with more complex expression patterns also tend to have longer intergenic sequences and, by inference, more *cis*-regulatory elements [2]. Iwama and Gojobori recently showed that genes encoding transcription factors, which are often expressed in complex ways, tend to have more conserved noncoding regions than other types of genes [3].

However, it is also conceivable that two genes with the same expression pattern can nevertheless have *cis*-regulatory regions of vastly different complexity. For example, a gene can have a highly sophisticated *cis*-regulatory region that integrates information from several regulatory pathways to realize its complex expression pattern, whereas another gene can have a very simple *cis*-regulatory region that accomplishes the same level of expression complexity if it is activated directly by the first gene. Thus, the sophistication of a gene's *cis*-regulatory region may correlate not only with the complexity of the gene's expression pattern, but also with where the gene lies in the regulatory pathway.

Here, we address this issue further by examining whether tissue-specific genes that lie in different places on the regulatory pathway have, on average, different levels of noncoding conservation. We first divided tissue-specific genes into two broad classes: regulators and effectors (Fig. 1). Regulators are defined as genes involved in tissue-specific transcriptional regulation. Effectors are defined as genes whose expression is controlled directly or indirectly by regulators, but who themselves do not play any appreciable role in transcriptional regulation. Effectors are typically genes that render the physiological properties characteristic of the cell type in which the gene is expressed (e.g., myosin genes in muscle cells or keratin genes in skin cells). As for regulators, they were further divided into four subgroups based on where they lie in the regulatory hierarchy (Fig. 1). These include (1) extracellular signals, which reside at the top of the hierarchy; (2) membrane-bound signal receptors, which are downstream of extracellular signals; (3) signal transducers, such as kinases, which lie at the next level; and (4) tissue-specific transcription factors, excluding nuclear receptors, that act on *cis*-regulatory regions of

* Corresponding author. Fax: +1 (773) 834 8470.
E-mail address: blahn@bsd.uchicago.edu (B.T. Lahn).

0888-7543/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2005.09.013

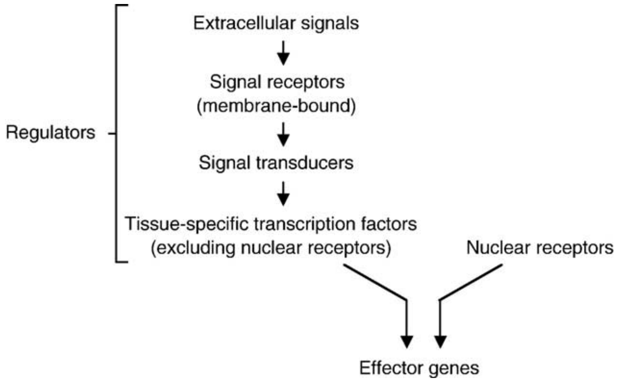

Fig. 1. Classification of tissue-specific genes.

downstream genes. We also included nuclear receptors as a separate subgroup of regulators, as they are transcription factors controlled directly by small molecular hormones (in this regard, nuclear receptors have the dual function of being both signal receptors and transcription factors).

For each subgroup of regulators, we compiled an extensive list of genes based on careful analysis of functional data pertaining to these genes in the literature (see Supplementary Table S1 for a complete list). We did not resort to Gene Ontology (GO) terms for the compilation of genes, because GO does not have the exact classification necessary for this study [4]. For effectors, we targeted two specific functional categories as well as including a nontargeted set. One targeted category is neurotransmitters, which were included to allow comparison with extracellular signals in the regulator class (both being secreted molecules). The other targeted category is neuroreceptors, which were included to allow comparison with signal receptors in the regulator class (both being membrane-bound receptors). The nontargeted set comprises all the other effectors (listed under “Other effectors” in Supplementary Table S1). We also compiled a set of ubiquitously expressed genes including general transcription factors, metabolic enzymes, and ribosomal proteins, which we used for comparison to tissue-specific genes.

We measured the conservation of noncoding regions for each gene. This approach, often referred to as phylogenetic footprinting, is predicated on the assumption that functional noncoding sequences tend to be preserved by purifying selection, whereas nonfunctional regions tend to diverge over evolutionary time [5–8]. We chose to use human–mouse conservation, as was done in many previous studies [3,9–12]. We found that if we used more distantly related species, such as human and chicken, a large fraction of the genes dropped out of the analysis due to the lack of any detectable conservation.

