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 \(^{a,b,c,\ast}\)

\(^{a}\) 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

A R T I C L E   I N F O

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

Keywords:
Promoter
Ultraconservation
Alternative splicing
Digital gene expression
Genomics

A B S T R A C T

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.

© 2009 Elsevier Inc. All rights reserved.

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).
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.

GO overrepresentation analysis [23] for the 1268 genes harboring SUPRs showed strong evidence of enrichment for transcription regulator activity (P<10−6), multicellular organismal development (P<10−8), 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, χ2-test). In general housekeeping genes are biased

### Table 1

<table>
<thead>
<tr>
<th>Exon</th>
<th>Intron</th>
<th>Non-coding</th>
<th>Promoter</th>
</tr>
</thead>
<tbody>
<tr>
<td>7-Way</td>
<td>11,708</td>
<td>17,373</td>
<td>31,094</td>
</tr>
</tbody>
</table>

### 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 (22.4%) | 745 (26.4%) |

### 70% Identity

| 7-Chicken | 10,330 (75.3%) | 12,094 (69.6%) | 20,290 (65.3%) | 1297 (45.0%) |
| 7-Frog | 3108 (22.6%) | 1930 (11.1%) | 3506 (11.3%) | 269 (9.5%) |
| 7-Pufferfish | 3993 (29.1%) | 983 (5.7%) | 1940 (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.
towards high-CpG promoters and show a lower conservation than low-CpG promoters [25–28], however transcription factors are also enriched in high-CpG promoters and show also a higher level of conservation [27,29], which is similar to the characteristics of SUPRs.

**Intragenic comparison of transcripts demonstrates higher expression for SUPR-associated transcripts**

In previous work, we showed that the SUPR-associated transcript of FBN1 had the highest expression level in luciferase reporter assays [17]. In order to investigate whether this is a general characteristic of SUPRs in the human and mouse genome, we used an approach based on CAGE tags, ESTs, and short reads from 5’ oligocapped cDNAs from MCF7 and HEK293 cells from the database of transcription start sites (DBTSS) [30] to estimate the level of transcription from SUPR-associated and non-SUPR-associated transcripts from genes with multiple promoters including at least one promoter with a SUPR. We repeated the above analysis for each individual library; significantly higher expression of the SUPR-associated transcripts was also found for many of the individual libraries (Table 2).

SUPR-associated transcripts in genes with alternative promoters had 2.2 times more mapped CAGE tags, 1.9 times as many mapped ESTs, and 2.7 times as many short reads than alternative 3’end transcripts of the same genes that are not associated with SUPRs (Table 2). A similar analysis was performed using mouse CAGE tag data. Using the same criteria to assign CAGE tags to transcripts within ±100 bp, SUPR-associated transcripts showed a mean expression level of 33.4 whereas non-SUPR-associated transcripts had 11.5 tags per transcript ($P=0.017$, Wilcoxon). Since relatively few mouse CAGE tags could be assigned to transcripts in this way (see Methods), the analysis was repeated using ±1000 bp criterion to assign CAGE tags to transcripts. Here, SUPR-associated transcripts showed a mean expression level of 18.1 whereas non-SUPR-associated transcripts had 6.8 tags per transcript ($P=10^{-8}$, Wilcoxon). The CAGE data were taken from 17 CAGE libraries obtained from different tissues, and the EST data from 60 EST libraries (see Methods).

We additionally compared tag counts from SUPR- and non-SUPR-associated transcripts of individual genes in order to investigate whether there is a general tendency for the SUPR transcripts to be more highly expressed than other transcripts from the same gene. Although this was not the case for all genes tested, the SUPR-associated transcripts were significantly more likely to be the highest expressed transcript in genes with multiple annotated transcripts (Table 3).

