Available online at www.sciencedirect.com
SCIENCE DIRECT®
Gene 321 (2003) 185–193
GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES
www.elsevier.com/locate/gene

mRNA 5′ region sequence incompleteness: a potential source of systematic errors in translation initiation codon assignment in human mRNAs☆

Raffaella Casadei, Pierluigi Strippoli*, Pietro D'Addabbo, Silvia Canaider, Luca Lenzi, Lorenza Vitale, Sandra Giannone, Flavia Frabetti, Federica Facchin, Paolo Carinci, Maria Zannotti

Center for Research into Molecular Genetics "Fondazione CARISBO", Institute of Histology and General Embryology, University of Bologna, Via Belmeloro, 8-40126 Bologna, Italy

Received 28 March 2003; received in revised form 20 June 2003; accepted 31 July 2003

Received by W. Makalowski

Abstract

The amino acid sequence of gene products is routinely deduced from the nucleotide sequence of the relative cloned cDNA, according to the rules for recognition of start codon (first-AUG rule, optimal sequence context) and the genetic code. From this prediction stem most subsequent types of product analysis, although all standard methods for cDNA cloning are affected by a potential inability to effectively clone the 5′ region of mRNA. Revision by bioinformatics and cloning methods of 109 known genes located on human chromosome 21 (HC 21) shows that 60 mRNAs lack any in-frame stop upstream of the first-AUG, and that in five cases (*DSCR1*, *KIAA0184*, *KIAA0539*, *SON*, and *TFF3*) the coding region at the 5′ end was incompletely characterized in the original descriptions. We describe the respective consequences for genomic annotation, domain and ortholog identification, and functional experiments design. We have also analyzed the sequences of 13,124 human mRNAs (RefSeq databank), discovering that in 6448 cases (49%), an in-frame stop codon is present upstream of the initiation codon, while in the other 6676 mRNAs (51%), identification of additional bases at the mRNA 5′ region could well reveal some new upstream in-frame AUG codons in the optimal context. Proportionally to the HC 21 data, about 550 known human genes might thus be affected by this 5′ end mRNA artifact.

Keywords: Genomics; 5′ UTR (5′ untranslated region); Full-length mRNA; Human genome; Human chromosome 21

1. Introduction

We will in general use the term "5′ end mRNA artifact" to refer to the incorrect assignment of the first AUG codon in an mRNA sequence, due to the incomplete determination of the mRNA 5′ end sequence (Fig. 1). The amino acid sequence of gene products is routinely deduced from the nucleotide sequence of the relative cloned cDNA, according to rules for recognition of start codon (first-AUG rule, optimal sequence context) and the genetic code (Kozak, 2002). The identification of a more complete mRNA 5′ end

Abbreviations: 5′ UTR, 5′ untranslated region; BLAST, Basic Local Alignment Search Tool; BLASTN, Blast nucleotide–nucleotide; BLASTP, Blast protein–protein; bp, base pairs; cDNA, DNA complementary to RNA; dbEST, EST database; dNTP, deoxyribonucleoside triphosphate; EST, expressed sequence tag; GSS, genomic sequence survey; HC 21, human chromosome 21; HTGS, high throughput genome sequence; mRNA, messenger RNA; ORF, open reading frame; PCR, polymerase chain reaction; RACE, 5′ rapid amplification of cDNA ends; RT-PCR, reverse transcription-PCR; SMART, Simple Modular Architecture Research Tool.

☆ GenBank accession numbers: KIAA0184, AF432263; KIAA0539, AF432264; SON, AF435977; TFF3, AF432265.

* Corresponding author. Tel.: +39-51-209-4100; fax: +39-51-209-4110.
E-mail address: pierluigi.strippoli@unibo.it (P. Strippoli).

0378-1119/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0378-1119(03)00835-7

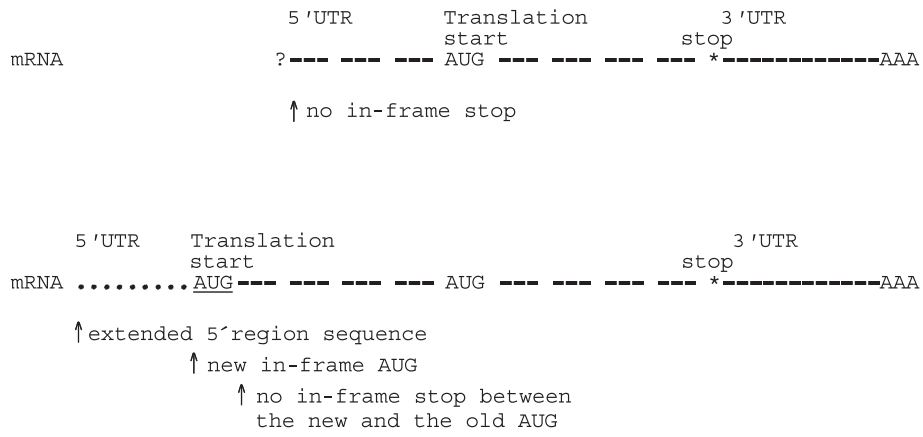

Fig. 1. The 5′ artifact problem. A schematic view of the conditions in which a known mRNA might lack coding sequences at its 5′ end (top) and might have a longer 5′ UTR, containing a new in-frame AUG and encoding a longer product.

could reveal an additional upstream AUG in-frame with the previously determined one and in the optimal context, thus extending the predicted amino terminus sequence of the product.