We aligned human and mouse orthologous genomic sequences and examined the extent of noncoding conservation in regions flanking genes of interest (see online supplementary materials). It is well known that *cis*-regulatory elements, especially tissue-specific enhancers, can act at great distances from the gene, either upstream or downstream [6]. However, it is also true that many genes have very short intergenic regions. This poses a technical challenge in selecting the right amount of flanking noncoding sequences to analyze. If the selected region is too small, many regulatory elements would fall outside of the region and would be missed in the analysis. If the region is too large, neighboring genes or their control elements may be erroneously included. We therefore tested several sizes of flanking sequences, including 10, 5, 2, or 1 kb. We found that the overall trends did not change significantly under these different sizes. That is to say, if one class of genes has more noncoding conservation than another class when a particular size of flanking sequences is used for the analysis, this trend will likely remain when a different size is used. However, when longer flanking sequences were chosen, many genes dropped out of the analysis because the flanking sequences contained neighboring genes, and this reduced the statistical power of the analysis. As a compromise, we decided to analyze 2 kb upstream and 1 kb downstream of each gene.

The definition of conservation also needs careful consideration. For human–mouse comparison, regions longer than 100 bp and that have greater than 70% sequence identity were often used in past studies as criteria to define conserved noncoding elements [6,9]. We tested several other criteria such as region size >50 bp and sequence identity >80% or >90%. None seemed to affect the overall conclusions, though small window size or high sequence identity reduced statistical power. We therefore chose to use the convention of >100 bp size and >70% identify as the definition for a conserved noncoding region.

We first confirmed that tissue-specific genes have significantly greater conservation than ubiquitously expressed genes ($p = 0$ by Wilcoxon rank sum test) (Fig. 2A and Supplementary Table S2). The p values that we mention here are for 5′ and 3′ together. We also showed that, as had been reported previously [3], tissue-specific transcription factors (including nuclear receptors) have significantly greater conservation than all other genes ($p = 0$ by Wilcoxon rank sum test) (Fig. 2A and Supplementary Table S2). Tissue-specific transcription factors also have much greater conservation than ubiquitously expressed general transcription factors ($p = 9 \times 10^{-10}$ by Wilcoxon rank sum test) (Fig. 2A and Supplementary Table S2).

We next compared the two major classes of tissue-specific genes: regulators and effectors. We found that regulators have about twice as much conservation as effectors ($p = 4 \times 10^{-11}$ by Wilcoxon rank sum test) (Fig. 2A and Supplementary Table S2). Of particular relevance is the comparison between extracellular signals in the regulator class and neurotransmitters in the effector class. Both are secreted peptide ligands expressed in a highly tissue-specific manner, yet one has regulatory functions, whereas the other does not. We found that extracellular signals show significantly more conservation than neurotransmitters ($p = 0.008$ by Wilcoxon rank sum test) (Fig. 2A and Supplementary Table S2). Somewhat unexpectedly, there is no significant difference between signal receptors in the regulator class and neuroreceptors in the effector class (Fig. 2A and Supplementary Table S2). Indeed, signal receptors have the least conservation among the various subgroups of regulators (see below).

We next compared the levels of conservation between the various subgroups of regulators. We found that tissue-specific transcription factors have the greatest conservation, whereas signal receptors have the least conservation (Fig. 2B). This

