Diversity of Translation Start Sites May Define Increased Complexity of the Human Short ORFeome*§

Masaaki Oyama‡§, Hiroko Kozuka-Hata‡, Yutaka Suzuki¶, Kentaro Semba||, Tadashi Yamamoto‡, and Sumio Sugano¶

Our previous proteomics analysis of small proteins expressed in human K562 cells provided the first direct evidence of translation of upstream ORFs in human full-length cDNAs (Oyama, M., Itagaki, C., Hata, H., Suzuki, Y., Izumi, T., Natsume, T., Isobe, T., and Sugano, S. (2004) Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. Genome Res. 14, 2048–2052). In the present study, we performed an in-depth proteomics analysis of human K562 and HEK293 cells using a two-dimensional nano-liquid chromatography-tandem mass spectrometry system. The results led to the identification of eight protein-coding regions besides 197 small proteins with a theoretical mass less than 20 kDa that were already annotated coding sequences in the curated mRNA database. In addition to the upstream ORFs in the presumed 5′-untranslated regions of mRNAs, bioinformatics analysis based on accumulated 5′-end cDNA sequence data provided evidence of novel short coding regions that were likely to be translated from the upstream non-AUG start site or from the new short transcript variants generated by utilization of downstream alternative promoters. Protein expression analysis of the GRINL1A gene revealed that translation from the most upstream start site occurred on a major alternative splicing transcript, whereas this initiation site was not utilized on the major mRNA, resulting in translation of the downstream ORF from the second initiation codon. These findings reveal a novel post-transcriptional system that can augment the human proteome via the alternative use of diverse translation start sites coupled with transcriptional regulation through alternative promoters or splicing, leading to increased complexity of short protein-coding regions defined by the human transcriptome. Molecular & Cellular Proteomics 6:1000–1006, 2007.

According to a widely accepted model of translation in eukaryotes, a 40 S ribosomal subunit is first recruited to the cap structure of mRNA and linearly scans the 5′-untranslated region (UTR)1 for the initiator ATG. When it recognizes the translation initiation site, it pauses until a large 60 S subunit joins, and the complete ribosomal complex starts translation (1). Thus, the most “upstream” ORF of the mRNA should be translated according to this model especially when there is a good context around its ATG codon (2).

Surprisingly large scale sequence analyses focusing on the 5′-UTRs of human full-length cDNA sequences showed that 41–49% of these cDNAs had at least one ATG codon upstream of the presumed coding sequence (3, 4). This means that there are potential coding regions in the 5′-UTRs of many genes if this classical model indeed represents a general mechanism of translation initiation. However, there have been only a limited number of reports on mammalian genes that investigated translation of the upstream ORFs (5). These previous studies referred to the upstream short ORF as a regulator of the translation of each downstream coding sequence (CDS) in which cases there was no direct evidence of translation of the upstream ORF in vivo (6, 7). Our recent proteomics analysis of small proteins expressed in human leukemia K562 cells led to identification of novel peptides corresponding to the upstream short ORFs in four kinds of human full-length cDNAs (8). That study clearly provided direct evidence of the translation of the upstream ORFs in the presumed 5′-UTRs within human cells.

To make a more comprehensive search for unpredicted short coding sequences from human mRNAs, we have constructed an automated two-dimensional (2D) nano-LC system coupled with a high resolution hybrid tandem mass spectrometer. Separation of peptide mixtures through on-line multidimensional LC enables us to perform large scale identification of highly complex biological samples, such as cell lysates (9–11). In the present study, we applied this technology for shotgun identification of the components in small protein-enriched fractions prepared from human cultured cells. We carried out analyses of small proteins expressed in human K562 and HEK293 cells and mapped the regions for all the identified peptides on RefSeq human mRNAs, which repre-

---

* The abbreviations used are: UTR, untranslated region; CDS, coding sequence; 2D, two-dimensional; DBTSS, DataBase of Transcriptional Start Sites; HEK, human embryonic kidney; aa, amino acids.
sent curated non-redundant transcripts (12) based on large scale collection of human full-length cDNA sequences (13–16). Here we describe the locations of short coding regions besides the RefSeq CDSs and analyze the diversity and the utilization of their probable translation start sites.

