Available online at www.sciencedirect.com
ScienceDirect
Gene 390 (2007) 153–165
GENE
SECTION
EVOLUTIONARY GENOMICS
www.elsevier.com/locate/gene

Repetitive sequence environment distinguishes housekeeping genes

C. Daniel Eller [a], Moira Regelson [a,1], Barry Merriman [a], Stan Nelson [a], Steve Horvath [a,b], York Marahrens [a,*]

[a] *UCLA Department of Human Genetics, David Geffen School of Medicine, Gonda Center, 695 E. Young Drive South, Los Angeles, California 90095-7088, USA*

[b] *UCLA Department of Biostatistics, School of Public Health, Box 951772, Los Angeles, California 90095-1772, USA*

Received 21 June 2006; received in revised form 18 September 2006; accepted 24 September 2006
Available online 5 October 2006
Received by M. Batzer

Abstract

Housekeeping genes are expressed across a wide variety of tissues. Since repetitive sequences have been reported to influence the expression of individual genes, we employed a novel approach to determine whether housekeeping genes can be distinguished from tissue-specific genes by their repetitive sequence context. We show that Alu elements are more highly concentrated around housekeeping genes while various longer (>400-bp) repetitive sequences ("repeats"), including Long Interspersed Nuclear Element-1 (LINE-1) elements, are excluded from these regions. We further show that isochore membership does not distinguish housekeeping genes from tissue-specific genes and that repetitive sequence environment distinguishes housekeeping genes from tissue-specific genes in every isochore. The distinct repetitive sequence environment, in combination with other previously published sequence properties of housekeeping genes, was used to develop a method of predicting housekeeping genes on the basis of DNA sequence alone. Using expression across tissue types as a measure of success, we demonstrate that repetitive sequence environment is by far the most important sequence feature identified to date for distinguishing housekeeping genes.

Keywords: Random forest; Alu; SINE; LINE; Repeat; Tissue-specific genes; Isochores

1. Introduction

Housekeeping genes perform the basic functions common to all dividing cells; they are widely expressed across tissues, and are associated with CpG islands (Bird, 1986). Housekeeping genes also typically have small introns (Eisenberg and Levanon, 2003) that lack repetitive sequences (Han et al., 2004). Housekeeping genes have been found to cluster together on the genome to some degree (Lercher et al., 2002), and to preferentially localize to GC-rich fractions of genomic DNA known as isochores on cesium sulfate gradients (Lercher et al., 2003).

We were interested in the relationship of housekeeping genes to repetitive sequences. Nearly half of the human genome consists of repetitive sequence, the majority of which is transposon-derived and widely considered to be "junk DNA". The major classes of repeats are LTR retrotransposons (7.9% of genome sequence), non-LTR retrotransposons (32.0%), DNA transposons (2.8%), satellite and satellite-related sequences (0.34%), low complexity repeats (0.54%), and simple sequence repeats (0.84%). The non-LTR retrotransposons consist primarily of the Long Interspersed Nuclear Element-1 (LINE-1, 15.6%) and the non-autonomous Alu element (10.1%). LINE-1 transposons encode the enzymatic activities required for both their own mobility and for the mobilization of Alu elements (Hagan and Rudin, 2002; Kajikawa and Okada, 2002; Dewannieux et al., 2003). Alu transposons differ from other SINEs in that they are not derived from tRNA genes, but rather from the 7SL RNA gene (Ullu and Tschudi, 1984; Quentin, 1994; Smit, 1996;

Abbreviations: HK, housekeeping gene; TS, tissue-specific gene; LINE-1, Long Interspersed Nuclear Element-1; SINE, Short Interspersed Nuclear Element; repeat, repetitive sequence.

* Corresponding author. UCLA Department of Human Genetics, Gonda Center, Room 4554b, 695 Charles E. Young Drive, Los Angeles, CA 90095, USA. Tel.: +1 310 267 2466; fax: +1 310 794 5446.
E-mail addresses: moira@Yahoo-inc.com (M. Regelson), ymarahrens@mednet.ucla.edu (Y. Marahrens).
[1] Current address: 3333 West Empire Ave, Burbank CA 91504, USA. Tel.: +1 818 524 3549.

0378-1119/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2006.09.018

Okada and Hamada, 1997; Terai et al., 1998; Lander et al., 2001) which encodes the RNA component of the signal recognition particle that mediates the translocation of nascent secretory and membrane proteins (Wild et al., 2004). Aside from favoring TT|AAAA target sequences (Feng et al., 1996; Jurka, 1997; Cost and Boeke, 1998), human Alu and LINE-1 elements have been reported to insert at random positions in the genome (Smit, 1999; Boissinot et al., 2001; Lander et al., 2001; Ovchinnikov et al., 2001; Gilbert et al., 2002; Myers et al., 2002; Symer et al., 2002; Szak et al., 2002; Jurka et al., 2004; Gilbert et al., 2005). However, there is some evidence for insertional hot spots (Cost and Boeke, 1998; Myers et al., 2002; Graham and Boissinot, in press). A popular idea is that the non-random distribution of these repetitive sequences arises from their loss via purifying selection (Boissinot et al., 2001; Myers et al., 2002; Graham and Boissinot, in press).

