Gene 429 (2009) 74–79

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Embryonic nervous system genes predominate in searches for dinucleotide simple sequence repeats flanked by conserved sequences

Donald E. Riley [a,b,*], John N. Krieger [a,b]

[a] Department, Urology, University of Washington, Seattle, WA 98195, USA
[b] Department, Research, VA Puget Sound Health Care System, Room 517, Building 1, 1660 South Columbian Way, Seattle, WA 98012, USA

ARTICLE INFO

Article history:
Received 1 April 2008
Received in revised form 18 September 2008
Accepted 23 September 2008
Available online 2 October 2008

Received by S.M. Mirkin

Keywords:
Highly conserved elements
HCEs
Simple sequence repeats
SSRs
Dinucleotide simple sequence repeats
diSSRs
Microsatellites
Nervous system
Development

ABSTRACT

To study evolution of dinucleotide simple sequence repeats (diSSRs) we searched recently available mammalian genomes for UTR-localized diSSRs with conserved upstream flanking sequences (CFS). There were 252 reported *Homo sapiens* genes containing the repeats (AC)*n*, (GT)*n*, (AG)*n* or (CT)*n* in their UTRs including 22 (8.7%) with diSSR-upstream flanking sequences conserved comparing divergent mammalian lineages represented by *Homo sapiens* and the marsupial, *Monodelphis domestica*. Of these 22 genes, 19 had known functions including 18 (95%) that proved critical for mammalian nervous systems (Fishers exact test, $P<0.0001$). The remaining gene, *Cd2ap*, proved critical for development of kidney podocytes, cells that have multiple similarities to neurons. Gene functions included voltage and chloride channels, synapse-associated proteins, neurotransmitter receptors, axon and dendrite pathfinders, a NeuroD potentiator and other neuronal activities. Repeat length polymorphism was confirmed for 68% of CFS diSSRs even though these repeats were nestled among highly conserved sequences. This finding supports a hypothesis that SSR polymorphism has functional implications. A parallel study was performed on the self-complementary diSSRs (AT)*n* and (GC)*n*. When flanked by conserved sequences, the self-complementary diSSR (AT)*n* was also associated with genes expressed in the developing nervous system. Our findings implicate functional roles for diSSRs in nervous system development.

Published by Elsevier B.V.

1. Introduction

Simple sequence repeats (SSRs), also termed short tandem repeats or microsatellites, represent some of the most recognizable noncoding sequences. Often, SSRs modulate their associated genes to provide a means for genetic variation with minimal genetic load (Fondon and Garner, 2004; Li et al., 2004; Kashi and King, 2006). Previously, we reported association of the dinucleotide SSRs (diSSRs) (AC)*n* and (GT)*n* with the 3′ untranslated regions (UTRs) of genes encoding membrane functions and transcription factors (Riley and Krieger, 2004b). (Sequences are written as they appear relative to the sense-strand of nearby coding sequence.) During database searches, we noticed that mRNAs encoding proteins of similar function (e.g., members of the aquaporin family) sometimes had 3′ UTRs with different diSSRs such as (AC)*n* or (AG)*n*. For this reason, we hypothesized that different diSSR sequences might serve similar functions. This hypothesis led to the prediction that diSSRs might replace one another during evolution of a given UTR.

Comparing orthologous UTRs in different species, dinucleotide SSRs (diSSRs) such as (AC)*n*, (GT)*n*, (AG)*n* and (CT)*n*, were found to frequently replace one another at the same position within the UTR supporting the hypothesis that these repeats function in similar structural roles (Riley and Krieger, 2004a, 2005). Folding algorithms predicted that the potential to form single-stranded loops represented a common structural feature because (AC)*n*, (GT)*n*, (AG)*n* and (CT)*n* all lack the ability to form canonical base pairs by themselves. We termed these, "weak-folding repeats."

The diSSRs are present in less than 0.15% of reported UTRs. Thus, the many precisely localized diSSR replacements that have been observed are unlikely to represent random events. While diSSR replacements proved common, we never observed a triplet repeat replacing a diSSR. Triplet repeats and diSSRs also have different genomic distributions, with the latter more common among noncoding sequences (Toth et al., 2000; Wren et al., 2000; Cordeiro et al., 2001). These observations support separate study of diSSRs since they might represent a distinct functional category.

