Genomics 94 (2009) 308–316

Contents lists available at ScienceDirect

Genomics

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

Short ultraconserved promoter regions delineate a class of preferentially expressed alternatively spliced transcripts

Christian Rödelsperger [a,b,c], Sebastian Köhler [a,c], Marcel H. Schulz [b,d], Thomas Manke [b], Sebastian Bauer [a], Peter N. Robinson [a,b,c,*]

[a] Institute for Medical Genetics, Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany
[b] Max Planck Institute for Molecular Genetics, Berlin, Germany
[c] Berlin-Brandenburg Center for Regenerative Therapies, Berlin, Germany
[d] International Max-Planck Research School for Computational Biology and Scientific Computing, Berlin, Germany

ARTICLE INFO

Article history:
Received 15 April 2009
Accepted 21 July 2009
Available online 4 August 2009

Keywords:
Promoter
Ultraconservation
Alternative splicing
Digital gene expression
Genomics

ABSTRACT

Ultraconservation has been variously defined to describe sequences that have remained identical or nearly so over long periods of evolution to a degree that is higher than expected for sequences under typical constraints associated with protein-coding sequences, splice sites, or transcription factor binding sites. Most intergenic ultraconserved elements (UCE) appear to be tissue-specific enhancers, whereas another class of intragenic UCEs is involved in regulation of gene expression by means of alternative splicing. In this study we define a set of 2827 short ultraconserved promoter regions (SUPR) in 5 kb upstream regions of 1268 human protein-coding genes using a definition of 98% identity for at least 30 bp in 7 mammalian species.

Our analysis shows that SUPRs are enriched in genes playing a role in regulation and development. Many of the genes having a SUPR-containing promoter have additional alternative promoters that do not contain SUPRs. Comparison of such promoters by CAGE tag, EST, and Solexa read analysis revealed that SUPR-associated transcripts show a significantly higher mean expression than transcripts associated with non-SUPR-containing promoters. The same was true for the comparison between all SUPR-associated and non-SUPR-associated transcripts on a genome-wide basis.

SUPR-associated genes show a highly significant tendency to occur in regions that are also enriched for intergenic short ultraconserved elements (SUE) in the vicinity of developmental genes. A number of predicted transcription factor binding sites (TFBS) are overrepresented in SUPRs and SUEs, including those for transcription factors of the homeodomain family, but in contrast to SUEs, SUPRs are also enriched in core-promoter motifs. These observations suggest that SUPRs delineate a distinct class of ultraconserved sequences.

Introduction

'Ultraconservation' was originally defined as perfect conservation of at least 200 bp between human, mouse, and rat [1]. Many ultraconserved elements (UCE) are located in non-coding DNA nearby to genes that are involved in developmental processes, and a number of such intergenic UCEs display enhancer activity as demonstrated in assays using transgenic animals [2–4]. Another class of UCEs overlaps coding sequences and encodes alternatively spliced exons containing in-frame stop codons that trigger nonsense-mediated decay. In essence, these UCEs regulate gene expression by coupling alternative splicing to mRNA decay [5,6]. Since many of the genes containing UCEs in their coding regions perform functions related to RNA binding and regulation of splicing [1], the extreme evolutionary conservation of these UCEs could be related to the evolutionary importance of maintaining tightly tuned homeostasis of RNA-binding protein levels [5]. Further studies showed that a large number of intergenic and intronic UCEs are themselves transcribed, that at least some of the transcribed UCEs are regulated by microRNAs, and that the expression of UCEs can be altered in cancer [7]. These observations suggest that UCEs represent a functionally heterogeneous family of DNA sequences.

