Genomics 100 (2012) 125–130

Contents lists available at SciVerse ScienceDirect

Genomics

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

Methods

Genome-scale analysis of human mRNA 5′ coding sequences based on expressed sequence tag (EST) database

Raffaella Casadei [1], Allison Piovesan [1], Lorenza Vitale, Federica Facchin, Maria Chiara Pelleri, Silvia Canaider, Eva Bianconi, Flavia Frabetti, Pierluigi Strippoli *

Center for Research in Molecular Genetics "Fondazione CARISBO", Department of Histology, Embryology and Applied Biology, University of Bologna, via Belmeloro 8, 40126 Bologna (BO), Italy

ARTICLE INFO

Article history:
Received 28 July 2011
Accepted 23 May 2012
Available online 31 May 2012

Keywords:
Human genome
Expressed sequence tag (EST)
5′ Untranslated region (5′ UTR)
mRNA 5′ coding sequence
Translation start codon

ABSTRACT

The "5′ end mRNA artifact" issue refers to the incorrect assignment of the first AUG codon in an mRNA, due to the incomplete determination of its 5′ end sequence. We performed a systematic identification of coding regions at the 5′ end of all human known mRNAs, using an automated expressed sequence tag (EST)-based approach. Following parsing of more than 7 million BLAT alignments, we found 477 human loci, out of 18,665 analyzed, in which an extension of the mRNA 5′ coding region was identified. Proof-of-concept confirmation was obtained by in vitro cloning and sequencing for *GNB2L1*, *QARS* and *TDP2* cDNAs, and the consequences for the functional studies of these loci are discussed. We also generated a list of 20,775 human mRNAs where the presence of an in-frame stop codon upstream of the known start codon indicates completeness of the coding sequence at 5′ in the current form.

1. Introduction

The term "5′ end mRNA artifact" refers to the incorrect assignment of the first AUG codon in an mRNA, due to the incomplete determination of its 5′ end sequence [1]. Since the '70s, the amino acid sequence of gene products has been routinely deduced from the nucleotide sequence of the relative cloned cDNA (DNA complementary to mRNA), according to rules for recognition of the start codon (first-AUG rule, optimal sequence context) and the genetic code [2]. All standard methods for the cloning of cDNA are affected by a potential inability to effectively clone the 5′ region of mRNA [3]. This is due to the reverse transcriptase failure to extend first-strand cDNA along the full length of the mRNA template toward its 5′ end [3]. These incomplete clone sequences consequently lead to the incorrect assignment of the first AUG codon. The identification of a more complete mRNA 5′ end could reveal an additional upstream AUG – in-frame with the previously determined one – thus extending the predicted amino terminus sequence of the product and avoiding subsequent relevant errors in the experimental study of the relative cDNA [1].

Methods to determine the full-length mRNA sequence on a large scale have been developed, such as 5′ cap trapping [4], cap analysis of gene expression (CAGE) [5], systematic empirical annotation of a set of transcript products by 5′ rapid amplification of cDNA ends (RACE) and high-density resolution tiling arrays [6]. However, they are experimentally labor-intensive and they have not been widely applied in comparison with the standard expressed sequence tag (EST) approach for fast characterization of cDNAs [7,8].

We previously used individual EST-based gene model refinement by classic in silico sequence analysis to revise the mRNA sequence of 109 human chromosome 21 protein-coding genes [1]. The success of this approach encouraged us to develop a piece of software ("5′_ORF_Extender" software) in order to automate the steps that were previously performed manually, applying it to the *Danio rerio* (zebrafish) genome [9].

The aim of this work was to perform a systematic identification of coding regions at the 5′ end of all human known mRNAs. However, it proved difficult to simply transfer the method used for *D. rerio* to *Homo sapiens*, due to the much larger size and complexity of RNA and EST sequence databases as well as the sequence analysis (BLAST, Basic Local Alignment Search Tool) results file. In order to overcome these problems, a fully revised computational biology strategy was adopted, which has been able to conclude the task for human mRNAs. We have thus been able to compile a database containing 477 loci, out of a total of 18,665 investigated (2.6%), where an extension of the RNA 5′ coding region has been identified. Proof-of-concept confirmation has been obtained by actual in vitro cloning and sequencing for *GNB2L1*, *QARS* and *TDP2* genes. The availability of the database with the results of the

* Corresponding author. Fax: +39 0512094110.
E-mail address: pierluigi.strippoli@unibo.it (P. Strippoli).
[1] These authors contributed equally to the work.