Compared to the weak-folding diSSRs, the diSSRs (AT)*n* and (GC)*n* have a somewhat different usage. The self-complementary diSSR (AT)*n*≥14 has single-stranded folding potential and was often replaced during evolution by palindromic sequences with similar folding

Abbreviations: CFS, conserved flanking sequence; HCE, highly conserved element; UTR, untranslated region; diSSRs, dinucleotide simple sequence repeats; us, upstream; ds, downstream.

* Corresponding author. Puget Sound Health Care System (VA) Room 517, Building 1; 1660 South Columbian Way, Seattle, WA 98108, USA. Tel.: +1 206 277 3608; fax: +1 206 764 2164.

0378-1119/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.gene.2008.09.025

potential (Riley et al., 2007). The diSSR (GC)$n \geq 14$ also has folding potential but is rarely used. There were no human UTRs reported with (GC)$n \geq 14$. Based on its replacement by other folding sequences (AT)n may function differently from the weak-folding diSSRs.

An excellent past study surveyed all SSRs including di-, tri-, and tetranucleotide repeats at all genomic locations (Toth et al., 2000). In contrast, we restricted study to highly selected SSR groups to allow for the possibility that heterogeneous functions might be obscured if the study's scope were too broad. The current investigation substantially extends this approach by examining UTR-localized diSSRs that have highly conserved flanking sequences. Highly conserved noncoding sequence elements have been termed HCEs (Siepel et al., 2005). We hypothesized that examination of UTR-localized diSSRs, flanked by HCEs, using recently available genome sequences as a resource, might provide new insights into the long-term evolution and function of UTR-localized diSSRs.

Like previous studies of SSRs, past studies of HCEs were broad, encompassing introns, intragenic spacers, gene deserts and UTRs (Siepel et al., 2005). 3′ UTRs accounted for an unexpectedly high proportion of HCE bases (11-fold enrichment). Presence of noncoding HCEs predicted genes implicated in development, differentiation and malignancies (Bejerano et al., 2004; Sandelin et al., 2004; Derti et al., 2006). However, there were no clear usage patterns or consistent functional contexts. One possible explanation is that previous broad studies of HCEs may involve heterogeneous functions obscuring usage patterns that might be present in more restricted groups of HCEs. For this reason, we obtained new data using a novel perspective provided by limiting our initial screen to diSSRs flanked by HCEs within mammalian UTRs.

2. Materials and methods

2.1. Database searches and the conserved upstream flanking sequence (CFS) strategy

The CFS strategy was designed to identify a set of diSSRs flanked by HCEs, then study evolution of these sites using recently available genome sequences from a variety of mammalian species. To obtain an initial set of CFS UTRs (those that have diSSRs flanked by HCEs) we used the weak-folding diSSR sequences (AC)n, (GT)n, (AG)n or (CT)n with $n \geq 14$ to search non-redundant 5′ and 3′ UTR databases (version, 2.2.1, 4-13-01, 368,154 sequences (http://www.ba.itb.cnr.it/BIG/Blast/BlastUTR.html)) for all *Homo sapiens* UTRs that contained such diSSRs. The non-redundant UTR database has been previously described (Pesole et al., 2002). For these searches, we used the Basic Local Alignment Search Tool (BLAST) with low-complexity filtering turned off and default parameters for Expect and Matrix. Using the same parameters, a parallel study of (AT)n, a "strong folding" diSSR, was done separately because we anticipated (AT)n might have different usage compared with weak-folding diSSRs. For reasons outlined below, trinucleotide and tetranucleotide repeats did not play major roles in the current study. For the initial search sequences, we accepted no base substitutions within the core repeat ($n \geq 14$) for two reasons: 1. We knew there were hundreds of perfect, core repeats in the UTR database. 2. We were concerned that accepting imperfections might inadvertently capture repeats that were evolving into non-repetitive sequences.