**EXPERIMENTAL PROCEDURES**

**Cell Culture**—Human erythroleukemia K562 cells were cultured in RPMI 1640 medium supplemented with 10% fetal calf serum. HEK293 cells were grown in Dulbecco’s modified Eagle’s medium supplemented with 10% fetal calf serum. After harvesting, the cells were washed three times with PBS. For proteasome inhibition analysis, HEK293 cells were labeled with stable isotopes of arginine according to a previous study (17). HEK293 cells labeled with either L-arginine, L-[U-15C6,14N4]arginine, or L-[U-15C6,15N4]arginine (2 × 10^{-cm dishes per condition) were treated with 20 μM MG-132 (Calbiochem) for 0, 2, and 5 h, respectively. After washing with PBS, the harvested cells were all combined for subsequent quantitative analysis.

**Protein Identification by 2D LC-MS/MS Analysis**—A small protein-enriched fraction was prepared from the harvested cells by acid extraction as described previously (8). Sixty micrograms of the protein-enriched fraction was digested overnight at 37 °C with 150 pmol of trypsin in 25 mM Tris-HCl (pH 8.8) and desalted using a ZipTip (C_{18}; Millipore). The digested peptide mixture was then analyzed on an automated 2D LC-MS/MS system. After applying the peptide mixture to an strong cation exchange column (500-μm inner diameter × 35 mm long), elution was done with discrete step gradients of ammonium acetate from 0 to 400 mM (0, 5, 10, 15, 20, 25, 30, 40, 50, 100, 200, 300, and 400 mM) at a flow rate of 3 μl/min. Peptides eluted by each step were temporarily trapped on a C_{18} column (800-μm inner diameter × 3 mm long) and desalted with 0.1% formic acid. Second dimensional reversed-phase separation of the captured peptides was done on a 3 mm long) and desalted with 0.1% formic acid. Second dimensional reversed-phase separation of the captured peptides was done on a column (150-μm inner diameter × 75 mm long) filled with HIC sil C_{18} (3-μm particles, 120-Å pores; KYA Technologies) using a direct nano-LC flow system (Dina, KYA Technologies). The peptides were eluted with a linear 4–41% gradient of acetonitrile containing 0.1% formic acid over 80 min at a flow rate of 200 nl/min and sprayed into a quadrupole time-of-flight tandem mass spectrometer (Q-ToF-2, Micromass). The acquired MS/MS spectra were processed against the RefSeq (National Center for Biotechnology Information (NCBI)) protein- and mRNA database (12) using the Mascot algorithm (Matrix Science) with the following parameters: variable modifications, oxidation (Met), N-acetylation, pyroglutamination (Gln); maximum missed cleavages, 3; peptide mass tolerance, 500 ppm; MS/MS tolerance, 0.5 Da. Only the peptide sequences that satisfied stringent peptide identification criteria with a Mascot score >40 were considered in this study. As to the search results for novel peptides, we performed further inspection to obtain the peptides with clear MS/MS spectra containing a sequence tag of at least four consecutive amino acids. The RefSeq databases were downloaded periodically from the NCBI ftp site (ftp.ncbi.nih.gov/refseq/). The list of the identified proteins/peptides was performed using AYUMS, which we have developed for automatic quantitation based on LC-MS/MS data (18).