0888-7543/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2012.05.012

whole analysis should help further to reduce the incidence of 5′ end mRNA artifacts when studying human gene structure and function in biomedical research.

2. Materials and methods

2.1. Database construction

The 5′_ORF_Extender software parses RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/) and EST sequence databases and makes calculations on these sequences. This is done following the import of BLAT (BLAST-like alignment tool) genome alignment data for human mRNAs and ESTs, in order to determine a list of genes with an incompletely described mRNA 5′ coding sequence (CDS). The algorithm, previously described for *D. rerio* [9], has been completely revised and improved for *H. sapiens* analysis. It has been developed using the FileMaker Pro 10 Advanced (FileMaker, Santa Clara, CA) database management system for both Windows and Macintosh operating systems. It is freely available as a stand-alone software (version 2.0) including the FileMaker runtime and a step-by-step user tutorial at http://apollo11.isto.unibo.it/software/.

The downloading, import and parsing of RefSeq and EST sequence databases as well as of the corresponding BLAT genome alignment data are described in detail in the software documentation.

2.2. Computational analysis

The 5′_ORF_Extender analysis script performs the following steps: extraction of the EST sequence stretch upstream of the matched RefSeq mRNA first base when BLAT alignment shows a 5′ extension of the EST compared with the known RefSeq sequence (following removal of introns from both EST and mRNAs genome-aligned sequences); a search in this EST stretch for the most upstream existent ATG (corresponding to AUG in RNA) in-frame with the described one in the RefSeq mRNA sequence entry; calculation of the new putative extended coding region by merging the EST extended stretch starting from the new ATG with the previously known 5′ UTR of the RefSeq mRNA sequence; confirmation of the coding potential of this new extended sequence by excluding the presence of any in-frame stop codon within it (Fig. 1). It can also be estimated whether or not the determined extended CDS is complete, by searching for any in-frame stop codon that might occur in the transcript upstream of the newly determined start codon.

As a final result, the software provides a list of genes whose mRNA possesses an extended 5′ CDS on the basis of EST comparison.

2.3. Quality control and summarization of results

Due to the very large size and high complexity of the human genome and of the human EST database, together with the unavailability of a systematic assignment of mRNA and EST sequences to a defined genomic locus (in the form of an official gene symbol) in the UCSC data, we have introduced an automated method of quality control of results compared with the previous version [9] of our software. This ex-ante control verifies if each investigated EST has been assigned by UniGene (http://www.ncbi.nlm.nih.gov/unigene) system to the same locus as the mRNA sequence for which the EST is a possible candidate for 5′ end extension. This has been made possible thanks to the availability of a UniGene parser (the “UniGene Tabulator”) able to produce a structured table including all UniGene updated text information [10]. This table is imported into the 5′_ORF_Extender as a first step, allowing analysis to be limited to the mRNAs and corresponding ESTs that are mapped to the same defined locus. Due to possible errors in the large text file generated by UniGene Tabulator data parsing, a quality control assessment of the completeness of the UniGene entries was made as described in the software guide of the TRAM tool [11].

2.4. In vitro cloning and sequencing of the mRNA 5′ region

We decided to confirm the sequence analysis predictions of three example genes of the 5′_ORF_Extender results list. We utilized a reverse transcription-polymerase chain reaction (RT-PCR) approach, based on the amplification of a stretch extended from the new putatively defined 5′ UTR to at least as far as the known exon 2, in order to prove that the amplified cDNA is derived from mRNA. The human RNA sources were: skeletal muscle, small intestine, ovary, brain and bone marrow total RNA purchased from Clontech (Palo Alto, CA).

Details concerning primer sequences, RT-PCR and amplicon sequencing are described in the Supplementary File available at http://apollo11.isto.unibo.it/suppl/.

2.5. Sequence analysis

In order to test whether the newly determined CDS at 5′ was conserved in different species, TBLASTN searches were performed using standard parameters, except the filter for low complexity regions was unchecked. Alignment of the protein products was made by ClustalW software (version 2.1 at: http://www.ebi.ac.uk/Tools/msa/clustalw2/).

In order to identify novel domains which were not present in the described gene products, the predicted extended amino acid sequences for the three example genes were searched for in domain databases such as the Simple Modular Architecture Research Tool (SMART, http://smart.embl-heidelberg.de/) and the Conserved Domains Database (CDD, http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml).