**Table 2**

<table>
<thead>
<tr>
<th>Tissue/Stage/Species Source</th>
<th>Tag number</th>
<th>SUPR</th>
<th>Non-SUPR</th>
<th>Wilcoxon p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adipose CAGE</td>
<td>1345</td>
<td>7.1</td>
<td>2.3</td>
<td>$&lt;10^{-6}$</td>
</tr>
<tr>
<td>Liver CAGE</td>
<td>4975</td>
<td>13.0</td>
<td>9.3</td>
<td>$&lt;10^{-6}$</td>
</tr>
<tr>
<td>Cerebrum CAGE</td>
<td>2570</td>
<td>7.8</td>
<td>2.9</td>
<td>$&lt;10^{-5}$</td>
</tr>
<tr>
<td>Undefined CAGE</td>
<td>2221</td>
<td>8.0</td>
<td>3.7</td>
<td>0.0004</td>
</tr>
<tr>
<td>Cecum CAGE</td>
<td>1792</td>
<td>7.7</td>
<td>2.4</td>
<td>0.0004</td>
</tr>
<tr>
<td>Kidney CAGE</td>
<td>582</td>
<td>2.9</td>
<td>1.0</td>
<td>0.0006</td>
</tr>
<tr>
<td>Large intestine CAGE</td>
<td>1899</td>
<td>8.5</td>
<td>2.6</td>
<td>0.0007</td>
</tr>
<tr>
<td>Small intestine CAGE</td>
<td>625</td>
<td>3.4</td>
<td>1.0</td>
<td>0.007</td>
</tr>
<tr>
<td>Prostate CAGE</td>
<td>52</td>
<td>1.0</td>
<td>0.3</td>
<td>0.008</td>
</tr>
<tr>
<td>Rectum CAGE</td>
<td>256</td>
<td>2.4</td>
<td>0.8</td>
<td>0.01</td>
</tr>
<tr>
<td>Epididymis CAGE</td>
<td>32</td>
<td>0.7</td>
<td>0.2</td>
<td>0.03</td>
</tr>
<tr>
<td>Total CAGE</td>
<td>16,879</td>
<td>35.9</td>
<td>16.7</td>
<td>7.3 $\times 10^{-9}$</td>
</tr>
<tr>
<td>Fetuse EST</td>
<td>1112</td>
<td>7.0</td>
<td>1.6</td>
<td>$&lt;10^{-8}$</td>
</tr>
<tr>
<td>Brain EST</td>
<td>3182</td>
<td>15.6</td>
<td>5.5</td>
<td>$&lt;10^{-7}$</td>
</tr>
<tr>
<td>Adult EST</td>
<td>2190</td>
<td>8.7</td>
<td>5.6</td>
<td>$10^{-6}$</td>
</tr>
<tr>
<td>Embryonic tissue EST</td>
<td>311</td>
<td>2.4</td>
<td>1.1</td>
<td>$&lt;10^{-5}$</td>
</tr>
<tr>
<td>Uterus EST</td>
<td>409</td>
<td>3.5</td>
<td>1.4</td>
<td>0.0002</td>
</tr>
<tr>
<td>Eye EST</td>
<td>365</td>
<td>3.5</td>
<td>1.6</td>
<td>0.0004</td>
</tr>
<tr>
<td>Muscle EST</td>
<td>220</td>
<td>2.5</td>
<td>0.7</td>
<td>0.0008</td>
</tr>
<tr>
<td>Trachea EST</td>
<td>170</td>
<td>2.4</td>
<td>0.8</td>
<td>0.002</td>
</tr>
<tr>
<td>Ovary EST</td>
<td>96</td>
<td>1.4</td>
<td>0.8</td>
<td>0.003</td>
</tr>
<tr>
<td>Neonate EST</td>
<td>46</td>
<td>1.3</td>
<td>0.6</td>
<td>0.004</td>
</tr>
<tr>
<td>Mouth EST</td>
<td>107</td>
<td>1.9</td>
<td>0.6</td>
<td>0.004</td>
</tr>
<tr>
<td>Vascular EST</td>
<td>98</td>
<td>2.0</td>
<td>0.8</td>
<td>0.006</td>
</tr>
<tr>
<td>Embryoid body EST</td>
<td>105</td>
<td>1.3</td>
<td>0.7</td>
<td>0.02</td>
</tr>
<tr>
<td>Infant EST</td>
<td>88</td>
<td>1.6</td>
<td>0.7</td>
<td>0.02</td>
</tr>
<tr>
<td>Prostate EST</td>
<td>163</td>
<td>1.4</td>
<td>0.9</td>
<td>0.02</td>
</tr>
<tr>
<td>Total EST</td>
<td>13,099</td>
<td>46.6</td>
<td>24.7</td>
<td>3.1 $\times 10^{-7}$</td>
</tr>
<tr>
<td>MCF7 DBTSS</td>
<td>159,499</td>
<td>256.2</td>
<td>110.0</td>
<td>1.3 $\times 10^{-4}$</td>
</tr>
<tr>
<td>Hek293 DBTSS</td>
<td>146,196</td>
<td>246.1</td>
<td>76.4</td>
<td>1.8 $\times 10^{-5}$</td>
</tr>
<tr>
<td>Mouse CAGE</td>
<td>9016</td>
<td>18.1</td>
<td>6.2</td>
<td>2.6 $\times 10^{-6}$</td>
</tr>
</tbody>
</table>