The evolutionary origin of vertebrate UCEs remains unclear; although homologs of many vertebrate protein-coding genes can be found in invertebrates, virtually none of the vertebrate UCEs have recognizable homologs in invertebrate genomes. In at least some cases, vertebrate-specific UCEs are derived from an ancient transposable element that has been exapted to acquire novel functions as enhancers or alternatively spliced exons that might be involved in regulating levels of the proteins they encode [8]. UCEs show a marked shift toward rare derived alleles [9], which is a characteristic of DNA regions under negative selection rather than a reduced mutation rate. However, the reasons for the extreme degree of sequence conserva-

* Corresponding author. Institute for Medical Genetics, Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany.
E-mail address: peter.robinson@charite.de (P.N. Robinson).

0888-7543/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2009.07.005

tion of UCEs remain unclear. Selection exerted on proteins does not normally result in near-identity of coding nucleotide sequences, splice signals at branch points or exon/intron junctions allow sequence variation [10], and the sequences bound by transcription factors show a high degree of degeneracy [11]. Also, deletion of several UCEs in mice failed to reveal notable abnormalities, indicating that the extreme sequence conservation does not necessarily reflect crucial functions required for viability [12].

The original definition of ultraconservation, sequence identity of at least 200 bp between human, mouse, and rat [1], is arbitrary. Many of the UCEs identified using this definition show nucleotide substitutions in the orthologous sequences of other mammals [13], and computational analysis by several groups has shown that sequences that are extremely conserved but either shorter or less than completely identical show similar properties to those of the UCEs [14–16]. Moreover, there are no apparent functional differences between intergenic UCEs and extremely conserved elements not satisfying the original definition of a UCE in enhancer assays in transgenic mice [13].

There are tens of thousands of shorter ultraconserved sequences in the human genome, many of which appear to be mammalian specific [1]. In previous work, we examined one such short ultraconserved element located in an alternate promoter of *FBN1*, and showed that it drove a much higher level of transcription than the three other alternative promoters of *FBN1*, which although conserved in opossum and other mammalian species, display a much lower degree of sequence identity [17]. This led us to develop a computational approach to investigate short ultraconserved elements in the human genome. We sought to characterize elements ultraconserved throughout the mammalian lineage. We therefore defined short ultraconservation to mean sequences with at least 98% identity over at least 30 aligned nucleotides in alignments of humans and six other mammals including the opossum (last common ancestor with humans ~180 million years ago). The arbitrary length threshold of 30 bp is still greater than the extent of sequence identity or near identity that can be explained by any known functional constraint.

Using this definition, we identified 2827 short ultraconserved promoter regions in 1268 human protein-coding genes. We showed that the higher expression associated with the alternative *FBN1* promoter that contains a short ultraconserved element is a general characteristic of promoters with ultraconserved sequences across the genome. Many of the genes associated with short ultraconserved promoter sequences are involved in development and are located in the vicinity of intergenic ultraconserved sequences, suggesting the possibility that whatever mechanisms are responsible for the extreme constraint found in ultraconserved enhancers surrounding developmental genes may also pertain to a subset of proximal promoters of these genes.

Results

SUPRs are present in 6% of human protein-coding genes

We combined pairwise alignments between the human genome and that of mouse and rat (last common ancestor 90 million years ago [Mya] [18]), dog, cow, horse (100 Mya [18]), as well as opossum (180 Mya, [18]) and used them to identify 65,002 short ultraconserved elements (SUE) in the human genome that display at least 98% nucleotide identity over at least 30 bp. Less than 1% of these SUEs showed at least 98% nucleotide identity in pufferfish, frog, and chicken. However, less stringent cutoffs demonstrated that 82% of the SUEs have homologous sequences in chicken (326 Mya, [19]), 58% with the frog (370 Mya [19]), and 32% with the pufferfish (476 Mya [19]) (Table 1).