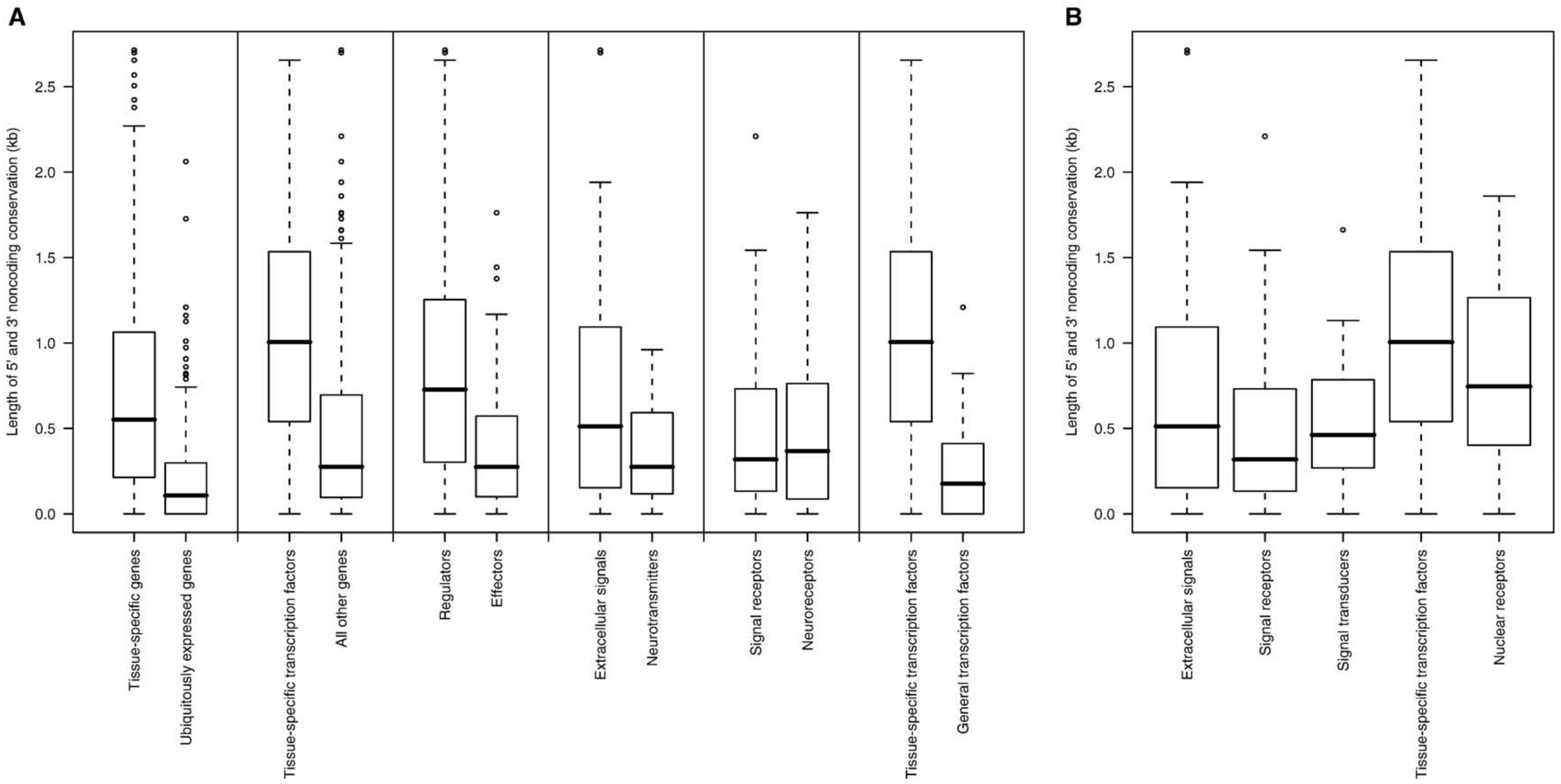

Fig. 2. Box plots of noncoding conservation for different categories of genes. Top and bottom of each box correspond to the first and third quartile, respectively, with the midline indicating the median. Bars extend to the most extreme data point within 1.5 quartile from the median. Outliers are plotted individually. (A) Two-way comparisons. Here, tissue-specific transcription factors include nuclear receptors. (B) Comparison among the various subgroups of regulators.

raises the possibility that the difference between regulators and effectors might be attributable to transcription factors alone. To address this possibility, we removed transcription factors (including nuclear receptors) from the analysis. We found that regulators continue to have significantly greater noncoding conservation than effectors ($p = 0.002$ by Wilcoxon rank sum test) (Supplementary Table S2). Thus, transcription factors do not account entirely for the different levels of noncoding conservation between regulators and effectors.

We noted that the levels of conservation in 5′ and 3′ flanking regions are correlated with levels of conservation in introns (data not shown). This suggests that the above analysis can perhaps be repeated using introns. The advantage of using introns is that they are clearly defined by intron–exon boundaries, which is unlike 5′ and 3′ flanking regions, whose boundaries are difficult to define. However, a major complicating factor in analyzing introns is the fact that there is a strong positive correlation between the amount of sequence conservation and the intron length (nonparametric correlation: Kendall's $\tau = 0.49$; $p < 10^{-15}$). Given such a correlation, genes with large introns will tend to have more conserved intronic sequences regardless of gene function. This correlation could be due to the fact that large genes do indeed have more *cis*-regulatory elements, or it could result from the fact that large genes tend to have proportionally more neutral conservation between human and mouse (assuming that a certain fraction of the genome remains conserved between these two species even in the absence of purifying selection). In the latter case, this correlation would cause gene size or exon number to confound our analysis if uncorrected.

