Available online at www.sciencedirect.com
ScienceDirect
Gene 407 (2008) 54–62
GENE
www.elsevier.com/locate/gene

Housekeeping and tissue-specific genes differ in simple sequence repeats in the 5′-UTR region

Mark J. Lawson, Liqing Zhang *

Department of Computer Science, Virginia Tech, United States

Received 29 July 2007; received in revised form 25 September 2007; accepted 26 September 2007
Available online 4 October 2007
Editor: I.B. Rogozin

Abstract

SSRs (simple sequence repeats) have been shown to have a variety of effects on an organism. In this study, we compared SSRs in housekeeping and tissue-specific genes in human and mouse, in terms of SSR types and distributions in different regions including 5′-UTRs, introns, coding exons, 3′-UTRs, and upstream regions. Among all these regions, SSRs in the 5′-UTR show the most distinction between housekeeping genes and tissue-specific genes in both densities and repeat types. Specifically, SSR densities in 5′-UTRs in housekeeping genes are about 1.7 times higher than those in tissue-specific genes, in contrast to the 0.8–1.2 times differences between the two classes of genes in other regions. Tri-SSRs in 5′-UTRs of housekeeping genes are more GC rich than those of tissue-specific genes and CGG, the dominant type of tri-SSR in 5′-UTR, accounts for 74–79% of the tri-SSRs in housekeeping genes, as compared to 42–57% in tissue-specific genes. 75% of the tri-SSRs in the 5′-UTR of housekeeping genes have 4–5 repeat units, versus the 86–90% in tissue-specific genes. Taken together, our results suggest that SSRs may have an effect on gene expression and may play an important role in contributing to the different expression profiles between housekeeping and tissue-specific genes.

Keywords: 5′-UTR; Gene expression; Gene function; Microsatellite; Tissue-specificity

1. Introduction

Simple sequence repeats (SSRs) are tandem repeat nucleotides (oftentimes defined as being between 1 and 6 base pairs) in DNA sequences. They can be found in both eukaryotes and prokaryotes and in both protein coding and non-coding regions. Historically, SSRs have been used mainly as genetic markers to study populations and ecology of various species. However, recent research has shown that SSRs play a more active role than previously thought in terms of development, gene regulation, and evolution (Li et al., 2004; Kashi and King, 2006). Fondon and Garner (2004) demonstrated this very effectively by showing that SSRs in the *ALX-4* and *Runx-2* genes were associated with the fast evolution of limb and skull morphology in different breeds of dogs. Hammock and Young (2005) showed that a polymorphic SSR upstream of the *vasopressin 1a receptor* gene affects the gene's expression level which in turn led to observed differences in socio-behavioral traits of the prairie vole.

SSRs have also been shown to be behind a variety of neurological diseases (Cummings and Zoghbi, 2000). Based on where a trinucleotide SSR resides in a gene, these diseases can be classified into two subclasses: non-coding and coding repeat diseases. One example of a non-coding trinucleotide repeat disease is myotonic dystrophy caused by the expansion of a $(CTG)_n$ repeat in the 3′-UTR of the *DMPK* gene (Mahadevan et al., 1992) which hinders normal transcription of *DMPK*. The coding repeat diseases are caused by expansion or shrinkage of a $(CAG)_n$ repeat in the coding regions, which can lead to protein misfolding or abnormal gene expression.

Our interest in SSRs consists mainly of how they may affect gene expression in general within eukaryotic genomes, specifically

Abbreviations: AD, average difference; GNF, Genomics Institute of the Novartis Research Foundation; SSR, simple sequence repeats; UTR, untranslated regions.

* Corresponding author. 2050 Torgersen Hall, Blacksburg, VA 24061-0106, United States. Tel.: +1 540 231 9413; fax: +1 540 231 6075.
E-mail address: lqzhang@vt.edu (L. Zhang).

0378-1119/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2007.09.017

in the 5′-UTR region. This is mainly because of two observations. First, it has been shown that the SSRs in 5′-UTR can affect gene expression in a number of genes which in turn can have a great effect on the organism itself. For instance, the $(CAG)_n$ SSR in the 5′-UTR of the *hCALM1* gene is necessary for *hCALM1*'s normal expression. Removal of this SSR causes a decrease in expression by 45% (Toutenhoofd et al., 1998). Second, our previous study of SSR distribution in different genic regions of the *Oryza sativa* and *Arabidopsis thaliana* genomes indicates that both organisms possess an abundance of SSRs in 5′-UTRs and the SSR densities in 5′-UTRs are much higher than in other regions such as coding exons, 3′-UTRs, and introns (Lawson and Zhang, 2006). Combining these observations, we speculate that SSRs in 5′-UTRs may differ among genes that show different patterns of gene expressions. Patterns of gene expression show the most contrast between housekeeping genes and tissue-specific genes: housekeeping genes are genes that are expressed constitutively in all tissues of an organism, whereas tissue-specific genes are expressed in only one tissue. Thus, they seem to be the most appropriate combination of genes that we can test the hypothesis on. Here we contrasted SSR distributions in different regions of housekeeping and tissue-specific genes to see whether there are any distinguishable differences between the two classes of genes.

