DNA Research **18**, 177–187, (2011)
Advance Access Publication: 17 May 2011
doi:10.1093/dnares/dsr010

Exploration for Functional Nucleotide Sequence Candidates within Coding Regions of Mammalian Genes

Rumiko Suzuki[1,2,†] and Naruya Saitou[1,2,*]

Department of Genetics, School of Life Science, Graduate University for Advanced Studies, Mishima 411-8540, Japan[1] and Division of Population Genetics, National Institute of Genetics, Mishima 411-8540, Japan[2]

*To whom correspondence should be addressed. Tel. +81 559-81-6790. Fax. +81 559-81-6789.
E-mail: saitounr@lab.nig.ac.jp

Edited by Katsumi Isono
(Received 12 September 2010; accepted 14 April 2011)

Abstract

The primary role of a protein coding gene is to encode amino acids. Therefore, synonymous sites of codons, which do not change the encoded amino acid, are regarded as evolving neutrally. However, if a certain region of a protein coding gene contains a functional nucleotide element (e.g. splicing signals), synonymous sites in the region may have selective pressure. The existence of such elements would be detected by searching regions of low nucleotide substitution. We explored invariant nucleotide sequences in 10 790 orthologous genes of six mammalian species (*Homo sapiens*, *Macaca mulatta*, *Mus musculus*, *Rattus norvegicus*, *Bos taurus*, and *Canis familiaris*), and extracted 4150 sequences whose conservation is significantly stronger than other regions of the gene and named them significantly conserved coding sequences (SCCSs). SCCSs are observed in 2273 genes. The genes are mainly involved with development, transcriptional regulation, and the neurons, and are expressed in the nervous system and the head and neck organs. No strong influence of conventional factors that affect synonymous substitution was observed in SCCSs. These results imply that SCCSs may have double function as nucleotide element and protein coding sequence and retained in the course of mammalian evolution.

Key words: mammal; protein coding; nucleotide conservation

1. Introduction

The neutral theory of molecular evolution[1,2] predicts that synonymous sites of codons are evolving faster than non-synonymous sites because of the smaller selective pressure. This is true in general; however, several factors are known to influence on a certain region of a coding sequence and suppress synonymous substitution.

One of the well-known factors is the codon bias towards optimum codons. Optimum codons reflect the composition of genomic tRNA pool.[3–5] Because optimum codons are advantageous for fast and accurate translation, highly expressed or biologically important genes would have more optimum codons than others.[6–8] Changes from an optimum codon to a non-optimum codon will be suppressed in such genes. Because optimum codons are similar in closely related species, highly expressed genes tend to show similar codon usage; therefore synonymous substitution is lowered. In fact, the requirement for translational efficiency or accuracy enhances the optimum codon usage and suppresses nucleotide changes through purifying selection.[5,6,9–11] Codon bias towards optimum codons is strong in fast-growing organisms such as *Escherichia coli* or *Saccharomyces cerevisiae*, but generally weak in

† Present address: Department of Environmental and Preventive Medicine, Oita University, Yufu 879-5593, Japan.

species with slow growth rate small population size.[12,13]

Another factor is exonic splicing enhancer or silencer, which are splicing signals embedded in exons.[14,15] Existence of such elements lowers the synonymous substitution.[16,17] In addition, ultraconserved elements (UCEs), which majorly reside in non-protein coding regions, sometimes extend to coding regions.[18] In mammals, UCEs are reported to exist near to or overlap with genes associated with nucleotide binding, transcriptional regulation, RNA recognition motif, zinc finger domain, and homeobox domain.[18–20] Hox genes also contain long conserved nucleotide regions other than UCEs outside the homeobox domain.[21]

Although the primary role of protein coding region is to encode amino acids, there may be also functional nucleotide elements embedded within coding regions. For example, transcription-factor-binding sites are found in coding regions,[22] messenger RNAs are targeted by various post-transcriptional regulations,[23] and the requirement for a specific secondary structure for RNA editing decreases synonymous substitution.[24,25]

Functional nucleotide elements are extensively explored in the non-coding regions,[26–30] but less studies have been done to explore probable functional elements within the coding regions. We extracted significantly conserved coding sequences (SCCSs) from orthologous genes of six mammalian species (human, rhesus macaque, mouse, rat, cow, and dog), and compared genes containing SCCSs and genes without SCCSs. Analyses on gene ontology (GO), InterPro codes, and KEGG pathways enlighten difference between the two gene groups. We also investigated RNA secondary structures, codon preference, GC content, exonic splicing signals, and gene expression of SCCSs to survey the influence of these factors.

2. Materials and methods

2.1. Genome data

We obtained peptide and nucleotide sequences of orthologous genes of six mammalian species (*Homo sapiens*, *Macaca mulatta*, *Mus musculus*, *Rattus norvegicus*, *Bos taurus*, and *Canis familiaris*) from Ensembl database[31] version 54 (http://May2009.archive.ensembl.org/index.html). These species are selected considering genome data quality and evolutionary diversity. We selected one-to-one type single copy orthologs and compiled 10 790 orthologous gene sets (Supplementary file S1). We then constructed multiple alignments of peptide sequences using ClustalW[32] and constructed nucleotide alignments based on the peptide alignments. From the nucleotide alignments, we extracted sequences that are invariant among all the species.