There are a number of reports of repetitive sequences influencing gene expression. For example, in fragile X patients expansion mutations of a tandem simple sequence repeat located in an intron of the *FMR1* gene result in the transcriptional silencing of the *FMR1* gene (Pieretti et al., 1991). Transposable elements in *Drosophila* and plants have been implicated in the transcriptional silencing of nearby genes by the spread of heterochromatin (Lippman et al., 2004; Sun et al., 2004) raising the possibility that transposons may also be capable of reducing expression if located near genes in humans. DNA methylation is an important feature of heterochromatin that silences gene expression (Stancheva, 2005). Tissue-specific differences in DNA methylation have been reported for various repetitive sequences including LINE-1 elements (Sano and Sager, 1982; Breznik et al., 1984; Nishioka, 1988; Mietz and Kuff, 1990; Allingham-Hawkins et al., 1996; Hassan et al., 2001; Chalitchagorn et al., 2004; Khodosevich et al., 2004), raising the expectation that repetitive sequences are more repressive to the expression of nearby genes in some tissues than in others. Here we sought to determine whether the repetitive sequence environments flanking housekeeping genes that are widely expressed and important across tissues are subject to unique constraints. We show that long (>400-bp) repeats including LINE-1 elements are excluded from the regions flanking housekeeping genes and that short repeats, in particular Alu elements, are particularly highly enriched around these genes. We demonstrate that repetitive sequence environment is by far the most important sequence feature identified to date for distinguishing housekeeping genes and speculate that Alu elements are advantageous for housekeeping genes.

2. Methods

2.1. Assembly of gene lists

For housekeeping genes, we used a published list of 575 genes that were expressed in all available tissues above 200 standard Affymetrix average-difference units on an Affymetrix U95A microarray chip containing 12,600 probes from 47 different human tissues and cell lines (Eisenberg and Levanon, 2003). We assembled a list of tissue-specific genes by combining the published lists of two studies (Warrington et al., 2000; Hsiao et al., 2001). Genes from all lists were identified by either Genbank RefSeq ID or Unigene ID and were converted to RefSeq ID via DAVID Tools (http://apps1.niaid.nih.gov/david) (Dennis et al., 2003). We then looked up each gene by its RefSeq ID in the UCSC Genome Browser (http://genome.ucsc.edu) and marked it as HK (housekeeping) or TS (tissue-specific), as appropriate. Genes with common transcription start or stop positions on the same chromosome which were considered to be the same gene were treated identically. Seventeen genes appeared in both the housekeeping and tissue-specific lists and were therefore excluded from both groups. After all conversions, we had 586 autosomal housekeeping genes and 468 autosomal tissue-specific genes.

2.2. Sequence characteristics

Human sequence information including repetitive sequences were obtained from the July 2003 assembly (hg16) UCSC annotation tracks in the *chrN_rmsk* tables (http://hgdownload.cse.ucsc.edu/downloads.html#human). For each gene, we initially defined a region of analysis extending from 100-kb upstream of the transcription start position (txStart) to 100-kb downstream of the transcription end position (txEnd). We excluded the transcribed gene regions from our analysis to avoid effects that can be attributed to displacement by the coding sequence or splicing elements, or to the interference of transcriptional elongation by repetitive sequences in introns (Han et al., 2004). Gaps in the DNA sequence were also omitted.

The total number of base pairs comprising each repeat by family was then calculated and divided by the total number of base pairs included in the region. These data were extracted from a copy of the UCSC Genome Browser Database running locally on MySQL (http://hgdownload.cse.ucsc.edu/downloads.html#human; http://www.mysql.com) (Regelson et al., 2006). Calculations were performed using the Perl scripting language (http://www.perl.org) and MySQL functions.

CpG islands were treated in the same manner as repetitive sequence composition: The total number of bases comprising CpG islands was divided by total bases in each region of analysis. CpG islands are defined by the UCSC Genome Browser as sequences that are at least 50 base pairs long and have at least 50% GC content. The extent of gene clustering was estimated by counting the number of transcription start positions found within each region of analysis. Sense and antisense orientation were determined relative to the gene being analyzed; repeats oriented in the same direction as the gene are designated “sense”, while genes in the opposite orientation are designated “antisense”.

The long-range distribution of each characteristic was measured in much the same way as above, except flanking regions extended 40-Mb in each direction, excluding the gene, rather than 100-kb. Also, rather than working with the entire flanking region at once, we divided the 80-Mb regions into 1-Mb segments and calculated the fraction of each segment comprising repeats, CpG islands and number of genes that have transcription start sites in that segment.

Isochores are defined as contiguous regions along a chromosome sharing a homogenous GC composition and are identified