One obvious cause of the incorrect assignment of the first AUG codon is the fact that all standard methods for cDNA cloning are affected by a potential inability to effectively clone the 5′ region of mRNA, owing to the low processivity of reverse transcriptase, while methods able to identify the actual mRNA 5′ end (e.g. primer extension analysis, oligo-capping) are not routinely performed along with cDNA cloning. A concurrent cause is often the lack, until recent years, of complete bioinformatics analysis, which could reveal clues as to possible 5′ end mRNA incompleteness by systematic genomic and expressed sequence tag (EST) databanks search, along with phylogenetic comparison.

The translation start prediction based on incomplete mRNA sequence may lead to incorrect prediction of the product amino acid sequence, and to subsequent errors in the experimental cloning and functional assay of the relative cDNA.

This study was undertaken in order to establish the number of genes that, in the absence of an in-frame stop codon upstream the translation initiation codon, are possible candidate for a refinement of their actual coding sequence. A simple software tool was developed to assess the presence or the absence of such a 5′ in-frame stop codon in the human mRNA sequences collected in RefSeq database (http://www.ncbi.nih.gov/RefSeq/). In particular, we focused on the bioinformatic and molecular biology analysis of the known, well characterized genes present in the original map of human chromosome 21 (HC 21), to establish the actual most 5′ translation codon in the mRNAs candidate for a possibly extended 5′ region.

We show that the 5′ end mRNA artifact is not a rare event in the analysis of the human genome, and that it is a potential and sometimes actual cause of errors in the post-genomic analysis of the human genome.

2. Materials and methods

2.1. Sequence analysis of the human mRNA 5′ UTR

Each of the RefSeq mRNA sequences described for the 109 loci catalogued as “category: known” by Hattori et al. (2000), and linked to at least one published report, was manually analyzed by DNAid (ftp://iubio.bio.indiana.edu/molbio/mac) software for the presence of an in-frame stop codon upstream of the described ATG. In 49 cases, such a stop codon was found. The sequence of 60 mRNAs in which bases in the 5′ untranslated region (5′ UTR) could be consistent with amino acid coding codons was systematically compared, with finished and unfinished genomic sequences and with the ESTs, at both nucleotide and amino acid levels. The comparisons were made by BLAST software (http://www.ncbi.nih.gov/BLAST/) in *Homo sapiens* databanks, as well as in all other species whose sequences were available.

2.2. Development of software for stop codon search

The search for the absence of any in-frame stop codon 5′ UTR of known mRNAs was extended to the mRNAs of the RefSeq databank (http://www.ncbi.nih.gov/RefSeq/). The RefSeq data file was downloaded via ftp (version September 12, 2001), then the file was treated by a word processor to remove all newline and carriage return control characters. At this point, the end record symbol “//” was replaced by a carriage return, to import each entry into a text field of a FileMaker Pro 5 (FileMaker, Santa Clara, CA) database. Some records were excluded from the subsequent analysis in these cases: the entry was redundant with respect to this

investigation in that the same gene symbol and the same 5′ UTR were present in more than one entry; the 5′ UTR was longer than 801 bases (340 genes lost, only 2% of the set with the advantage of greatly simplified calculations; mean 5′ UTR had been estimated at 240 bases—Lander et al., 2001); the sequence was of DNA type or did not include an AUG initiation codon. Following this cleaning, 13,124 out of the original 13,997 entries were then analyzed. Calculation fields were programmed via the provided text functions so as to extract each in-frame group of three letters (bases) upstream of the AUG indicated in the "CDS" line of the entry, as far as the minus 801 base. Finally, a calculation via an "if" instruction gave a "stop" result if any stop codon was found from −3 position to −801, and "no stop" in the absence of any such stop codon. The FileMaker Pro 5 template which includes these calculation scripts is available upon request from the corresponding author.

2.3. In vitro cloning of the mRNA 5′ region

We designed an in vitro cloning approach to confirm the sequence analysis predictions. We followed a reverse transcription-polymerase chain reaction (RT-PCR) approach, mainly based on amplification of a more complete AUG-containing first exon, extended from the new putatively defined 5′ UTR and exon 2, in order to be sure that amplified cDNA derives from mRNA. The human RNA sources were placenta total RNA and heart total RNA purchased from Clontech (Palo Alto, CA). The RT-PCR standard conditions used were as previously described (Strippoli et al., 2000a). A total of 20 genes out of a 60-gene set were recognized as possible candidates for extension on the basis of bioinformatic data analysis and were subjected to RT-PCR based on the presence of possible in-frame upstream AUG (data not shown; see Section 3.2).