2.2. Identification of SCCSs

We performed permutation simulation to identify SCCSs, or abbreviated as SCCSs, which are invariant longer than 10 codons. This length is set to confine the permutation run time within a feasible range. For an *N*-codon long alignment, we generated a non-redundant series of random numbers from 1 to *N*, and permuted codon columns (rows of codons in the same site of the alignment) according to the generated random numbers. In this process, gap sites are fixed and the rest of the sites are permuted. Then the length and numbers of invariant sequences in the permuted alignment are counted. We repeated this process 500 000 times per ortholog set and took averages of the frequency of invariant sequences. We used the length and averaged frequency of invariant sequences obtained from the permutation results as random expectation, and evaluated the probability of invariant sequences in the original alignment based on the expectation. This approach helps identify sequences whose conservation is rare to occur in the substitution background of each alignment. Multiple testing correction of p-values is done by FDR (false discovery rate).[33] Then we identified invariant sequences longer than 10 codons and $P < 0.01$ as SCCSs.

2.3. Analysis on GO, InterPro, and KEGG pathways

Protein coding genes that contain at least one SCCS is named SCCS genes and those that do not contain an SCCS is named non-SCCS genes. We used Fatigo web service (http://babelomics.bioinfo.cipf.es/functional.html) to identify GO terms, InterPro codes, and KEGG pathways that are significantly enriched with the SCCS gene group or with the non-SCCS gene group. Fatigo accepts a list of Ensembl gene IDs as input and provides *P*-values for enrichment of the above terms. *P*-values are calculated by Fisher's exact test and corrected by FDR. We used Ensembl gene IDs of *H. sapiens* as input and performed two-tailed comparison between SCCS genes and non-SCCS genes.

2.4. Analysis on preferred codons and average codon degeneracy of SCCSs

We defined preferred codons as the most frequently used codons for a given amino acid referring to Codon Usage Database (http://www.kazusa.or.jp/codon/index.html) provided by Kazusa DNA Research Institute. Because the codon usage pattern was similar among the six species we used, the codon

table of human was used as the representative. We counted the number of preferred codons in a SCCS and divided it by the codon length of the SCCS, and then used the quotient as the ratio of preferred codons. The average codon degeneracy is calculated by summing up the degeneracy of each codon and dividing it by the codon length of the SCCS.

2.5. *Prediction of RNA secondary structures*

We computationally predicted secondary structures and free folding energy of SCCSs using Vienna RNA software package[34] (http://www.tbi.univie.ac.at/~ivo/RNA/). Because folding free energy varies depending on the sequence length, we constructed free energy distribution by 1000 randomly chosen sequences for each length (33–246 nucleotides). The *P*-value for a given free energy was evaluated based on these distributions. Multi testing correction is done by FDR.

2.6. *Evaluation of exonic splicing enhancer density*

We obtained 238 hexamers from RESCUE-ESE Web Server[34] as candidates of exonic splicing enhancers. We counted the number of the hexamers in the SCCSs and the rest of the coding regions of the 10 790 genes (human sequences). We also measured the total nucleotide length of the SCCS and the other regions and applied the chi-square test for the ratio of hexamers.

2.7. *Analysis on gene expression*

We used EGenetics (http://www.nhmrc.gov.au/your_health/egenetics/index.htm) to investigate gene expression of SCCS genes and non-SCCS genes. Human anatomical system data, which give information about in which organs a gene is expressed, were obtained from EGenetics database by Ensemble Biomart. We counted how many of SCCS genes and non-SCCS genes are expressed in each organ and divided the numbers by the total number of SCCS genes and non-SCCS genes, respectively. Then we performed the Fisher's exact test to evaluate whether the difference between SCCS and non-SCCS genes is significant. All *P*-values were corrected by FDR.

3. Results

3.1. *Identification of SCCSs*

If an alignment has a high ratio of conservation, long invariant sequences may occur easily, and vice versa. Therefore, the rareness of invariant sequences differs depending on the conservation background in each alignment. To compensate this, we performed permutation simulation. The idea of permutation is to count the length and frequency of invariant sequences after random change of loci and use the result as the random expectation.

We used the frequency distribution of invariant sequences in the permuted alignments as the random expectation and evaluated probability of invariant sequences in the original alignments, and defined invariant sequences longer than 10 codons and whose probability is below 0.01 (corrected by FDR) as SCCSs.

In total, 4150 SCCSs (192 306 bp) were obtained from 2273 alignments of 10 790 orthologous gene sets (Supplementary file S2). This occupies 0.94% of the coding region of the 10 790 genes. Table 1 shows the number of SCCSs per gene and the number of genes that contain that number of SCCSs. Figure 1 is a graph of lengths and numbers of SCCSs (grey bars). Black dots indicate the random expectation obtained from the permuted alignments.

In the permuted alignments, there are 141 sequences (their total length is 5550 bp) whose probability is below 0.01, which means the region size of SCCS is 35-fold larger than this expectation (Supplementary file S3). The χ^2 test between SCCSs

Table 1. Number of SCCSs in coding genes

	Number of SCCSs (per gene)	Number of genes
SCCS genes	1	1366
	2	475
	3	219
	4	105
	5	42
	>5	66
Non-SCCS genes		8517

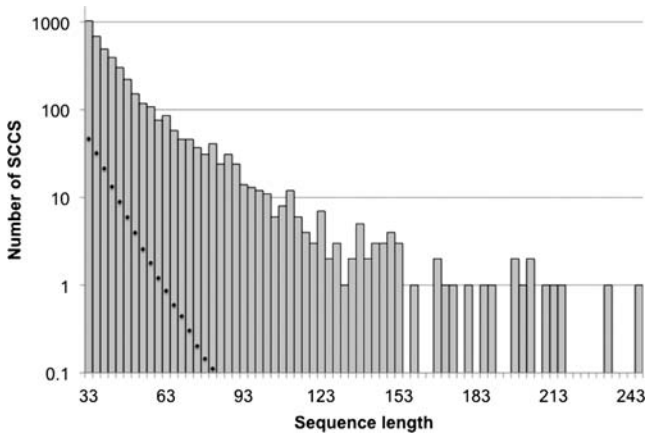

Figure 1. The length and number of SCCSs. X and Y-axes represent the length and frequency of SCCSs, respectively. Grey bars show the length and the number of SCCSs. Black dots indicate the number of sequences with probability below 0.01 in the permuted alignments.

and the permutation result showed a significant difference ($P < 2.2E-16$).