3. Results

3.1. Database construction and computational analysis

The processing by 5′_ORF_Extender of 30,909 human RefSeq mRNA sequences assigned by UniGene to a defined locus (out of a total of 31,903) revealed the presence of an in-frame stop codon upstream of the known start codon in 20,775 cases. 10,134 sequences had a CDS which was putatively further extendable at their 5′ end. 159,378 UCSC EST-to-genome alignments, for the EST candidate to potentially extend the mRNA CDS at its 5′ in these 10,134 selected human mRNAs, were then processed to identify positive final results. Following calculations executed by the software, it was possible to obtain candidate extended coding regions at 5′ end from 2505 ESTs (Table 1).

3.2. Summarization of results

The final set of 2505 ESTs corresponded to 477 distinct human loci (2.6% of all studied genes with a RefSeq sequence) (Table 1). The mean number of EST sequences that allowed the extension of one mRNA sequence was 4.1, with 298 different mRNAs extended by at least two distinct EST sequences. In particular, the ESTs extending 270 out of these 298 mRNAs were not derived from the same library. The mean size of the additional open reading frames (ORF) stretch was 178.5 bases, with a standard deviation of 134.8 bases (range: 3–1014 bases) (Table 1).

For 224 genes (46.96%) it can be estimated that the determined extended CDS is complete, due to the presence of an in-frame stop codon upstream of the newly determined start codon.

3.3. In vitro cloning and sequencing of the mRNA 5′ region

The predicted additional coding region was cloned for each of the three example genes: *GNB2L1*, *QARS* and *TDP2* (Supplementary Table available at http://apollo11.isto.unibo.it/suppl). The nucleotide sequences of the extended coding regions determined exactly between