The primer pairs for the extended cDNAs were (5′–3′ direction): #1-GGTTTCAAGAGCAGCTTAATACTC (forward) and #2-CAAGAAGACTCCACAGCTGC (reverse), #3-CTGTCTGACTGCCTTGGATACAAC (forward) and #4-CACGCCATGCCACAGACCAGCATC (reverse), #5-GGTACTCAGAAGCTGAAACATTAAC (forward) and #6-GCCAGTTGCACTGGAACTGACGC (reverse) for *KIAA0184*; #7-CAGGAGTGTCTTCCGCTGTCATTC (forward) and #8-GTAAAATCTTCTGATCTAGGAC (reverse) for *KIAA0539*; #9-GACTAGCGAGGAGGAGTTGGAG (forward) and #10-GATGTCTTATGAGCGGTCTATG (reverse) for *SON*; #11-CTAGGAGGGCAATTGACACACATC (forward) and #12-CAAAGCAGCAGCCCCGGTTGTTG (reverse) for *TFF3*.

2.4. Amino acid sequence analysis

The predicted extended amino acid sequences for the five genes on HC 21 were searched in several domain databases (SMART—http://www.smart.embl-heidelberg.de/; CDD—http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) to identify novel domains not present in the described gene product. Alignment of the human and mouse DSCR1 products was made by ClustalW software (version 1.81).

3. Results and discussion

3.1. Translation initiation codon assignment in DSCR1 mRNA

Our first observation was casually made in studying the *DSCR1* gene (Fuentes et al., 1995, 1997), which encodes a predicted protein of 197 amino acids. While studying the corresponding gene family (Strippoli et al., 2000a,b), whose members present a domain of interaction with calcineurin (Rothermel et al., 2000), we noted that the human and mouse 5′ untranslated region (5′ UTR) sequences described were very similar, and in both cases they might be compatible with an unstopped coding sequence. Although multiple attempts to clone the extended 5′ UTR from several human tissues failed in our hands, analysis of independent sequences deposited in the context of cDNA cloning projects clearly shows that in humans the first initiation codon is located 165 bases earlier than described, thus predicting a polypeptide of 252 amino acids instead of the 197 reported (Table 1 and Fig. 2). It should be noted that the deposited extended sequences were mostly obtained by large-scale cDNA sequencing project teams employing techniques for specific enrichment in 5′-cap and 5′ ends, such as NEDO Human cDNA Sequencing Project (http://www.nedo.go.jp/bio-e/) and the "Mammalian Gene Collection" (http://www.mgc.nci.nih.gov/). The high homology of the extended amino terminus between human and mouse at both protein and DNA sequence level (85% DNA identity, consistent with data for coding sequence: Makalowski and Boguski, 1998; Larizza et al., 2002), together with the presence of a Kozak consensus sequence surrounding the new AUG, strongly suggests that this region is actually coding (Fig. 2b). This finding extends the previous amino acid sequence by 22%. In addition, the two-hybrid tests published for this protein have employed the reported incomplete cDNA both for human and mouse (Rothermel et al., 2000; Fuentes et al., 2000).

Given the domain architecture of most proteins and the reproducibility of the results of two-hybrid screening in other species (Kingsbury and Cunningham, 2000; Gorlach et al., 2000), it is unlikely that testing the whole cDNA for *DSCR1* would change the basic results, although the new amino acid sequence could be responsible for new interactions. The possibility of designing molecules with pharmacological activity based on binding to DSCR1-like proteins (Rothermel et al., 2000) only emphasizes the importance of knowing the actual protein structure, whereas the error in determination of the highly similar corresponding murine *DSCR1* ortholog (Strippoli et al., 2000b) underlines that a bias deriving from the original data went into the interpretation of the murine DSCR1 product sequence.