If invariant and variant sites distributed randomly in the original alignment, the frequency of invariant sequences in the permuted alignments would show similar frequency to the original alignment because of the randomness at the start point. The difference before and after the permutation suggests that the distribution of invariant sites in the original alignments is rather clustered than being random.

3.2. *GO, InterPro, and KEGG pathways enriched in SCCS containing genes*

We used Fatigo web service to investigate difference in GO, InterPro codes, and KEGG pathways between the SCCS genes and non-SCCS genes. The difference was evaluated by the two-tailed Fisher's test, and *P*-values were corrected by FDR. Tables 2 and 3 show GO terms, InterPro codes, and KEGG pathways significantly enriched ($P < 0.01$) in SCCS genes.

The all terms in GO Biological process section is related to developmental process. One term of

Table 2. GO terms significantly ($P < 0.01$) enriched with SCCS genes

Terms	*P**	In SCCS containing genes (%)[a]	In non-SCCS containing genes (%)[b]	Fold[c]
Biological process				
GO:0035136: Forelimb morphogenesis	7.75E−05	0.53	0.04	13.25
GO:0035115: Embryonic forelimb morphogenesis	2.50E−04	0.48	0.04	12.00
GO:0060070: Canonical Wnt receptor signalling pathway	2.11E−03	0.4	0.04	10.00
GO:0035137: Hindlimb morphogenesis	4.88E−04	0.53	0.06	8.83
GO:0001702: Gastrulation with mouth forming second	1.78E−03	0.44	0.05	8.80
GO:0009954: Proximal/distal pattern formation	3.80E−04	0.57	0.07	8.14
GO:0031128: Developmental induction	1.03E−03	0.53	0.07	7.57
GO:0048593: Camera-type eye morphogenesis	1.22E−05	0.88	0.13	6.77
GO:0021510: Spinal cord development	2.37E−04	0.75	0.13	5.77
GO:0031016: Pancreas development	1.85E−03	0.62	0.12	5.17
Cellular components				
GO:0014704: Intercalated disc	8.43E−03	0.35	0.04	8.75
GO:0043198: Dendritic shaft	2.38E−03	0.48	0.06	8.00
GO:0030425: Dendrite	2.38E−03	2.29	1.1	2.08
GO:0043025: Neuronal cell body	2.38E−03	2.24	1.11	2.02
GO:0015629: Actin cytoskeleton	1.15E−03	3.39	1.8	1.88
GO:0043005: Neuron projection	1.15E−03	4.09	2.32	1.76
Molecular function				
GO:0035254: Glutamate receptor binding	3.41E−03	0.35	0.02	17.50
GO:0005072: Transforming growth factor beta receptor, cytoplasmic mediator activity	8.92E−03	0.31	0.02	15.50
GO:0004843: Ubiquitin-specific protease activity	3.71E−03	0.48	0.07	6.86
GO:0031625: Ubiquitin protein ligase binding	6.44E−03	0.62	0.14	4.43
GO:0003725: Double-stranded RNA binding	6.44E−03	0.62	0.14	4.43
GO:0042054: Histone methyltransferase activity	8.64E−03	0.57	0.13	4.38
GO:0050825: Ice binding	6.50E−11	2.86	0.76	3.76
GO:0004221: Ubiquitin thiolesterase activity	4.18E−04	1.01	0.27	3.74
GO:0005199: Structural constituent of cell wall	1.34E−03	0.92	0.25	3.68
GO:0003682: Chromatin binding	4.08E−08	2.15	0.59	3.64

[a]The percentages of SCCS genes that have the GO term.
[b]The percentages of non-SCCS genes that have the GO term.
[c]The fold of 'a' to 'b'. Terms are listed in the descending order of the fold difference. Terms with the highest 10-folds are shown for Biological Process and Molecular function.
*Probability for enrichment of the GO term in the SCCS group.

Table 3. InterPro and KEGG terms significantly ($P < 0.01$) enriched with SCCS genes

Terms	P*	In SCCS containing genes (%)[a]	In non-SCCS containing genes (%)[b]	Fold[c]
InterPro				
IPR003619: MAD homology 1, Dwarfin-type	1.55E−04	0.44	0.01	44.00
IPR001827: Homeobox protein, antennapedia type, conserved site	1.77E−03	0.44	0.04	11.00
IPR000569: HECT	1.26E−05	0.75	0.07	10.71
IPR002077: Voltage-dependent calcium channel, alpha-1 subunit	5.65E−04	0.53	0.05	10.60
IPR010982: Lambda repressor-like, DNA-binding	1.49E−03	0.48	0.05	9.60
IPR002343: Paraneoplastic encephalomyelitis antigen	9.91E−05	0.7	0.08	8.75
IPR018359: Bromodomain, conserved site	9.75E−04	0.57	0.07	8.14
IPR001487: Bromodomain	1.94E−06	0.97	0.12	8.08
IPR017995: Homeobox protein, antennapedia type	7.36E−04	0.62	0.08	7.75
IPR004088: K Homology, type 1	7.36E−04	0.62	0.08	7.75
KEGG				
hsa03018: RNA degradation	1.05E−03	0.92	0.24	3.83
hsa04340: Hedgehog signalling pathway	6.75E−03	0.88	0.29	3.03
hsa04520: Adherens junction	4.55E−03	1.06	0.37	2.86
hsa05211: Renal cell carcinoma	9.88E−03	0.92	0.33	2.79
hsa04120: Ubiquitin-mediated proteolysis	1.05E−03	1.5	0.57	2.63
hsa04310: Wnt signalling pathway	2.22E−04	2.07	0.79	2.62
hsa04360: Axon guidance	3.47E−04	1.8	0.69	2.61
hsa04810: Regulation of actin cytoskeleton	1.07E−03	2.07	0.94	2.20
hsa04010: MAPK signaling pathway	3.23E−04	3.08	1.5	2.05

[a]The percentages of SCCS genes that have the InterPro or KEGG term
[b]The percentages of non-SCCS genes that have the InterPro or KEGG term.
[c]The fold of 'a' to 'b'. Terms are listed in the descending order of the fold difference.
*Probability for enrichment of the InterPro or KEGG terms in the SCCS group.