A simple solution to this problem is to present the data in a bivariate plot of conserved intron length against total intron length. Such a plot would allow comparison of not only the total noncoding conservation in introns, but also the amount of noncoding conservation under any given intron length. A regression model can then be applied to test the statistical significance that two groups of genes have distinct levels of intron conservation (see supplementary materials). Using this approach, we showed that, consistent with the analysis of upstream and downstream regions, tissue-specific genes have greater intronic conservation than ubiquitously expressed genes ($p = 4 \times 10^{-19}$), tissue-specific transcription factors have greater conservation than all other genes ($p = 3 \times 10^{-15}$), regulators have greater conservation than effectors ($p = 5 \times 10^{-8}$), and extracellular signals have greater conservation than neurotransmitters ($p = 1 \times 10^{-4}$) (Supplementary Fig. S1 and Supplementary Table S2).

A potential criticism of the above conclusions is that differences in the amount of conservation might in part be due to differences in local mutation rate [13]. We note that if we added K_s—the rate of synonymous substitutions, which is often used to approximate local mutation rate—as a covariate to our multiple regression model, the results do not change appreciably.

In conclusion, our study shows that different classes of tissue-specific genes have different levels of noncoding conservation (i.e., different density of noncoding conservation in the region of interest). Regulators generally have significantly more noncoding conservation than effectors. Furthermore, regulators that reside in different positions in the regulatory hierarchy tend to have different levels of noncoding conservation, with tissue-specific transcription factors having the most conservation and signal receptors having the least. These findings significantly extend previous studies showing that genes with more complex expression patterns tend to have longer intergenic sequences [2] and that tissue-specific transcription factors tend to have highly conserved noncoding regions [3]. Our study should therefore contribute to a holistic understanding of how the regulatory network of gene expression is constructed inside of the cell.

Acknowledgments

We thank Su Yeon Kim, Omar De la Cruz, Ryan Carr, and Peter McCullagh for suggestions on the data analysis.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ygeno.2005.09.013.

References

[1] E.H. Davidson, D.R. McClay, L. Hood, Regulatory gene networks and the properties of the developmental process, Proc. Natl. Acad. Sci. USA 100 (2003) 1475–1480.

[2] C.E. Nelson, B.M. Hersh, S.B. Carroll, The regulatory content of intergenic DNA shapes genome architecture, Genome Biol. 5 (2004) R25.

[3] H. Iwama, T. Gojobori, Highly conserved upstream sequences for transcription factor genes and implications for the regulatory network, Proc. Natl. Acad. Sci. USA 101 (2004) 17156–17161.

[4] M.A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, et al., The Gene Ontology (GO) database and informatics resource, Nucleic Acids Res. 32 (2004) D258–D261 (Database issue).

[5] R.C. Hardison, J. Oeltjen, W. Miller, Long human–mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome, Genome Res. 7 (1997) 959–966.

[6] L.A. Pennacchio, E.M. Rubin, Genomic strategies to identify mammalian regulatory sequences, Nat. Rev. Genet. 2 (2001) 100–109.

[7] W.W. Wasserman, A. Sandelin, Applied bioinformatics for the identification of regulatory elements, Nat. Rev. Genet. 5 (2004) 276–287.

[8] A. Woolfe, M. Goodson, D.K. Goode, P. Snell, G.K. McEwen, T. Vavouri, S.F. Smith, P. North, H. Callaway, K. Kelly, et al., Highly conserved noncoding sequences are associated with vertebrate development, PLoS Biol. 3 (2005) e7.

[9] G.G. Loots, R.M. Locksley, C.M. Blankespoor, Z.E. Wang, W. Miller, E.M. Rubin, K.A. Frazer, Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons, Science 288 (2000) 136–140.

[10] W.W. Wasserman, M. Palumbo, W. Thompson, J.W. Fickett, C.E. Lawrence, Human–mouse genome comparisons to locate regulatory sites, Nat. Genet. 26 (2000) 225–228.

[11] Z. Zhang, M. Gerstein, Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements, J. Biol. 2 (2003) 11.

[12] B. Lemos, J.A. Yunes, F.R. Vargas, M.A. Moreira, A.A. Cardoso, H.N. Seuanez, Phylogenetic footprinting reveals extensive conservation of Sonic Hedgehog (SHH) regulatory elements, Genomics 84 (2004) 511–523.

[13] H. Ellegren, N.G. Smith, M.T. Webster, Mutation rate variation in the mammalian genome, Curr. Opin. Genet. Dev. 13 (2003) 562–568.