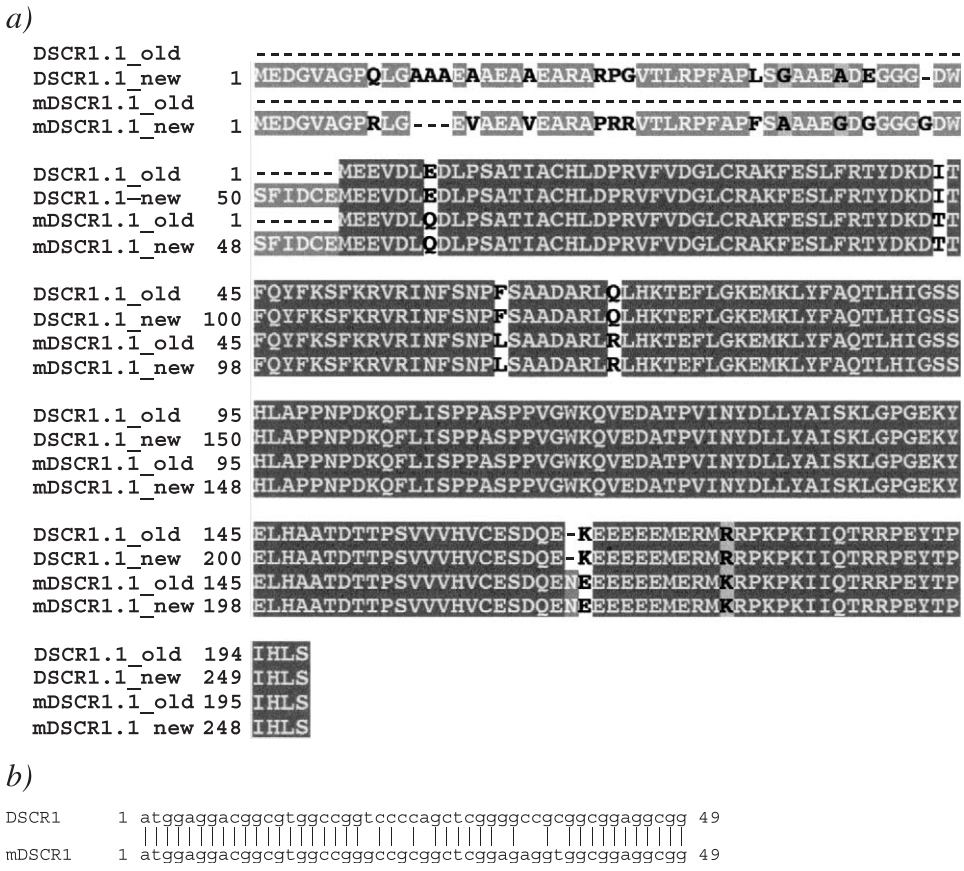

Fig. 2. Alignment of old and new DSCR1s. (a) ClustalW alignment of human and mouse (m) DSCR1 products, derived from described (old) and extended (new) mRNAs. (b) BLAST alignment of the first 49 coding bases in DSCR1 extended mRNAs.

3.2. Translation initiation codon assignment in HC 21 mRNAs

We wondered if errors consistent with the 5′ end artifact might be present in other described cDNAs, and decided to systematically analyze a large set of genes located on HC 21, for the following reasons: location of *DSCR1* in this chromosome, availability of the HC 21 finished genomic sequence (Hattori et al., 2000), observation that the predictions and statistical analyses based on HC 21 data proved to be a good projection on the subsequently released human whole-genome sequence (e.g. predicting a lower gene number than expected in the human genome: Hattori et al., 2000; Lander et al., 2001), and, finally, the need for precise knowledge of the HC 21 gene encoded proteins for understanding Down Syndrome (trisomy 21) in terms of molecular pathophysiology. We first searched for the presence of an in-frame stop codon upstream of the described initiation codon in the mRNA sequences of 109 HC 21 loci, selected according to their presence in the "known" category according to Hattori et al. (2000), and the existence of at least one published report linked to the sequence. When a stop was present (49 cases), one could rule out any incompleteness of the coding sequence (Kozak, 2002), and thus select a set of 60 mRNAs in which, on the contrary, the possibility that the recorded 5′ UTR sequence might actually be part of a longer coding sequence could not be excluded (Fig. 1). Each candidate mRNA was then accurately analyzed by means of systematic in-depth bioinformatics analysis and by in vitro cDNA cloning. A total of 19 genes (in addition to DSCR1) out of the 60-gene set were recognized as possible candidates for extension on the basis of bioinformatic data analysis and were subjected to RT-PCR based on the presence of possible in-frame upstream AUG (namely ABCG1, COL6A2, GCFC, ETS2, FTCD, KIAA0184, KIAA0539, KIAA0653, KIAA0958, NDUFV3, PDXK, PRED22, PRED28, PRKCBP2, SLC19A1, SON, SRra4, TFF3, WDR4; data not shown).

As a final result, the sequence of *KIAA0184* (Nagase et al., 1996), *KIAA0539* (Nagase et al., 1998), *SON* (Khan et al., 1994) and *TFF3* (Podolsky et al., 1993) mRNAs was extended, as in the case of *DSCR1*, demonstrating a new coding tract and thus identifying an open reading frame (ORF) with amino acids extra to those reported by Hattori et al. (2000) (Table 1). In these cases, both of the following conditions occurred: an extension of described exon 1 predicted new coding codons upstream of the known AUG; and a novel AUG was present upstream of these codons, in-frame with the previously described AUG and without any intervening stop codon. Following the rules of

Table 1
Genes on human chromosome 21 with extended cDNA 5′ region and deduced protein

Gene	Error type[a]	GenBank[b] No., human	Genomic clone	No. of new amino acids (% of reference length)	Kozak sequence old (top)/new (bottom). Consense[c]: gccGCCRCCATGG	GenBank No., non-human[d]
DSCR1	1	AL538796 (EST) AL551657 (EST) BE882755 (EST) AL536447 (EST) AK092184 (PRI)[e] BM741496 (EST) BM743740 (EST)	#AP001720	55 (28%)	gacTGCGAGATGG ggcGCGGGGATGG	m BE287582 (EST) m AW121699 (EST) m BE954084 (EST) m BE954961 (EST) m BY276032 (EST) m BY274253 (EST) m BY739662 (EST) m CB248298 (EST)
KIAA0184	1, 2	AF432263 (PRI)[f] BC038443 (PRI)[e] AK056738 (PRI)[e] AK098236 (PRI)[e] AF490768 (PRI)[g] AA232480 (EST) BG253602 (EST) BG422067 (EST) AL040229 (EST)	#AP001760	722 (84%)	ggtCAGGTGATGC ggcCTGGCCATGG	m BB193613 (EST) m BF584364 (EST) p BI341957 (EST) p BG834389 (EST)
KIAA0539	1, 2	AF432264 (PRI)[f] AF231919 (PRI)[g] BU659585 (EST) BF334583 (EST) BM458654 (EST) BG007646 (EST)	#AP001714	1449 (176%)	tttTTAAAGATGC ctcTCGGCCATGG	m AK050735 (HTC) m BB658427 (EST) m BB654946 (EST) m BB438600 (EST) m AW825146 (EST) m BI854164 (EST) m BY717282 (EST) p BE234186 (EST) p BE234833 (EST)
SON	1, 2	AF435977 (PRI)[f] AY026895 (PRI)[g] AF380183 (PRI)[g] BC005337 (PRI)[e] BE935931 (EST) BE843288 (EST) AW450807 (EST) BG620557 (EST) AA602113 (EST) AW964378 (EST) BG943677 (EST) AI929819 (EST)	#AP001717	968 (68%)	gaaCGTTCCATGA gcgGACGCCATGG	m AF193607 (ROD) m BC046419 (ROD) m BG084043 (EST) m BG071049 (EST) m BE949693 (EST) r BE112986 (EST) r AA926049 (EST) b AV617554 (EST)
TFF3	1	AF432265 (PRI)[f] BU731867 (EST) BM743413 (EST) BM921123 (EST) BM708674 (EST) BM856919 (EST) BM272999 (EST)	#AP001746	57 (77%)	gcgCTCTGCATGC gacAAAGGCATGC	–