In this work, we define the promoter region of transcripts of protein-coding genes to be the 5 kb upstream to 50 bp downstream region of an annotated transcription start site (TSS). Using this definition, we identified 2827 SUEs within promoters that show no overlap with any coding sequence. In the following, we will denote the SUEs located within this region as short ultraconserved promoter regions (SUPR). Many promoters contain multiple SUPRs, some of which are separated from one another by short, less conserved sequences. Thus, the 2827 SUPRs form 2304 clusters of SUPRs separated by ≤20 nucleotides (nt) from one another. We assigned each SUPR to the gene (transcript) with the nearest TSS. This set represents 1268 genes, which contain 2688 annotated alternative 5′ exons. At least one SUPR can be identified in the promoter sequences associated with 1404 of the 2688 5′ exons. The set of 1268 SUPR-associated genes corresponds to about 6% of the approximately 21,400 [20] protein-coding genes in the human genome.

Table 1
Short ultraconserved elements in mammals.

	Exon	Intron	Non-coding	Promoter
7-Way	13,708	17,373	31,094	2827
Alignment overlap (≥20nt)				
7+Chicken	12,177 (88.9%)	14,390 (82.8%)	24,644 (79.3%)	1761 (62.3%)
7+Frog	11,203 (81.7%)	9,717 (55.9%)	15,308 (49.2%)	1531 (54.2%)
7+Pufferfish	8779 (64.0%)	4556 (26.2%)	6831 (22.0%)	745 (26.4%)
70% Identity				
7+Chicken	10,330 (75.3%)	12,094 (69.6%)	20,290 (65.3%)	1297 (45.9%)
7+Frog	3108 (22.6%)	1930 (11.1%)	3506 (11.3%)	269 (9.5%)
7+Pufferfish	3993 (29.1%)	983 (5.7%)	1949 (6.3%)	164 (5.8%)
98% Identity				
7+Chicken	1848 (13.5%)	4026 (23.2%)	6208 (20.0%)	287 (10.2%)
7+Frog	239 (1.7%)	519 (3.0%)	900 (2.9%)	46 (1.6%)
7+Pufferfish	50 (0.3%)	34 (0.2%)	59 (0.2%)	6 (0.2%)

65,002 SUEs were tested for overlap with various classes of genomic sequences. The row *7-Way* shows the number of SUEs located in four classes of genomic regions. About 4% of all SUEs are located in promoters. The other three sections of the table show how many of these elements can be identified in non-mammalian species according to increasingly stringent criteria. The analysis was performed by adding an additional pairwise alignment to the 7-way human-centric multiple alignment using the pairwise blastz alignment from UCSC [21,22], and counting how many sequences in chicken (*7+Chicken*), frog (*7+Frog*) or pufferfish (*7+Pufferfish*) displayed an aligned sequence with ≥20 nt overlap, or 70% or 98% identity.

GO overrepresentation analysis [23] for the 1268 genes harboring SUPRs showed strong evidence of enrichment for *transcription regulator activity* ($P<10^{-69}$), *multicellular organismal development* ($P<10^{-19}$), *RNA metabolic process* ($P<10^{-7}$), and *pattern specification process* ($P=0.002$) (Supplementary Table S1). Analysis of protein domains and motifs using Pfam [24] showed a significant enrichment of the HOX domain ($P<10^{-19}$) and HLH domain ($P<10^{-5}$). There was no significant enrichment for the RNA recognition motif RRM, which had been reported for the group of all genes harboring intragenic UCEs ≥200 bp [1].

SUPRs are enriched in extremely conserved high-CpG promoters

It was previously shown that alternative promoters display more sequence conservation than single promoters, and within each class, CpG-poor promoters are more highly conserved than CpG-rich promoters [25]. We compared the conservation of different promoter sets by calculating the percentage identity in the 400 bp upstream to 50 bp downstream region around the TSS in the 7-way alignments. In comparison to the most highly conserved promoter class [25], which consists of CpG-poor alternative promoters (17.6% mean percent identity [25]), SUPR-containing promoters display a significantly higher mean percent identity of 30.3% ($P<10^{-16}$, Wilcoxon; Supplementary Figure S1). This is surprising, since SUPR-containing promoters tend to have a high-CpG content ($P<10^{-5}$, χ^2-test). In general housekeeping genes are biased