CAGE tags, ESTs, and Solexa reads were counted in ±100 bp around the annotated TSS of SUPR-associated and non-SUPR transcripts. For both groups mean tag numbers are shown. 16 of the 17 CAGE tag libraries and 50 of 60 EST libraries demonstrated a higher expression of the SUPR-associated transcripts. The table shows the libraries in which the difference was statistically significant. $P$-values for enrichment of tag numbers were calculated by a Wilcoxon-test with Bonferroni correction. For all data sets SUPR-associated transcripts show the higher expression level. The columns SUPR and non-SUPR display the mean number of tags per promoter.

**Table 3**

<table>
<thead>
<tr>
<th>Data source</th>
<th>SUPR</th>
<th>Non-SUPR</th>
<th>$E_{SUPR}$</th>
<th>$E_{non-SUPR}$</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAGE</td>
<td>136</td>
<td>91</td>
<td>94.6</td>
<td>132.5</td>
<td>$1.4 \times 10^{-7}$</td>
</tr>
<tr>
<td>EST</td>
<td>69</td>
<td>59</td>
<td>53.0</td>
<td>75.0</td>
<td>0.06</td>
</tr>
<tr>
<td>MCF7</td>
<td>163</td>
<td>154</td>
<td>130.8</td>
<td>186.2</td>
<td>0.01</td>
</tr>
<tr>
<td>HEK</td>
<td>181</td>
<td>159</td>
<td>144.6</td>
<td>195.4</td>
<td>6.6 $\times 10^{-7}$</td>
</tr>
</tbody>
</table>

The columns show the number of genes, for which either a SUPR-associated transcript or a non-SUPR-associated transcript has the highest tag count. These values were compared to expected values using a $\chi^2$-test (see Methods). For all datasets, there was a preference for SUPR-associated transcripts to be more highly expressed than non-SUPR expressed transcripts from the same genes. The difference was statistically significant for the CAGE and the Solexa read data.
highly conserved non-coding elements (HCNEs) that putatively regulate a target gene. These target genes themselves show an enrichment in transcription factors involved in embryonic development and differentiation. However, GRBs also include a number of so called ‘bystander’ genes which are functionally unrelated to target genes and show a much broader expression. We found that 134 of 270 of target genes that were identified by Akalin et al. [29] are equivalent to SUPR-associated genes ($P < 10^{-16}$, Fisher’s exact test). In order to investigate whether SUPR-associated genes in general share certain features with the target genes of GRBs, we compared the tissue-specific expression pattern of SUPR-associated transcripts to those of neighboring genes by counting the number of CAGE tags for each transcript and tissue and defined the breadth of a transcript as the number of tissues, in which its tag count is higher than the average tag count for this tissue. There was a minor trend towards more tissue-specific expression of SUPR-associated transcripts which had a mean breadth of 4.0 in comparison to transcripts of neighboring genes (mean 4.4; $P = 0.02$ Wilcoxon). The equivalent test for ESTs was not significant.