Cellular components (GO:0014704) mediate mechanical and electrochemical integration between cardiomyocytes and the rest of the five (GO:0043198, GO:0030425, GO:0043025, GO:0015629, and GO:0043005) have an association with the neuron. In the molecular function category, three terms (GO:0004843, GO:0031625, and GO:0004221) are related to the ubiquitin system, two (GO:00042054 and GO:00003682) are associated with chromatin. Ubiquitins are known to be involved not only with protein degeneration but also with signal transduction, chromatin modification, and cell cycle.

Of the ten Interpro codes listed in Table 3, IPR001827 is related to ubiquitin and IPR002077 represents calcium channel and other eight are all associated with DNA or RNA-binding functions that mediate transcriptional regulation or chromatin modification.

Tow KEGG pathways (hsa04340 and hsa04310) in Table 3 are developmental signalling pathways, hsa04120 is related to the ubiquitin system, and hsa04360 is involved in axon guidance, which well corresponds with the GO and Interpro terms.

Table 4 shows the terms that are significantly scarce in SCCS genes. In contrast to Tables 2 and 3, majority of the terms are involved with metabolic processes. Fatigo can also explore enrichment of micro RNA target and transcription-factor-binding sites; no significant item was found.

3.3. *Overlap with UCEs*

UCEs are defined as nucleotide sequences that are absolutely conserved longer than 200 bp between orthologous regions of the human, rat and mouse.[18,19] UCEs are found in both coding and non-coding regions. Precedent studies report that genes with low synonymous substitution or genes overlapped with UCEs are associated with DNA binding, RNA binding, transcription activity, and Homeobox.[18,19] In our 10 790 genes, 4009 bp in

Table 4. GO, InterPro and KEGG terms significantly ($P < 0.01$) scarce in SCCS genes

Terms	P*	In SCCS containing genes (%)[a]	In non-SCCS containing genes (%)[b]	Fold[c]
Biological process				
GO:0043039: tRNA aminoacylation	6.15E−03	0.09	0.63	0.14
GO:0006725: Cellular aromatic compound metabolic process	7.31E−03	0.48	1.3	0.37
GO:0051186: Cofactor metabolic process	6.05E−03	0.66	1.59	0.42
GO:0055114: Oxidation-reduction process	5.86E−04	2.11	3.87	0.55
GO:0019752: Carboxylic acid metabolic process	3.54E−04	2.42	4.34	0.56
GO:0006082: Organic acid metabolic process	3.54E−04	2.42	4.35	0.56
GO:0044255: Cellular lipid metabolic process	8.60E−03	3.3	4.98	0.66
Cellular component				
GO:0019866: Organelle inner membrane	2.38E−03	0.66	1.72	0.38
GO:0044429: Mitochondrial part	1.27E−03	1.63	3.24	0.50
GO:0005615: Extracellular space	2.87E−03	2.29	3.93	0.58
Molecular function				
GO:0001595: Angiotensin receptor activity	8.64E−03	0	0.4	0.00
GO:0016616: Oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	6.42E−03	0.18	0.82	0.22
GO:0016614: Oxidoreductase activity, acting on CH-OH group of donors	8.92E−03	0.26	0.97	0.27
Interpro				
IPR002198: Short-chain dehydrogenase/reductase SDR	8.96E−03	0	0.46	0.00
Kegg				
hsa04060: Cytokine–cytokine receptor interaction	2.86E−03	0.57	1.56	0.37

[a]The percentages of SCCS genes that have the GO, InterPro or KEGG term.
[b]The percentages of non-SCCS genes that have the GO, InterPro or KEGG term.
[c]The fold of 'a' to 'b'. Terms are listed in the ascending order of the fold difference.
*Probability for enrichment of the GO, InterPro or KEGG terms in the SCCS group.

29 genes are found to overlap with UCEs. In the 4009 bp region, 2835 bp in 22 genes overlap with SCCSs (Supplementary file S4). Because SCCSs are conserved in six mammalian species, including the three species referred for UCEs, the 2835 bp reflect conservation in other three species (macaque, cow, and dog). We surveyed nucleotide sites in the 10 790 genes that are conserved among human, rat, and mouse, and evaluated how many of them are also conserved in macaque, cow, and dog. The resulted ratio is 0.751; therefore the expected conservation is 3035 bp. This matches well with the observation. The regions overlapped with UCEs make only 1.47% of the entire SCCS regions. Other SCCSs convey shorter but deeper conservation than UCEs.

3.4. *SCCSs that form stable RNA secondary structures*

There are cases that a secondary structure of mRNA conveys functions.[24,25] We examined secondary structures and free energy of the SCCSs using Vienna RNA package. We found three SCCSs whose folding energy were significantly low (Table 5).

Polg encodes a catalytic subunit of mitochondrial DNA polymerase POLG. The POLG protein is the only polymerase known to be involved in replication of mtDNA. *Gal3st3* encodes a member of the galactose-3-O-sulfotransferase protein family. This protein exists on the membrane of Golgi apparatus. *Smarcd3* encodes a protein of SWI/SNF family, whose members display helicase and ATPase activities. This protein is thought to regulate transcription of target genes by altering the chromatin structure around those genes.