[a] (1) Extended exon 1; (2) new exons. Details about positions of new coding sequences and deduced amino acid product sequences are given as a Web supplement at address: http://apollo11.isto.unibo.it/suppl/casadei2003/index.htm.

[b] GenBank sequence: in EST (Expressed Sequence Tag), PRI (primate), ROD (rodent) or HTC (unfinished high-throughput cDNA sequencing) division.

[c] The two most conserved positions (Kozak, 2002) are underlined; start codon, double underline.

[d] m = murine; p = porcine; r = rat; b = bovine.

[e] Deposited in the context of large-scale cDNA sequencing project.

[f] Experimentally determined in the present work.

[g] Determined from other authors.

translation initiation (reviewed by Kozak, 2002), the actual coding sequence should be considered as that included between the novel "first-AUG" and the known stop. The use of "internal" AUGs, enabling additional initiation events at downstream AUG codons in some mRNAs, may occur only in three well-defined circumstances (Kozak, 2002): re-initiation, which doesn't apply to the considered mRNAs because the newly determined AUG is not part of a small upstream ORF separated from the main ORF by a stop codon; context-dependent leaky scanning, which may be excluded because we considered the concordance with the Kozak sequence (Pesole et al., 2000; Kozak, 2002) for the novel AUGs, observing full (sometimes better) compatibility with the use of the novel AUG (Table 1); and presence of internal ribosome entry site sequence modules (IRES), that are found only in some known viral mRNAs.

In four out of five cases, additional evidence came from conserved coding sequences in other species (Table 1). In Table 2, we show the comparison of the 48 nucleotides upstream and downstream of the first AUG in both human and mouse. It is shown that in both species there is a dropout of similarity just before the newly identified AUG, as it is expected by the percentage of human/mouse identities (Makalowski and Boguski, 1998) in noncoding and coding sequences. In contrast, the same type of analysis centered on the previously known AUG codon indicates a lack of a noncoding/coding boundary, as it may be judged by a similar percentage of identities upstream and downstream of the AUG codon. The coding nature of these upstream bases is further evidenced by the comparison at the amino acid level of the same sequences (Table 2). We conclude that the predicted product for these five genes on chromosome 21 should be redefined for functional studies.

Table 2
Comparison between human and mouse of the 48 nucleotides upstream (−48/−1) and downstream (+1/+48) the previously (OLD) and the newly (NEW) determined AUG start codon, at nucleotide (Nt) and amino acid (AA) level

Nt	−48/−1 identities	+1/+48 identities
DSCR1 OLD	38/48 (79%)	45/48 (94%)
DSCR1 NEW	18/48 (38%)	41/48 (85%)
KIAA0184 OLD	39/48 (81%)	42/48 (88%)
KIAA0184 NEW	21/48 (44%)	44/48 (92%)
KIAA0539 OLD	41/48 (85%)[a]	35/48 (73%)
KIAA0539 NEW	16/48 (33%)	37/48 (77%)
SON OLD	48/48 (100%)	45/48 (94%)
SON NEW	47/48 (98%)[b]	47/48 (98%)
TFF3 OLD	14/48 (29%)	26/48 (54%)
TFF3 NEW	no similarity	no similarity
AA	**−48/−1 identities**	**+1/+48 identities**
DSCR1 OLD	11/16 (69%)	16/16 (100%)
DSCR1 NEW	02/16 (13%)	13/16 (81%)
KIAA0184 OLD	16/16 (100%)	12/16 (75%)
KIAA0184 NEW	03/16 (19%)	16/16 (100%
KIAA0539 OLD	12/16 (75%)[a]	10/16 (63%)
KIAA0539 NEW	01/16 (6%)	12/16 (75%)
SON OLD	16/16 (100%)	16/16 (100%)
SON NEW	15/16 (94%)[b]	13/16 (81%)
TFF3 OLD	00/16 (0%)	10/16 (38%)
TFF3 NEW	no similarity	no similarity

The complete alignment data is available as a Web supplement at: http://apollo11.isto.unibo.it/suppl/casadei2003/table2.htm.