**Effect of alternative definitions**

As described in the Introduction, a number of different definitions of ultraconservation have been proposed. The threshold between conservation and ultraconservation is clearly arbitrary, and one will identify different numbers of “ultraconserved” sequences depending on the definition used. For instance, the original definition of 200 bp and 100% identity in human, mouse and rat identified 481 ultraconserved sequences [1], and we identified 96 regions of ≥200 bp with 100% identity in all seven genomes. Our definition of short ultraconservation is clearly also arbitrary, and different numbers of SUEs are identified using different thresholds (Table 1; Fig. 1). In order to estimate the effect of different definitions of short ultraconservation on the main findings of this work, we repeated the expression analysis as described above for alternative definitions of short ultraconservation.

Higher expression levels were observed for SUPR-associated transcripts using a broad range of definitions of short ultraconservation. Interestingly, the difference between SUPR-associated and non-SUPR-associated transcripts was most pronounced if SUPRs are restricted to lie within 200 bp of the annotated TSS and for higher percent identity thresholds (Fig. 2). In contrast, the minimum length threshold did not have a significant effect within the range 30–100 bp (Fig. 2).

**Co-occurrence of SUPRs and intergenic SUEs**

As noted above, GO analysis suggests that genes containing SUPRs tend to be enriched for functions such as transcription factor activity and DNA binding, which is similar to genes located in the vicinity of UCEs in non-coding regions [1]. Such intergenic UCEs can occur as clusters that span hundreds of kilobases in gene-poor regions around their presumptive target genes, suggesting that UCEs act as essential long-range modulators of gene expression [14]. Since GO-annotations of SUPR-associated genes overlap with those of genes in the vicinity of intergenic UCEs [1,14,15], we asked whether SUPRs and intergenic SUEs tend to co-occur in the promoters and neighboring regions of certain transcription factors.

We identified a highly significant co-occurrence of SUEs and SUPRs. Non-overlapping windows of 2 Mb were slid across all human chromosomes and the counts of intergenic SUEs (Table 1) and SUPRs were recorded. Fig. 3 shows the distribution of SUEs and SUPRs on human chromosome 7 (Similar figures for all chromosomes are shown as Supplementary Figure S2). The genome-wide Spearman correlation was $\rho = 0.28$ ($P < 10^{-16}$). We repeated the analysis for window sizes between 100 kb and 50 Mb. In each case, a significant correlation was found. The highest correlation was for a window size of 20 Mb ($\rho = 0.56$; $P < 10^{-14}$).

**Overrepresented transcription factor binding sites in SUPRs and intergenic SUEs**

We searched for overrepresented predicted transcription factor binding sites (TFBS) in the SUPRs and intergenic SUEs, by comparing occurrences of predicted binding sites against a background of adjacent, less conserved genomic sequences (see Methods). Sequences matching 56 TRANSFAC matrices were significantly enriched in SUEs (Supplementary Table S2) and 97 were significantly enriched in SUPRs (Supplementary Table S3), 34 matrices were enriched in both sets. We analyzed the protein sequences of the corresponding transcription factors for common
motifs and found that the homeobox domain (Pfam PF00046) was significantly enriched in both SUPRs ($P < 10^{-2}$) and SUEs ($P < 10^{-5}$), which is in agreement with previous studies on non-exon UCEs [29,33].

In order to investigate the differences between SUPRs and intergenic SUEs, we tested for enriched motifs in SUPRs relative to SUEs. In total we found 106 TRANSFAC matrices to be enriched in SUPRS. The most significant enrichment was found for E2F and SP1, as well as for members of the basal transcription factor machinery such as Inr and TFIll. (Supplementary Table S4).