Table 5. Genes containing SCCS with significantly ($P < 0.001$) low free folding energy

	Gene	Length	Free energy
polg	DNA polymerase subunit gamma-1	36	−19.9
gal3st3	Galactose-3-O-sulfotransferase 3	36	−22.6
smarcd3	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily D member	39	−23.9

The gene names are represented by those of human.

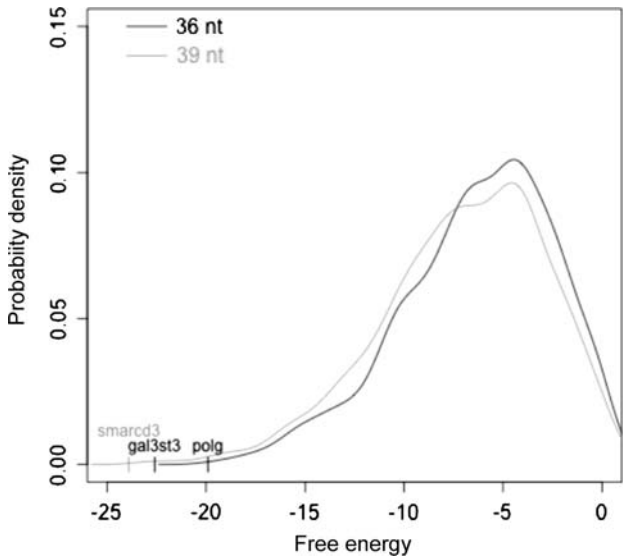

Figure 2. The probability density of the folding free energy. This graph shows the probability density of the folding free energy for 36 and 39 nucleotide-long sequences. The three SCCSs with significantly low free energy are indicated on the graph. Probability density was created using software package R.[38]

Figure 2 shows the probability density of the folding free energy constructed by randomly extracted sequences. Each line shows the free energy of the sequences of the same length as the above three SCCSs. Gene names on the lines represent free energy of the SCCSs. This figure suggests that these three SCCSs have extremely low free energy and that these regions will form stable secondary structures. These SCCSs may be conserved because of the requirement for the secondary structures. However, substitution restriction of this type might be modest because a secondary structure can be retained by another combination of nucleotides as far as the complementarity is maintained.

3.5. *The density of exonic splicing enhancers in SCCSs and in the other coding regions*

One of the well-known functional nucleotide elements in the coding region is splicing signals. We obtained 238 hexamers from RESCUE-ESE Web server as candidates of exonic splicing enhancers, and counted the number of hexamers in SCCSs and non-SCCS regions of the 10 790 genes (Table 6). Splicing signals in non-SCCS regions are counted on human sequences. We observed 20 420 hits of signals in 192 306 bp of SCCS regions and 2 183 544 hits in 205 666 520 of non-SCCS regions. The number of the signals per nucleotide is both 0.106 for SCCS and non-SCCS, and there was no significant difference ($P = 0.99$, χ^2 test).

Table 6. Splicing signals in SCCS and non-SCCS regions of 10 790 genes

	Region size (bp)	#Splicing signals	Per nucleotide
SCCS	192 306	20 420	0.106
Non-SCCS	20 566 520	2 183 544	0.106

Splicing signals in non-SCCS regions are counted on human sequences.

3.6. *Gene expression*

We investigated the difference of gene expression between SCCS genes and non-SCCS genes referring to anatomical system data of EGenetics, which give qualitative information about in what organs a gene is expressed. We counted the number of SCCS genes and non-SCCS genes expressed in the organs and performed the Fisher's exact test as described in Materials and method section.

We compared the percentages of genes that are expressed in each organ. Table 7 shows organs in which significantly higher percentage of SCCS genes are expressed compared with non-SCCS genes. In general, SCCS genes are expressed in a wider variety of organs. This observation agrees with a previous study.[35] Only in medulla oblongata and trophoblast, non-SCCS genes showed significantly higher percentage than SCCS genes (data not shown).

Seven organs (amygdala, spinal cord, cerebellum cortex, cerebellum, frontal lobe, pituitary grand, and sympathetic chain) in Table 7 are related with the nervous system and six organs (cochlea, trabecular meshwork, hypopharynx, larynx, tongue, and thyroid) are associated with head and neck.

3.7. *Preferred codons, GC content, and codon degeneracy of SCCSs*

Codon usage biases towards optimum codons are known to suppress synonymous substitution. Optimum codons reflect the composition of the genomic tRNA pool and are advantageous for translation efficiency or accuracy. As the approximate index of optimum codons, we used preferred codons, or most frequently used codon for an amino acid, and evaluated the preferred codon fraction in SCCSs. In SCCS regions, 28 188 of 64 102 codons are preferred codons and in non-SCCS regions, 2 961 482 of 6 855 505 codons are preferred codons. The ratio of preferred codons in SCCS and non-SCCS regions are 0.440 and 0.432, respectively. The difference is significant at the 0.05 significance level ($P = 0.013$) but the difference of the ratios is merely 0.008. We also observed that the ratio of preferred codon decreases as the length of SCCS increases (Fig. 3). Judging from these results, SCCSs are unlikely to be retained solely by codon preference.