[a] In the alignment of the previously determined coding sequences, the first ATG was not present in mouse.

[b] Although this upstream region shows a high percentage of identities, possibly explained by 5′ UTR selective conservation, it includes an in-frame stop codon in both human and mouse.

Following the original gene map by Hattori et al., new genes were recently added at the list of verified, characterized HC 21 genes while this work was in progress (Reymond et al., 2001, 2002; Gardiner et al., 2002; Vitale et al., 2002; Takamatsu et al., 2002). In some cases, the 5′ end sequence of known mRNAs was also refined and deposited in GenBank, confirming our previously released sequence data for *KIAA0539* (Gardiner et al., 2002; in this work we show a further extension) and confirming and extending our previously released sequence data for *KIAA0184* (Gardiner et al., 2002). The correct *SON* cDNA 5′ end sequence was subsequently released in GenBank in 2001 by Sun et al. (2001), our group, and Reymond et al. (2001), with agreement among respective data.

The 19 new genes described by Reymond et al. (2002), were screened for the 5′ in-frame stop codon by the authors themselves, and only two of them are candidates for a future extension of the coding sequences (*MRPS6* and *MCM3*). Our group described the *CYYR1* gene (Vitale et al., 2002), which is a possible candidate for a longer open reading frame; however, only a single human cDNA sequence from a neuroblastoma tumor supports the extension of the 5′ region, and multiple attempts in our laboratory to clone the extended cDNA end from normal tissues RNA failed (data unshown). It should also be noted that in this and other cases the alternative use of different promoters, as well as alternative splicing at the 5′ UTR (Kozak, 2002) or alternative start codon selection, could be consistent with the coexistence of "long" and "short" isoforms, possibly differentially expressed in specific tissues or cell conditions.

Finally, both *DSCR9* and *DSCR10* mRNAs described by Takamatsu et al. (2002) have an in-frame stop codon upstream the translation initiation codon.

3.3. Translation initiation codon assignment in human mRNAs and its consequences

If the same percentage of mRNA 5′ end cloning error in HC 21 were to apply to the whole human genome, we would predict that 556 mRNA coding sequences are incomplete in their 5′ ends in the studied set; the summary of data is shown in Table 3, which is linked to a Web supplement with the complete gene list. While human genome draft data are currently being widely rediscussed as to the actual number of genes, with special regard to the thousands of predicted and uncloned genes, little attention

Table 3
Percentage of incidence of the 5′ end artifact in human chromosome 21 and estimation on human RefSeq mRNAs

	Human chromosome 21	Human RefSeq
Genes[a]	109	13,124
mRNAs[a]		
with in-frame stop in 5′ UTR	49 (45%)	6448 (49%)
without in-frame stop in 5′ UTR	60 (55%)	6676 (51%)
mRNA with 5′ UTR revised as cds	5	(556 estimated)
percentage of all genes	4.6%	
percentage of mRNA without stop	8.3%	

[a] The complete lists are available as Web supplements at: http://apollo11.isto.unibo.it/suppl/casadei2003/index.htm.

has been given to the fact that the mRNA sequence of many well known genes, at least for the longest isoform, could not be correct. Although a recent statistical report has also underlined a comparable frequency in the presence of new in-frame upstream AUGs in a sample of 954 human 5′ UTRs, following screening of oligo-capping 5′ end enriched libraries (Suzuki et al., 2000), the biological consequences of the product sequence change were not analyzed.

Actually, several reports have recently demonstrated, as anecdotal evidence randomly found for single genes, the presence of a further extension of mRNA 5′ end sequence with consequent correction of the previously accepted predicted product. For example, mRNA coding sequence was extended in this way for RANBP9/RanBPM on 6p23 (230 new amino acids; Nishitani et al., 2001), NFE2L3 on 7p15–p14 (sequence #AB010812.1, 2174 bp, from Kobayashi et al. (1999), has been replaced by #AF134891.1, 2618 bp, adding 294 new amino acids), SP2 on 17q21.32 (#M97190, 2063 bp, from Kingsley and Winoto (1992), has been replaced by #D28588.1, 3288 bp, from Nomura et al. (1994), adding 111 new amino acids). These three mRNAs, whose coding nature of the extension is also supported by very high similarity with the respective murine hortologs, are all in our list of candidate for mRNA extension (Table 3, linked gene list), further underlining the importance of a systematic revision of the human mRNAs sequences to discover all the incompletenesses in the coding sequences. Due to the large body of needed work, this task could be best accomplished by each group originally reporting the cloned cDNA, in the absence of positive proof of cDNA completeness (e.g. if it was obtained by oligo-capping or determination of actual transcription start).

A list of effects generated by the 5′ end artifact in interpreting data concerning five HC 21 genes is presented in Table 4. It will be noted that along with errors in predicting the human protein sequence and the negative consequences thereof on the study of product structure and function—leading to the possibility of vast amounts of work being based on incorrect starting data—the 5′ artifact could also cause: failure to recognize genomic sequences as genes, keeping the search space for novel genes artificially expanded (overlapping genes are a very rare exception in the human genome); errors in promoter prediction and analysis, if sequences annotated as promoters are actually part of a longer mRNA (as in the case of the now revised *TFF3* mRNA, whose potential promoter was previously analyzed by Ribieras et al., 2001); chain errors in the prediction of orthologs in other species; failure to identify functionally remarkable protein domain sequences; possible underestimation of alternative splicing at the 5′ terminus of genes; inaccurate mutation screening design for coding sequences (which may sometimes explain the failure to find expected mutations in candidate genes), followed by possibly inaccurate genotype/phenotype correlations.