**Discussion**

Non-coding intergenic UCEs have been identified in the vicinity of genes with prominent roles in development. Many of these UCEs have been shown to function as developmentally relevant enhancers in the central nervous system and other tissues [34,35]. The unusually high degree of sequence conservation of these UCEs has suggested that some feature of vertebrate developmental regulation may impose an extraordinarily strong selection pressure on a subset of non-coding regulatory sequences that cannot be...
explained by previously known mechanisms such as conservation of transcription factor binding sites [36]. In the current work, we have characterized a new class of ultraconserved sequence termed SUPR. These sequences are present in the promoters of about 6% of human genes. These genes, as well as the individual transcripts associated with the SUPRs, show a number of distinguishing characteristics. SUPRs are enriched in genes playing a role in regulation and development and are enriched in extremely conserved high-CpG promoters. SUPR-associated genes show a highly significant tendency to occur in regions that are also enriched for intergenic UCEs in the vicinity of developmental genes. Some of the above observations about SUPRs are similar to observations made about other ultraconserved sequences [1–6]. The main novel finding of our study is that SUPRs are associated with a significantly higher expression of the associated transcripts. This is true both for genes with multiple promoters, some of which have SUPRs, as well as for the genome-wide average, and holds even after correction for CpG class, total number of isoforms per gene, and for a number of alternative SUPR-definations, and was most marked for the class of SUPRs with higher percentage identity that are located close to the TSS. Additionally, SUPRS display a distribution of predicted TFBS that differs from that of intergenic SUEs, because SUPRs are enriched for core-promoter motifs such as Inr in addition to predicted Hox binding sites. We interpret these observations as indications that SUPRs represent a novel class of ultraconserved sequence.

The fact that SUPR-associated transcripts show higher levels of expression in both humans and mice, whose last common ancestor lived about 90 million years ago [18], suggests that a functional correlate of many SUPRs, viz. stronger expression of the SUPR-associated transcripts, has also been conserved during evolution, although clearly more data from other species will be needed to confirm this observation. Usage of alternative promoters can be associated with different levels of transcription initiation, different tissue specificities, and in some cases generation of protein isoforms differing at their amino terminus. Thus, alternative promoters may be utilized to achieve different expression levels in different tissues or in response to specific signals [37]. This suggests the possibility that, in some cases, the evolutionary constraints maintaining ultraconservation in promoters may be related to a need for strong expression of certain 5’ alternate transcripts.

A continuum of (ultra)conservation can be observed in multiple genome alignments (Fig. 1). It remains unknown whether the characteristics of the SUPRs described in this work, and indeed of ultraconserved sequences in general [1–6], merely represent exaggerated versions of the characteristics of all conserved sequences, or whether there is a fundamental mechanistic difference between ultraconserved and other conserved sequences. To our knowledge, none of the research published to date on this topic provides a compelling answer to the question why ultraconserved sequences should show a degree of conservation that is so much higher than that typically seen on the basis of constraints related to splicing, transcription factor binding, or protein-coding sequences. As one possibility, we note that some ultraconserved sequences have been suggested or shown to underlie multiple constraints [2,38,39]. In this light, it is interesting that there is a highly significant tendency of SUPRs towards co-occurrence with intergenic SUEs (Fig. 3), many of which are developmental enhancers or silencers [2–4]. We speculate that a similar selective pressure might act on developmentally crucial enhancers as on the promoters that are activated by the enhancer resulting from long-range interactions between these two types of ultraconserved sequences, whereby mammalian development would be controlled by a multiply interconnected core network of mutually regulating ultraconserved promoters and enhancers.

Methods

7-way human-centric multiple genome alignments

We downloaded pairwise blastz [22] alignments from the UCSC [21] website and used the MULTIZ program [40] to create an alignment of 6 mammalian genomes (mouse (mm8), rat (rn4), dog (canFam2), cow (bosTau3), horse (equCab1), and opossum (monDom4)) to the human (hg18) genome. The alignment order was defined by the following subtree of the UCSC 28-way alignment [41]: (((human (mouse rat)) ((dog horse cow)) opossum). We also retrieved the UCSC known genes [42] and isoform annotation together with the repeat-masked genomic DNA sequences for human (hg18) and mouse (mm8) from UCSC. We additionally downloaded pairwise blastz alignments between human and chicken (Gallus gallus, galgal3), frog (Xenopus tropicalis, xenTro2), and pufferfish (Fugu rubripes, fr2) and extracted all aligned blocks (≥20 nt) and all conserved blocks with ≥70% and ≥98% identity. These blocks were then compared with the locations in the human genome of the SUEs defined by the 7-way alignment (Table 1).

Definition of SUEs and SUPRs

We defined short ultraconserved elements (SUE) as consecutive blocks of l ≥ 30 bp in all species and a total identity I ≥ 98%. The identity I of a multiple alignment block A of length l and n = 7 species is given by

\[ I_{A,n,l} = \frac{1}{l} \sum_{i=1}^{l} \frac{x(A_i)}{n}, \]  

where \( x(A_i) \) denotes the number of occurrences of the most frequent base at column \( A_i \).