**DNA Construction and Transfection**—Each of the GRINL1A variant 1 cDNAs from the 5′-end shown in Fig. 4B to the 3′-end of the downstream longest ORF was amplified and cloned in-frame 5′ to the V5 epitope in the pcDNAs.2/V5-DEST vector (Invitrogen) as described previously (19). The constructs were verified by sequence analysis, and subconfluent HEK293 cells were transfected with each plasmid using FuGENE 6 reagent (Roche Applied Science). Forty-eight h post-transfection HEK293 cells were then lysed in lysis buffer (10 mM Tris-HCl at pH 7.8, 150 mM NaCl, 1 mM EDTA, 1% Nonidet P-40) supplemented with protease inhibitor mixture Complete (mini; Roche Applied Science) and clarified by centrifugation for 10 min at 12,000 rpm at 4 °C.

**Antibody Generation and Immunoblotting**—For detection of endogenous proteins encoded by the GRINL1A transcript variants, we generated antibodies against the upstream ORF (86 aa) and the downstream ORF (368 aa) of the transcript variant 1 (NM_015532) (all accession numbers cited are from GenBank™). Antisera were raised in rabbits using a synthetic peptide extending from 64 to 77 residues (GNVEAPGETFAQRK) of the upstream ORF or from 357 to 368 residues (CRDEDDDWSSDF) of the downstream ORF followed by purification by affinity chromatography. Harvested human cells were lysed as described above, and Western blotting of each cell lysate (15 μg) was performed using anti-rabbit Ig-alkaline phosphatase conjugated antibody (Promega; 1:5000 dilution) for detection of anti-GRINL1A antibody (1:1000 dilution). The lysates were also analyzed with mouse anti-β-tubulin antibody (Sigma; 1:200 dilution) as a protein loading control.

**RESULTS**

**Identification of Small Proteins Using a 2D Nano-LC-MS/MS System**—To carry out a proteomics analysis of small proteins expressed in human cells, we prepared a small protein-enriched fraction for mass spectrometric analysis by performing acid extraction. After 60 μg of the acid extract was digested with trypsin, the peptide mixture was analyzed using a 2D nano-LC-MS/MS system. For the samples prepared from human K562 and HEK293 cells, fully automated 2D LC separation of the tryptic peptides led to the generation of 11,310 and 11,127 MS/MS spectra, respectively. As a result of a database search against the RefSeq human protein database (as of July 3, 2006), 1859 peptides were identified from the sample of human K562 cells, and 2337 peptides were identified from that of HEK293 cells. The identified peptides were ultimately assigned to 350 and 341 protein sequences, respectively. They included 134 and 162 proteins with a theoretical mass of less than 20 kDa, resulting in 197 non-redundant short proteins (Supplemental Table 1).

**Identification of Protein-coding Regions besides the Annotated CDSs of Human Curated mRNAs**—Next to search for protein-coding regions besides the annotated CDSs in the RefSeq human database, the acquired MS/MS data were searched against all three reading frames of RefSeq mRNAs (as of July 3, 2006). As a result, 10 peptides were identified from regions besides the annotated CDSs (Table I). Five peptides were mapped on four upstream ORFs in the 5′-UTRs of the corresponding mRNAs, while three other peptides were located in two downstream short ORFs in the 3′-UTRs (Fig. 1). The remaining two peptides were derived from reading frames with no initiator ATG in the corresponding mRNAs (NM_006442 and NM_152758).
Analysis of the Translation Start Sites of the Novel Coding Regions with No Initiator ATG Codon—Recent large scale analyses of 5'-end sequences of human cDNAs have provided in-depth information for identifying the precise transcriptional start sites of human mRNAs on a genome-wide scale (20–22). To investigate the probable translation start sites of the novel protein-coding regions with no initiator ATG in the corresponding RefSeq mRNAs (NM_006442 and NM_152758), we analyzed their 5'-end sequence data accumulated in the DataBase of Transcriptional Start Sites (DBTSS) (Release 5.2) (dbtss.hgc.jp/) in which we have entered 1.4 million 5'-end sequence data of human cDNAs derived from various kinds of 5'-oligo-capped cDNA libraries (23).