Table 4
Consequences of the correction of 5′ end mRNA artifact in five human chromosome 21 genes

Gene	Genomic annotation	Ortholog identification	Domain identification	Protein function studies
DSCR1	exon 1 extended	correction of the reported murine DSCR1 product sequence (Strippoli et al., 2000b)	–	two-hybrid tests were performed with truncated cDNAs (Rothermel et al., 2000; Fuentes et al., 2000)
KIAA0184	exon 1 extended, 17 exons added (82,895 further bp annotated as gene)	–	new "Acyl-CoA synthetases (AMP-forming)/AMP-acid ligases II" domain (amino acids 324–653)	–
KIAA0539	exon 1 extended, 23 exons added, (46,462 further bp annotated as gene)	–	–	–
SON	exon 1 extended, 2 exons added, (7835 further bp annotated as gene)	–	internal repeats, new collagen domain (586–645)	–
TFF3	exon 1 extended, 170 further bp annotated as transcript (previously annotated as promoter)	comparison of human/mouse TFF3 promoters (Ribieras et al., 2001) should be revised	–	–

4. Conclusions

The 5′ end mRNA artifact is less likely to be systematically produced from now on, in view of the growing sensitization to start large-scale projects for full-length mRNA sequencing and to perform in-depth bioinformatics analysis of genomic data, in particular by methods aimed at the identification of the first exon. However, some careful individual revision of the thousands of mRNA sequences described in the pre-human genome epoch would be advisable in order to avoid potentially significant pitfalls, due to the 5′ end mRNA artifact, in the transition from genomics to post-genomics.

Acknowledgements

This work has been supported by MURST ex 60% grants to PS and PC. We thank Gabriella Mattei for her excellent technical assistance in DNA sequencing.

References

Fuentes, J.J., Pritchard, M.A., Planas, A.M., Bosch, A., Ferrer, I., Estivill, X., 1995. A new human gene from the Down Syndrome Candidate Region encodes a proline-rich protein highly expressed in fetal brain and heart. Hum. Mol. Genet. 4, 1935–1944.

Fuentes, J.J., Pritchard, M.A., Estivill, X., 1997. Genomic organization, alternative splicing, and expression patterns of the *DSCR1*, (Down syndrome candidate region 1) gene. Genomics 44, 358–361.

Fuentes, J.J., Genesca, L., Kingsbury, T.J., Cunningham, K.W., Perez-Riba, M., Estivill, X., de la Luna, S., 2000. DSCR1, overexpressed in Down syndrome, is an inhibitor of calcineurin-mediated signaling pathways. Hum. Mol. Genet. 9, 1681–1690.

Gardiner, K., Slavov, D., Bechtel, L., Davisson, M., 2002. Annotation of human chromosome 21 for relevance to Down syndrome: gene structure and expression analysis. Genomics 79, 483–833.

Gorlach, J., Fox, D.S., Cutler, N.S., Cox, G.M., Perfect, J.R., Heitman, J., 2000. Identification and characterization of a highly conserved calcineurin binding protein, CBP1/calcipressin, in *Cryptococcus neoformans*. EMBO J. 19, 3618–3629.

Hattori, M., et al., 2000. The DNA sequence of human chromosome 21. Nature 405, 311–319.

Khan, I.M., Fisher, R.A., Johnson, K.J., Bailey, M.E., Siciliano, M.J., Kessling, A.M., Farrer, M., Carritt, B., Kamalati, T., Buluwela, L., 1994. The SON gene encodes a conserved DNA binding protein mapping to human chromosome 21. Ann. Hum. Genet. 58, 25–34.

Kingsbury, T.J., Cunningham, K.W., 2000. A conserved family of calcineurin regulators. Genes Dev. 14, 1595–1604.

Kingsley, C., Winoto, A., 1992. Cloning of GT box-binding proteins: a novel Sp1 multigene family regulating T-cell receptor gene expression. Mol. Cell. Biol. 12, 4251–4261.

Kobayashi, A., Ito, E., Toki, T., Kogame, K., Takahashi, S., Igarashi, K., Hayashi, N., Yamamoto, M., 1999. Molecular cloning and functional characterization of a new Cap'n' collar family transcription factor Nrf3. J. Biol. Chem. 74, 6443–6452.

Kozak, M., 2002. Pushing the limits of the scanning mechanism for initiation of translation. Gene 99, 1–34.

Lander, E.S., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Larizza, A., Makalowski, W., Pesole, G., Saccone, C., 2002. Evolutionary dynamics of mammalian mRNA untranslated regions by comparative analysis of orthologous human, artiodactyl and rodent gene pairs. Comput. Chem. 26, 479–940.

Makalowski, W., Boguski, M.S., 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2820 orthologous rodent and human sequences. Proc. Natl. Acad. Sci. U. S. A. 95, 9407–9412.