We defined the promoter region of each transcript to be the 5000 bp upstream and 50 bp downstream of the annotated TSS of an annotated known gene [42]. A short ultraserving promoter region (SUPR) denotes a SUE overlapping the promoter region and not overlapping with any coding sequence. We assigned each SUPR to the transcript with the nearest TSS, the corresponding promoters are denoted as SUPR-containing promoters if they harbor one or more SUPRs. SUEs lying within or overlapping with exons were classified as ‘exon’ SUEs, and SUEs contained completely in introns were classified as ‘intron’ SUEs. All remaining SUEs are assigned to the non-coding category.

Definition of promoter classes

In order to investigate whether CpG content and the presence of ultraconserved sequences is correlated with the number of isoforms and alternative promoters, we adapted a definition of promoter classes from [31]. First, CpG content was determined by the ratio of observed to expected CpG dinucleotides in sliding 500-bp windows with 5-bp offsets:

\[ \text{CpG ratio} = \frac{\text{CpG}s \times \text{number of bp}}{\text{G}s \times \text{C}s} \]  

The number of base pairs for a 500 bp window is taken to be 500 minus the number of uncalled bases (Ns) or repeats called by RepeatMasker [43]. High-CpG promoters (HCP) were defined as those containing at least one 500 bp area with CpG ratio above 0.75 and GC content above 55%. Low-CpG promoters (LCP) do not contain even a single 500-bp area with a CpG ratio above 0.48. Intermediate CpG promoters (ICP) are neither HCPs or LCPs.
Gene ontology analysis

We downloaded Gene Ontology Association file for human from the European Bioinformatics Institute (EBI) site and used Ensembl Biomart to map the UCSC known gene identifiers to Uniprot IDs. In order to avoid biases due to genes giving rise to multiple proteins, we restricted the number of proteins for each gene to 1. In cases where multiple proteins were annotated for one gene, we randomly picked one representative. We then used the Ontologizer [44] with Parent–Child intersection analysis and Bonferroni correction in order to perform GO term overrepresentation analysis.

Analysis of protein domain content

We retrieved the protein sequences for the 43,141 known genes using the known gene2pep-table [45]. We used InterProScan [46] to test for an enrichment of protein domains of our test set using the $\chi^2$-test with Bonferroni correction.

CpG-bias

We defined promoter regions as being the 5000 nucleotides upstream and 50 nucleotides downstream from annotated transcription start sites (TSS) in the known gene dataset [42]. We divided all promoters into high-CpG (HCP), intermediate-CpG (ICP), and low-CpG (LCP) classes [31], and compared the distribution of SUPR-containing promoters to the genome-wide average using a $\chi^2$-test.

Expression analysis

We downloaded CAGE and EST sequences and annotation from the FANTOM3 database [47] and UniGene [48]. In addition we downloaded mapped short reads from 5’-oligocapped cDNAs for MCF7 and HEK293 cells [30] and treated these data in the same way as the CAGE tags.

We extracted ±100 bp TSS regions and exon annotations for each annotated protein-coding gene and ran BLAT [49] to map the tags against the target sequences, whereby we extracted only 100% identical hits. For the EST data we used the exon annotation from the UCSC database [21] and assigned each exon to one or more TSSs. Only exons that could be unambiguously assigned to a single TSS were used for the EST expression analysis.

We filtered out genes that do not have orthologs in all 6 species (Ensembl52 [20]). CAGE tags from mouse were also obtained from FANTOM3 [47]. We used the program Eland (Illumina) to map those reads to the mouse genome (version mm8 from UCSC). In total 191,332 tags mapped uniquely to the ±100 bp regions of all annotated TSSs (known Gene from UCSC) and 377,667 to the ±1000 bp regions. However only 6154 tags mapped to ±100 bp TSS regions of SUPR-associated transcripts, and 9016 tags mapped to ±1000 bp regions. This dataset was used for the analysis of the expression of alternative promoters as described for the human data.