Table 7. Organs in which significantly ($P < 0.001$) higher percentage of SCCS genes are expressed compared with non-SCCS genes

Organ	P*	SCCS		Non-SCCS		Fold[c]
		#Expressed	Percentage[a]	#Expressed	Percentage[b]	
Amygdala	1.32E−11	201	9.86	318	5.37	1.84
Cochlea	1.24E−22	436	21.39	723	12.21	1.75
Small intestine	5.28E−03	49	2.40	86	1.45	1.66
Amnion	1.03E−04	102	5.00	183	3.09	1.62
Amniotic fluid	9.12E−07	175	8.59	321	5.42	1.58
Spinal cord	7.21E−05	129	6.33	243	4.10	1.54
Artery	9.80E−05	133	6.53	254	4.29	1.52
Cerebellum cortex	2.74E−03	83	4.07	160	2.70	1.51
Cerebellum	3.62E−08	277	13.59	543	9.17	1.48
Trabecular meshwork	1.38E−05	199	9.76	399	6.74	1.45
Frontal lobe	1.63E−28	899	44.11	1803	30.45	1.45
Hypopharynx	1.75E−06	246	12.07	497	8.39	1.44
Pituitary gland	1.67E−09	421	20.66	877	14.81	1.39
Sympathetic chain	9.65E−06	271	13.30	575	9.71	1.37
Breast	5.95E−30	1133	55.59	2430	41.04	1.35
Larynx	9.05E−13	642	31.50	1385	23.39	1.35
Tongue	4.55E−07	377	18.50	816	13.78	1.34
Smooth muscle	1.36E−05	307	15.06	670	11.32	1.33
Thyroid	3.36E−23	1076	52.80	2375	40.11	1.32
Adrenal gland	4.51E−06	362	17.76	801	13.53	1.31

[a]The percentages of SCCS genes that are expressed in the organ.
[b]The percentages of non-SCCS genes that are expressed in the organ.
[c]The fold of 'a' to 'b' Terms are listed in the descending order of the fold difference.
*Probability for enrichment of the expressed genes in the SCCS group.

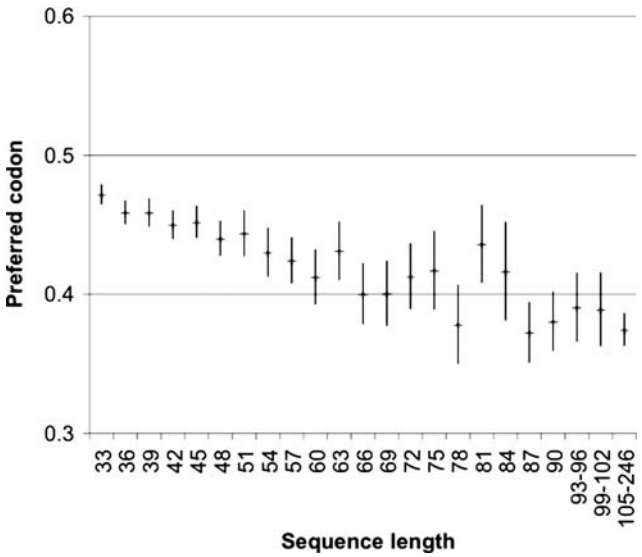

Figure 3. The ratio of preferred codons in SCCSs. *X*-axis represents the length of SCCSs, and *Y*-axis represents the ratio of preferred codon of the sequences. Classes whose sample size < 20 were combined. Error bars represent 1 SE.

In mammals, of the influence GC content on nucleotide change as a result of CpG hyper mutability. We investigated GC content of SCCSs. GC content in the first (GC1), second (GC2), and third position (GC3) of codons show different patterns along the sequence length (Fig. 4). GC1 is mostly constant but GC2 increases while GC3 decreases as the length of SCCS increases. Because mammalian genomes prefer GC-ending codons, the decrease of GC3 corresponds to the decrease of preferred codons. The decrease of GC3 seems to be complementary with the increase of GC2 because GC content as a whole is constant (Supplementary Fig. S1).

Conservation of SCCSs may occur by chance in a region where amino acid constraint is strong and codon degeneracy is low. We investigated codon degeneracy of SCCSs to examine this possibility. The average degeneracy is between three and four and increases as the sequence length increases (Fig. 5). This result suggests that even if the first and second positions of codons are restricted, the third position has enough freedom to change. Therefore, it is unlikely that SCCSs are conserved because of the amino acid constraint combined with the low degeneracy of codons.