Nagase, T., Seki, N., Ishikawa, K., Tanaka, A., Nomura, N., 1996. Prediction of the coding sequences of unidentified human genes: V. The coding sequences of 40 new genes (KIAA0161–KIAA0200) deduced by analysis of cDNA clones from human cell line KG-1. DNA Res. 3, 17–24.

Nagase, T., Ishikawa, K., Miyajima, N., Tanaka, A., Kotani, H., Nomura, N., Ohara, O., 1998. Prediction of the coding sequences of unidentified human genes: IX. The complete sequences of 100 new cDNA clones from brain which can code for large proteins in vitro. DNA Res. 5, 31–39.

Nishitani, H., Hirose, E., Uchimura, Y., Nakamura, M., Umeda, M., Nishii, K., Mori, N., Nishimoto, T., 2001. Full-sized RanBPM cDNA encodes a protein possessing a long stretch of proline and glutamine within the N-terminal region, comprising a large protein complex. Gene 272, 25–33.

Nomura, N., Nagase, T., Miyajima, N., Sazuka, T., Tanaka, A., Sato, S., Seki, N., Kawarabayasi, Y., Ishikawa, K., Tabata, S., 1994. Prediction of the coding sequences of unidentified human genes: II. The coding sequences of 40 new genes (KIAA0041–KIAA0080) deduced by analysis of cDNA clones from human cell line KG-1. DNA Res. 1, 223–229.

Pesole, G., Gissi, C., Grillo, G., Licciulli, F., Liuni, S., Saccone, C., 2000. Analysis of oligonucleotide AUG start codon context in eukariotic mRNAs. Gene 261, 85–91.

Podolsky, D.K., Lynch-Devaney, K., Stow, J.L., Oates, P., Murgue, B., DeBeaumont, M., Sands, B.E., Mahida, Y.R., 1993. Identification of human intestinal trefoil factor. Goblet cell-specific expression of a peptide targeted for apical secretion. J. Biol. Chem. 268, 6694–6702.

Reymond, A., Friedli, M., Henrichsen, C.N., Chapot, F., Deutsch, S., Ucla, C., Rossier, C., Lyle, R., Guipponi, M., Antonarakis, S.E., 2001. From PREDs and open reading frames to cDNA isolation: revisiting the human chromosome 21 transcription map. Genomics 78, 46–54.

Reymond, A., Camargo, A.A., Deutsch, S., Stevenson, B.J., Parmigiani, R.B., Ucla, C., Bettoni, F., Rossier, C., Lyle, R., Guipponi, M., de Souza, S., Iseli, C., Jongeneel, C.V., Bucher, P., Simpson, A.J., Antonarakis, S.E., 2002. Nineteen additional unpredicted transcripts from human chromosome 21. Genomics 79, 824–832.

Ribieras, S., Lefebvre, O., Tomasetto, C., Rio, M.C., 2001. Mouse Trefoil factor genes: genomic organization, sequences and methylation analyses. Gene 266, 67–75.

Rothermel, B., Vega, R.B., Yang, J., Wu, H., Bassel-Duby, R., Williams, R.S., 2000. A protein encoded within the Down syndrome critical region is enriched in striated muscles and inhibits calcineurin signaling. J. Biol. Chem. 275, 8719–8725.

Strippoli, P., Lenzi, L., Petrini, M., Carinci, P., Zannotti, M., 2000a. A new gene family including DSCR1 (Down Syndrome Candidate Region 1) and ZAKI-4: characterization from yeast to human and identification of DSCR1-like 2, a novel human member (DSCR1L2). Genomics 64, 252–263.

Strippoli, P., Petrini, M., Lenzi, L., Carinci, P., Zannotti, M., 2000b. The murine DSCR1-like (Down syndrome candidate region 1) gene family: conserved synteny with the human orthologous genes. Gene 257, 223–232.

Sun, C.T., Lo, W.Y., Wang, I.H., Lo, Y.H., Shiou, S.R., Lai, C.K., Ting, L.P., 2001. Transcription repression of human hepatitis B virus genes by negative regulatory element-binding protein/SON. J. Biol. Chem. 276, 24059–24067.

Suzuki, Y., Ishihara, D., Sasaki, M., Nakagawa, H., Hata, H., Tsunoda, T., Watanabe, M., Komatsu, T., Ota, T., Isogai, T., Suyama, A., Sugano, S., 2000. Statistical analysis of the 5′ untranslated region of human mRNA using "Oligo-Capped" cDNA libraries. Genomics 64, 286–297.

Takamatsu, K., Maekawa, K., Togashi, T., Choi, D.K., Suzuki, Y., Taylor,

T.D., Toyoda, A., Sugano, S., Fujiyama, A., Hattori, M., Sakaki, Y., Takeda, T., 2002. Identification of two novel primate-specific genes in DSCR. DNA Res. 9, 89–97.

Vitale, L., Casadei, R., Canaider, S., Lenzi, L., Strippoli, P., D'Addabbo, P., Giannone, S., Carinci, P., Zannotti, M., 2002. Cysteine and tyrosine-rich 1 (CYYR1), a novel unpredicted gene on human chromosome 21 (21q21.2), encodes a cysteine and tyrosine-rich protein and defines a new family of highly conserved vertebrate-specific genes. Gene 290, 141–151.