2. Materials and methods

2.1. Gene expression data

We used the gene expression data of Su et al. (2004) to determine housekeeping and tissue-specific genes in human and mouse. Gene expression data and annotation files linking Ensembl transcripts to their representative probes were downloaded from GNF (Genomics Institute of the Novartis Research Foundation) from their SymAtlas website (http://symatlas.gnf.org/SymAtlas/). The data consists of probe expression levels (AD values) for 79 human tissues using the Affymetrix U133A probeset and a custom-made probeset, and also for 61 mouse tissues using a custom-made probeset. Two samples for each tissue were obtained in each species. We used the average AD for the two samples in each tissue. We averaged out all probe values for a gene within a given tissue and excluded those with probe IDs that end with "_s_at" or "_x_at" unless those are the only probes associated with a gene. Housekeeping genes are formally defined as genes that are expressed constitutively in all tissues and tissue-specific genes are genes that are expressed in only one tissue. For microarray data, because an AD level of 200 corresponds to about 3–5 copies per cell (Su et al., 2002), the AD level of a gene in comparison with 200 has been commonly used to define computationally whether a gene is expressed in a tissue or not (e.g. Zhang and Li, 2004; Yang et al., 2005; Zhang and Li, 2005). Therefore, similar to what has been done before, we defined housekeeping genes as the genes that had AD levels of higher than 200 in all 79 human tissues or all 61 mouse tissues. Tissue-specific genes are those that had AD values above 200 in only one tissue. All genic regions of each gene, including 5′-UTR, 3′-UTR, coding exons, introns, and 500 bp upstream of the gene, were downloaded from Ensembl (Hubbard et al., 2007) and stored in a MySQL database. Ensembl is a well-known, continuously updated database of sequence annotation and its gene structure predictions are based on experimental evidence whenever available in UniProt/Swiss-Prot, UniProt/TrEMBL, NCBI RefSeq, as well as cDNA entries from EMBL. The annotated gene structures in both human and mouse should be of high quality due to the availability of large-scale full length cDNAs in both species (e.g. Carninci et al., 2005).

2.2. Determining SSRs

SSRs were determined using the mreps program (Kolpakov et al., 2003). We have used this program in our previous analysis of plant SSRs (Lawson and Zhang, 2006) and found it to be robust and efficient in locating SSRs. We considered only perfect SSRs with a length of longer than 10 bp (Lawson and Zhang, 2006). We then wrote perl scripts to extract information from the mreps output files, which was then linked to further genic data that was stored in a MySQL database. A repeat type includes all of its circular permutations and their complements as done in Zhang et al. (2004). For example, for mononucleotide SSRs, the type $(A)_n$ represents both $(A)_n$ and $(T)_n$; for dinucleotide SSRs, $(AC)_n$ represents $(AC)_n$, $(CA)_n$, $(TG)_n$, and $(GT)_n$. This was done due to the double-stranded nature of DNA and the fact that the start site of a SSR can be considered arbitrary (Jurka and Pethiyagoda, 1995).

2.3. Statistical analysis

Chi-square goodness-of-fit tests with 1 degree of freedom were applied to test whether SSR density is significantly different between housekeeping and tissue-specific genes in the 5′-UTRs, coding exons, introns, 3′-UTRs, and 500 bp upstream regions. Specifically, in a particular region, if we assume the density is the same between housekeeping and tissue-specific genes, we can calculate the expected number of SSRs using the following formula:

$$E_i = \frac{N}{L} * L_i$$

where E_i is the expected number of SSRs of either housekeeping (E_H) or tissue-specific (E_T) genes, N is the total number of SSRs in the two classes of genes, L is the total length in base pairs of the two classes of genes, and L_i is the length in base pairs of either housekeeping (L_H) or tissue-specific (L_T) genes.

3. Results

We studied a total of 1914 housekeeping and 275 tissue-specific genes in human, and 1597 housekeeping and 890 tissue-specific genes in mouse. In human, the housekeeping genes cover 91.3 MB (mega bases) containing 56017 SSRs; the tissue-specific genes cover 16.1 MB containing 8331 SSRs. In mouse, the housekeeping genes cover 12.2 MB containing 9234 SSRs; the tissue-specific genes cover 8.5 MB containing 6415 SSRs.