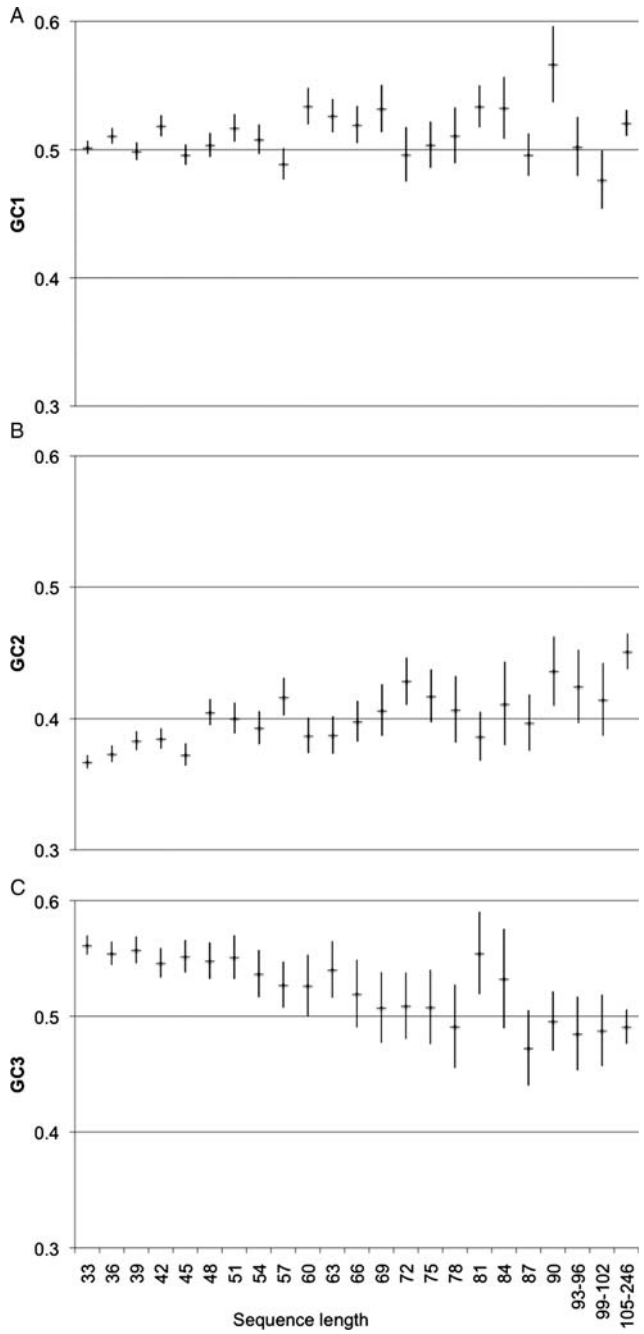

Figure 4. GC content of the first (GC1), the second (GC2), and the third (GC3) position of codons in SCCSs: (**A**) GC1, (**B**) GC2, (**C**) GC3. *X*-axis represents the length of SCCSs, and *Y*-axis represents GC content of the sequences. Classes whose sample size < 20 were combined. Error bars represent 1 SE.

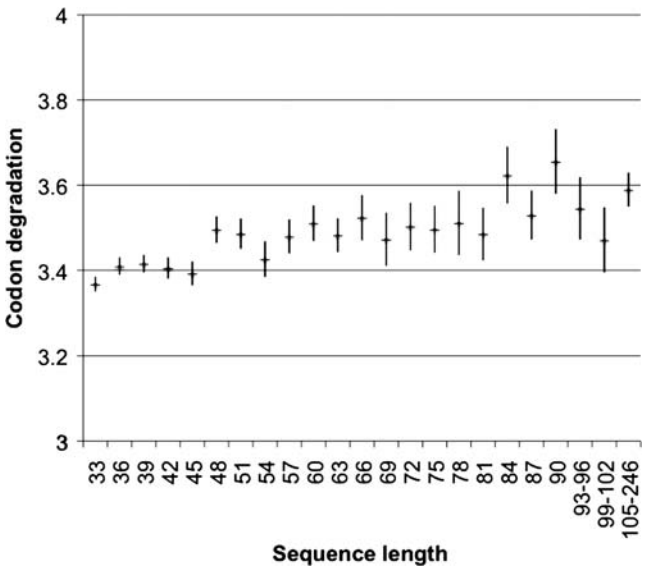

Figure 5. Codon degeneracy of SCCSs. *X*-axis represents the length of SCCSs and *Y*-axis represents the averaged codon degeneracy of the sequences. Classes whose sample size < 20 were combined. Error bars represent 1 SE.

4. Discussion

GO terms, InterPro codes, and KEGG pathways enriched with SCCS genes show a strong commitment to the developmental process, transcriptional regulation, and the neurons. Genes associated with transcriptional regulation or the neurons are known to have a low synonymous substitution ratio. This phenomenon is discussed in relation with codon biases to improve translational efficiency or accuracy. However, our analysis on the preferred codons in SCCSs suggests that codon preference is not likely the major factor influencing on the conservation of SCCSs.

Analyses on the ratio of preferred codons, GC content, and codon degeneracy enlighten the characteristics of SCCSs. The ratio of preferred codons decreases as the SCCS length increases. Drummond and Wilke[36] investigated the correlation between synonymous substitution rate (dS) and fraction of optimum (Fop) codons, and detected negative correlations between dS and Fop in rodents (mouse and rat) and positive correlation in human–dog comparison. If the SCCSs have the same trend as the rodents of the previous study, the ratio of preferred codons in SCCSs should be high; however, our result is not. There was no factor that would lower nucleotide substitution in GC content and codon degeneracy.

Methodological difference is that our research focused on local and complete conservation of nucleotides instead of the dS in the entire region of a gene and that we investigated conservation among the six mammalian species instead of a pair-wise comparison. The difference of results may suggest that factors underlying local and strong conservation such as SCCS differ from the factors working on the gene-wide conservation.

Makalowski *et al.*[37] showed a correlation between synonymous substitution rate (dS) and non-synonymous substitution rate (dN). Such correlations may occur when the constraint on a certain nucleotide sequence is so strong that dN is also lowered.

The usage of relatively rare codons and strong local conservation of SCCSs may be preferable as regulatory signals. The fraction of SCCSs in the coding region of the 10 790 genes is 0.94%. This fraction is so small that it would not have an influence on conventional evolutionary analysis. Although the fraction is small, or because the fraction is small, SCCSs may have potential as regulatory elements.

Acknowledgements: We thank Drs Kenta Sumiyama, Kiyoshi Ezawa, and Hiroshi Akashi for their insightful suggestion and discussion.

Supplementary data: Supplementary Data are available at www.dnaresearch.oxfordjournals.org.

Funding

This study was supported partly by Grant-in-Aid for scientific research from the Ministry of Education, Culture, Sports, Science, and Technology of Japan to N.S.