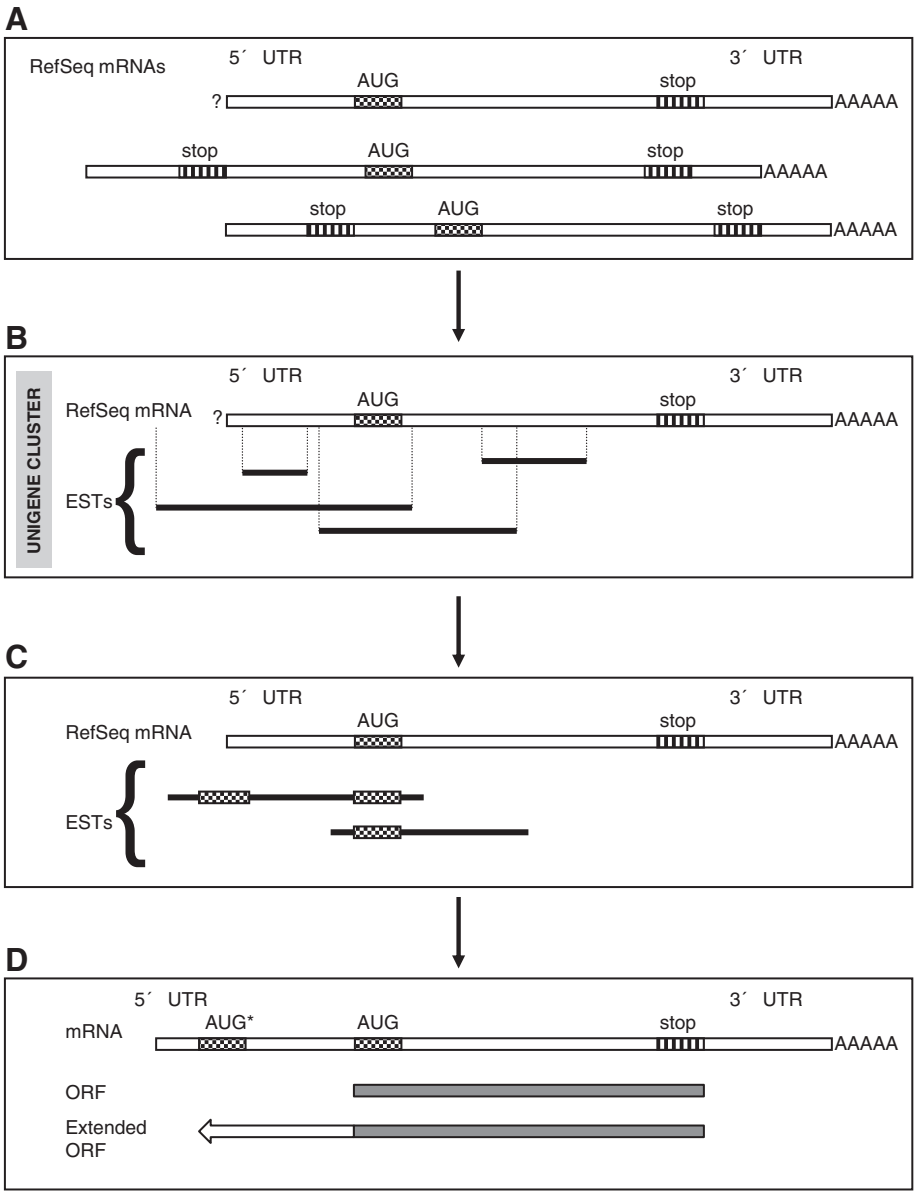

Fig. 1. Pipeline of the 5′_ORF_Extender software version 2.0 approach. Sequence comparisons exploit BLAT-pre-computed UCSC genomic coordinates of the RefSeq and EST sequences. Detailed explanation in the text. A. Identification of RefSeq mRNA sequences without a known in-frame stop codon upstream of the described initiation codon (and thus candidates for further extension of their CDS at 5′). B. The parsed and embedded UniGene database allows the determination of those EST sequences that cluster with each RefSeq mRNA sequence and that are possible candidates for extending their 5′ coding region. C. Identification of EST sequences with an upstream in-frame AUG codon and absence of any stop codon between the previously and the newly determined AUG codons. D. Calculation of the new extended open reading frame (new AUG codon indicated with an *).

the 3′ end of the primer pairs for *GNB2L1*, *QARS* and *TDP2* cDNAs have been deposited in the GenBank database under accession nos. JN104586, JN104585 and JN104587, respectively.

3.4. Sequence analysis

The extended coding sequences for *GNB2L1*, *QARS* and *TDP2* were analyzed using the TBLASTN program to compare them with known nucleotide sequences deposited in the NCBI databases. This confirmed that no human matching sequence had been previously deposited in the "mRNA" (molecular type) division of GenBank, except two sequences (#AK302867 and #AK298699) relating to *QARS* and *TDP2*, respectively. Although these sequences are not present in the GenBank EST division, they were generated in the context of the NEDO large-scale cDNA sequencing project [12] and the relative entries were not tagged with the corresponding gene symbol as well as their predicted proteins (classified as "unnamed protein product"). They were not used by the genome browsers NCBI Map Viewer [13] and University of California at Santa Cruz (UCSC) Genome Browser [14] to build mRNA models with the extended CDS. mRNA models including the extended CDS reported here for *QARS* (Ensembl Entry ENST00000420147) and *TDP2* (Ensembl Entry ENST00000545995), but not for *GNB2L1*, were available at the European Bioinformatics Institute (EBI) Ensembl genome browser [15]. These CDSs were not however included in the entries containing coding sequences (Ensembl CCDS) available for the two genes, respectively, and the mRNA models were mainly based on mRNA sequences. These include the aforementioned "mRNA" sequences relating to *QARS* and *TDP2*, with limited support from available ESTs (2 ESTs out of the 24 identified by 5′_ORF_Extender in the case of *QARS* and 2 out of the 12 in the case of *TDP2*). In addition, as stated in the Ensembl

Table 1
Summary of computational analysis. CDS, coding sequence. Length is given in nucleotides.

Summary of analysis	
Human loci analyzed	18,665
Human Reference mRNAs (RefSeq) analyzed	31,903
Human RefSeq mRNA sequences assigned by UniGene to a defined locus	30,909
mRNAs with CDS not extendable at 5′ end (in-frame stop codon located upstream of the known start codon)	20,775
mRNAs with CDS possibly further extendable at 5′ end	10,134
ESTs assigned to the same locus of the 10,134 mRNAs possibly further extendable at 5′ end	7,166,113
EST-to-genome alignments for the EST candidates to potentially extend the mRNA CDS at their 5′ end	159,378
Final set of results	
ESTs with putative CDS extension	*2505*
mRNAs with putative extension of their known CDS at 5′ end	*615*
Loci with putative extension of their known CDS at 5′ end	*477*
Mean number of ESTs with extended sequence per mRNA	4.1
Mean length of extended 5′ CDS	178.5
Standard deviation of the extended 5′ CDS length	134.8
Minimum length of extension	3
Maximum length of extension	1014
mRNAs with CDS extension supported by more than one EST	298
mRNAs with CDS extension supported by more than one EST not derived from the same library	270
Loci with CDS extension supported by more than one EST	232
Loci with CDS extension supported by more than one EST not derived from the same library	213

genome annotation documentation, EST alignments are displayed on the website but are not usually used as supporting evidence in the gene-building process. The nucleotide and amino acid analysis data are summarized in the Supplementary Table.