Regarding the transcripts of DRA1 (NM_006442) and YTHDF3 (NM_152758), most of the 5'-end sequence data

**TABLE I**

<table>
<thead>
<tr>
<th>Coding sequences besides each main CDS identified by searching against RefSeq human mRNAs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Length</strong>&lt;sup&gt;a&lt;/sup&gt; (bp)</td>
</tr>
<tr>
<td>NM_003426</td>
</tr>
<tr>
<td>NM_006442</td>
</tr>
<tr>
<td>NM_015332</td>
</tr>
<tr>
<td>NM_019048</td>
</tr>
<tr>
<td>NM_024099</td>
</tr>
<tr>
<td>NM_152758</td>
</tr>
<tr>
<td>NM_194846</td>
</tr>
<tr>
<td>NM_207368</td>
</tr>
</tbody>
</table>

<sup>a</sup> Each data entry is based on the sequence information of the corresponding RefSeq mRNA (as of July 3, 2006). Frame indicates the relationship between the identified peptide and the corresponding RefSeq CDS.

<sup>b</sup> The peptides identified from each cell lysate are indicated (C).
showed the same exon structure as the corresponding RefSeq transcript with diverse transcriptional start sites as reported previously (20). Even the longest 5’-UTR contained no ATG codon upstream of the novel coding region, suggesting the existence of their non-AUG translation start sites. Fig. 2 shows the 5’-end structure of the YTHDF3 transcripts accumulated in DBTSS (Release 5.2) (dbtss.hgc.jp). As previous studies indicated that recognition of a nonstandard ACG, CUG, or GUG start codon by ribosomes required a strong Kozak’s context (24), the only GUG codon flanked by a purine at position −3 and a guanine at position +4 was considered to be a putative translation start site in the corresponding YTHDF3 transcripts expressed in K562 cells (Fig. 2). For DRAP1, the upstream CUG codon that satisfied Kozak’s context was also predicted to be a novel translation start site in the same reading frame as the newly identified coding region (data not shown).

Analysis of the Translation Start Sites of the Novel Coding Regions Downstream of the Already Annotated CDS—As shown in Fig. 1, we identified novel protein-coding regions downstream of the annotated CDS in the RefSeq transcripts of C11orf48 (NM_024099) and LOC348262 (NM_207368). Large scale identification of transcription start sites of human genes also provided evidence of frequent utilization of alternative promoters at the genomic level (21, 22). For the C11orf48 gene, analysis of the accumulated 5’-end cDNA sequence data in DBTSS (Release 5.2) (dbtss.hgc.jp) provided evidence of short transcript variants harboring the identified coding region in the longest ORF (Fig. 3A). The curated CDS that was deduced from the short transcript showed 83% identity and 96% similarity with the translated product of the mouse RIKEN cDNA 18100009A15 gene (NP_079739; hypothetical protein LOC66276), indication functional constraint on this novel CDS (Fig. 3B). Although there was only one 5’-end cDNA data entry for LOC348262 (see the sequence data at DBTSS (Release 5.2); dbtss.hgc.jp), it also represented a short transcript form that was considered to be generated from the putative downstream transcriptional start site in the 3’-UTR of the RefSeq transcript. These results indicated that translation of these small proteins was likely to occur from the novel translation start sites on the short transcript variants generated through their downstream alternative promoters.