References

1. Kimura, M. 1983, *The Neutral Theory of Molecular Evolution*. Cambridge University Press: Cambridge.
2. Nei, M. 1987, *Molecular Evolutionary Genetics*. Columbia Univ. Press: New York.
3. Ikemura, T. 1982, Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs, *J. Mol. Biol.*, **158**, 573–97.
4. Ikemura, T. 1981, Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system, *J. Mol. Biol.*, **151**, 389–409.
5. Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y., and Ikemura, T. 2001, Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis, *J. Mol. Evol.*, **53**, 290–8.
6. Akashi, H. 1994, Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy, *Genetics*, **136**, 927–35.
7. Kurland, C.G. 1991, Codon bias and gene expression, *FEBS Lett.*, **285**, 165–9.
8. Hershberg, R., and Petrov, D.A. 2008, Selection on codon bias, *Annu. Rev. Genet.*, **42**, 287–99.
9. Ikemura, T. 1985, Codon usage and tRNA content in unicellular and multicellular organisms, *Mol. Biol. Evol.*, **2**, 13–34.
10. Sharp, P.M., and Li, W.H. 1987, The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias, *Mol. Biol. Evol.*, **4**, 222–30.
11. Akashi, H. 2003, Translational selection and yeast proteome evolution, *Genetics*, **164**, 1291–303.
12. Eyre-Walker, A.C. 1991, An analysis of codon usage in mammals: selection or mutation bias? *J. Mol. Evol.*, **33**, 442–9.
13. Sharp, P.M., Averof, M., Lloyd, A.T., Matassi, G., and Peden, J.F. 1995, DNA sequence evolution: the sounds of silence, *Philos. Trans. R Soc. Lond. B Biol. Sci.*, **349**, 241–7.
14. Reed, R. 1996, Initial splice-site recognition and pairing during pre-mRNA splicing, *Curr. Opin. Genet. Dev.*, **6**, 215–20.
15. Blencowe, B.J. 2000, Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem. Sci.*, **25**, 106–10.
16. Takahashi, A. 2009, Effect of exonic splicing regulation on synonymous codon usage in alternatively spliced exons of Dscam, *BMC Evol. Biol.*, **9**, 214.
17. Parmley, J.L., and Hurst, L.D. 2007, Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals, *Mol. Biol. Evol.*, **24**, 1600–3.
18. Bejerano, G., Pheasant, M., Makunin, I., et al. 2004, Ultraconserved elements in the human genome, *Science*, **304**, 1321–5.
19. Schattner, P., and Diekhans, M. 2006, Regions of extreme synonymous codon selection in mammalian genes, *Nucleic Acids Res.*, **34**, 1700–10.
20. Lareau, L.F., Inada, M., Green, R.E., Wengrod, J.C., and Brenner, S.E. 2007, Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements, *Nature*, **446**, 926–9.
21. Lin, Z., Ma, H., and Nei, M. 2008, Ultraconserved coding regions outside the homeobox of mammalian Hox genes, *BMC Evol. Biol.*, **8**, 260.
22. Lampe, X., Samad, O.A., Guiguen, A., et al. 2008, An ultraconserved Hox-Pbx responsive element resides in the coding sequence of Hoxa2 and is active in rhombomere 4, *Nucleic Acids Res.*, **36**, 3214–25.
23. Licatalosi, D.D., and Darnell, R.B. 2010, RNA processing and its regulation: global insights into biological networks, *Nat. Rev. Genet*, **11**, 75–87.
24. Bhalla, T., Rosenthal, J.J., Holmgren, M., and Reenan, R. 2004, Control of human potassium channel inactivation by editing of a small mRNA hairpin, *Nat. Struct. Mol. Biol.*, **11**, 950–6.
25. Delgado, M.D., Gutierrez, P., Richard, C., Cuadrado, M.A., Moreau-Gachelin, F., and Leon, J. 1998, Spi-1/PU.1 proto-oncogene induces opposite effects on monocytic and erythroid differentiation of K562 cells, *Biochem. Biophys. Res. Commun.*, **252**, 383–91.
26. Donehower, L.A., Slagle, B.L., Wilde, M., Darlington, G., and Butel, J.S. 1989, Identification of a conserved sequence in the non-coding regions of many human genes, *Nucleic. Acids Res.*, **17**, 699–710.
27. Lipman, D.J. 1997, Making (anti)sense of non-coding sequence conservation, *Nucleic Acids Res.*, **25**, 3580–3.

28. Levy, S., Hannenhalli, S., and Workman, C. 2001, Enrichment of regulatory signals in conserved non-coding genomic sequence, *Bioinformatics*, **17**, 871–7.
29. Sandelin, A., Bailey, P., Bruce, S., et al. 2004, Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes, *BMC Genomics*, **5**, 99.
30. Kryukov, G.V., Schmidt, S., and Sunyaev, S. 2005, Small fitness effect of mutations in highly conserved non-coding regions, *Hum. Mol. Genet.*, **14**, 2221–9.
31. Hubbard, T.J., Aken, B.L., Beal, K., et al. 2007, Ensembl 2007, *Nucleic Acids Res.*, **35**, D610–7.
32. Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–80.
33. Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., and Golani, I. 2001, Controlling the false discovery rate in behavior genetics research, *Behav. Brain Res.*, **125**, 279–84.
34. Fairbrother, W.G., Yeh, R.F., Sharp, P.A., and Burge, C.B. 2002, Predictive identification of exonic splicing enhancers in human genes, *Science*, **297**, 1007–13.
35. Plotkin, J.B., Robins, H., and Levine, A.J. 2004, Tissue-specific codon usage and the expression of human genes, *Proc. Natl. Acad. Sci. USA*, **101**, 12588–91.
36. Drummond, D.A., and Wilke, C.O. 2008, Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution, *Cell*, **134**, 341–52.
37. Makalowski, W., and Boguski, M.S. 1998, Synonymous and nonsynonymous substitution distances are correlated in mouse and rat genes, *J. Mol. Evol.*, **47**, 119–21.
38. Ihaka, R., and Gentleman, R. 1996, R: a language for data analysis and graphics, *J. Comp. Graph. Stat.*, **5**, 299–314.