Sequence comparison also showed the presence of high conservation of the extended stretch with predicted proteins in non human primates, a finding consistent with the coding nature of these regions (Fig. 2).

The amino acid sequences predicted at the amino terminus of these three genes did not show new known functional domains through database searches.

4. Discussion

The continuous incorporation of information derived from individual and large-scale cDNA sequencing projects (including those specifically designed to characterize mRNA 5′ end [4,16,17]) in the last few years led to continuous improvement of completeness of mRNA reference sequences (e.g., RefSeq), and also to the corresponding protein coding sequences. However, genome browsers do not appear to systematically extract useful information from the ever-increasing vast quantity of EST data. To date, EST data remain invaluable due to significantly longer continuous RNA sequences they may provide in comparison with the very short fragments typically deposited in current high-throughput nucleotide sequencing databases. We first showed in zebrafish that EST analysis by 5′_ORF_Extender software could extend the currently known mRNA CDS [9], thereby differing from other methods, which do not incorporate prediction of the putative CDS extension (e.g., [18]).

In this work, we have presented a modified strategy that was able to analyze the much more numerous human sequences. Firstly, we fully revised the software algorithm by using pre-computed coordinates of the UCSC-downloaded RefSeq and EST genome alignment data (rather than the results of a large scale BLAST comparison), and specific UCSC-downloaded EST sequence entries. Rather than GenBank EST raw entries, these are EST sequence entries in which nucleotides which are unaligned to the genome are removed, and undetermined ('N') or mismatched nucleotides are replaced by the corresponding nucleotides

A

```
GNB2L1_HUMAN    MPCNFPLPFALHGAAILSRNVSWGSPFCMVERVFPVPAGGFSLSLSLQGGGRRGCGASFS 60
GNB2L1_PANTR    MPCNFPLPFALHGAAILSRNVSWGSPFCMVERVFPVPAGGFSLSLSLQGGGRSGCGAGFS 60
GNB2L1_PONAB    MPCNFPLPFALHGAAILSRSVSWGSPFCMVERVFPVPAGGF-XSLSLQGGGGSGCGASFS 59
GNB2L1_HILLE    MPYNFPLPFALHGAAILSRSVSWGSPFCMVERVFPVPAGGF-LSLSLQGGGGSGCGASFS 59
GNB2L1_MACMU    MPCNFPLPFALHGAAILSRSVSWGSPFYLVEWVFPVPAEGF--SLSLQGGGGSGCVARFS 58
GNB2L1_CALJA    MRCNFPSPFAIQGAAILNRGVGKGSPFCLVVRVFPVPEGGG-SLSSTQGGGGRSCGAVFS 59
                *  *** ***::*****.*.*. **** :*  *****  *     * ****  .* * **

GNB2L1_HUMAN    KPSSAILVAAATHALAAAMTEQMTLRGTLKGHNGWVTQ 98
GNB2L1_PANTR    KPSSAILVAAATHALAAAMTEQMTLRGTLKGHNGWVTQ 98
GNB2L1_PONAB    KPSSAILVAAATHALAAAMTEQMTLRGTLKGHNGWVTQ 97
GNB2L1_HILLE    KLSSAILVAAATHALAAAMTEQMTLRGTLKGHNGWVTQ 97
GNB2L1_MACMU    EPSSAILVAAATHALAAAMTEQMTLRGTLKGHNGWVTQ 96
GNB2L1_CALJA    EPSSAILVGAVVHALSAAMTEQMTLRGTLKGHNGWVTQ 97
                : ******.*..***:**********************
```

B

```
QARS_HUMAN      MPTCRLGPKFLLVSGVSAMAALDSLSLFTSLGLSEQKA 38
QARS_PANTR      MPTCRLGPKFLLVSGVSAMAALDSLSLFTSLGLSEQKA 38
QARS_MACMU      MPTCRRGPKFLLVSGVSAMAALDSLSLFTGLGLSEQKA 38
                ***** **********************.********
```