Genewise comparison of transcript expression

For each gene with at least one SUPR-associated transcript as well as at least one non-SUPR-associated transcript, we tested whether the SUPR-associated transcript had the highest tag number amongst all transcripts of this gene.

We calculated the number of times a SUPR-associated transcript would be expected to be the highest expressed transcript purely at random as follows. For a gene $i$ with $N$ TSSs of which $n$ are associated with at least one SUPR, $K$ denotes the number of TSSs with the highest tag number (typically $K = 1$). The probability $p_i$ of choosing $k \geq 1$ times the TSS with the highest tag number can be computed by a hypergeometric distribution:

$$p_i(K, n, N) = \sum_{k=1}^{n} \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}.$$  

If there is only one SUPR-associated transcript, and only one transcript with the highest expression, then the probability simplifies to $1/N$. Eq. 3 will also compute the correct probability of a SUPR-associated transcript being one of the most highly expressed transcripts in case of ties or in case more than one transcript of a gene is associated with a SUPR. To test for enrichment we compute an expected number of successes $pm$, where $m$ is the number of genes with multiple promoters including at least one that is associated with a SUPR. Then $p = \frac{m}{\pi} \sum_{i=1}^{m} p_i$ is the mean chance probability of observing at least one SUPR-associated transcript being the most highly expressed. These values are used to perform a $\chi^2$-test.

Genome-wide expression analysis

In order to investigate the level of expression for all SUPR-associated transcripts (single promoter and alternative promoter), we compared this set to three other promoter classes, defined by CpG content [31] (see above).

We counted the BLAT [49] hits of 1,020,853 CAGE tags, 40,615 ESTs, and 30,889,015 short reads from 5’-oligocapped cDNAs to the ±100 bp TSS regions and unique exons in case of ESTs for all transcripts and compared the tag distributions using a Wilcoxon-test.

Broadness of expression

To evaluate the breadth of expression between SUPR-associated transcripts and the transcripts of neighboring genes, we defined a breadth for each transcript as the number of tissues in which the transcript had a higher tag count than the lower 50% quantile of all transcripts. We compared these numbers between SUPR-associated transcripts and the transcripts of neighboring genes using a Wilcoxon-test.

Effect of alternative definitions

We repeated the above analysis for higher expression of SUPR-associated transcripts using other definitions with varying sizes of the upstream region (200, 1000, and 5000 bp), minimum SUPR length (30, 40, 50, and 100 bp), and different percentage identity thresholds (80, 90, 98, and 100%). For each combination of upstream region, minimum SUPR length, and percentage identity an enrichment was computed as the ratio between mean tag number for SUPR-associated and non SUPR-associated transcripts. Fig. 2 displays the sum of enrichments for ESTs and CAGE tags.

Overrepresentation of transcription factor binding sites

We looked for overrepresented motifs in SUPRs and SUEs by comparing predicted binding sites from the TRANSFAC 10.3 MATCH [50] program in the ultraconserved sequences and five flanking nucleotides on each side of the ultraconserved block to a background set of sequences located 100 bp upstream and downstream sequences not overlapping and coding sequence or another SUE. We tested for enrichments of matrices using a $\chi^2$-test with Bonferroni correction (585 matrices). We performed a similar enrichment analysis using a set of random sequences created by a first order Markov chain with a dinucleotide.
distribution identical to that of the original sequences. Matrices that were found to be significantly enriched in only one of the two analyses were filtered out.

Authors contributions

C.R. and P.N.R. designed the study. C.R., S.K., M.H.S., T.M. and S.B. carried out the analyses, calculations and performed the statistical analyses. C.R. and P.N.R. prepared the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the Bundesministerium für Bildung und Forschung (0313911) and the Deutsche Forschungsgemeinschaft (SFB 760). P.N.R. gratefully acknowledges additional support from the National Marfan Foundation of the USA. The authors acknowledge the work of the Horse Genome Project and the Broad Institute, which conducted the horse whole genome sequencing under the auspices of the National Human Genome Research Institute (NHGRI).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at do:10.1016/j.ygeno.2009.07.005.

References


C. Rödelsperger et al. / Genomics 94 (2009) 308–316