Analysis of the Translation Start Sites of the GRINL1A Alternative Transcripts—In the case of the GRINL1A gene, recent intensive analysis of its transcript forms has revealed that this gene locus generates a wide variety of transcript variants (25). The two peptides identified in our analysis corresponded to three kinds of splicing variants whose expression was regulated by the same promoter but which were differentially spliced (Fig. 4A). Transcript variant 1 (NM_015532) and variant 2 (NM_001018102) harbor the most upstream ORFs (86 aa) followed by the downstream longest ORFs as long as 368 and 211 aa, respectively. Transcript variant 3 (NM_001018103) is generated by the use of the alternative donor site of the first exon containing an ORF (148 aa) with the same N- and C-terminal amino acid sequences as the most upstream ORF of variant 1. These transcript variants share the same 5’-end structure in which the common first ATG is located in a good Kozak’s consensus sequence, whereas the second partially meets the criteria (2) (Fig. 4B). As RT-PCR analysis showed ubiquitous expression of variants 1 and 3 in various human tissues as well as human K562 cells and HEK293 cells (data not shown), we performed Western blot analysis of their translated products in vivo. Our result indicated the existence of the proteins encoded by the upstream ORF of variant 3 and the downstream longest ORF of variant 1 in the human cells investigated (Fig. 4C). Although HEK293 cells that were transfected with cDNA constructs corresponding to GRINL1A transcript variant 1 generated the product encoded by the upstream ORF (Fig. 4C) as well as the downstream ORF (Fig. 4D), translation of this upstream ORF was not performed in the non-transfected control human cells investi-
FIG. 3. A, structure of the alternative transcript presumed to encode a novel short protein downstream of the RefSeq CDS in the human C11orf48 locus. The open areas indicate the exon structures of each transcript. The shaded and filled areas denote their longest ORFs, respectively. B, alignment of the novel CDS with its mouse ortholog (NP_079739; hypothetical protein LOC66278) using ClustalW (31). An asterisk (*) and a dot (.) indicate identity and similarity in amino acid sequence, respectively. DBTSS, dbtss.hgc.jp/ (23).

FIG. 4. A, genomic organization of the three alternative GRINL1A transcripts. The open areas indicate the exon structures of each GRINL1A transcript variant. Transcript variants 1 and 2 are each composed of four exons, whereas variant 3 consists of two exons. In each transcript, the filled area shows the most upstream ORF, whereas the shaded area denotes the longest ORF that starts from the second ATG. Note that these downstream ORFs are not out of the reading frame of the corresponding upstream ORF. B, 5′-end structure of the GRINL1A transcripts. The first two ATGs are indicated as uppercase. The two major transcriptional start sites are indicated 14 and 5 nucleotides upstream of the first ATG as a result of the sequence analysis of their 5′-end “oligo-capped” clones obtained from HEK293 cells (data not shown; according to the previously described method in Refs. 30 and 32). C, Western blot of human cells using antibodies against both of the two overlapping ORFs of NM_015532. Lanes pC1, pC2, and pCdel, HEK293 cells transfected with the GRINL1A variant 1 cDNA construct with the 5′-end as described in B. D, Western blot of HEK293 cells transfected with each construct using anti-V5 epitope antibody. The V5 epitope was tagged at the C terminus of the downstream ORF of NM_015532. The lysates were also analyzed with anti-β-tubulin antibody as a protein loading control. WB, Western blot; TSS, transcription start site.
gated (Fig. 4C). These results demonstrated alternative utilization of translation start sites of co-expressed GRINL1A splicing variants to generate two independent proteins within the cells.

**DISCUSSION**

Our large scale proteomics analysis of small proteins expressed in human K562 and HEK293 cells enabled us to find novel short CDSs from more diverse regions of human mRNAs than we had expected. In addition to novel upstream ORFs with the corresponding initiator ATG codon (Fig. 1), our study also revealed translation of the presumed 5’-UTRs of DRAP1 (NM_006442) and YTHDF3 (NM_152758) despite the lack of a standard ATG start site (Fig. 2). There were some previous studies that showed translation initiation at an upstream non-AUG codon such as ACG, CUG, or GUG to generate N-terminal extended protein isoforms (2, 24). Our results indicate that out-of-frame non-AUG start codons would also be available for ribosomes to translate short proteins that are independent of the main products. Furthermore our analysis led to identification of the translated products derived from the downstream short ORFs in the presumed 3’-UTRs of C11orf48 (NM_024099) and LOC348262 (NM_207368). Recent genome-wide analyses of transcriptional start sites of mammalian genes have revealed the prevalence of new transcripts generated through the use of alternative promoters in the CDSs or 3’-UTRs of already known transcripts (22). Although short transcripts generated through downstream promoters are considered to be potential noncoding transcripts (22, 26), our results suggest the possibility that an unexpectedly large number of small proteins could be encoded by such short transcripts recently unveiled.