C

```
TDP2_HUMAN      MRERHDTGACAEPRVGLLFRLKGRCRGGRKMELGSCLEGGREAAEEEGEP 50
TDP2_PANTR      MRERHGTGACAEPRVGLLFRLKGRCRGGKKMELGSCLEGGREAAEEEGEP 50
TDP2_PONAB      MRKRHGTGACAEPRVGLLFRLKGRCRGGKKMELGSFLEGGREAAEEEGEP 50
TDP2_HILLE      MRERHGTGACAEPRVGLLFRLKGRCTGGKKMELGSSLEGGREAAEEEGEP 50
TDP2_MACMU      MRERRGAGACAEPTVGLLFRLKGRCGSGKKMELGSCLG----AAEEEGEP 46
TDP2_CALJA      MRGRRSAGACAEPGVGFLFRLKGRCGSGKTMELGGCLEGGTEAAGEEGEP 50
                ** *:.:****** **:******** .*:.****. *     ** *****
```

Fig. 2. ClustalW alignment of GNB21L (A), QARS (B) and TDP2 (C) protein sequences from different species. Human sequences are derived from the original cDNA sequencing data presented here. The methionine corresponding to the previously determined start codon in the human mRNA reference sequence is underlined, followed by the first 20 amino acids of the reference protein sequence. HUMAN: *Homo sapiens*, PANTR: *Pan troglodytes*, PONAB: *Pongo abelii*, HILLE: *Nomascus leucogenys*, MACMU: *Macaca mulatta*, CALJA: *Callithrix jacchus*. Asterisk: residue conserved in all sequences; colon: conservative substitution; dot: less conservative substitution.

present on the genome. This key change significantly improved a number of areas: the software speed of analysis, sensitivity (due to the implementation of management of sequence in 'complement' orientation with respect to the genome recorded DNA strand, with consequent identification of previously undetected mRNA extensions thank to ESTs in opposite orientation to the corresponding mRNA), specificity (due to the use of EST sequence entries processed by UCSC as described above, thereby avoiding false positive identification of start codons in the EST sequence, and possibly false negatives too, thus further improving sensitivity), and usability (due to the removal of all steps previously requiring Unix functions, such as local running of BLAST and manipulations of large text files). Furthermore, we adopted an original quality filter which was able to test if each single EST candidate with sequence information of possible use for extending a known mRNA, was attributed to the same locus of that mRNA by an updated, complete and embedded version of UniGene. Lastly, we automated data summarization for an analyzed genome.

Following these improvements, 5′_ORF_Extender recognized a total of 477 loci, out of the 18,665 human loci represented in the mRNA reference set, as bona fide candidates for extension. The percentage of genes with an estimated incomplete mRNA 5′ coding sequence (2.6%) is in the lower range compared with previous estimates (in the range of 2–5%), which were based on more limited samples of sequences [1,16,17]. The sensitivity of the method depends on the size of the ever-growing EST repertoire available. Although EST single-pass sequencing itself is prone to experimental errors, we strongly suggest that the mRNAs for which more than one EST was found, deriving from two independent cDNA libraries and leading to the same prediction, possess a longer CDS than the one described so far.

The identification of the most upstream currently definable AUG start codon in an mRNA sequence cannot itself formally exclude that in some cases a downstream AUG codon may also be used by the ribosome, due to the phenomenon of alternative translation [19]. In addition, due to the availability of a large number of tissue- or stage-specific EST data, the EST-based extended CDS and/or the mRNA with the incomplete ORF could possibly derive from alternative transcription starting sites and/or splicing at the investigated locus. Nevertheless, the protein-coding nature of additional nucleotides at the 5′ of the locus is highlighted, and in the results each distinct alternative RefSeq mRNA isoform mapping to the same locus is associated only with the EST-based extended CDS with which it is compatible.

As a proof-of-concept, we have experimentally confirmed the EST-based models showing an extended coding region at 5′ for three randomly chosen mRNAs: *GNB2L1* (guanine nucleotide binding protein (G protein), beta polypeptide 2-like 1), *QARS* (glutaminyl-tRNA synthetase) and *TDP2* (tyrosyl-DNA phosphodiesterase 2) (Supplementary Table). In these three cases, cross-species comparison at amino acid level indicated a very high grade of conservation of the extended sequence among primates (Fig. 2). Therefore, the predicted product for these three human genes should be redefined for functional studies. A detailed analysis of these extended sequences is provided in the Supplementary File.