Upstream flanking sequences, consisting of 150 to 350 bp, from each of the 252 UTRs identified were then used to search the opossum genome using BLASTn at the same website. We assumed that upstream flanking sequences conserved comparing a primate and a marsupial would likely recover sequences conserved across a broad range of mammals. Presence of a diSSR in a human UTR, and conservation of the upstream flanking sequence in opossum, represented the only selection criteria applied during CFS strategy searches.

2.2. Coding potential score

The diSSR site sequences were annotated as noncoding sequences in Genbank. However, to allow for possible alternative, unreported transcripts, we used a variation of CSTMiner (Castrignano et al., 2004) to evaluate all diSSR-flanking sequences, both upstream and downstream, for coding potential. The initial CSTMiner screen identifies sequences conserved comparing two species, such as human and mouse, to identify sequences likely to function. The algorithm then assigns coding potential scores (CPS) based on quantification of synonymous and non-synonymous substitutions at the nucleotide level and conservative changes vs. non-conservative ones at the protein level. We followed the classification scheme of Castrignano et al., (2004) who categorized sequences as either noncoding sequences (CPS<6.74; probability of coding $<1\%$), possible coding sequences ($6.74 \leq$ CPS ≤ 7.71) or coding sequences (CPS>7.71; probability of coding $>99\%$). For each UTR-localized diSSR conserved in human and opossum, at least 200 bases of upstream and downstream SR flanking sequence were pasted into the search window at http://pentagramma.caspur.it/GenoMinerNew/. Then, the sequences were compared with the mouse genome and CPS scores obtained. Conservation comparing human and mouse, while less stringent

Table 1
Genes recovered with UTR-localized weak-folding diSSRs and CFS criteria

Category	Gene	Function	CPS
Neuronal membrane proteins	*Vsnl1*	Calcium sensing; CNS injury biomarker (Mathisen et al., 1999; Laterza et al., 2006)	<6.74
	Sema6D	Axon targeting in cerebral cortex (Chen et al., 2005)	<6.74
	Glp1r	Learning and neuroprotection (During et al., 2003)	us, 6.74–7.71
	Pou4f2	Axon guidance (Samady et al., 2006)	<6.74
	Zfhx1a	Essential for axon development (Eppig et al., 2005)	us, 6.74–7.71
Neuronal transcription factors	*Ches1*	Strong expression CNS development (Tribioli et al., 2002)	us>7.71
	Ppargc1a	CNS hyperactivity; receptor coactivator; (Lin et al., 2004)	<6.74
	Rreb1	Potentiator of neurogenic factor NeuroD (Ray et al., 2003)	<6.74
	Kcnip	Transcriptional repressor involved in pain modulation (Eppig et al., 2005)	<6.74
	Nlk	Midbrain patterning (Thorpe and Moon, 2004)	<6.74
Neuronal kinases or phosphatases	*Camk2n1*	Frontal cortex, hippocampus, colliculus (Chang et al., 2001)	<6.74
	Ublcp1	Expressed in embryonic brain and other (Eppig et al., 2005)	<6.74
Neuronal microtubule motor	*Kif1b*	Brain size/synaptic vesicle development (Miki et al., 2001; Zhao et al., 2001; Mok et al., 2002)	<6.74
Neuronal apoptosis	*Faslg*	Adult motor neuron degeneration/apoptosis (Su et al., 2003; Landau et al., 2005; Martin et al., 2005)	<6.74
Synapse formation and function	*Nrxn1 (imp)*	Neurotransmitter release and synapse formation (Puschel and Betz, 1995; Graf et al., 2004; Chubykin et al., 2005)	<6.74
	Clcn3 (imp)	Chloride channel; loss of hippocampus; motor chord deficit (Stobrawa et al., 2001; Yoshikawa et al., 2002)	<6.74
	Rab3A	Neuromuscular synapse and learning (Shirataki et al., 1993; Castillo et al., 1997; Sons and Plomp, 2006)	<6.74
Dendritic membrane protein	Igsf9	Dendrite development/arborization/learning (Doudney et al., 2001; Shi et al., 2004; Falls, 2005)	<6.74
Kidney podocyte development	*Cd2ap*	Essential for podocyte extension/development (Eppig et al., 2005; Huber et al., 2006)	<6.74