In addition to alternate transcription initiation, RNA editing through alternative splicing also contributes to diversification of an ORF. Recent intensive analysis of the human GRINL1A transcripts (25) has revealed multiple exon structures corresponding to the identified peptides (Figs. 1 and 4A). As described in Fig. 4B, the flanking sequence of their common first ATG satisfies Kozak’s consensus, suggesting that the presence of the upstream ORF in the GRINL1A transcript variant 1 (NM_015532) and variant 2 (NM_001018102) exerts an inhibitory effect on translation initiation of the downstream ORF. Western blot analysis of human cells, however, indicated the existence of an ~50-kDa protein that was considered to be encoded by the downstream ORF of the GRINL1A transcript variant 1 (Fig. 4C). Transfection of the corresponding cDNA constructs demonstrated translation of this downstream ORF despite the presence of the upstream ORF and excluded the possibility that translational frameshift might enable ribosomes to produce a fused protein from the two ORFs (Fig. 4D) (27). Interestingly our results showed that the upstream first ATG was recognized as a translation start site on transcript variant 3 (NM_001018103), although the same ATG codon was ignored by ribosomes on variant 1 in HEK293 and K562 cells (Fig. 4C). It is possible that selection of the translation start sites of these two GRINL1A variants is regulated by some mechanisms including leaky scanning or internal ribosome entry site-dependent translation (7, 28). Considering that recent intensive analysis of 5’-end cDNA sequences showed high diversity of the first exon structures of human genes as well as their transcriptional start sites on a genomewide scale (21), it is possible that selective utilization of translation start sites on diverse first exons greatly increases the complexity of short protein-coding regions that are located upstream of each major CDS. Detailed analysis of the underlying mechanism that regulates the generation of two independent proteins from the GRINL1A locus will help to clarify the rules for the utilization of translation start sites on the 5’-ends of mRNAs.

Our results revealed evidence for a novel landscape of complex regulation that led to the production of a wide variety of short proteins that have not been mentioned in conventional gene prediction. This study, however, resulted in the identification of only a relatively small number of novel short protein-coding sequences from already known transcripts. As the ubiquitin-proteasome system is known to be responsible for regulating the amount of a variety of short lived proteins, we tried investigating the effect of proteasome inhibition to estimate the diversity of unknown translated products expressed in HEK293 cells. Although ubiquitin-associated proteins such as valosin-containing protein were relatively accumulated by blocking proteasome activity (29), there were no novel proteins that accumulated and were identified in our analysis, suggesting that widespread translation of unannotated short ORFs might not occur within the cells (Supplemental Table 2). As shown in Fig. 3B, comparative sequence analysis of the upstream ORF of NM_015532 showed 85% identity and 95% similarity with the mouse ortholog counterpart, indicating functional constraint on this sequence. In addition, the upstream ORF of NM_019048 was also highly conserved with 66% identity and 91% similarity. As none of the novel ORFs identified in our analysis have known motifs in their amino acid sequences, detailed analysis is needed to reveal the functions of these translated products.

Establishment of a comprehensive mRNA dataset including rare transcripts generated by the use of alternative promoters or alternative splicing will be essential to achieve precise definitions of a variety of unpredicted short CDSs based on accumulating MS/MS data. The accumulation of data on short coding regions in the total set of human transcripts, in combination with detailed analysis of the translational regulation of each gene, will result in a global view of a highly elaborate translation system for augmenting the human proteome.

**Acknowledgments**—We thank K. Kudo for technical support in operating the nanoflow LC system. We are also thankful to T. Adachi and A. Saito for processing large scale proteomics data. We are grateful to E. Nakajima for critical reading of the manuscript.
REFERENCES