In conclusion, while genomic browsers continuously scan deposited sequences and try to build mRNA models employing different methods, they appear not to systematically address the issue of completeness of the coding region at mRNA 5′ end. Our approach has been able to generate, on a genome scale, 477 EST-driven original extended CDSs of human mRNAs, which are now available to researchers interested in these loci. In addition, software users can access a list of 20,775 human mRNAs in which the presence of an in-frame stop codon upstream of the known start codon indicates completeness of the coding sequence at 5′ in the current form.

Acknowledgments

This work was funded by "RFO" grants from Alma Mater Studiorum — University of Bologna to P.S., F. Fr., S.C., R.C., L.V. and F. Fa. The 5′_ORF_Extender software was executed on the Apple Mac Pro "Multiprocessor Server" available at the Center for Research in Molecular Genetics "Fondazione CARISBO", Bologna, and funded by "Fondazione CARISBO". We are grateful to Gabriella Mattei and Michela Bonaguro for their excellent technical assistance with cDNA sequencing. We are also grateful to the anonymous manuscript reviewers for their significant help in improving the software and the manuscript. In particular, the idea of using precomputed BLAT UCSC genome coordinates as the starting point of the algorithm was suggested by one of them.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2012.05.012.

References

[1] R. Casadei, P. Strippoli, P. D'Addabbo, S. Canaider, L. Lenzi, L. Vitale, S. Giannone, F. Frabetti, F. Facchin, P. Carinci, M. Zannotti, mRNA 5′ region sequence incompleteness: a potential source of systematic errors in translation initiation codon assignment in human mRNAs, Gene 321 (2003) 185–193.
[2] M. Kozak, Pushing the limits of the scanning mechanism for initiation of translation, Gene 99 (2002) 1–34.
[3] J. Sambrook, D.W. Russell, Rapid amplification of 5′ cDNA Ends, in: J. Sambrook, D.W. Russell (Eds.), Molecular Cloning — A Laboratory Manual, Cold Spring Harbor Laboratory Press, New York, 2001, pp. 8.54–8.60.
[4] P. Carninci, A. Westover, Y. Nishiyama, T. Ohsumi, M. Itoh, S. Nagaoka, N. Sasaki, Y. Okazaki, M. Muramatsu, C. Schneider, Y. Hayashizaki, High-efficiency full-length cDNA cloning by biotinylated CAP trapper, Genomics 37 (1996) 327–336.
[5] R. Kodzius, M. Kojima, H. Nishiyori, M. Nakamura, S. Fukuda, M. Tagami, D. Sasaki, K. Imamura, C. Kai, M. Harbers, Y. Hayashizaki, P. Carninci, CAGE: cap analysis of gene expression, Nat. Methods 3 (2006) 211–222.
[6] F. Denoeud, P. Kapranov, C. Ucla, A. Frankish, R. Castelo, J. Drenkow, J. Lagarde, T. Alioto, C. Manzano, J. Chrast, S. Dike, C. Wyss, C.N. Henrichsen, N. Holroyd, M.C. Dickson, R. Taylor, Z. Hance, S. Foissac, R.M. Myers, J. Rogers, T. Hubbard, J. Harrow, R. Guigó, T.R. Gingeras, S.E. Antonarakis, A. Reymond, Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions, Genome Res. 17 (2007) 746–759.
[7] M.D. Adams, J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde, R.F. Moreno, et al., Complementary DNA sequencing: expressed sequence tags and human genome project, Science 252 (1991) 1651–1656.
[8] M.S. Boguski, T.M. Lowe, C.M. Tolstoshev, dbEST—database for "expressed sequence tags", Nat. Genet. 4 (1993) 332–333.
[9] F. Frabetti, R. Casadei, L. Lenzi, S. Canaider, L. Vitale, F. Facchin, P. Carinci, M. Zannotti, P. Strippoli, Systematic analysis of mRNA 5′ coding sequence incompleteness in *Danio rerio*: an automated EST-based approach, Biol. Direct 2 (2007) 34.
[10] L. Lenzi, F. Frabetti, F. Facchin, R. Casadei, L. Vitale, S. Canaider, P. Carinci, M. Zannotti, P. Strippoli, UniGene Tabulator: a full parser for the UniGene format, Bioinformatics 22 (2006) 2570–2571.
[11] L. Lenzi, F. Facchin, F. Piva, M. Giulietti, M.C. Pelleri, F. Frabetti, L. Vitale, R. Casadei, S. Canaider, S. Bortoluzzi, A. Coppe, G.A. Danieli, G. Principato, S. Ferrari, P. Strippoli, TRAM (Transcriptome Mapper): database-driven creation and analysis of transcriptome maps from multiple sources, BMC Genomics 12 (2011) 121.
[12] H.T. Yudate, M. Suwa, R. Irie, H. Matsui, T. Nishikawa, Y. Nakamura, D. Yamaguchi, Z.Z. Peng, T. Yamamoto, K. Nagai, K. Hayashi, T. Otsuki, T. Sugiyama, T. Ota, Y. Suzuki, S. Sugano, T. Isogai, Y. Masuho, HUNT: launch of a full-length cDNA database from the Helix Research Institute, Nucleic Acids Res. 29 (2001) 185–188.
[13] E.W. Sayers, T. Barrett, D.A. Benson, E. Bolton, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, S. Federhen, M. Feolo, I.M. Fingerman, L.Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D.J. Lipman, Z. Lu, T.L. Madden, T. Madej, D.R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, L. Phan, K.D. Pruitt, G.D. Schuler, E. Sequeira, S.T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T.A. Tatusova, L. Wagner, Y. Wang, W.J. Wilbur, E. Yaschenko, J. Ye, Database resources of the National Center for Biotechnology Information, Nucleic Acids Res. 39 (2011) D38–D51.
[14] J.Z. Sanborn, S.C. Benz, B. Craft, C. Szeto, K.M. Kober, L. Meyer, C.J. Vaske, M. Goldman, K.E. Smith, R.M. Kuhn, D. Karolchik, W.J. Kent, J.M. Stuart, D. Haussler, J. Zhu, The UCSC Cancer Genomics Browser: update 2011, Nucleic Acids Res. 39 (2011) D951–D959.
[15] P. Flicek, M.R. Amode, D. Barrell, K. Beal, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. Kähäri, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, P. Larsson, I. Longden, W. McLaren, B. Overduin, B. Pritchard, H.S. Riat, D. Rios, G.R. Ritchie, M. Ruffier, M. Schuster, D. Sobral, G. Spudich, Y.A. Tang, S. Trevanion, J. Vandrovcova, A.J. Vilella, S. White, S.P. Wilder, A. Zadissa, J. Zamora, B.L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X.M. Fernández-Suarez, J. Herrero, T.J. Hubbard, A. Parker, G. Proctor, J. Vogel, S.M. Searle, Ensembl 2011, Nucleic Acids Res. 39 (2011) D800–D806.
[16] Y. Suzuki, D. Ishihara, M. Sasaki, H. Nakagawa, H. Hata, T. Tsunoda, M. Watanabe, T. Komatsu, T. Ota, T. Isogai, A. Suyama, S. Sugano, Statistical analysis of the 5′

untranslated region of human mRNA using "Oligo-Capped" cDNA libraries, Genomics 64 (2000) 286–297.
[17] B.M. Porcel, O. Delfour, V. Castelli, V. De Berardinis, L. Friedlander, C. Cruaud, A. Ureta-Vidal, C. Scarpelli, P. Wincker, V. Schachter, W. Saurin, G. Gyapay, M. Salanoubat, J. Weissenbach, Numerous novel annotations of the human genome sequence supported by a 5′-end-enriched cDNA collection, Genome Res. 14 (2004) 463–471.
[18] N. Kitagawa, T. Washio, S. Kosugi, T. Yamashita, K. Higashi, H. Yanagawa, K. Higo, K. Satoh, Y. Ohtomo, T. Sunako, K. Murakami, K. Matsubara, J. Kawai, P. Carninci, Y. Hayashizaki, S. Kikuchi, M. Tomita, Computational analysis suggests that alternative first exons are involved in tissue-specific transcription in rice (*Oryza sativa*), Bioinformatics 21 (2005) 1758–1763.
[19] G.A. Bazykin, A.V. Kochetov, Alternative translation start sites are conserved in eukaryotic genomes, Nucleic Acids Res. 39 (2011) 567–577.